Вероятность, статистика и прикладные исследования в аграрном университете

Серия основана в 2012 году

Министерство сельского хозяйства Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ»

МОДЕЛИ И МЕТОДЫ ПРИКЛАДНЫХ СИСТЕМНЫХ ИССЛЕДОВАНИЙ (ПРАКТИКУМ)

Учебное пособие

ДОПУЩЕНО МИНИСТЕРСТВОМ СЕЛЬСКОГО ХОЗЯЙСТВА РФ
В КАЧЕСТВЕ УЧЕБНОГО ПОСОБИЯ ДЛЯ СТУДЕНТОВ ВЫСШИХ АГРАРНЫХ УЧЕБНЫХ ЗАВЕДЕНИЙ, ОБУЧАЮЩИХСЯ ПО НАПРАВЛЕНИЯМ
«ЭКОНОМИКА», «БИЗНЕС-ИНФОРМАТИКА»,
«ПРИКЛАДНАЯ ИНФОРМАТИКА», «ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ»

Серия: Вероятность, статистика и прикладные исследования в аграрном университете
Под редакцией: профессора А. И. Трубилина, профессора И. А. Кацко

КРАСНОДАР
2014

УДК 519.257+681.3
ББК 38.81я73
М74

Рецензенты:
кафедра математической статистики, эконометрики и актуарных расчётов РГЭУ (РИНХ, г. Ростов н/Дону),
заведующая кафедрой – заслуженный деятель науки РФ,
доктор экономических наук,
профессор Л. И. Ниворожкина;

М. К. Беданоков – доктор экономических наук, профессор кафедры высшей математики МГТУ (г. Майкоп)

Авторский коллектив:
Л.С. Болотова (23), Г. В. Горелова (22), Г. В. Гудков (13), А. Е. Жминько (3,4), Ю. Н. Захарова (8,10), И. А. Кацко (8-12, 14, 20, 24), С. А. Кацко (6, 7), Е.В. Луценко (25), Ю. И. Лыпарь (26), А. И. Орлов (21), Н. Б. Паклин (15-19), А. Е. Сенникова (5, 12), С. Г. Чефранов (1,2).
Ответственный за выпуск И. А. Кацко.

М74 **Модели** и методы прикладных системных исследований (практикум): учеб. пособие / Под ред. А. И. Трубилина, И. А. Кацко. – Краснодар: КубГАУ, 2014. – 449 с., илл. (Серия: Вероятность, статистика и прикладные исследования в аграрном университете)

ISBN 978-5-94672-760-0

В учебном пособии кратко излагаются современные теоретические и практические основы представления и обработки данных и знаний для поддержки принятия решений при изучении сложных объектов и процессов. Первый раздел посвящен методам формализованного описания объектов на основании структурированных данных, реализующих подход «снизу вверх». В первой части предлагаются работы с использованием MS Excel. Вторая часть посвящена системе Statistica (© Statsoft), которая охватывает стандартные разделы многомерного статистического анализа (прикладной статистики или анализа данных). Третья часть позволяет дать представление о нелинейных методах анализа данных на примере оценки сложности временных рядов. Четвертая часть посвящена технологии интеллектуального анализа данных, реализованной в двух российских системах: 1) Data Mining системе PolyAnalyst (© Megaputer), 2) аналитической платформе Deductor (© BaseGroup Labs) – интеллектуальной информационной системы класса KDD для создания законченных прикладных решений в области Data Mining. В Deductor реализованы практически все современные технологии анализа табличных данных: Data Warehouse – хранилища данных, OLAP – многомерный анализ данных, Data Mining – добыча данных, Knowledge Discovery in Databases – обнаружение знаний в базах данных.

Второй раздел посвящен моделям и методам формализации знаний на основании представления и обработки знаний экспертов, реализующим подход «сверху вниз».

Материалы, представленные в настоящем пособии Г.В. Гореловой и И.А. Кацко разработаны при поддержке гранта РФФИ № 14-01-90401.

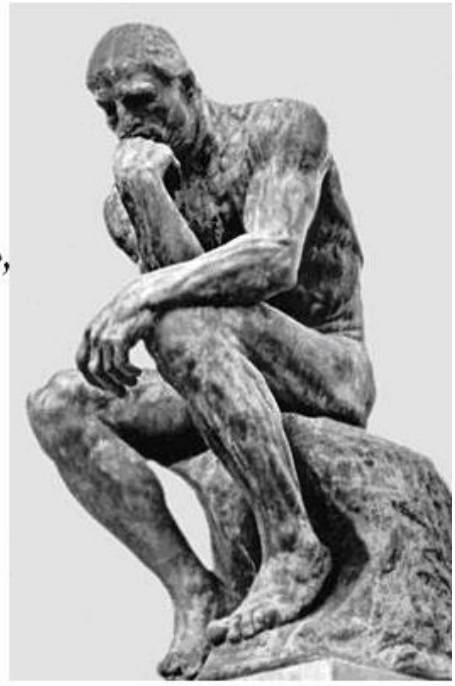
УДК 519.257+681.3
ББК 38.81я73

ISBN 978-5-94672-760-0

СОДЕРЖАНИЕ

DATA

Нам говорят «безумец» и «фантаст»,
Но, выйдя из зависимости грустной,
С годами мозг мыслителя искусный
Мыслителя искусственно создаст.
И. Гёте



IT KNOWLEDGE

Вид КубГАУ из космоса

Задаваться вопросом о том, как использовать компьютер на фирме, коротко говоря, неверно. Лучше спросить, как управлять фирмой в компьютерный век?

…Высшая задача управления – разработка курса дальнейшего развития фирмы. В этом смысле самой неотложной проблемой, стоящей перед нами, является проблема взаимоотношения человека и машины… Компьютер предлагает человеку инструмент, который превращает его в равного человеку партнёра…

Компьютер – нечто такое, что может быть использовано как дополнительные лобные отделения нашего мозга. Тут ожидается, а в некоторых случаях так оно и будет, некоторое слияние человека с машиной – их симбиоз…

Поведение есть функция законов управления, с помощью которых управление может быть организовано так, чтобы и мозг, и компьютер работали в согласии.

Стаффорд Бир. «Мозг фирмы», 1993

ВСТУПИТЕЛЬНОЕ СЛОВО

Функционирование организационных систем, их взаимодействие с окружающей средой невозможно представить в виде традиционных формальных количественных взаимосвязей, так как оно характеризуется наличием неопределенности, описанием на качественном уровне, неоднозначностью последствий тех или иных решений. Наличие таких условий позволяет отнести проблемы управления организационными системами к слабоструктурированным. Как известно, проблемы принято классифицировать как структурированные (I), слабоструктурированные (II) и неструктурированные (III).

Структурированные (или хорошо структурированные: well-structured) проблемы, это такие проблемы, в которых существенные зависимости ясно выражены и могут быть представлены в числах или символах. Это проблемы «количественно выраженные»; для решения проблем этого класса обычно используют методологию исследования операций.

Неструктурированные – это проблемы, выраженные главным образом в качественных признаках и характеристиках и не поддающихся количественному описанию и численным оценкам. Исследование таких проблем возможно только эвристическими методами анализа, ибо здесь отсутствует возможность применения логически упорядоченных процедур отыскания решений.

Слабоструктурированные проблемы характеризуются наличием как качественных, так и количественных элементов. Неопределенные, не поддающиеся количественному анализу зависимости, признаки и характеристики имеют тенденцию доминировать в этих смешанных проблемах. К этому классу проблем относится большинство наиболее сложных задач экономического, технического, политического, военно-стратегического характера.

В настоящее время реализация управления сложными системами и ситуациями привела к необходимости создания систем поддержки управленческих решений в условиях всех видов названных проблем.

Принятие решений – это наиболее сложный и ответственный этап деятельности человека в различных организационных структурах. Поэтому компьютерное моделирование процесса принятия решений сегодня становится центральным направлением автоматизации деятельности лица, принимающего решение (ЛПР).

Разрабатываются автоматизированные управленческие организационные системы (УПС).

Рисунок 1 –Укрупнённая схема предприятия: производство, организация и управление

Опыт свидетельствует о том, что системы поддержки повышают производительность лиц, принимающих решения. Улучшение качества решений возможно

потому, что ЛПР рассматривает альтернативы решения перед тем, как его принять, используя для этого модели формирования решений и их оценки.

Как известно, информационные системы на предприятиях выполняют функции сбора, обработки, хранения, передачи и представления информации. Между тем недостаточно уделяется внимания возможностям обработки имеющихся данных с использованием компьютера: проведению анализа, построению прогнозов и сценариев развития.

Традиционно считается, что система управления предприятием может быть представлена в виде некоторой иерархической организационной структуры, изображённой на рисунке 1 (Э.А. Трахтенгерц. Компьютерная поддержка принятия решений.– М.: СИНТЕГ, 1998.). Производственное предприятие (прямоугольник) состоит из трёх частей (блоков): подготовка и обслуживание производства, собственно производство, сбыт готовой продукции. Система управления предприятием изображается в виде треугольника, внутри и снаружи функционируют информационные потоки. В основании треугольника находятся информационные системы (сбора, обработки, хранения, передачи и представления информации), представляющие собой информационную модель предприятия. На вершине треугольника находятся руководители предприятия (системы принятия решений), которые принимают решения по управлению предприятием в соответствии с некоторыми целями. (Одна из целей – это получение наибольшей прибыли.)

Среднее звено системы управления (системы поддержки принятия решений – СППР) – среднее звено специалистов, которое на основании данных информационных систем проводит многовариантные расчёты для получения прогнозов и сценариев развития, оптимизации заданных параметров производства и т.д.

Считается, что описанная трёхуровневая схема управления является универсальной и позволяет исследовать деятельность любой сложной системы: от индивидуума, до правительства.

Для оказания поддержки управленческой деятельности при наличии больших объёмов информации в решении проблем I и II типа используют компьютерные методы, основанные на применении методов исследования операций, прикладной статистики, а также интеллектуального анализа данных.

В практикуме рассматриваются компьютерные методы, использующие основные направления обработки табличных данных большого объёма для решения задач анализа и прогнозирования. Это прикладная статистика, Data Warehouse – хранилища данных, OLAP – многомерный анализ данных, Data Mining – добыча данных, Knowledge Discovery in Databases – обнаружение знаний в базах данных.

Кроме того, в качестве примера использования нелинейных методов в анализе данных приводится идеология оценки сложности временных рядов.

Указанные направления компьютерной поддержки принятия решений при анализе табличных данных являются наиболее востребованными на практике. При расширении классов решаемых задач используются:

– генетические алгоритмы (комбинаторные задачи и задачи оптимизации);

– нечёткая логика (задачи управления в сложных системах);

– когнитивные карты (задачи первичного анализа сложных организационных систем) и т.д.

Второго апреля 2014 года проводился учредительный съезд «Российской ассоциации статистиков», участники которого говорили о необходимости привлечения современных информационных технологий в процесс обучения студентов. Авторы надеятся, что настоящий практикум, в известной степени, послужит реализации этой идеи.

Достоинством практикума является объединение в одном ключе основных методов анализа табличных данных и методов представления и обработки знаний, иллюстрированных применением в различных областях деятельности.

Для овладения искусством представления и обработки данных и знаний необходимо формулировать и проверять различные гипотезы о природе и структуре данных и знаний, изменять модели и т.д. То есть научиться проводить разведочный анализ, которому, собственно, и посвящён настоящий практикум. Хочется верить, что книга будет полезна в учебном процессе и послужит одним из шагов в подготовке нового поколения специалистов.

Кафедра статистики и прикладной математики КубГАУ выражает глубокую признательность и благодарность известным российским специалистам в области анализа данных и знаний, поддержавшим идею создания практикума по курсу модели и методы прикладных системных исследований и представившим свои материалы:

- д.т.н., проф. Л.С. Болотову (ВШЭ, г. Москва),
- д.т.н., проф. Г.В. Горелову (ЮФУ, г. Таганрог),
- д.м.н., проф. Г.В. Гудкова (КГМУ, г. Краснодар),
- д.т.н., проф. Ю.И. Лыпаря (СПбГПУ, г. Санкт-Петербург),
- д.т.н., д.э.н., проф. А.И. Орлова (МГТУ им. Баумана, г. Москва),
- к.т.н., доц. Н.Б. Паклина (BaseGroup Labs, г. Рязань),
- д.э.н., проф. С.Г. Чефранова (МГТУ, г. Майкоп).

Ответственный за выпуск
И.А. Кацко

– Вы можете, – продолжал Герман, – составить счастье моей жизни, и оно ничего не будет вам стоить: я знаю, что вы можете угадать три карты сряду…

А.С. Пушкин «Пиковая дама»

ПРЕДИСЛОВИЕ

Изучение систем в окружающем мире – это сложная задача, которая решается либо экспертно, либо статистически (в идеале оба подхода должны комбинироваться).

Пусть в результате ежегодных наблюдений за некоторым объектом, например, с/х предприятиями Краснодарского края, мы наблюдаем ряд переменных x_j – для *i*-го предприятия это будет наблюдения x_{ij}. Таким образом, все наблюдения – исходные статистические данные – можно представить в виде так называемых панелей, или иначе матриц, строки которых соответствуют объектам, а столбцы – наблюдениям:

$$X_{mk} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{pmatrix}.$$

Если $T_{1n}= (t_1, t_2, \dots, t_n)$ – вектор-строка, обозначающая *n* лет наблюдений, то исходные данные с помощью произведения Кронекера можно представить в виде блочной матрицы $T \otimes X$ размерности $k \times mn : T \otimes X = (\mathrm{t_1X}, \mathrm{t_2X}, \dots, \mathrm{t_nX})$.

Графически произведение Кронекера в данном случае можно представить как трехмерный куб (рисунок П1). В настоящее время существует несколько подходов к изучению подобных структур:

1) Рассмотрение срезов куба в пространстве и во времени. Отсюда практически все методы многомерного статистического анализа (прикладной статистики) ориентируются на решение трёх типов задач:

- ✓ выявления сходства между объектами – строками матрицы (одномерная классификация объектов – простая или комбинированная группировка; многомерная классификация – кластерный и дискриминантный анализ);
- ✓ анализ взаимодействия между признаками – столбцами матрицы (дисперсионный анализ, корреляционно-регрессионный анализ, ковариационный анализ, факторный и компонентный анализ, путевой анализ и т.д.);
- ✓ выявление закономерностей (трендов, сезонностей, циклов) изменения признаков предприятия – элементов x_{ij} во времени (анализ одномерных и многомерных временных рядов).

2) Применение оператора векторизации, преобразующего матрицу в вектор, позволяет получить матрицу размерности $nm \times k$, которую можно представить в виде модели ковариационного анализа.

3) Рассмотрение моделей панельных данных, предполагающих изучение зависимостей и в пространстве, и во времени.

4) Представление данных в виде многомерной модели *OLAP*–куба (Рисунок 2) с возможностями свёртки (обобщения одного или нескольких измерений и агрегировании соответствующих показателей); развёртки (получения подробной информации об одном или нескольких измерениях); расщепления и разрезания (развёртка на один уровень вниз по одному или нескольким измерениямдля ограниченного количества элементов); построения кросс-таблиц, кросс-диаграмм, что для небольших объёмов информации доступно в *Excel* (Данные – Сводная таблица).

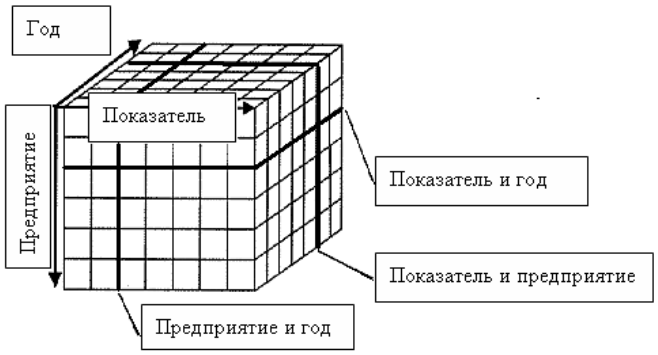


Рисунок 2–Представление данных в виде куба *OLAP*

5) Представление данных в виде пространственной базы данных с привязкой к некоторой базовой системе координат (например, земной поверхности) и использование в географических информационных системах (ГИС) для решения задач визуализации (нанесения информации на географическую карту в виде различных векторных слоёв с информацией о земельных участках, экологическом районировании, почвах, социальных, экономических показателях и.т.д.), тематического поиска, анализа местоположения, топографического анализа, анализа потоков (связность, кратчайший путь), пространственного анализа (поиск шаблонов, центров, автокорреляций), измерения (расстояний, периметра, очертания, направления).

Первые три подхода рассматриваются в рамках прикладной статистики и эконометрики. Последние два подхода относятся к информационным технологиям многомерных баз данных и прикладной статистики.

Цель решения задач выявления сходства между объектами, анализа взаимодействия между признаками (в пространстве и во времени), выявления закономерностей – получение описания объектов в виде конечных формул для решения задач управления и предсказания.

Опыт применения рассмотренных выше эконометрических моделей (подходы 1-3) показывает, что часто они неадекватно описывают реальную социально-экономическую ситуацию. Под адекватностью эконометрической модели мы будем понимать достижение целей моделирования (получение моделей, объясняющих имеющиеся данные, моделей для предсказания, моделей для управления).

Изучая проблему адекватности эконометрических методов и моделей, следует указать на ряд философских течений, опираясь на которые современная наука обосновывает методологию и методы анализа данных: механицизм, учение о при-

чинности, детерминизм, случайность (стохастичность), «когнитивизм», детерминированный хаос и синергетика, мягкие вычисления. Одно из основных философских течений, оказавших большое влияние сначала на физические и технические науки, а затем и на социально-экономические, – это механицизм. Он предполагает, что весь окружающий мир – это гигантский механизм, части которого взаимодействуют между собой. Учение о причинности предполагает, что, например, климатические условия, количество внесенных удобрений и т.д. влияют на урожайность с/х культур. Таким образом, с точки зрения учения о причинности, все социально-экономические явления – это следствия определенных причин.

В сельском хозяйстве это уменьшение численности работников вследствие миграции в города; снижение урожайности вследствие отсутствия денег на удобрения и т. д. Причинность предполагает наличие связи, посредством которой причина порождает следствие. Один из основных принципов планирования экспериментов основывается на методе различий, который позволяет установить причинную связь. Например, если при сохранении всех условий проведения полевого опыта, отличие для нескольких делянок только в дозе внесения удобрений либо в способе обработки почвы, то различие урожайности по делянкам обусловливается в силу метода различий только одной из перечисленных выше причин.

Практически все известные меры связи строятся на основе сопряженности признаков, – т.е. на мере неизменности следования в среднем (обычно нормированной). Следующее философское учение, которое было основным в XIX веке – детерминизм. Как отмечал Р.Декарт, «следствие отстает во времени от причины из-за ограниченности чувственных восприятий человека» (*causa sive ratio* – причина есть не что иное, как разум, М.Клайн). Поэтому предполагается, что существуют некоторые детерминированные (функциональные) зависимости. Этот подход к описанию окружающего мира был отвергнут после того, как в 1927г Гейзенберг подверг критике и причинность, и детерминизм, сформулировав принцип неопределенности. Он утверждал, что на уровне микромира нет причинно-следственных связей и детерминизма, здесь главенствуют лишь вероятностные законы.

В этом контексте следует отметить, что до сих пор не оценены по достоинству понятия «жесткой» и «мягкой» модели, введенные академиком В. И. Арнольдом в 1997году и соответствующие представлениям о знаниях в интеллектуальных системах [20]. В. И. Арнольд на примере дифференциальных уравнений показал, что «мягкие» модели (модели, поддающиеся изменениям) могут учитывать НЕ-факторы (неопределенность, неоднозначность путей развития) [2]. «Жесткие» модели не вариативны – в них все предопределено априорными условиями и предположениями.

По нашему мнению, основная проблема разочарования практиков в экономико-математических моделях – «жесткость» технологий математического программирования и математической статистики (эконометрики). В основе математических моделей должен лежать принцип учета НЕ-факторов при использовании показателей и управляющих воздействий. В этом контексте следует отметить, что нетнеобходимости полностью отвергать «жесткие» модели, которые должно теперь рассматривать как возможное «идеальное» состояние системы в близко прогнози-

руемом будущем. «Мягкие» модели в смысле В.И.Арнольда – это дифференциальные модели и синергетика… как основа теоретических построений.

Между тем с точки зрения одного из классиков теории искусственного интеллекта, создателя теории нечетких множеств и автора термина *Soft Computing* (мягкие вычисления, 1994г) Лотфи Заде, основой для построения «мягких» моделей могут быть модели, полученные с помощью средств «вычислительного интеллекта» (эволюционного моделирования, нейронных сетей ит.д.) [21].

Несмотря на неутихающие споры в науке о том, «существуют ли закономерности объективно» или«…носят локальный характер и с их помощью можно прогнозировать состояние системы (объекта) в соответствующей (локальной) области», одна из основных целей современной науки – это поиск связей между переменными принадлежащим одинаковым или разным типам шкал и установление соответствующих закономерностей в рамках границ, обусловленных неопределённостью.

Современные ученые (И.И. Елисеева, В.О. Рукавишников и др.) указывают, что методы изучения связей внутренне противоречивы. Теоретическая обоснованность прямых методов приводит к идеализации связей и введению жестких детерминированных зависимостей (регрессионный анализ). Косвенные методы основываются на измерении сопряженности варьирования переменных и не могут напрямую содержательно интерпретироваться (корреляционный анализ). Цель корреляционного анализа – на основании сопряженности установление тесноты предполагаемой связи, что часто подвергается критике (проблема ложной корреляции).

В настоящее время негласно в экономической науке преобладает подход, предполагающий наличие причинно-следственных связей, обусловливающих динамическое и статическое состояние системы и дающих возможность повысить эффективность управления производством (организацией). Знание взаимодействующих факторов, количественных мер их влияния создает основу для практического воздействия (управления), делает возможным научно обосновать прогнозирование процессов и управление производством, фирмой и т.д.

Общая цель методов анализа данных – это свёртка имеющейся информации для решения прикладных задач: анализа и объяснения особенностей функционирования изучаемой системы, управления, прогнозирования. При этом практические задачи в переводе на научный язык интерпретируются как проблемы разведочного анализа данных, сводящиеся к первичной обработке и визуализации, исследованию и построению зависимостей, классификации и снижению размерности.

В последние десятилетия в связи с развитием информационных технологий к ним добавились задачи поиска ассоциаций, последовательностей, паттернов в данных и т.д. Очевидно, с развитием человека и окружающих его технологий будут появляться новые задачи иновые методики их решения.

В настоящее время решение задач построения моделей на основе статистической информации основывается на нескольких основных подходах:

1) вероятностном – обычно с предположением нормальности распределения изучаемых величин (математическая статистика),

2) геометрическом – данные не имеют вероятностной природы и образуют в многомерном пространстве структуры с определенными свойствами,

3) содержательном, предполагающем достижение целей моделирования.

Первые два подхода реализуются в прикладной статистике, третий – в интеллектуальном анализе данных. И первый, и второй подходы постулируют тот факт, что имеет место некоторая модель, обычно линейная, и наша цель – найти для неё оптимальные в определенном смысле параметры. Методы интеллектуального анализа с помощью нейронных сетей, методов эволюционного программирования и других методов машинного обучения итеративно подбирают модель, в определённом смысле наилучшим образом описывающую исходные данные. Следует отметить, что анализ данных – это процесс движения по спирали от простых методов к более сложным. И если простая (детерминированная, вероятностная) модель позволяет решать наши задачи (анализа, прогнозирования, управления), нет смысла искать более сложные методы. Здесь можно вспомнить методологический принцип, который называют законом достаточного основания, или законом экономии, – «не изобретать сущностей сверх необходимого». Это изречение приписывают древнему философу Оккаму и называется "Бритвой Оккама". Смысл "бритвы" заключается в том, что во всяком рассуждении следует избегать придумывания новых понятий, терминов, слов и тому подобного, если без них можно обойтись.

В нашем случае, если данные можно объяснить простыми моделями, не стоит пытаться объяснять их более сложно, то есть начинать решение следует с простейших методов до достижения целей анализа (например, с минимума объясняющих переменных, простейшей модели и т.д.). Как любят говорить научные руководители своим аспирантам, «лучшее – враг хорошего». В соответствии с подобной идеологией в настоящее время рабочая группа *BaseGroup Labs* рассматривает возможности построения системы анализа, ориентируясь на основные задачи анализа данных, по принципу «бритвы оккама»: от простейших методов к более сложным до достижения целей анализа (моделирования).

В статье М. Киселева и Е.Соломатина указывается [26], что «…основным сдерживающим фактором развития сферы аналитических услуг являлся низкий спрос. Неплатежеспособные потребители были слабо заинтересованы в получении эффективных решений – экономическая коньюктура позволяла получать прибыль другими способами. Сейчас, казалось бы, изменился инвестиционный климат, появился спрос на аналитический консалтинг, более того, производителям и продавцам программного обеспечения есть что предложить – рынок вроде бы созрел и снизу, и сверху. Проблема, однако, не в том, чтобы предложить нужный инструмент – оказалось, что потенциальные пользователи не могут его взять». Ведь средства анализа данных – не просто рыночный продукт; это идеология, способ мышления. Спустя 10 лет можно сказать, что ситуация практически не изменилась. До сих пор имеет место большой разрыв между разработчиками современных средств статистического и интеллектуального анализа данных и пользователями.

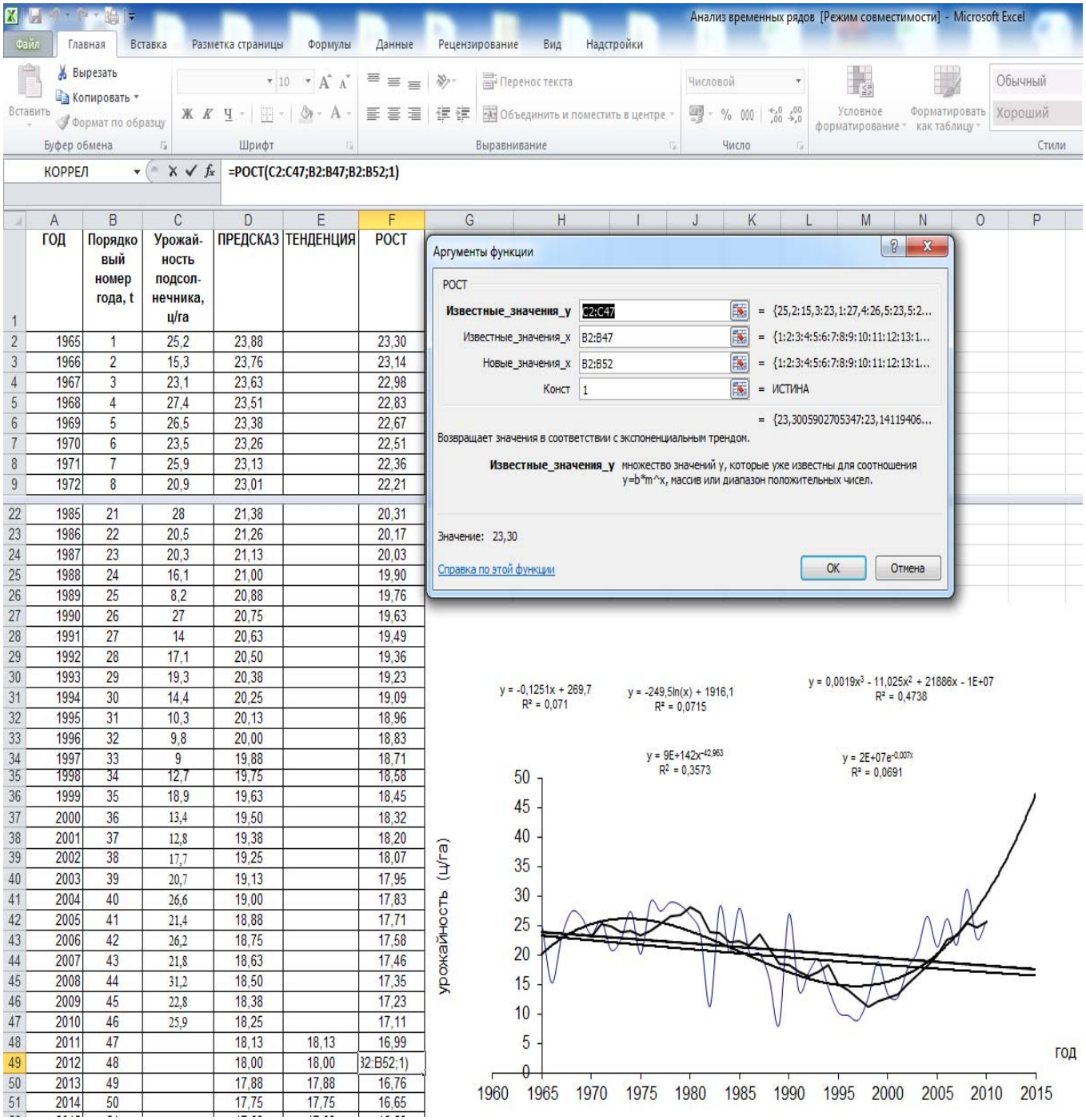
Именно поэтому предлагаемое учебно-методическое пособие состоит из трех частей, посвященных простейшим методам работы с данными в Excel, методам прикладной статистики и интеллектуального анализа – вместе представляющими

собой современную идеологию (методологию) анализа данных. Пособие можно использовать для системного, последовательного изучения большой области знаний – анализа структурированных данных, от статистики до хранилищ данных и DataMining.

Авторы

РАЗДЕЛ 1. ФОРМАЛИЗАЦИЯ ДАННЫХ – (ПОДХОД СНИЗУ – ВВЕРХ)

ЧАСТЬ I. РАБОТА С ДАННЫМИ В MS Excel



	A	B	C	D	E	F
1	ГОД	Порядковый номер года, t	Урожайность подсолнечника, ц/га	ПРЕДСКАЗ	ТЕНДЕНЦИЯ	РОСТ
2	1965	1	25,2	23,88		23,30
3	1966	2	15,3	23,76		23,14
4	1967	3	23,1	23,63		22,98
5	1968	4	27,4	23,51		22,83
6	1969	5	26,5	23,38		22,67
7	1970	6	23,5	23,26		22,51
8	1971	7	25,9	23,13		22,36
9	1972	8	20,9	23,01		22,21
22	1985	21	28	21,38		20,31
23	1986	22	20,5	21,26		20,17
24	1987	23	20,3	21,13		20,03
25	1988	24	16,1	21,00		19,90
26	1989	25	8,2	20,88		19,76
27	1990	26	27	20,75		19,63
28	1991	27	14	20,63		19,49
29	1992	28	17,1	20,50		19,36
30	1993	29	19,3	20,38		19,23
31	1994	30	14,4	20,25		19,09
32	1995	31	10,3	20,13		18,96
33	1996	32	9,8	20,00		18,83
34	1997	33	9	19,88		18,71
35	1998	34	12,7	19,75		18,58
36	1999	35	18,9	19,63		18,45
37	2000	36	13,4	19,50		18,32
38	2001	37	12,8	19,38		18,20
39	2002	38	17,7	19,25		18,07
40	2003	39	20,7	19,13		17,95
41	2004	40	26,6	19,00		17,83
42	2005	41	21,4	18,88		17,71
43	2006	42	26,2	18,75		17,58
44	2007	43	21,8	18,63		17,46
45	2008	44	31,2	18,50		17,35
46	2009	45	22,8	18,38		17,23
47	2010	46	25,9	18,25		17,11
48	2011	47		18,13	18,13	16,99
49	2012	48		18,00	18,00	32:B52;1)
50	2013	49		17,88	17,88	16,76
51	2014	50		17,75	17,75	16,65

Практическое занятие № 1

1.1 Анализ вариационных рядов в Excel

Цель работы: ознакомиться с возможностями методик анализа вариационных рядов, парного корреляционно-регрессионного анализа, получить навыки анализа данных в Excel.

Теоретические сведения

В реальных социально-экономических системах нельзя проводить активные эксперименты, поэтому данные обычно представляют собой наблюдения за происходящим процессом, например: курс валюты на бирже в течение месяца, урожайность пшеницы в хозяйстве за 30 лет, производительность труда рабочих за смену и т.д.

Результаты наблюдений − это, в общем случае, ряд чисел, расположенных в беспорядке, который для изучения необходимо упорядочить (проранжировать).

Операция, заключенная в расположении значений признака по возрастанию, называется ранжированием опытных данных.

После операции ранжирования опытные данные можно сгруппировать так, чтобы в каждой группе признак принимал одно и то же значение, которое называется вариантом (X_i). Число элементов в каждой группе называется частотой варианта (n_i).

Размахом вариации называется число $W=x_{max} - x_{min}$,
где, x_{max}– наибольший вариант;
x_{min}– наименьший вариант.

Сумма всех частот равна определенному числу n, которое называется объемом совокупности:

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_к = n$$

.

Отношение частоты данного варианта к объему совокупности называется относительной частотой или частостью этого варианта:

$$\widehat{p} = \frac{n_i}{n},$$

$$\sum_{i=1}^{k} \widehat{p}_i = \sum_{i-1}^{k} \frac{n_i}{n} = \frac{\sum_{i=1}^{k} n_i}{n} = \frac{n}{n} = 1$$

.

Последовательность вариант, расположенных в возрастающем порядке, называется вариационным рядом (вариация - изменение).

Вариационные ряды бывают дискретными и непрерывными. Дискретным вариационным рядом называется ранжированная последовательность вариант с соответствующими частотами и (или) частостями.

Пример 1.1. В результате тестирования группа из 24 человек набрала баллы: 4, 0, 3, 4, 1, 0, 3, 1, 0, 4, 0, 0, 3, 1, 0, 1, 1, 3, 2, 3, 1, 2, 1, 2. Построить дискретный вариационный ряд.

Проранжируем исходный ряд, подсчитаем частоту и частость вариант:
0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4.

В результате получим дискретный вариационный ряд (табл.1).

Таблица 1.1 – Ранжированный ряд успеваемости студентов

Балл, x_i	Число студентов, n_i	Относительная частота, $\hat{p}_i$
0	6	6/24
1	7	7/24
2	3	3/24
3	5	5/24
4	3	3/24
∑	24	1,000

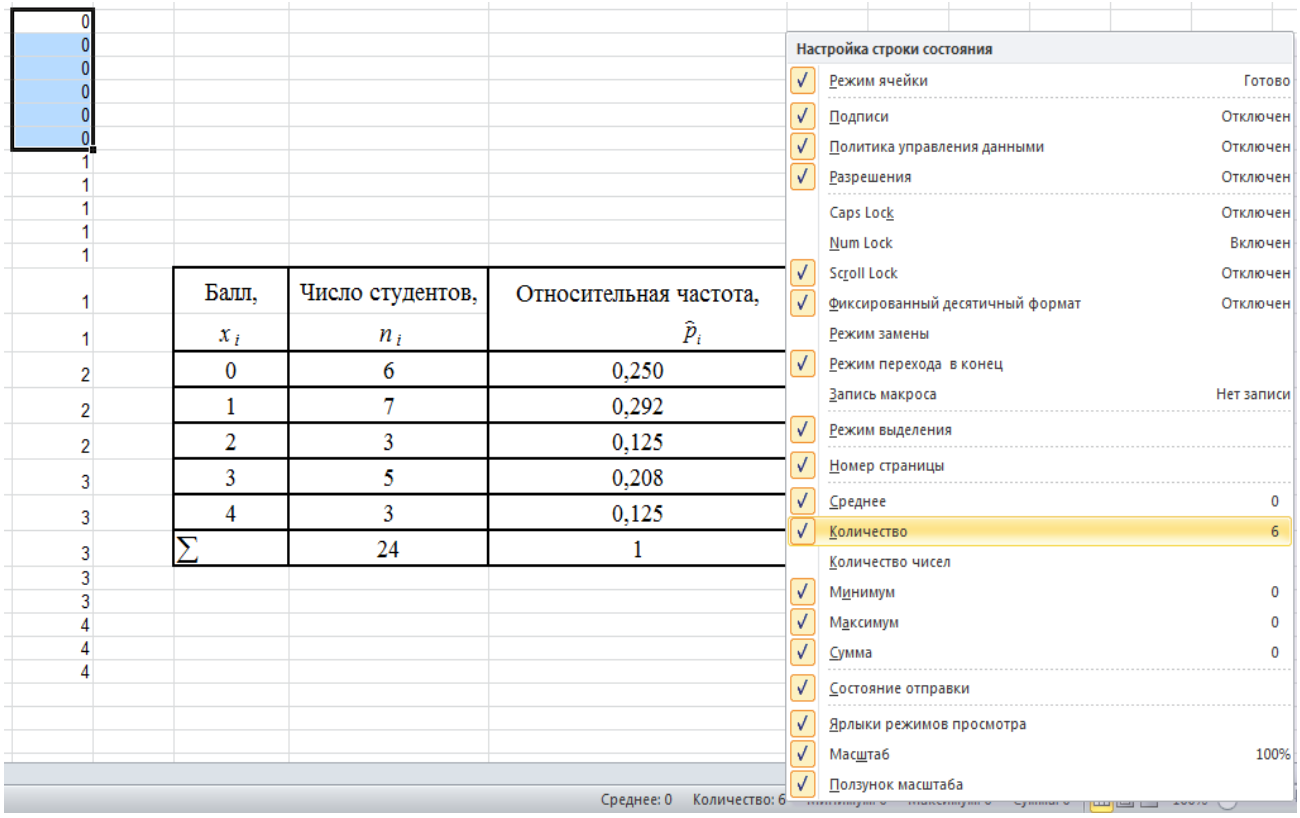


Рисунок 1.1 – Контекстное меню строки состояния

В *Excel*. Проранжируем исходный ряд. Для этого введём все данные в диапазон A1:A24 и воспользуемся кнопкой (Сортировка по возрастанию).

Подсчитаем частоту и частость вариант. Построим таблицу в диапазоне D2:G7 (рис.1.1). Рассмотрим два варианта подсчёта частот:

1) Выделим диапазон *A1:A6* – в котором находятся нули. Щёлкнем в нижней правой части окна *Excel* правой кнопкой мыши и выберем в контекстном меню вид итога, который по умолчанию будет появляться в итоговой строке при выделении произвольного диапазона (см. рис.1.1) – количество. Таким образом, последовательно выделяя диапазоны с одинаковыми значениями вариант, мы получим все частоты.

2) Выполним команду *Данные – Анализ данных – Гистограмма*. Заполним диалоговое окно в соответствии с рисунком 1.2.

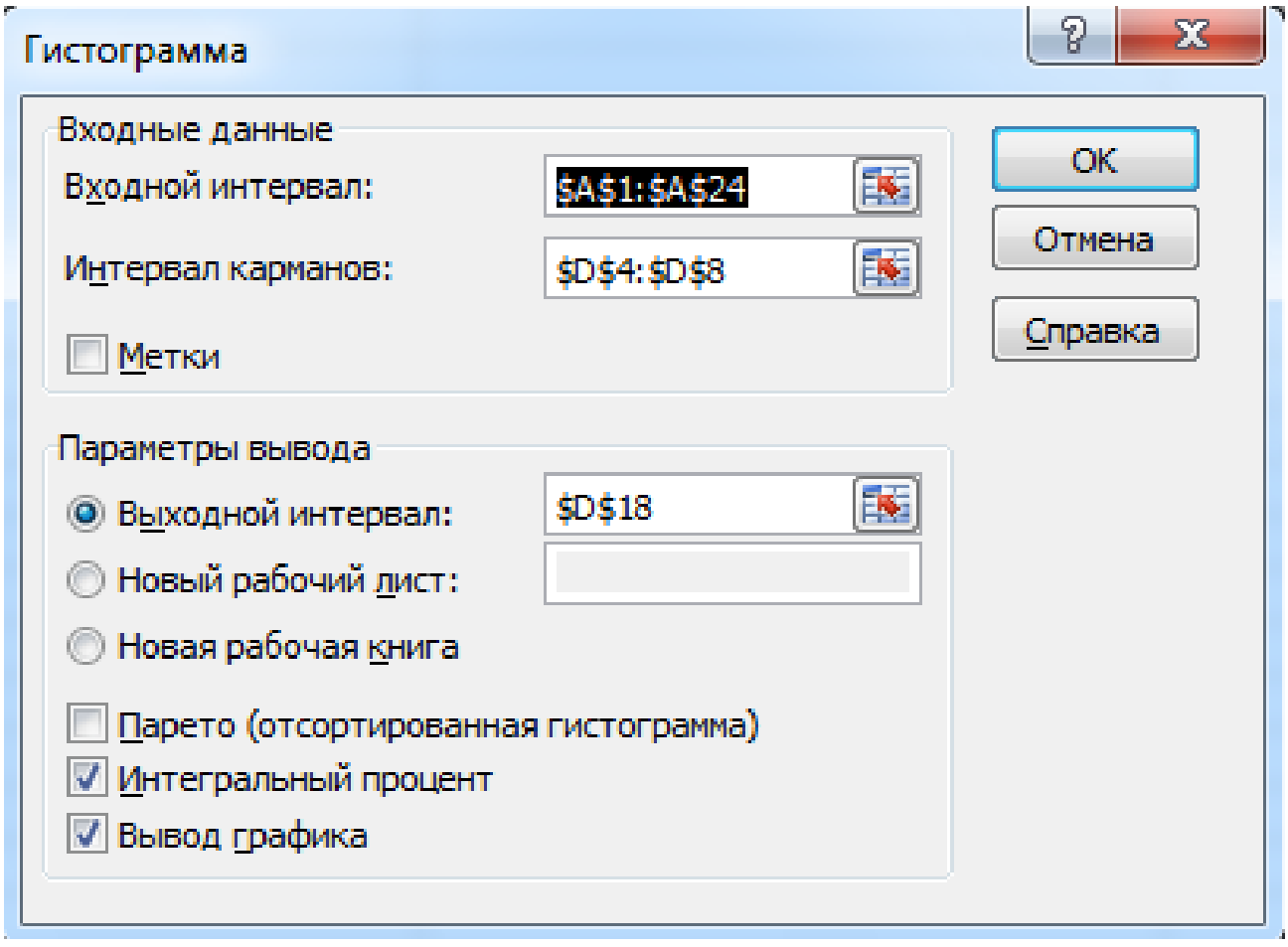
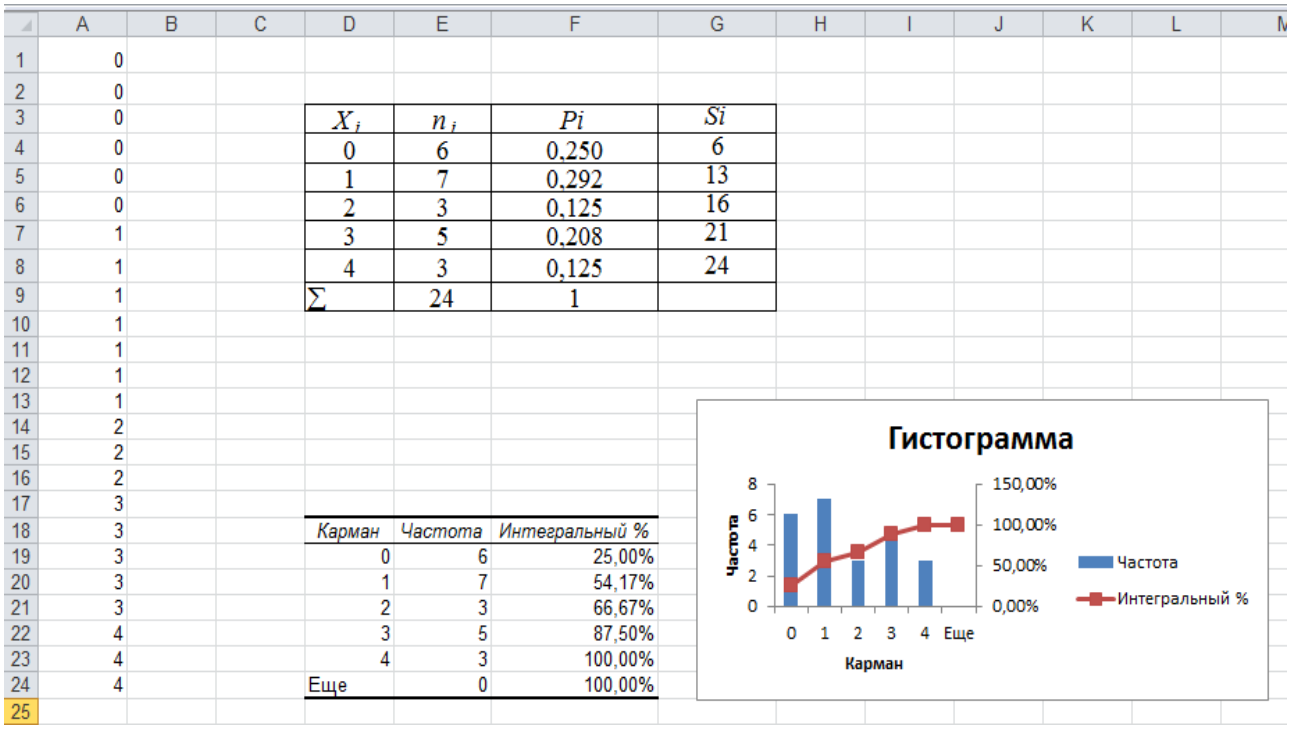


Рисунок 1.2 – Диалоговое окно инструмента пакета анализа *Гистограмма*

В результате получим таблицу с частотами вариантов и соответствующий график (рис.1.3).

Найдём объём выборки, заполнив все частоты вариант в диапозоне E4:E8, выделим в его левой кнопкой мыши и щёлкнем по кнопке Σ (автосумма).



X_i	n_i	Pi	Si
0	6	0,250	6
1	7	0,292	13
2	3	0,125	16
3	5	0,208	21
4	3	0,125	24
Σ	24	1	

Карман	Частота	Интегральный %
0	6	25,00%
1	7	54,17%
2	3	66,67%
3	5	87,50%
4	3	100,00%
Еще	0	100,00%

Рисунок 1.3 – Результаты применения инструмента Гистограмма

В ячейку F4 введём формулу «=E3/$E$8», за маркер заполнения (крест в правом нижнем углу ячейки) с помощью мыши скопируем до F8 и выберем кнопку автосумма, в результате мы получим частоты вариантов и их сумму (1).

В ячейку G4 введём частоту варианта 0 – цифру 6 (или ссылку на ячейку её содержащую – E4), в ячейку G5 введём формулу «=G4+E5» и скопируем её до ячейки G8, в результате получим накопленные частоты. Таким образом, мы получили дискретный вариационный ряд.

Частоты необходимо округлить, но так, чтобы их сумма равнялась 1,000. Для этого выделим левой кнопкой мыши диапазон частот (F4:F8), щёлкнув по правой кнопке, откроем контекстное меню и выполним команду Формат ячеек – Числовой – Число знаков 3 – ОК.

Преобразовав обозначения, получим дискретный вариационный ряд, изображённый в таблице 1.2.

Таблица 1.2 – Дискретный вариационный ряд

Балл, x_i	Число студентов, n_i	Относительная частота, $\hat{p}_i$	Сумма накопленных частот, S_i
0	6	0,250	6
1	7	0,292	13
2	3	0,125	16
3	5	0,208	21
4	3	0,125	24
Итого	24	1,000	-

Вариационные ряды изображают графически с помощью полигона и гистограммы.

Полигон частот – это ломаная, отрезки которой соединяют точки

$$(x_1;n_1),(x_2;n_2),...(x_k;n_k).$$

Полигон относительных частот – это ломаная, отрезки которой соединяют точки:

$$(x_1;\frac{n_1}{n}),(x_2;\frac{n_2}{n}),...(x_k;\frac{n_k}{n}).$$

Изобразим ряд графически. Для этого построим полигон частот с помощью мастера диаграмм - .

Выделим диапазон D3:E8, выполним команду *Вставка – Диаграммы - Точечная с прямыми отрезками и маркерами* (рис.1.4).

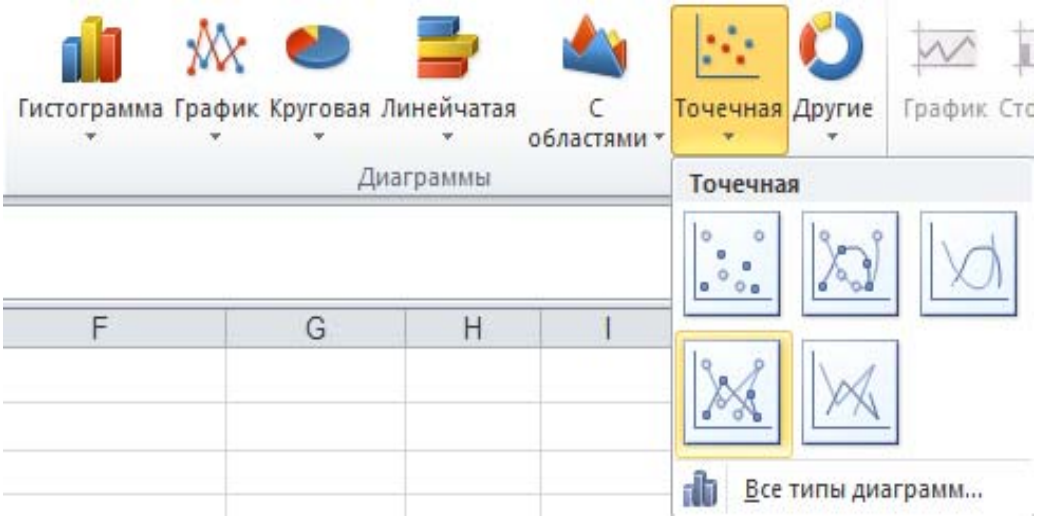


Рисунок 1.4 – Построение точечной диаграммы

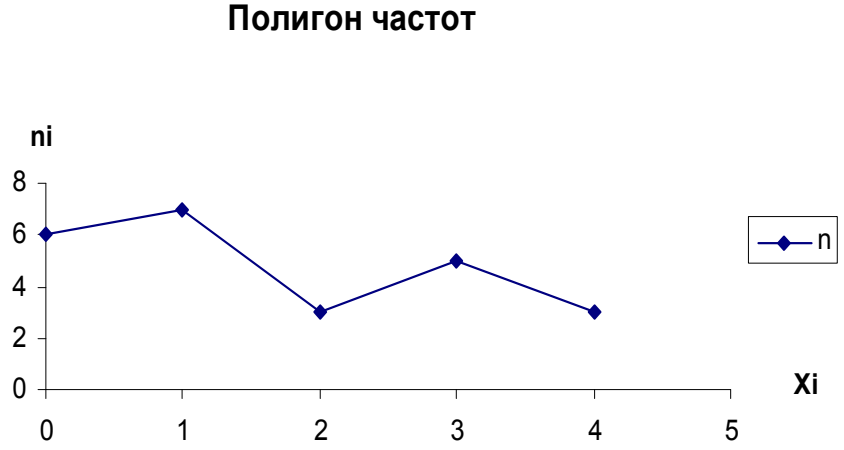


Рисунок 1.5 - Полигон частот

Получившийся график полигона частот желательно преобразовать, открыв с помощью контекстного меню, Формат области построения (уберите границы и заливку диаграммы). В результате получится полигон частот, изображённый на рисунке 1.5.

Постройте самостоятельно полигон относительных частот и кумуляту (рис. 1.6, 1.7). Для этого можно выделить в Excel всю таблицу (диапазон D3:G8, см. рис.1.3) и уже на графике удалить лишнее.

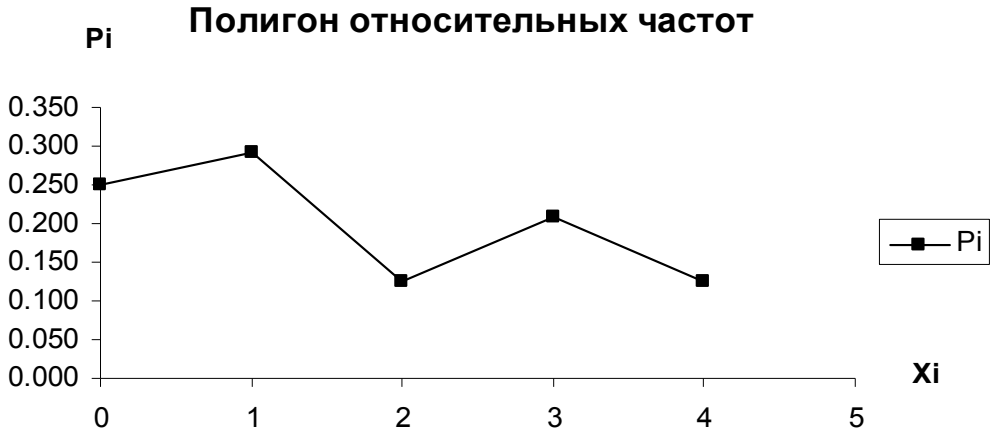


Рисунок 1.6 - Полигон относительных частот

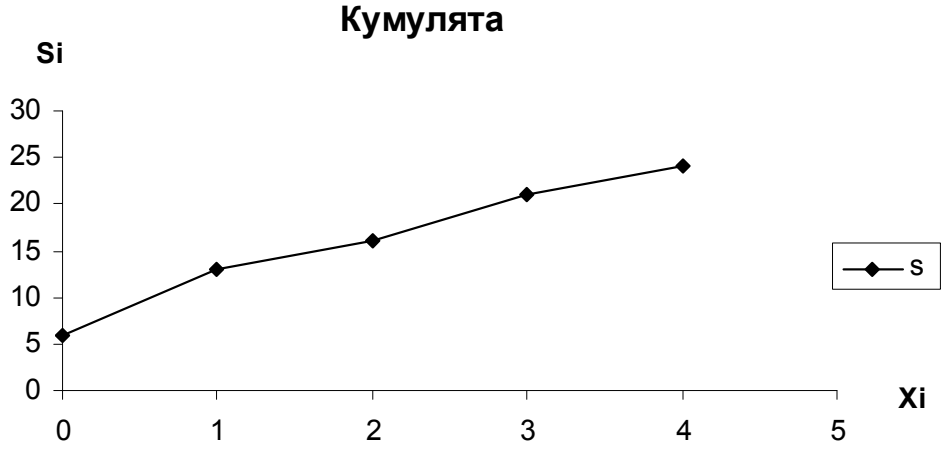


Рисунок 1.7 - Кумулята

Построение дискретного вариационного ряда нецелесообразно, если число значений признака велико. В этом случае следует построить интервальный вариационный ряд. Для построения такого ряда промежуток изменения признака разбивается на ряд отдельных интервалов и подсчитывается количество значений величины в каждом из них.

Будем считать, что отдельные (частичные) интервалы имеют одну и ту же длину. Число интервалов (k), в случае нормально распределённой совокупности, можно определить по формуле Стерджесса:

$$k = 1 + 3{,}322\lg(n)$$

или приближённо: $k \in [6,12]$.

Пример 1.2 Пусть дан ряд распределения хозяйств по количеству рабочих на 100 га с/х угодий (n=60):

12	6	8	6	10	11	7	10	12	8	7	7	6	7	8	6	11	9	11
9	10	11	9	10	7	8	8	8	11	9	8	7	5	9	7	7	14	11
9	8	7	4	7	5	5	10	7	7	5	8	10	10	15	10	10	13	12
11	15	6																

Построить интервальный вариационный ряд.

Для определения числа групп подставим значение n=60 в формулу Стерджесса:

$$k = 1 + 3{,}322\lg 60 \approx 6{,}907; \quad k = 7.$$

Длина частичного интервала определяется по формуле:

$$h = \frac{W}{k} = \frac{x_{\max} - x_{\min}}{k} = \frac{15-4}{7} \approx 1{,}6.$$

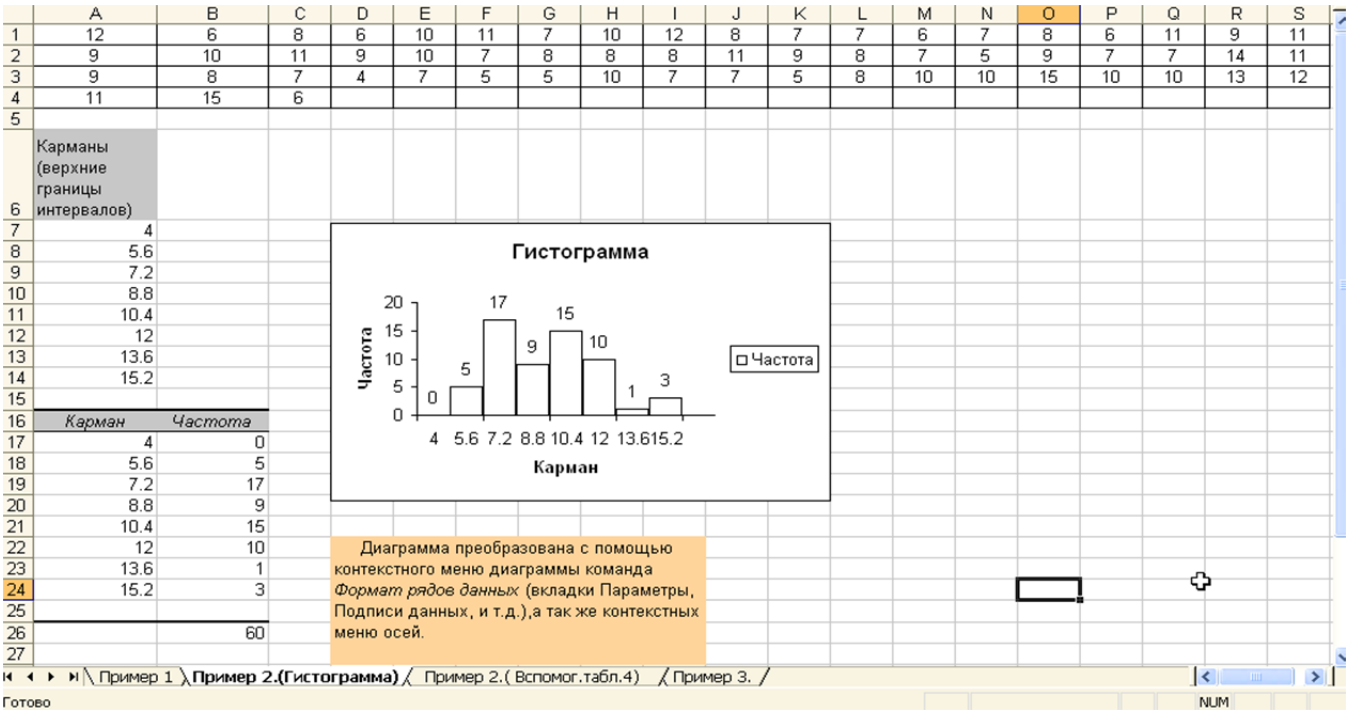
Построим интервальный вариационный ряд, для этого в качестве начального значения используем x_{min}.

Разобьем интервал вариации признака X на k=7 частичных интервалов (табл. 3) с шагом h=1,6 (4,0; 5,6; 7,2; 8,8; 10,4; 12,0; 13,6; 15,2).

Таблица 1.3 – Группировка хозяйств по численности работников на 100 га сельхозугодий

Группы хозяйств по численности работников на 100га с/х угодий	Число хозяйств в группе (n_i)	Накопленное число хозяйств (S_i)	Относительная частота ($\hat{P}_i$)
4,00 - 5,60	5	5	5/60
5,61 - 7,20	17	22	17/60
7,21 - 8,80	9	31	9/60
8,81 - 10,40	15	46	15/60
10,41 - 12,00	10	56	10/60
12,01 - 13,60	1	57	1/60
13,61 -15,20	3	60	3/60
Итого:	60	-	1,000

Далее подсчитаем количество рабочих на 100 га сельскохозяйственных угодий в каждом интервале с использованием инструмента *Гистограмма* пакета анализа (рис. 1.8, 1.9).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	12	6	8	6	10	11	7	10	12	8	7	7	6	7	8	6	11	9	11
2	9	10	11	9	10	7	8	8	8	11	9	8	7	5	9	7	7	14	11
3	9	8	7	4	7	5	5	10	7	7	5	8	10	10	15	10	10	13	12
4	11	15	6																
5																			
6	Карманы (верхние границы интервалов)																		
7	4																		
8	5.6																		
9	7.2																		
10	8.8																		
11	10.4																		
12	12																		
13	13.6																		
14	15.2																		
15																			
16	*Карман*	*Частота*																	
17	4	0																	
18	5.6	5																	
19	7.2	17																	
20	8.8	9																	
21	10.4	15																	
22	12	10																	
23	13.6	1																	
24	15.2	3																	
25																			
26		60																	
27																			

Рисунок 1.8 – Подсчёт частот и построение гистограммы для интервального вариационного ряда

Гистограммой частот называется фигура, состоящая из прямоугольников с основанием h и высотами n_i. Для гистограммы относительных частот в качестве высоты рассматривают n_i/n. Гистограмма относительных частот является аналогом дифференциальной функции случайной величины.

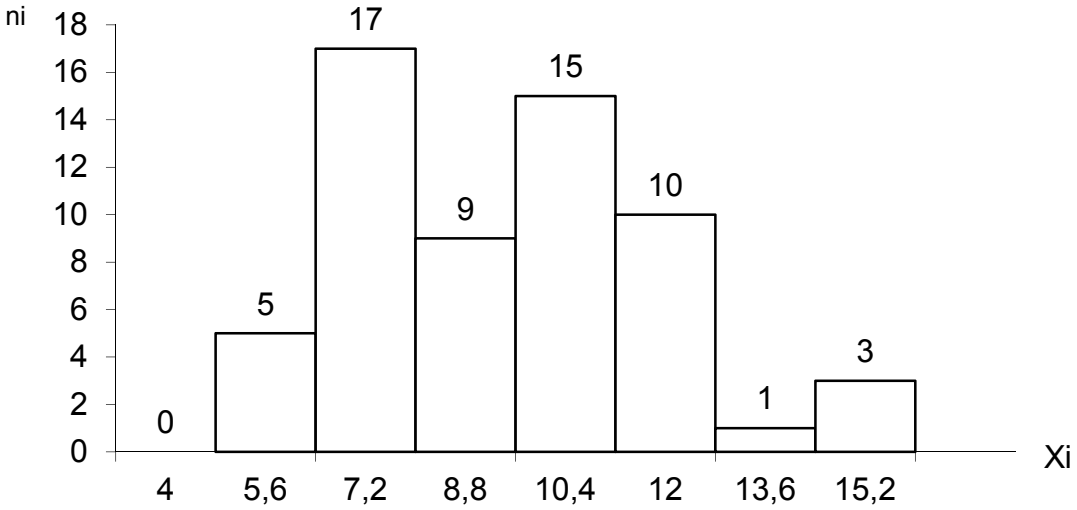


Рисунок 1.9 – Гистограмма частот

Построим гистограмму частот для примера 1.2. (рис.1.9). Гистограмма построена в *Excel* (Пакет анализа – инструмент Гистограмма) подпись под прямоугольником означает верхнюю границу интервала, над прямоугольником – соответствующую частоту.

График гистограммы относительных частот можно получить из графика рисунка 1.8 сжатием в 60 раз вдоль оси ординат.

Для примера 1.2 можно подобно примеру 1.1, построить полигон частот и полигон относительных частот.

1.2 Парная регрессия и корреляция

Пример 1.3 По имеющимся данным требуется построить график зависимости между переменными, по которому необходимо подобрать модель уравнения регрессии. Используя следующие функции:

- линейную;
- степенную;
- экспоненциальную;
- показательную.

Замечание. Важнейшим методом анализа данных является визуализация (представление данных в виде таблиц, диаграмм, кросс-таблиц, кросс-диаграмм, графиков).

Таблица 1.4 - Фондообеспеченность и производство продукции

№	Фондообеспеченность на 1 га сельхозугодий, тыс. руб.; (x)	Стоимость валовой продукции на 1 га сельхозугодий, тыс. руб.; (y)
1	38,4	62,3
2	24,2	30,1
3	29,2	47,3
4	23,0	29,9
5	18,2	37,2
6	33,2	46,1
7	14,1	22,3
8	26,2	43,0
9	20,1	34,1
10	35,0	49,2
11	31,7	41,4
12	24,4	37,4
13	18,9	28,2
14	27,1	37,0
15	17,0	26,1

Рассмотрим применение диаграммы рассеяния. Выделим в *Excel* диапазон переменных *x* и *y* (таблица 1.4), выполним команду1: Вставка - Точечная – Точечная с маркерами. В результате получим рисунок 1.10.

Важность графического представления данных заключается в возможности увидеть возможные ошибки, допущенные при вводе данных (артефакты – объекты созданные человеком) или неоднородные значения признаков - выбросы – явно не принадлежащие изучаемой совокупности.

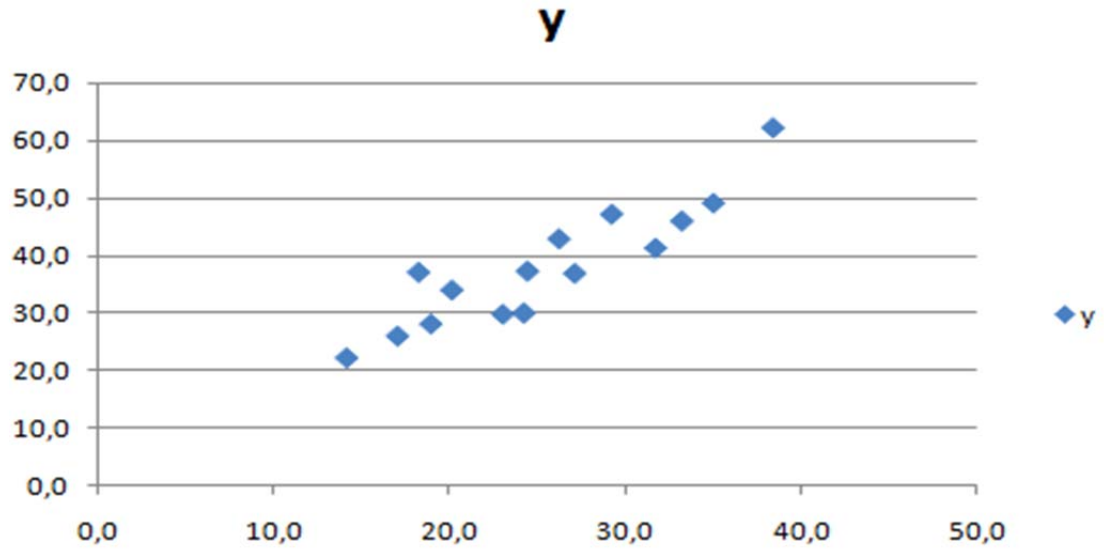


Рисунок 1.10 - Диаграмма рассеяния

Например, при вводе исходных данных мы вместо 62,3 ввели 623. Диаграмма рассеяния примет вид, отраженный на рисунке 1.11 из которого видно, что есть наблюдение, отличающееся от других данных.

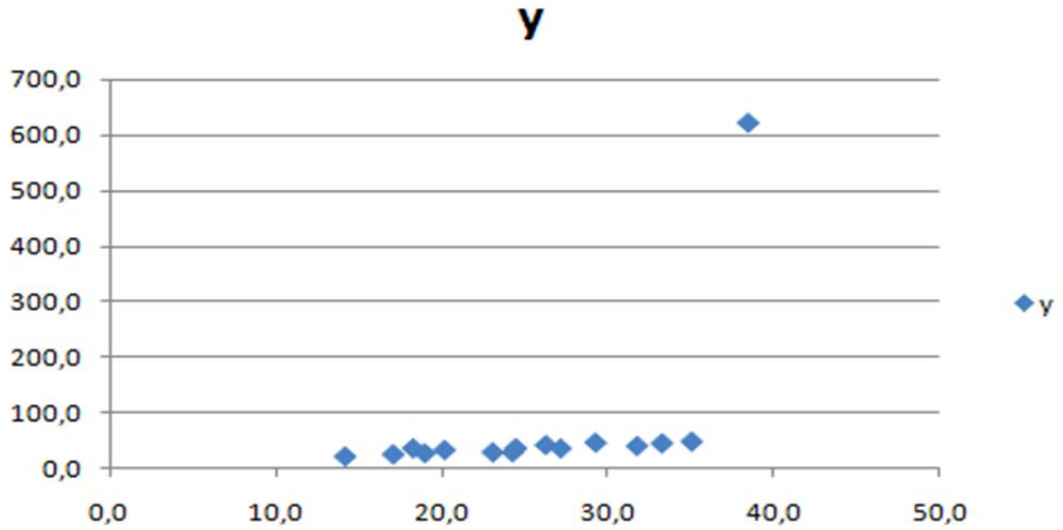


Рисунок 1.11 – Диаграмма рассеяния с артефактом (или выбросом)

[1]Словосочетание *Выполнить команду* (*выбрать команду*) означает, что необходимо установить на вкладку Ленты (*Excel* 2007) указатель и щёлкнуть левой кнопкой мыши

Важным методом анализа данных в *Excel* являются диаграммы. Выделим на рисунке 1.12 щелчком левой клавиши мыши маркеры наблюдений; с помощью правой клавиши откроем контекстное меню (рисунок 1.12) и выберем одну из перечисленных линий трендов (рисунок 1.13):

- Линейная;
- Логарифмическая;
- Полиномиальная;
- Степенная;
- Экспоненциальная;
- Линейная фильтрация (Скользящая средняя).

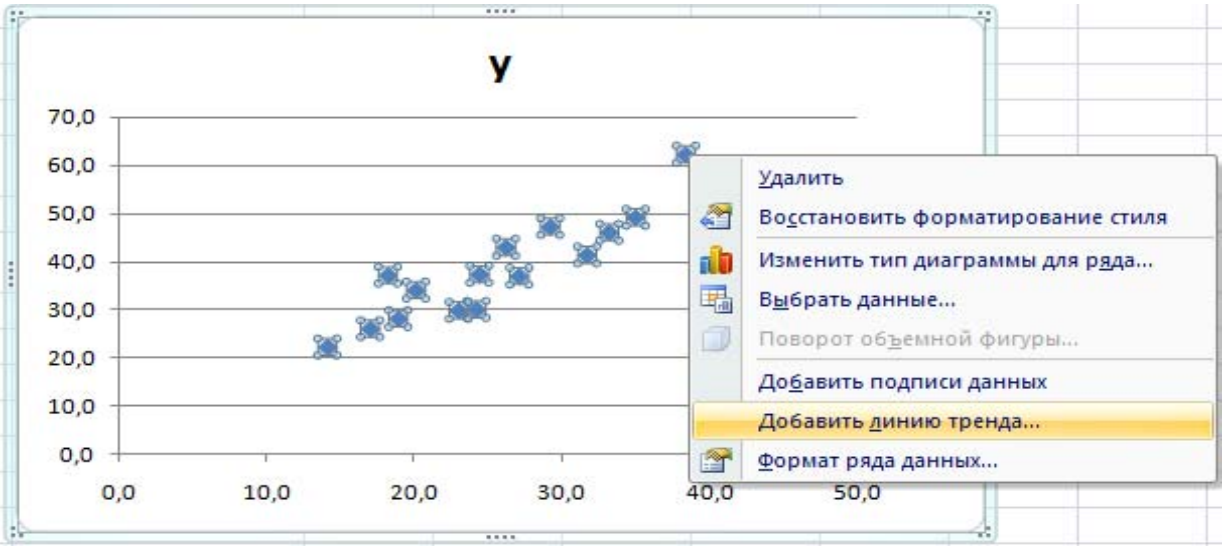


Рисунок 1.12 – Контекстное меню выделенных точек наблюдений

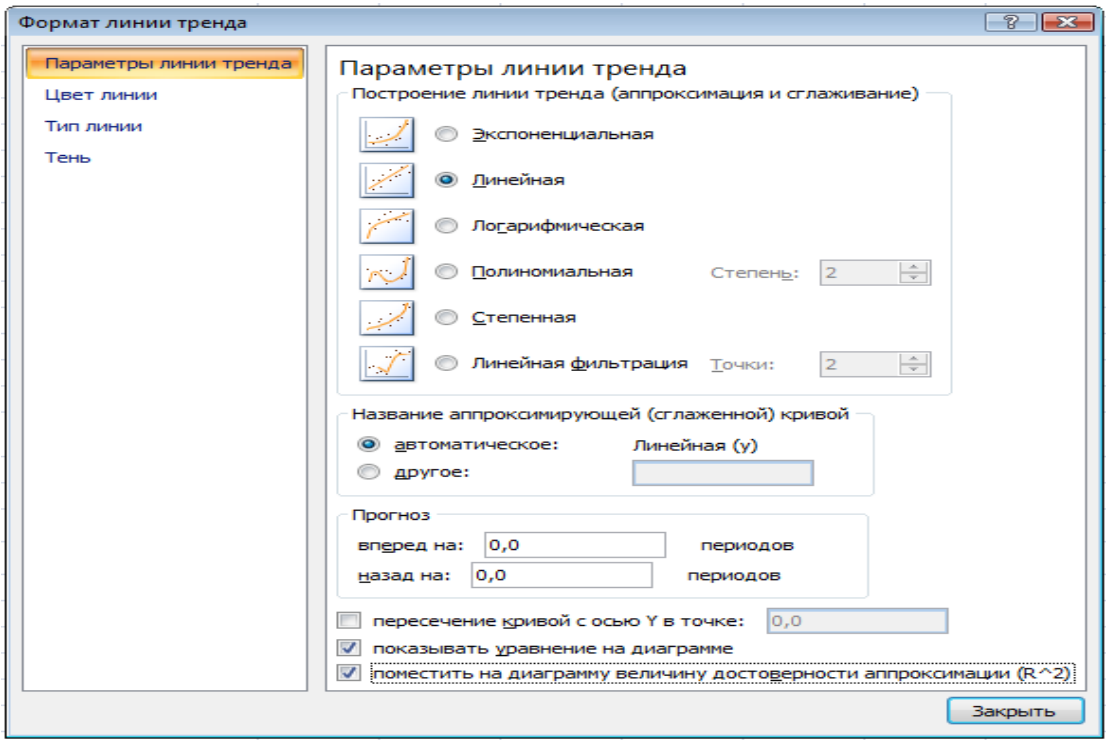


Рисунок 1.13 – Диалоговое окно выбора линии тренда

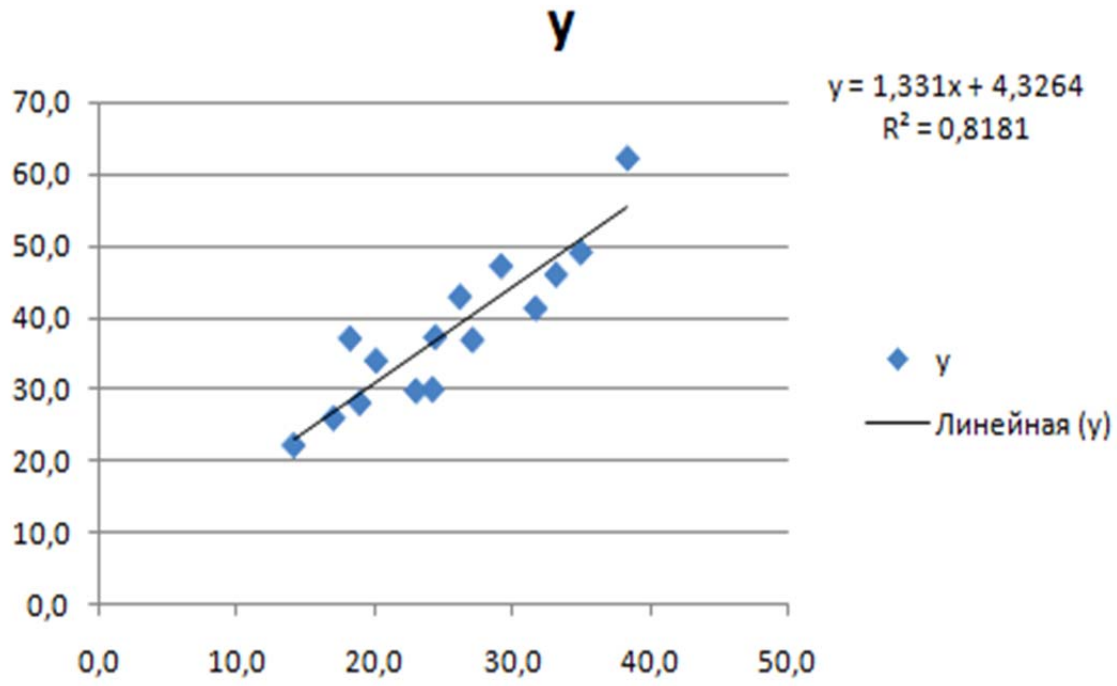


Рисунок 1.14 – График линейного уравнения

После выбора одного из трендов, например, линейного - выберем и заполним вкладку Параметры диалогового окна (рисунок 1.13). Можно выбрать название (назвать тренд самостоятельно) или оставить автоматически предлагаемое Excel; для прогноза согласно выбранной линии тренда на 5 лет вперёд выберем соответствующее значение в диалоговом окне; для отображения на диаграмме уравнения тренда и коэффициента детерминации отметим соответствующие элементы. Далее выберем ОК.

Задания

1. Построить вариационный ряд по заданию преподавателя.

2. Построить график точечной зависимости $y=2x+3$ для x от 1 до 50.

Наложить на переменную y «шум» (например, случайную величину, подчиняющуюся равномерному закону распределения), рассмотреть несколько вариантов с учетом уравнения тренда и коэффициента детерминации. Сделать выводы.

Вопросы для самоконтроля

1. Определение и виды вариационных рядов.
2. Способы графического изображения вариационных рядов.
3. Как определяется число групп и величина интервала при построении интервального вариационного ряда.
4. Основные этапы корреляционно-регрессионного анализа.
5. Расчет показателей тесноты связи между двумя признаками.

Практическое занятие № 2

2.1 Множественный корреляционно-регрессионный анализ в Excel

Цель работы: ознакомиться с возможностями корреляцинно-регрессионного анализа данных с использованием *Excel*. Провести анализ данных сельскохозяйственных организаций.

Теоретические сведения

В экономических исследованиях результативный признак *Y* формируется, как правило, под влиянием нескольких факторных признаков $X_1, X_2, \ldots, X_p$.

Уравнение множественной регрессии имеет вид $y = f(x_1, x_{21}, \ldots, x_p)$.

При построении уравнения множественной регрессии обычно используются следующие функции:

- линейная: $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon$;
- степенная: $y = b_0 \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \ldots \cdot x_p^{b_p} \cdot \varepsilon$;
- экспонента: $y = e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon}$.

Часто применяются и другие виды функций, например

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_5x_1x_2 + \varepsilon.$$

Если уравнение нелинейное, то оно вначале приводится к линейному. Параметры линейного уравнения множественной регрессии находятся методом наименьших квадратов, для чего строится и решается следующая система нормальных уравнений:

$$\begin{cases} \Sigma y = nb_o + b_1\Sigma x_1 + b_2\Sigma x_2 + \ldots + b_p\Sigma x_p, \\ \Sigma yx_1 = b_o\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + \ldots + b_p\Sigma x_1x_p, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Sigma yx_p = b_o\Sigma x_p + b_1\Sigma x_1x_p + b_2\Sigma x_2x_p + \ldots + b_p\Sigma x_p^2. \end{cases}$$

Множественный коэффициент регрессии b_i показывает, на сколько единиц изменяется в среднем результативный признак *Y*, если *i* –ый признак *X* увеличить на единицу, при условии, что все другие факторы в линейной модели закреплены на постоянном, обычно среднем, уровне.

Уравнение множественной регрессии может быть построено в стандартизованном масштабе, когда единицей измерения признаков принимается их среднее квадратическое отклонение:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots \beta_p t_{x_p},$$

где,

$$t_y = \frac{y - \bar{y}}{\sigma_y}, t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$$

β_i- стандартизованные коэффициенты регрессии,

σ – среднее квадратическое отклонение.

Для оценки тесноты связи между признаками применяются парные, частные и множественные коэффициенты (индексы) корреляции и детерминации. Если изучение зависимости проводится по выборочным данным, то оценивается значимость коэффициентов регрессии, корреляции и всего уравнения множественной регрессии в целом.

Пример 2.1Рассмотрим пример построения линейного уравнения регрессии между валовой продукцией на 1 га с/х угодий (Y), энергообеспеченностью на 100 га с/х угодий, л.с. (X_1), фондообеспеченностью на 100 га с/х угодий, тыс. руб. (X_2), затратами на производство продукции на 1 га с/х угодий, тыс. руб. (X_3) по данным приложения. В книге *Excel* **Группировка** вставим новый лист «Регрессия» и введём исходные данные (рис. 2.1).

Microsoft Excel - Группировка

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

B3 =ФормаlE2/ФормаlD2*100

	A	B	C	D	E	F	G	I	K	M
1	Валовая продукция на 1 га с/х угодий, тыс.руб.	Энергообеспеченность на 100 га с/х угодий, л.с.	Фондообеспеченность на 100 га с/х угодий, тыс. руб.	Затраты на производство продукции на 1 га с/х угодий, тыс.руб.						
2	Y	X1	X2	X3						
3	9,923057913	284,2326386	611,0140868	9,166817695						
4	4,630444412	217,0087111	950,5684335	4,571091097						
5	5,791561278	178,2708669	801,5750747	4,37525868						
6	6,096240136	254,4097168	408,0457992	4,89494043						
7	8,881634486	295,4737315	837,6919623	7,449753031						
8	7,640543881	296,9262464	1438,29419	6,889163576						
9	11,70316678	391,9634061	879,408867	8,766924701			Y	X1	X2	X3
10	6,938591118	349,9387443	1782,618683	7,162940276						
11	8,855745038	345,1635449	1109,854627	7,819681297		Среднее	8,511279	289,7707	1180,536	7,457692
12	9,759330618	264,5271386	1036,177474	6,791258395		Стандартная ошибка	0,412123	9,935091	64,48195	0,324782
13	9,255029057	240,9924005	1662,606169	7,627626285		Медиана	7,640544	293,2475	1071,671	7,071205
14	4,905933746	205,7007645	752,28249	5,467710229		Мода	#Н/Д	#Н/Д	#Н/Д	#Н/Д
15	5,994445679	302,0662075	1723,128194	5,839591202		Стандартное отклонение	3,271127	78,85734	511,8096	2,577877
16	13,39236437	337,5753516	1333,784327	9,962089752		Дисперсия выборки	10,70027	6218,481	261949,1	6,64545
17	8,290749601	315,7575758	1028,516746	7,594577352		Эксцесс	0,410778	-0,82352	-0,264	0,268729
18	9,903981043	316,7772512	842,985782	7,922654028		Асимметричность	1,029776	0,086946	0,562511	0,891476
19	12,78053239	307,734491	1347,389466	9,036558894		Интервал	12,94959	310,5475	2126,823	10,72132
20	6,486149584	190,1939058	901,0156971	5,3388735		Минимум	4,043331	148,5877	341,5221	3,580128
21	6,413850301	255,4945055	456,9044163	5,295044578		Максимум	16,99292	459,1351	2468,345	14,30145
22	4,878780367	160,7585523	341,5220625	3,715418939		Сумма	536,2106	18255,55	74373,77	469,8346
23	4,470726496	390,4059829	568,1837607	3,580128205		Счет	63	63	63	63
60	7,855538611	265,31263	1013,1568	5,494939692						
61	4,72593801	267,3191952	1127,351822	5,215878195						
62	6,916878225	203,5050364	510,5888134	8,265498321						
63	4,04333089	153,7193111	864,9322096	3,961982411						
64	5,388862218	236,4378301	1147,191551	5,745919347						
65	8,275821458	356,1182428	1681,450948	7,629372277						
66										

Регрессия | Форма | Лист1 | Сводная таблица | Группировочная таблица

Рисунок 2.1 - Показатели шестидесяти организаций Краснодарского края

Таблица 2.1 - Описательная статистика

	Y	X_1	X_2	X_3
Среднее	8,511279	289,7707	1180,536	7,457692
Стандартная ошибка	0,412123	9,935091	64,48195	0,324782
Медиана	7,640544	293,2475	1071,671	7,071205
Мода	#Н/Д	#Н/Д	#Н/Д	#Н/Д
Стандартное отклонение	3,271127	78,85734	511,8096	2,577877
Дисперсия выборки	10,70027	6218,481	261949,1	6,64545
Эксцесс	0,410778	-0,82352	-0,264	0,268729
Асимметричность	1,029776	0,086946	0,562511	0,891476
Интервал	12,94959	310,5475	2126,823	10,72132
Минимум	4,043331	148,5877	341,5221	3,580128
Максимум	16,99292	459,1351	2468,345	14,30145
Сумма	536,2106	18255,55	74373,77	469,8346
Счет	63	63	63	63
Наибольший(1)	16,99292	459,1351	2468,345	14,30145
Наименьший(1)	4,043331	148,5877	341,5221	3,580128

Для расчета необходимых показателей Y, X_1, X_2, X_3 введём в ячейки диапазона A2:D2 соответствующие формулы. Выделим диапазон A2:D2 и с помощью **маркера заполнения** скопируем формулы до строки 64. В результате получим исходные данные для регрессионного анализа, изображенные на рис. 2.1.
Применим инструмент **Описательная статистика**. Исследуем корреляцию факторов, используя инструмент Пакета анализа **Корреляция**. Из полученной корреляционной матрицы следует, что факторы X_1, X_2 коррелированны между собой, поэтому МНК применять нельзя.

Таблица 2.2 - Корреляционная матрица

	Y	X1	X2	X3
Y	1			
X1	0,594210765	1		
X2	0,58585918	0,565050507	1	
X3	0,937325544	0,549271689	0,630009324	1

Однако, так как коэффициент корреляции менее 0,65, то этот факт можно игнорировать.

В противном случае один из факторов необходимо отбросить, либо использовать другие методы оценки параметров регрессии.

Заполним диалоговое окно инструмента Пакета анализа **Регрессия** (рис.2.2). В результате выбора **ОК** получим таблицу итогов регрессионного анализа.

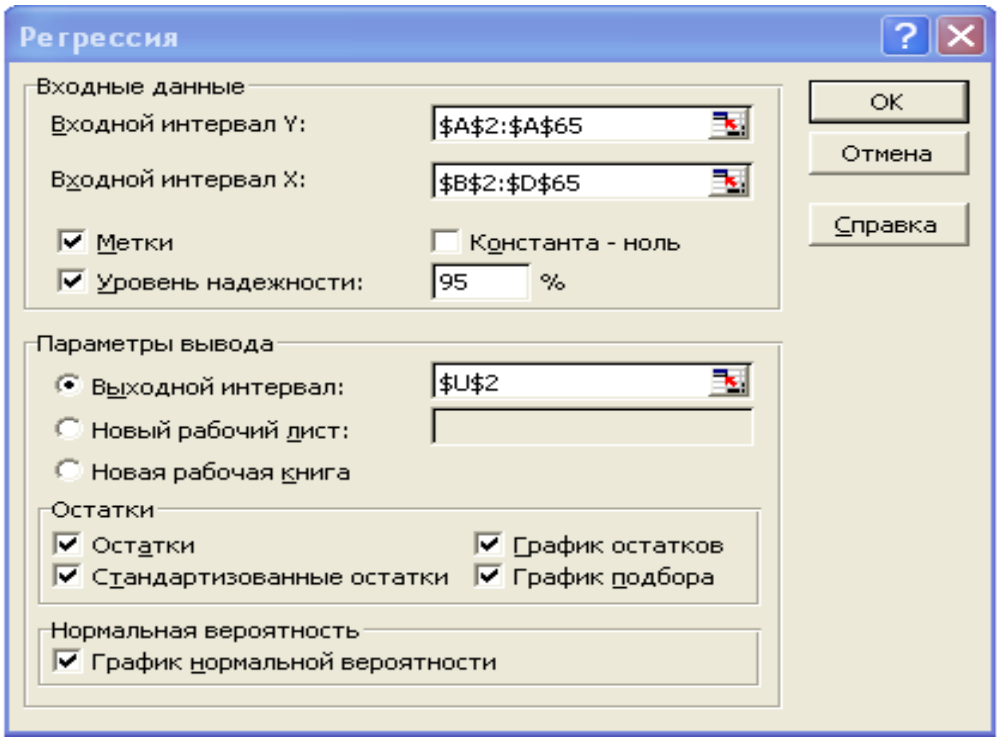


Рисунок 2.2 - Диалоговое окно Регрессия

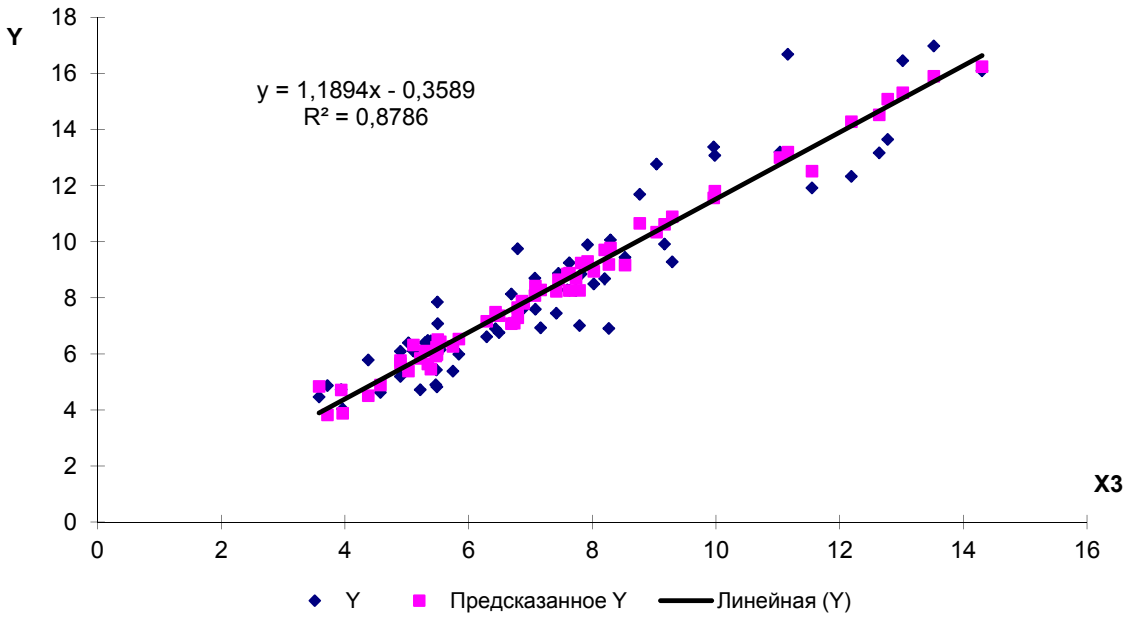


Рисунок 2.3 - График подбора

Таблица 2.3 – Регрессионная статистика

Регрессионная статистика	
Множественный R	0,94299
R-квадрат	0,88923
Нормированный R-квадрат	0,88360
Стандартная ошибка	1,11601
Наблюдения	63

ДА	df	SS	MS	F	Значимость F
Регрессия	3	589,932532	196,6441774	157,8845	3,7865E-28
Остаток	59	73,4841563	1,245494175		
Итого	62	663,416689			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	-1,1649987	0,56045999	-2,0786473	0,04200737	-2,286477719	-0,0435197
X1	0,00543403	0,002284902	2,3782336	0,02065306	0,000861948	0,01000611
X2	-0,0003535	0,000378819	-0,9330884	0,35457729	-0,001111488	0,00040454
X3	1,14230254	0,07425756	15,3829794	2,59445E-22	0,993713347	1,29089173

Дисперсионный анализ показывает, что уравнение является значимым при уровне значимости α=3,7865E-28. Множественный коэффициент корреляции R равен 0,942991, то есть полученное уравнение достаточно хорошо описывает изучаемую взаимосвязь между факторами. Коэффициент детерминации R^2 равен 0,889 - это означает, что 88,9% вариации результативного признака (Y) объясняется вариацией факторных переменных (X_1, X_2, X_3).

Согласно таблице 7, искомое уравнение регрессии имеет вид:

$$Y=0{,}0054X_1 - 0{,}00035X_2 + 1{,}1423X_3 - 1{,}165.$$

Причём доверительный интервал при уровне значимости 5%:

для коэффициента при X_1: (0,00086; 0,010);

для коэффициента при X_2: (-0,00111; 0,00040);

для коэффициента при X_3: (0,99371; 1,29089);

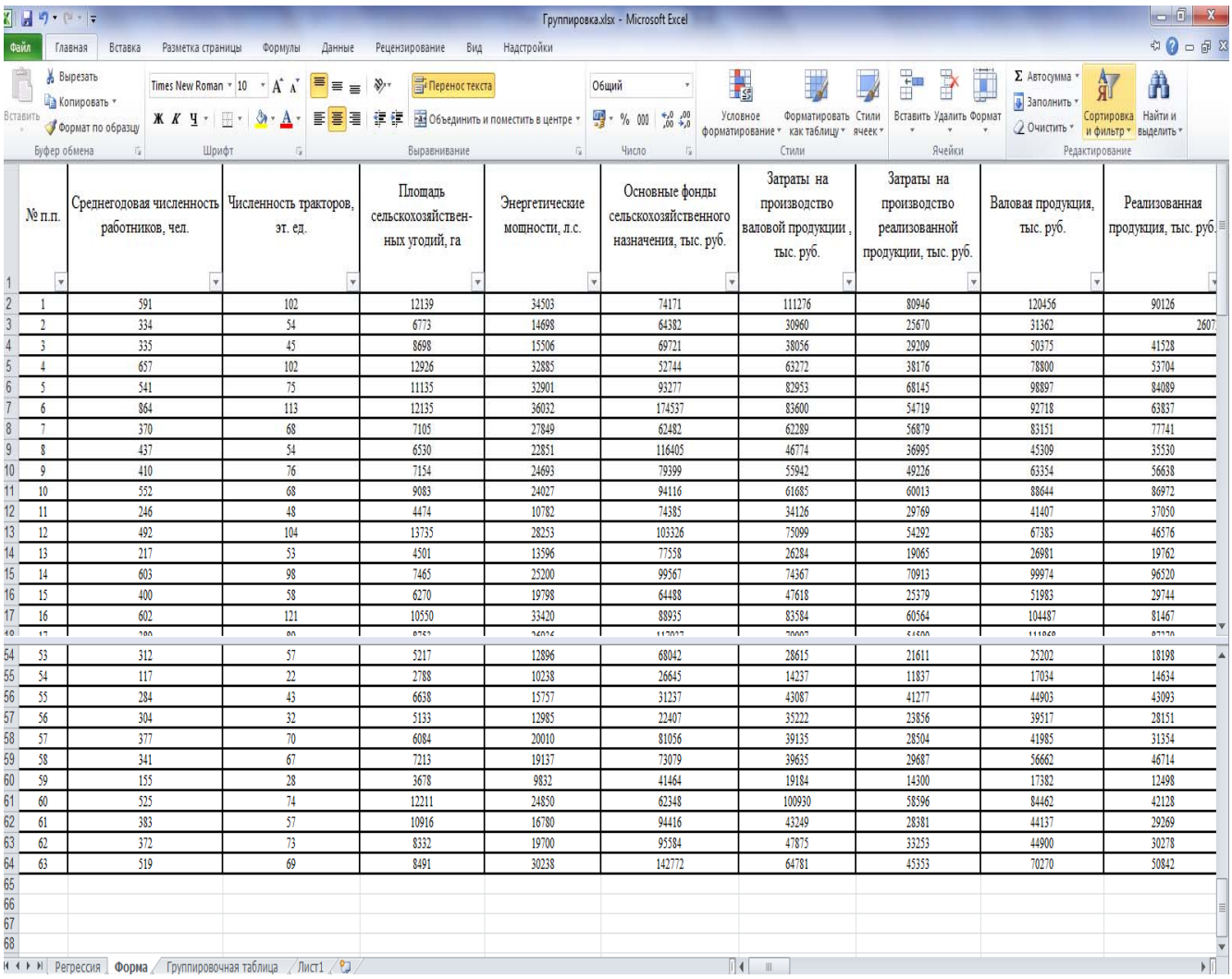
для свободного члена: (-2,28647; -0,04352).

Excel позволяет также анализировать парные регрессионные зависимости (остатки и многое другое). Так в нашем примере зависимость Y от X_3 имеет вид, представленный инструментом **Регрессия** на рис. 2.3. Коэффициенты уравнения регрессии показывают на сколько изменится Y при изменении одной из факторных переменных на единицу (при условии, что остальные переменные не изменяются).

2.2 Группировка в Excel (Сводные таблицы)[1]

Ценность данных определяется не их объёмом, а возможностью их преобразования в значимую – релевантную информацию (релевантный <англ. *relevant* - существенный, уместный, относящийся к делу). Согласно современным представлениям наиболее удобным способом хранения, организации и поиска информации являются базы данных (БД) (рисунок 2.4).

БД – это фактически любой набор данных: телефонный справочник, список книг в библиотеке, данные о курсе доллара по дням в разных банках, урожайность различных культур в сельском хозяйстве по годам, список лиц работающих в коллективе (год рождения, состав семьи, адрес, стаж работы, телефон, *e-mail*). Создание баз данных упрощает обработку данных и их анализ. Для этого в верхнюю строку необходимо ввести заголовки столбцов, а под ними без пропусков в каждую ячейку соответствующие данные. При большом их количестве, для редактирования или отбора по некоторому критерию, удобно воспользоваться командой *Данные-Фильтр*.



№ п.п.	Среднегодовая численность работников, чел.	Численность тракторов, эт. ед.	Площадь сельскохозяйственных угодий, га	Энергетические мощности, л.с.	Основные фонды сельскохозяйственного назначения, тыс. руб.	Затраты на производство валовой продукции , тыс. руб.	Затраты на производство реализованной продукции, тыс. руб.	Валовая продукция, тыс. руб.	Реализованная продукция, тыс. руб.
1	591	102	12139	34503	74171	111276	80946	120456	90126
2	334	54	6773	14698	64382	30960	25670	31362	2607
3	335	45	8698	15506	69721	38056	29209	50375	41528
4	657	102	12926	32885	52744	63272	38176	78800	53704
5	541	75	11135	32901	93277	82953	68145	98897	84089
6	864	113	12135	36032	174537	83600	54719	92718	63837
7	370	68	7105	27849	62482	62289	56879	83151	77741
8	437	54	6530	22851	116405	46774	36995	45309	35530
9	410	76	7154	24693	79399	55942	49226	63354	56638
10	552	68	9083	24027	94116	61685	60013	88644	86972
11	246	48	4474	10782	74385	34126	29769	41407	37050
12	492	104	13735	28253	103326	75099	54292	67383	46576
13	217	53	4501	13596	77558	26284	19065	26981	19762
14	603	98	7465	25200	99567	74367	70913	99974	96520
15	400	58	6270	19798	64488	47618	25379	51983	29744
16	602	121	10550	33420	88935	83584	60564	104487	81467
53	312	57	5217	12896	68042	28615	21611	25202	18198
54	117	22	2788	10238	26645	14237	11837	17034	14634
55	284	43	6638	15757	31237	43087	41277	44903	43093
56	304	32	5133	12985	22407	35222	23856	39517	28151
57	377	70	6084	20010	81056	39135	28504	41985	31354
58	341	67	7213	19137	73079	39635	29687	56662	46714
59	155	28	3678	9832	41464	19184	14300	17382	12498
60	525	74	12211	24850	62348	100930	58596	84462	42128
61	383	57	10916	16780	94416	43249	28381	44137	29269
62	372	73	8332	19700	95584	47875	33253	44900	30278
63	519	69	8491	30238	142772	64781	45353	70270	50842

Рисунок 2.4 – База данных

[1] Подготовлено при участии Е.В. Кремянской

Пример 2.2 Используя данные приложения 1, создадим Лист1 с исходными данными для группировки (таблица 2.4). Сохраним рабочую книгу под названием *Группировка*.

Важной задачей является задача разбиения на группы, удовлетворяющие определённым критериям. Она может быть решена в *Excel* различными способами: команда *Данные - Фильтр; команда Данные – Группировать* и т.д. Важным средством *Excel* при решении задачи разбиения на группы являются Сводные таблицы (команда *Вставка – Сводная таблица*).

Таблица 2.4 – Расчетные показатели для группировки

A	B	C	D	E	F	G
Фондообеспеченность на 100 га с/х угодий,тыс.руб.	Энергообеспеченность на 100 га с/х угодий, л.с.	Энергетические мощности, л.с.	Основные фонды сельскохозяйственного назначения, тыс. руб.	Площадь сельскохозяйственных угодий, га	Валовая продукция, тыс. руб.	Среднегодовая численность работников, чел.

Рассмотрим процесс построения группировки по шагам.

Шаг 1. Выберем команду *Вставка – Сводная таблица – ОК*. Если курсор мыши находится в поле таблицы, то диапазон задается автоматически (рисунок 2.5).

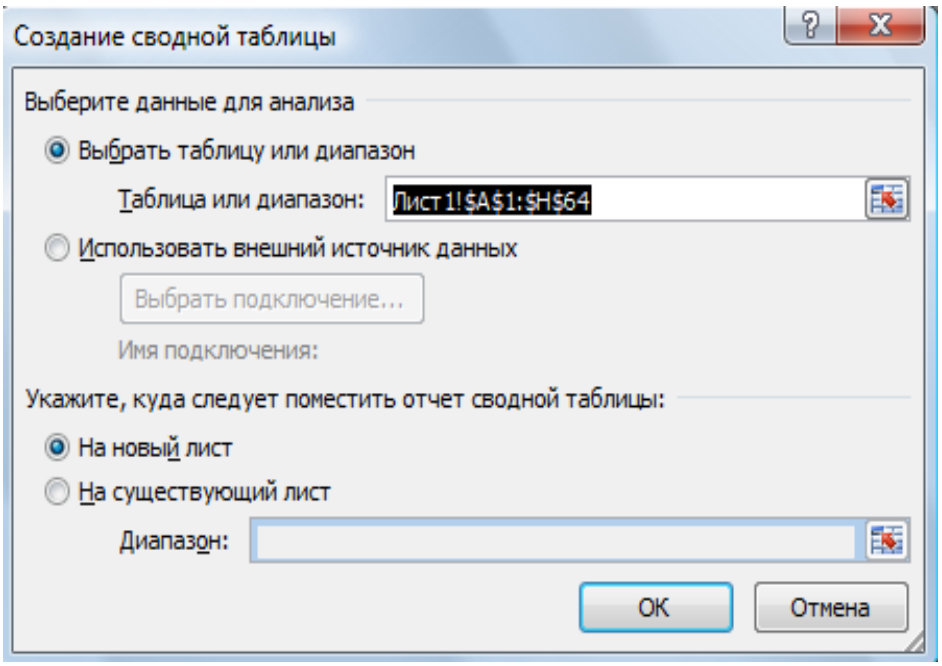


Рисунок 2.5 – Создание сводной таблицы

Шаг 2. Для группировки исходных данных по признаку энергообеспеченность на 100 га сельхозугодий предварительно заполним макет сводной таблицы (рисунки 2.6-2.7).

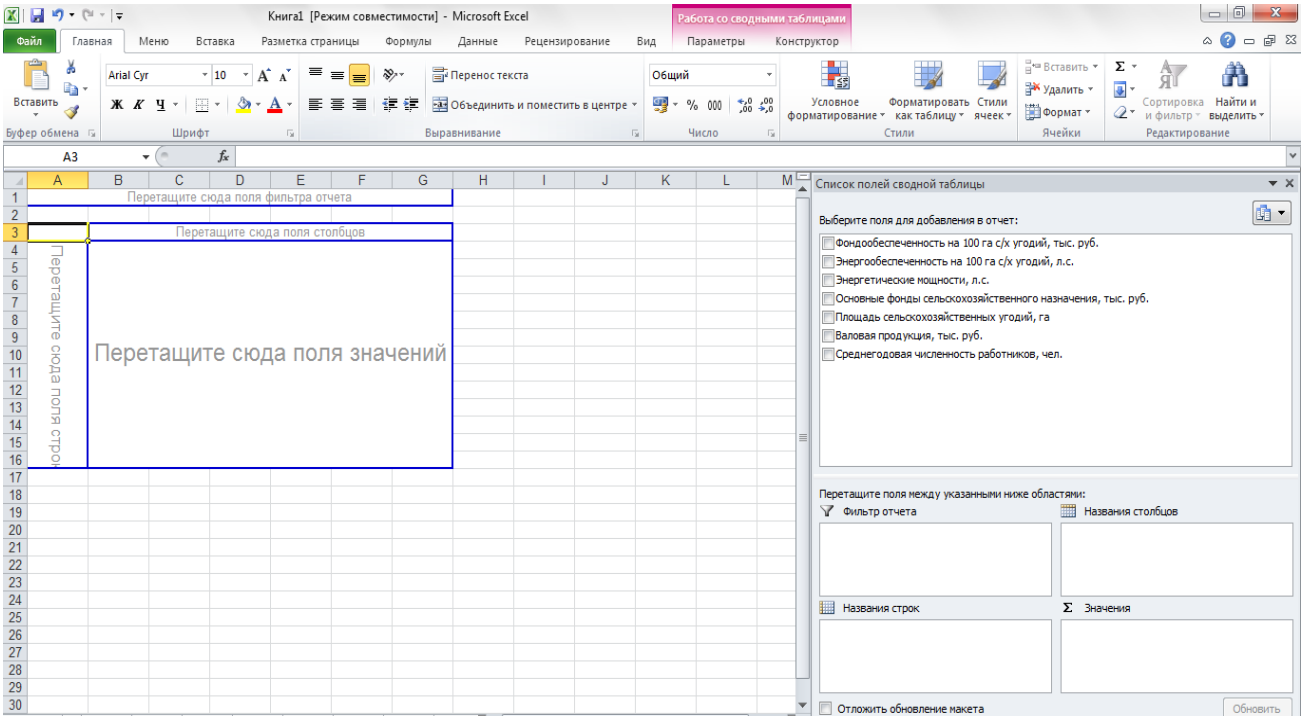


Рисунок 2.6 – Макет сводной таблицы

Шаг 3. Для группировки щёлкнем правой кнопкой мыши по полю «Энергообеспеченность» и выберем из контекстного меню команду «Группировать» с шагом 103,62 (для разбиения на три группы) (рисунок 2.8). В результате получим группировку по указанному признаку (рисунок 2.9).

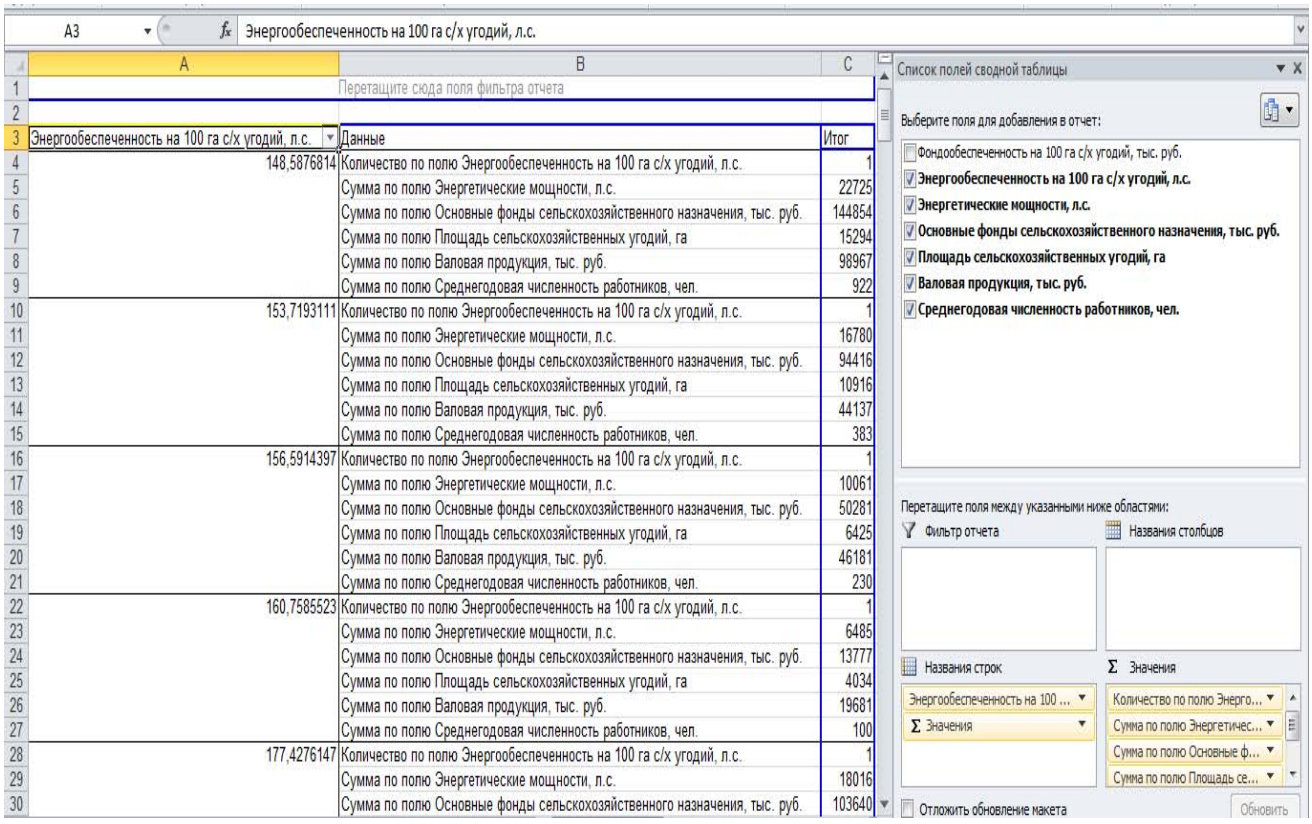


Рисунок 2.7 – Пример заполнения макета сводной таблицы

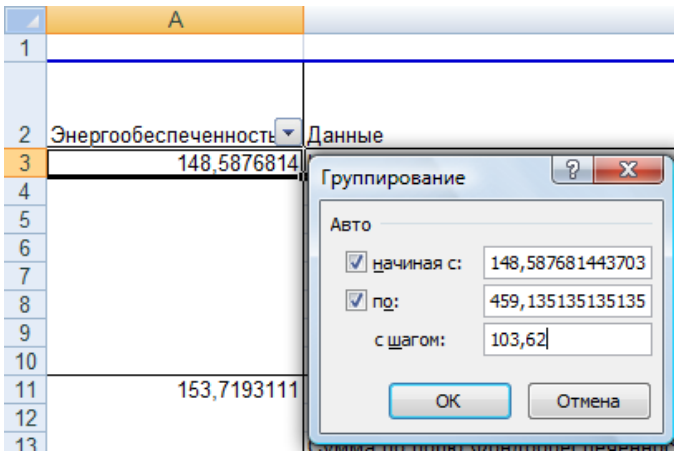


Рисунок 2.8 – Окно мастера «Группирование»

Энергообеспеченность на 100 га с/х угодий, л.с.	Данные	Итог
148,587681443703-252,207681443703	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	21
	Сумма по полю Энергетические мощности, л.с.	377617
	Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	1870700
	Сумма по полю Площадь сельскохозяйственных угодий, га	188897
	Сумма по полю Валовая продукция, тыс. руб.	1276704
	Сумма по полю Среднегодовая численность работников, чел.	8942
252,207681443703-355,827681443703	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	28
	Сумма по полю Энергетические мощности, л.с.	698922
	Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	2494482
	Сумма по полю Площадь сельскохозяйственных угодий, га	231549
	Сумма по полю Валовая продукция, тыс. руб.	1884415
	Сумма по полю Среднегодовая численность работников, чел.	12922
355,827681443703-459,447681443703	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	14
	Сумма по полю Энергетические мощности, л.с.	506458
	Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	2215299
	Сумма по полю Площадь сельскохозяйственных угодий, га	124356
	Сумма по полю Валовая продукция, тыс. руб.	1614693
	Сумма по полю Среднегодовая численность работников, чел.	8429
Итог Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.		63

Рисунок 2.9 – Сводная таблица после группировки по полю «Энергообеспеченность»

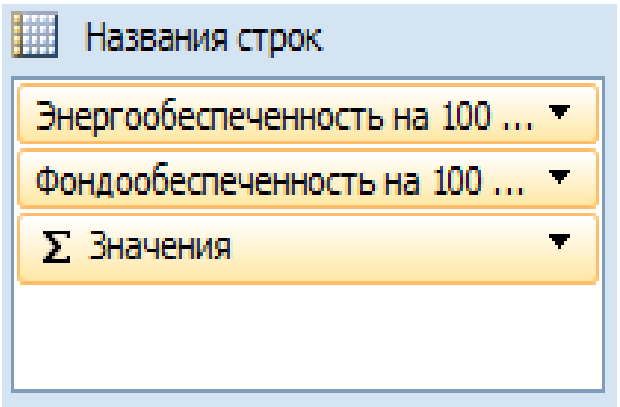


Рисунок 2.10 – Детализация списка полей сводной таблицы для вторичной группировки

Шаг 4. Для вторичной группировки перетащим в поле «Название строк» показатель «Фондообеспеченность», сделаем для него группировку с шагом 1063,45.

Энергообеспеченность на 100 га с/х у	Фондообеспеченность на 100 га с/х угодий, тыс. ру	Данные	Итог
⊟148,587681443703-252,207681443703	341,522062469013-1404,97206246901	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	18
		Сумма по полю Энергетические мощности, л.с.	331761
		Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	1503268
		Сумма по полю Площадь сельскохозяйственных угодий, га	167879
		Сумма по полю Валовая продукция, тыс. руб.	1094807
		Сумма по полю Среднегодовая численность работников, чел.	7748
	1404,97206246901-2468,42206246901	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	3
		Сумма по полю Энергетические мощности, л.с.	45856
		Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	367432
		Сумма по полю Площадь сельскохозяйственных угодий, га	21018
		Сумма по полю Валовая продукция, тыс. руб.	181897
		Сумма по полю Среднегодовая численность работников, чел.	1194
148,587681443703-252,207681443703 Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.			21
148,587681443703-252,207681443703 Сумма по полю Энергетические мощности, л.с.			377617
148,587681443703-252,207681443703 Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.			1870700
148,587681443703-252,207681443703 Сумма по полю Площадь сельскохозяйственных угодий, га			188897
148,587681443703-252,207681443703 Сумма по полю Валовая продукция, тыс. руб.			1276704
148,587681443703-252,207681443703 Сумма по полю Среднегодовая численность работников, чел.			8942
⊟252,207681443703-355,827681443703	341,522062469013-1404,97206246901	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	22
		Сумма по полю Энергетические мощности, л.с.	530889
		Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	1570015
		Сумма по полю Площадь сельскохозяйственных угодий, га	179094
		Сумма по полю Валовая продукция, тыс. руб.	1493595
		Сумма по полю Среднегодовая численность работников, чел.	9625
	1404,97206246901-2468,42206246901	Количество по полю Энергообеспеченность на 100 га с/х угодий, л.с.	6
		Сумма по полю Энергетические мощности, л.с.	168033
		Сумма по полю Основные фонды сельскохозяйственного назначения, тыс. руб.	924467

Рисунок 2.11 – Сводная таблица после вторичной группировки

Полученную сводную таблицу скопируем на новый лист («*Группировочная таблица*»), используя команду *Правка-Специальная вставка-Значения*, чтобы иметь возможность редактирования полей и итогов. Достроим сводную таблицу, введя показатели, которые необходимо рассчитать на основании итогов сводной таблицы. Введем соответствующие формулы. С помощью контекстных меню строк и столбцов скроем строки и столбцы с лишними расчетами, отформатируем. В результате получим комбинационную группировку влияния энергообеспеченности и фондообеспеченности сельхозугодий на эффективность производства.

D14 f_x

	A	B	C	D	E	F	G
1	Группы хозяйств по энергообеспеченности на 100 га с/х угодий, л.с	Подгруппы по фондообеспеченности на 100 га с/х угодий, тыс. руб.	Численность хозяйств в группе	Энергообеспеченность в среднем по группе, л.с.	Фондообеспеченность в среднем по группе, тыс. руб.	Валовая продукция на 100 га с/х угодий, тыс. руб.	Валовая продукция на 1 работника, тыс. руб.
2	148,6-252,2	341,5-1405,0	18	197,6	895,4	652,1	141,3
3		1405,0-2468,4	3	218,2	1748,2	865,4	152,3
4	Итого и в среднем по группе	X	21	199,9	990,3	675,9	142,8
5	252,2-355,8	341,5-1405,0	22	296,4	876,6	834,0	155,2
6		1405,0-2468,4	6	320,3	1762,4	745,1	118,5
7	Итого и в среднем по группе	X	28	301,8	1077,3	813,8	145,8
8	355,8-459,4	341,5-1405,0	5	380,6	1015,8	952,9	178,0
9		1405,0-2468,4	9	414,8	1999,2	1396,7	194,4
10	Итого и в среднем по группе	X	14	407,3	1781,4	1298,4	191,6
11	Всего и в среднем	X	63	290,6	1207,9	876,6	157,7

Рисунок 2.12 – Комбинационная группировка

Вывод: Расчеты показали, что с ростом энергообеспеченности эффективность сельскохозяйственного производства повышается. Так, если в первой группе хозяйств со средней энергообеспеченностью 199,9 л. с. в исследуемом году в расчете на 100 га сельскохозяйственных угодий было получено 675,9 тыс. руб. валовой продукции, а в расчете на одного работника – 142,8 тыс. руб., то в третьей группе хозяйств со средней энергообеспеченностью 407,3 л. с. стоимость валовой продукции в расчете на 100 га и одного работника превысила показатели первой группы соответственно на 622,5 и 48,8 тыс. руб.

Аналогичная связь прослеживается между эффективностью сельскохозяйственного производства и фондообеспеченностью: с ростом последней эффективность производства повышается. При этом в подгруппах с более высоким уровнем фондообеспеченности объем производства валовой продукции в расчете на 100 га сельскохозяйственных угодий и на одного работника, как правило, превышает средние по соответствующим группам показатели.

В среднем по всей исследуемой совокупности сельскохозяйственных организаций энергообеспеченность составила 290,6 л. с., а фондообеспеченность – 1207,9 тыс. руб. Средний объем производства валовой продукции в расчете на 100 га и одного работника достиг соответственно 876,6 и 157,7 тыс. руб.

Итак, повышение уровня энергообеспеченности и фондообеспеченности сельскохозяйственных организаций является важным фактором роста эффективности аграрного производства.

Вопросы для самоконтроля

1. Что характеризуют коэффициенты регрессии, корреляции, эластичности, детерминации.
2. Виды статистических группировок.
3. Выбор группировочного признака, определение числа групп, расчет величины интервала, нижних и верхних границ интервалов.
4. Способы построения вторичных группировок

Практическое занятие № 3

Пакет анализа

Цель работы: ознакомиться с возможностями пакета анализа данных в *Excel*

Теоретические сведения

Однофакторный дисперсионный анализ позволяет статистически обосновать существенность влияния факторного признака *A* на результативный *F*.

Замечание. Нет выделения моделей дисперсионного анализа (ДА) по виду факторов.

Однофакторный дисперсионный анализ используется для проверки гипотезы о сходстве средних значений двух или более выборок, принадлежащих одной и той же генеральной совокупности.

Этот метод распространяется также на тесты для двух средних (к которым относится, например, *t*-критерий). То есть если для разных уровней фактора *A* средние отличаются незначительно, то следует принять гипотезу о несущественном влиянии факторного признака *A* на результативный *F*.

Для активизации Пакета анализа необходимо выполнить команду кнопка *Office*-параметры *Excel*-Надстройки и выбрать Пакет анализа, после этого на ленте во вкладке Данные появится строка Анализ данных[1].

Пакет анализа позволяет решать в диалоговом режиме 19 задач, наиболее часто встречающихся в классической математической статистике (рис. 3.1) (хотя это не мешает подходить к ним творчески).

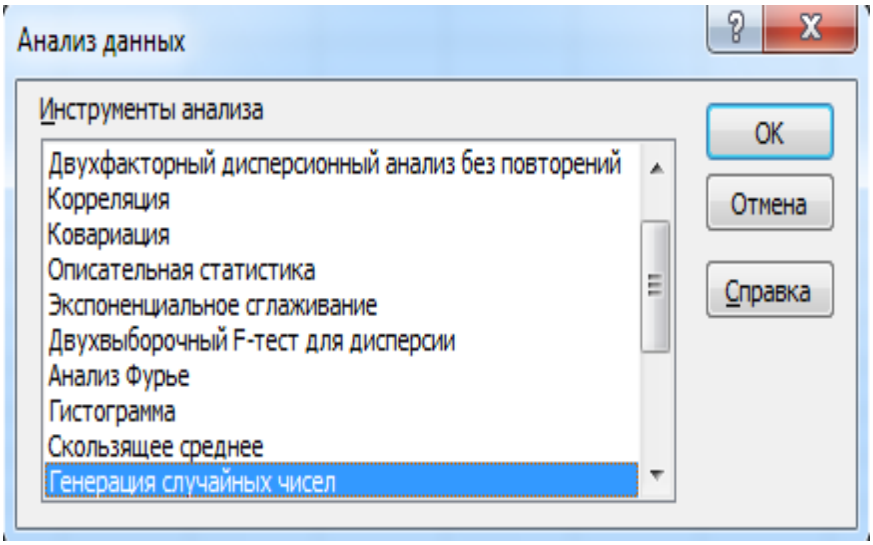


Рисунок 3.1– Инструменты анализа данных

[1] Аналогично устанавливается Поиск решения – инструмент для решения задач линейного программирования. Примеры*C:\ProgramFiles\MicrosoftOffice\Office12\SAMPLES\SOLVSAMP.xlsx*

Пример 3.1 Проверить статистическую существенность влияния катализатора *A* на химическую реакцию, результаты измерений при 5 уровнях фактора *A* приведены в таблице.

A1	*A2*	*A3*	*A4*	*A5*
3,2	2,6	2,9	3,7	3
3,1	3,1	2,6	3,4	3,4
3,1	2,7	3	3,2	3,2
2,8	2,9	3,1	3,3	3,5
3,3	2,7	3	3,5	2,9
3	2,8	2,8	3,3	3,1

Введём исходные данные в диапазон *A1:A7* листа Excel. Заполним параметры диалогового окна (рис 3.2).

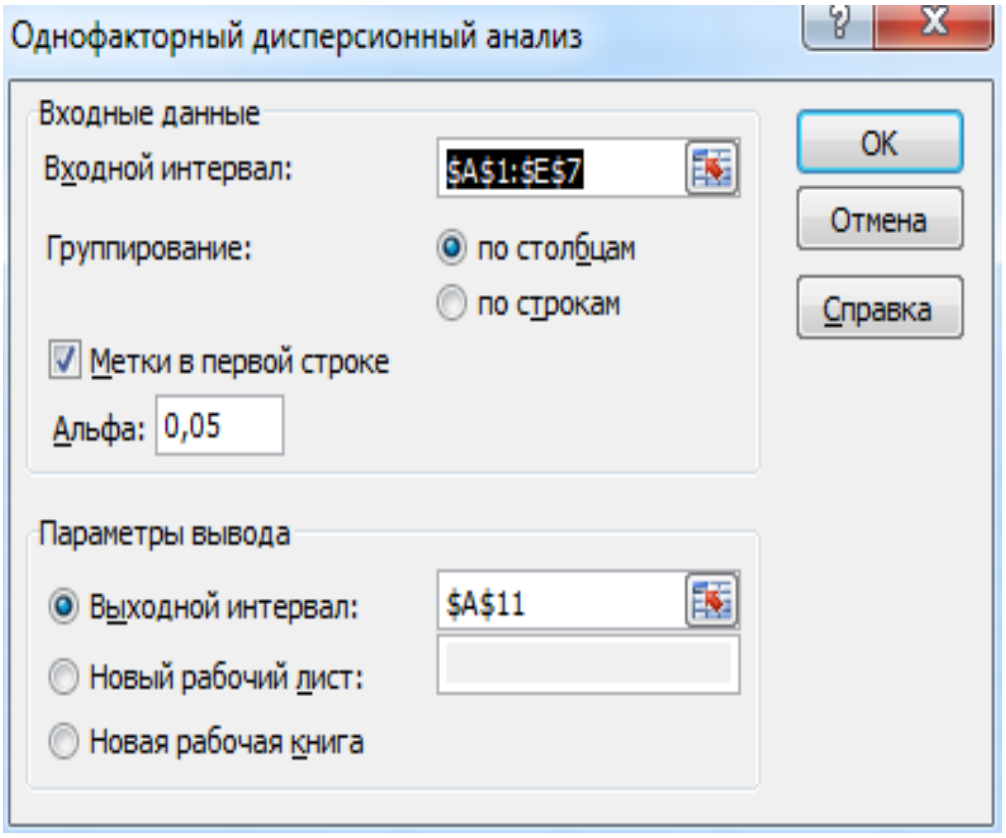


Рисунок 3.2 - Диалоговое окно однофакторного дисперсионного анализа

Группирование: по столбцам - именно такое расположение имеют уровни фактора *A*; отметим Метки в первой строке (там расположены уровни фактора *A*); в Выходном интервале достаточно отметить левую верхнюю ячейку выходного интервала $A$11.

После нажатия кнопки OK, получим таблицу однофакторного дисперсионного анализа (табл. 3.1).

Таблица 3.1 - Однофакторный дисперсионный анализ

Группы	Счет	Сумма	Среднее	Дисперсия
A1	6	18,5	3,083333333	0,029666667
A2	6	16,8	2,8	0,032
A3	6	17,4	2,9	0,032
A4	6	20,4	3,4	0,032
A5	6	19,1	3,183333333	0,053666667

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	1,342	4	0,335	9,35408	9,16424E-05	2,75871
Внутри групп	0,89667	25	0,035			
Итого	2,238667	29				

$F_{расч.}$=9,3540>$F_{кр.}$=2,7587, поэтому гипотезу о несущественном влиянии фактора *A* на результат следует отвергнуть.

Двухфакторный дисперсионный анализ с повторениями. Он представляет собой более сложный вариант однофакторного анализа, включающего более чем одну выборку для каждой группы данных. Двухфакторный дисперсионный анализ позволяет статистически обосновать существенность влияния факторных признаков *A* и *B* и взаимодействия факторов (*A* и *B*) на результативный фактор *F*.

Пример 3.2 У 60 рабочих фиксировалась среднечасовая выработка в натуральных единицах продукции. Данные обследования отражены в табл. 3.2.

Таблица 3.2 – Исходные данные

	A	*B*	*C*	*D*
1	Стаж	Возраст		
2		от25 до 35лет	от35 до 45лет	от 45 до 55 лет
3	от 1 до 4 лет	19	19	18
4		20	20	19
5		20	20	20
6		20	23	21
7		22	25	23
8	от 4 до 7 лет	30	20	19
9		31	29	25
10		32	30	25
11		32	31	26
12		34	31	26
13	от 7 до 10 лет	35	36	24
14		35	40	24
15		39	41	24
16		40	42	25
17		41	45	25
18	свыше 10 лет	40	28	20
19		40	31	24
20		41	35	25
21		41	36	31
22		42	40	32

Оценить существенность влияния возраста и стажа на производительность труда.

Данные в табл. 3.2 приведены так, как они должны выглядеть на листе *Excel*. Заполним диалоговое окно согласно рис. 3.3.

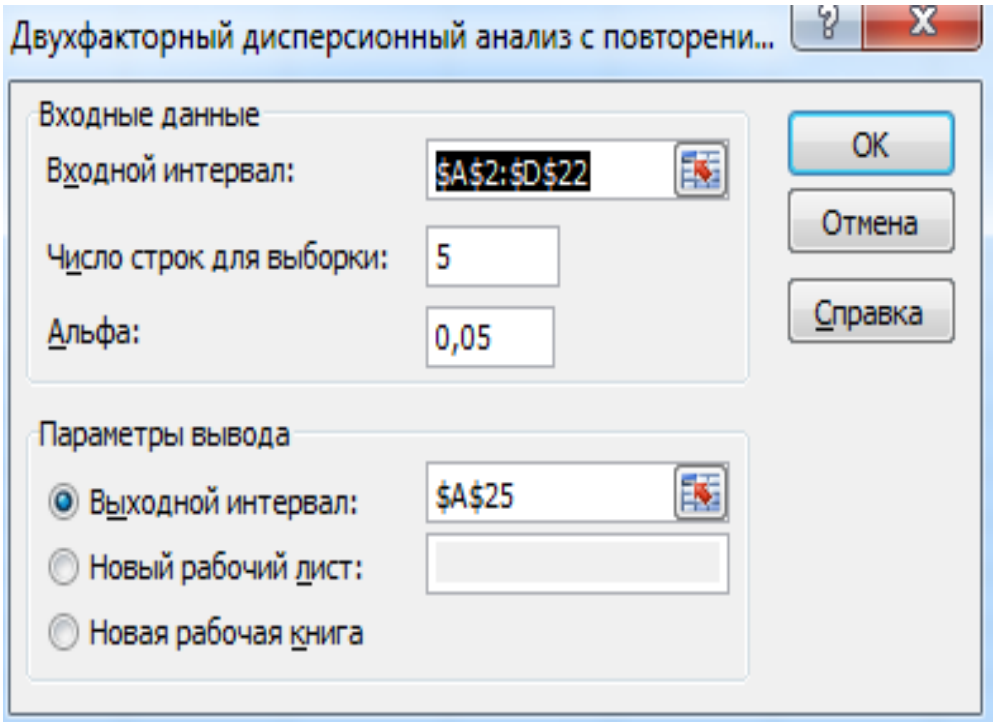


Рисунок 3.3 - Диалоговое окно двухфакторного дисперсионного анализа с повторениями

Следует отметить, что в исходных данных должно быть одинаковое число строк (повторений наблюдений), в противном случае рекомендуется ввести оценки пропущенных наблюдений, выбрав их таким образом, чтобы минимизировать остаточную дисперсию (иначе нужно ввести среднюю других наблюдений в ячейках), причём нельзя включать эти оценки при подсчёте соответствующих степеней свободы.

В случае пропуска данных в дисперсионном анализе без повторений необходимы более сложные методы. Как отмечали Кокрен и Кокс (*Cochran and Cox*): «единственное решение проблемы выпавших данных - не иметь их» [Л.II, 16].

После нажатия на кнопку **OK** мы получим итоговую таблицу 3.3.

Дисперсионный анализ с повторениями позволяет оценить существенность влияния фактора *A* (стажа), *B* (возраста), и их взаимодействия (факторов *A* и *B*) на среднечасовую выработку продукции в натуральных единицах.

Так как: $F_{A\text{расч.}}=66{,}8189>F_{A\text{ кр.}}=2{,}7980$;
$F_{B\text{расч.}}=48{,}9791>F_{B\text{ кр.}}=3{,}1907$;
$F_{AB\text{расч.}}=9{,}7456>F_{AB\text{ кр.}}=2{,}2945$,

Таблица 3.3 - Двухфакторный дисперсионный анализ с повторениями

ИТОГИ	от 25 до 35 лет	от 35 до 45 лет	от 45 до 55 лет	Итого
от 1до 4лет				
Счет	5	5	5	15
Сумма	101	107	101	309
Среднее	20,2	21,4	20,2	20,6
Дисперсия	1,2	6,3	3,7	3,54286
от 4 до 7 лет				
Счет	5	5	5	15
Сумма	159	141	121	421
Среднее	31,8	28,2	24,2	28,0667
Дисперсия	2,2	21,7	8,7	19,6381
от 4 до 7 лет				
Счет	5	5	5	15
Сумма	190	204	122	516
Среднее	38	40,8	24,4	34,4
Дисперсия	8	10,7	0,3	60,4
свыше 10 лет				
Счет	5	5	5	15
Сумма	204	170	132	506
Среднее	40,8	34	26,4	33,7333
Дисперсия	0,7	21,5	25,3	50,6381
Итого				
Счет	20	20	20	
Сумма	654	622	476	
Среднее	32,7	31,1	23,8	
Дисперсия	68,53684211	66,621053	13,3263158	

Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P*-Значение	*F* критическое
Выборка	1842,53333	3	614,177778	66,819	3,70226E-17	2,798
Столбцы	900,4	2	450,2	48,9791	2,56035E-12	3,190
Взаимодействие	537,466667	6	89,5777778	9,74554	5,19758E-07	2,294
Внутри	441,2	48	9,19166667			
Итого	3721,6	59				

то следует признать статистически значимым влияние стажа (фактор *A*) возраста (фактор *B*) и их взаимодействия (факторы *A* и *B*) на производительность труда рабочих. Итоговая таблица позволяет более детально рассмотреть свойства отдельных групп (например, возраст от 35до 45 лет и стаж от 4 до 7 лет). В *Excel* имеется хорошая возможность наглядного представления изучаемых процессов или явлений с помощью мастера диаграмм. Для этого: предварительно выделим данные табл.10; нажмём кнопку **Мастер диаграмм** панели **Стандартная**; выберем тип диаграммы Стандартные – График. В результате, после преобразований, получим рис. 3.4.

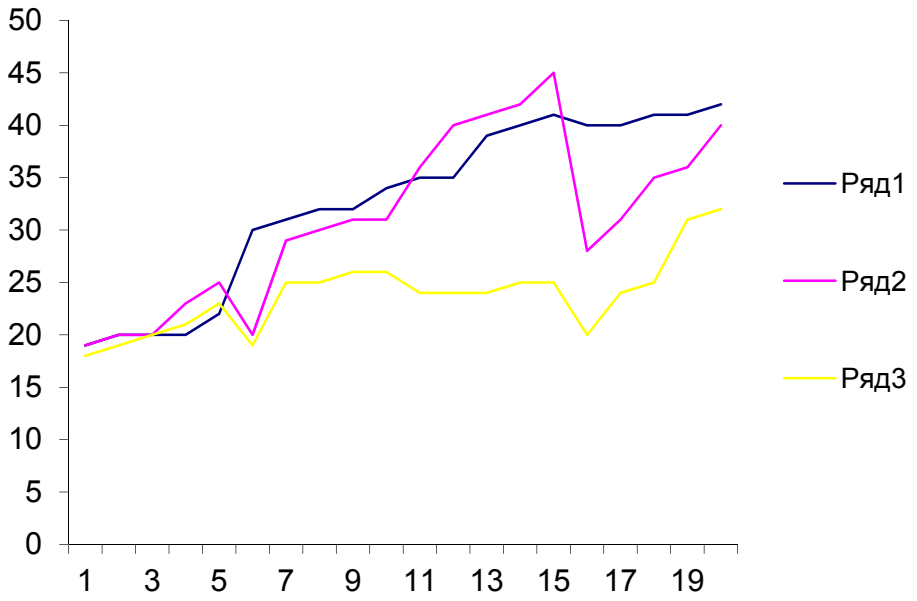


Рисунок 3.4 - Изменение среднечасовой выработки по категориям

Двухфакторный дисперсионный анализ без повторений позволяет оценить существенность воздействия факторов *A* и *B* на результирующий фактор без учёта воздействия взаимодействия факторов *A* и *B*.

Пример 3.3 Рассмотрим решение примера (данные из примера 3.2), взяв в каждой ячейке среднее значение наблюдений.

Стаж	Возраст		
	от 25 до 35 лет	от 25 до 35 лет	от 25 до 35 лет
от 1 до 4 лет	20,2	21,4	20,2
от 4 до 7 лет	31,8	28,2	24,2
от 7до 10 лет	38	40,8	24,4
свыше 10 лет	40,8	34	26,4

В результате заполнения диалогового окна (аналогичного окну примера 3.3) получим итоговую таблицу двухфакторного дисперсионного анализбезповторений.

Таблица 3.4 - Двухфакторный дисперсионный анализ без повторений

ИТОГИ	Счет	Сумма	Среднее	Дисперсия
от 1 до 4 лет	3	61,8	20,6	0,48
от 4 до 7 лет	3	84,2	28,06667	14,4533333
от 7 до 10 лет	3	103,2	34,4	76,96
свыше 10 лет	3	101,2	33,73333	51,8933333
от 25 до 35 лет	4	130,8	32,7	83,5866667
от 25 до 35 лет	4	124,4	31,1	68,3333333
от 25 до 35 лет	4	95,2	23,8	6,74666667

Дисперсионный анализ

Источник вариации	SS	df	MS	F	P-Значение	F критическое
Строки	368,507	3	122,8356	6,85636319	0,02295241	4,757055194
Столбцы	180,08	2	90,04	5,02580005	0,05222744	5,143249382
Погрешность	107,493	6	17,91556			
Итого	656,08	11				

Данные дисперсионного анализа свидетельствуют о том, что фактор *A* (стаж) существенно влияет на производительность труда (так как $F_{A\text{расч.}} = 6{,}8563 > F_{A\ \text{кр.}} = 4{,}7570$), а фактор *B* (возраст) статистически существенного влияния не оказывает (так как $F_{B\text{расч.}} = 5{,}0258 < F_{B\text{кр.}} = 5{,}1432$).

Различие в выводах примеров 3.2 и 3.3 можно объяснить тем, что в примере 3.3 мы не учитывали конкретные наблюдения в ячейках, а рассматривали лишь их среднее значение. Поэтому результат дисперсионного анализа с повторениями является более значимым, что соответствует и самому смыслу задачи.

Так как очевидно – на производительность труда влияют и стаж, и возраст, и их взаимодействие (стаж и возраст). Отсюда следует сделать вывод, что при организации наблюдений необходимо, для каждого уровня факторов, рассматривать, возможно, большее количество элементов в ячейках.

Корреляция используется для количественной оценки взаимосвязи двух наборов данных с помощью коэффициента корреляции.

Коэффициент корреляции выборки представляет собой ковариацию двух наборов данных, делённую на произведение их стандартных отклонений (модуль 4).

Ковариация дает возможность установить, ассоциированы ли наборы данных по величине. Например, большие значения из одного набора данных связаны с большими значениями другого набора (положительная ковариация), или, наоборот, малые значения одного набора связаны с большими значениями другого (отрицательная ковариация), или данные двух диапазонов никак не связаны (ковариация близка к нулю).

Описательная статистика. Генерирует статистический отчёт для массива чисел: характеристики положения и вариации, наибольшее и наименьшее значение, коэффициенты асимметрии и эксцесса, сумму, предельную ошибку довери-

тельного интервала (уровень надёжности (1-α)·100%) для средней при заданном уровне надежности (Модуль 3).

Экспоненциальное сглаживание. Предназначается для предсказания значения на основе прогноза для предыдущего периода, скорректированного с учетом погрешностей в этом прогнозе.

Использует константу сглаживания, по величине которой определяет, насколько сильно влияют на прогнозы погрешности в предыдущем прогнозе (Модуль 4). **Двухвыборочный тест для дисперсии.** Решает задачу сравнения дисперсий двух генеральных совокупностей при заданном уровне значимости α.

Анализ Фурье. Позволяет анализировать периодические данные посредством быстрого преобразования Фурье, которое, однако, предполагает, что количество наблюдений равно 2^{2n} ($n \in N$).При большом количестве данных (порядка нескольких тысяч) рекомендуется добавлять нули для достижения 2^{2n} (причём наибольшее значение в Excel для 2^{2n}=4096).

Гистограмма. Вычисляет выборочные и накопленные частоты попадания ряда данных в интервалы, а также может выводить соответствующие графики (модуль 3). Границы интервалов (карманы) можно указать (обычно задают верхние границы интервалов). Если не задать границы, то Excel сам разобьёт анализируемый ряд на интервалы с шагом h:

$$h = \frac{x_{max} - x_{min}}{\sqrt{n}},$$

где x_{max} -наибольший вариант,

x_{min} - наименьший вариант,

n-ряда (количество членов).

Гистограмму можно построить автоматически для этого достаточно выделить необходимый диапазон и нажать F11.

Скользящее среднее. Прогнозирует значения временного ряда на основе предшествующих значений с помощью простой скользящей средней.

Генерация случайных чисел. Заполняет диапазон случайными числами, заданными по одному из законов: равномерному; нормальному; Бернулли (СВ X принимает значение 1 с вероятностью p и 0 с вероятностью (1-p) (индикаторная СВ)); биномиальному; Пуассона; модельному (позволяющему генерировать последовательности случайных чисел от a до b с шагом c, и возможностью повторения каждого числа и последовательности); дискретному (решающему задачу, получения по имеющемуся распределению, новых значений того же распределения).

Замечание. Инструмент генерации случайных чисел позволяет решать целый ряд задач: численных методов (например, приближённого вычисления определённых интегралов методом статистических испытаний - методом Монте-Карло); имитационного моделирования изучаемых процессов и т.д.

Пример3.4 Вычислить определённый интеграл $J = \int_0^1 x^2 dx$ методом Монте-Карло.

Вычисление определенного интеграла J равносильно нахождению площади D криволинейной трапеции функции $Y=f(X)=X^2$ (рис. 3.5).

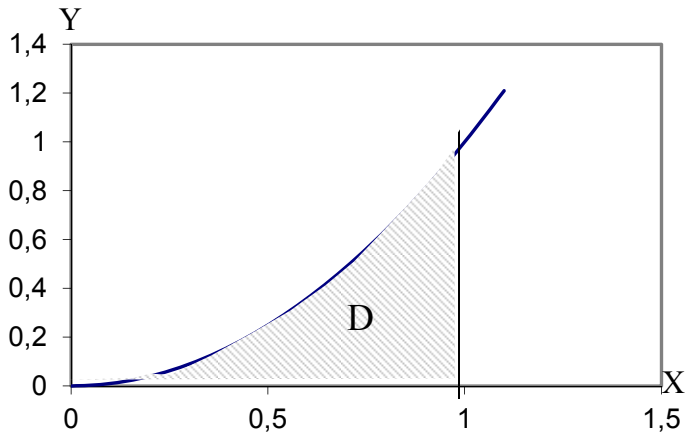


Рисунок 3.5 – Область D

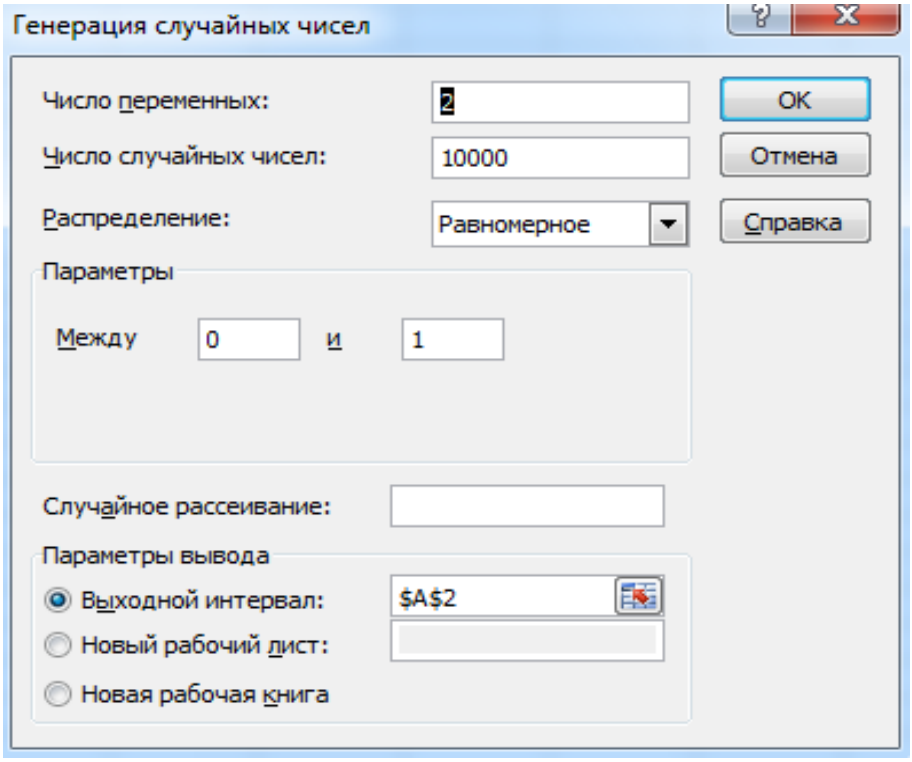


Рисунок 3.6 - Диалоговое окно генерации двумерных, равномерно распределённых случайных величин

Рассмотрим систему двумерных равномерно распределенных случайных величин (X, Y) на интервале от 0 до 1. При достаточно большом числе опытов N, площадь D будет приближенно равна относительной частоте попадания точек $M_i(x_i, y_i)$ в область D (в силу закона больших чисел):

$$J \approx \frac{n}{N}.$$

Для генерации системы двух равномерно распределенных на интервале от 0 до 1 случайных величин используем инструмент **Генерация случайных чисел**. Заполним диалоговое окно для генерации 10000 пар указанных случайных чисел (рис. 3.6). В результате в диапазоне A1:B10001 получим искомые пары случайных чисел. В ячейке C2 введем формулу: = A1^2; в ячейке D2: = Если (B2>C2;0;1). По-

следняя формула присваивает ячейке значение 0, если точка M_i не попадает в область D и значение 1 в противном случае. Выделим диапазон C2:D2 и скопируем вниз до строки 10001. Найдем сумму значений в диапазоне D2:D10001, в результате получим, что $n=\sum m_i=3337$ (рис. 3.7).

D2 =ЕСЛИ(B2>C2;0;1)

	A	B	C	D
1	Xi	Yi	f(Xi)	mi
2	0,226477859	0,658253731	0,051292221	0
3	0,716666158	0,462477493	0,513610382	1
4	0,016022217	0,819635609	0,000256711	0
5	0,003845332	0,202764977	1,47866E-05	0
6	0,241431928	0,048799097	0,058289376	1
7	0,867763298	0,15396588	0,753013142	1
8	0,167027802	0,709494308	0,027898287	0
9	0,246986297	0,911038545	0,061002231	0
10	0,276345103	0,046998505	0,076366616	1
9999	0,809625538	0,928403577	0,655493512	0
10000	0,989379559	0,969939268	0,978871911	1
10001	0,791222877	0,778283029	0,626033641	0
10002	Итого	-	-	3337

Рисунок 3.7– Результат применения метода Монте-Карло

Отсюда $J \approx \frac{n}{N} = \frac{3337}{10000} = 0{,}3337.$

Применяя неравенство Чебышева, имеем

$$P\left(\left|\frac{n}{N} - J\right| < \varepsilon\right) \geq 1 - \frac{J(1-J)}{\varepsilon^2 \cdot N} \geq 1 - \frac{1}{4\varepsilon^2 N}.$$

Если мы зададим уровень значимости α, то неравенство приведенное выше будет всегда верно с гарантийной вероятностью p=1-α, при $\alpha = \frac{1}{4\varepsilon^2 N}$.

При заданных значениях ε и α можно определить необходимое число испытаний

$$N = \frac{1}{4\varepsilon^2 \alpha}.$$

В силу того, что неравенство Чебышева дает нижнюю оценку вероятности значение N будет завышено, например, в нашем случае при $\varepsilon = 0{,}001$ и $\alpha = 0{,}01$: N = 25000000. Точное значение $J = 0{,}(3)$. В рассматриваемом примере точность ε=0,001 достигается уже при 10000 испытаний. (Существуют более точные методы оценки N[1], основывающиеся на предельных теоремах теории вероятностей).

[1] Демидович Б.П., Марон И.А. Основы вычислительной математики. - М.:Физматгиз. 1963. - 660 с.

Ранг и персентиль. Выводит таблицу с порядковым и процентным рангом для каждого значения в ряде данных.

Регрессия. Подбирает линейную функцию с помощью метода наименьших квадратов (МНК). Основные параметры диалогового окна: входной интервал Y - диапазон анализируемых зависимых данных (диапазон должен состоять из одного столбца); входной интервал X - диапазон независимых данных (находящихся в соседних столбцах), подлежащих анализу, Excel располагает независимые переменные этого диапазона слева направо в порядке возрастания.

Максимальное число входных диапазонов равно 16. При необходимости можно сформировать и нелинейную функцию. Например, для случая двух переменных

$$Y(X_1, X_2) = a_0 + a_1X_1 + a_2X_2 + a_3X_1^2 + a_4X_2^2 + a_5X_1X_2 \text{ -}$$

формируется диапазон, содержащий X_1^2, X_2^2, X_1X_2, рядом с диапазоном переменных X_1, X_2.

Пример 3.5 Исследовать влияние фондовооруженности и энергообеспеченности на стоимость валовой продукции.

Результативным признаком (Y) являются стоимость валовой продукции на 100 га сельхозугодий, тыс. руб., которая характеризует затраты на производство продукции.

Факторные признаки:

X_1 – фондовооруженность 1-го работника, характеризующая стоимость основных фондов, в расчете на одного работника (тыс. руб.);

X_2 – энергообеспеченность на 100 га сельхозугодий, л.с., выражающая мощность энергетических ресурсов на 100 га сельхозугодий.

Требуется определить:

- параметры множественного уравнения регрессии в натуральной и стандартизованной форме;
- средние коэффициенты эластичности для каждого фактора;
- коэффициенты частной и множественной корреляции;
- общий и частные критерии F – Фишера.

Уравнение регрессии примет вид: $y=b_0+b_1x_1+b_2x_2+\varepsilon$

Рассмотрим применение пакета анализа в *Excel MS Office* 2007 (2010) для решения данной задачи. Исходные данные для анализа введем на листе *MS Excel* в виде, представленном на рисунке 3.8.

Для проведения анализа предварительно установим пакет анализа, выполнив последовательно действия: кнопка *Office* (в *MS Office* 2010 – файл) – *Параметры Excel – Надстройки – Пакет анализа – Перейти* (выделим в окне доступных надстроек *Пакет анализа*), после этого во вкладке *Данные* ленты появится инструмент *Пакет анализа*.

	A	B	C
1	Стоимость валовой продукции на 100 га сельхозугодий, тыс. руб.	Фондовооруженность 1-го работника , тыс.руб.	Энергообеспеченность на 100 га сельхозугодий, л.с.
2	Y	X1	X2
3	2715	315	255
4	4190	479	309
5	4219	467	315
6	3193	355	180
7	4699	450	255
8	3574	347	229
9	3806	503	307
10	4891	529	343
11	4120	581	199
12	2989	370	205
13	3722	477	243
14	4544	542	321
15	3005	404	212
16	5364	541	320
17	4216	425	246

Рисунок 3.8 – Вид исходных данных в *MS Excel*

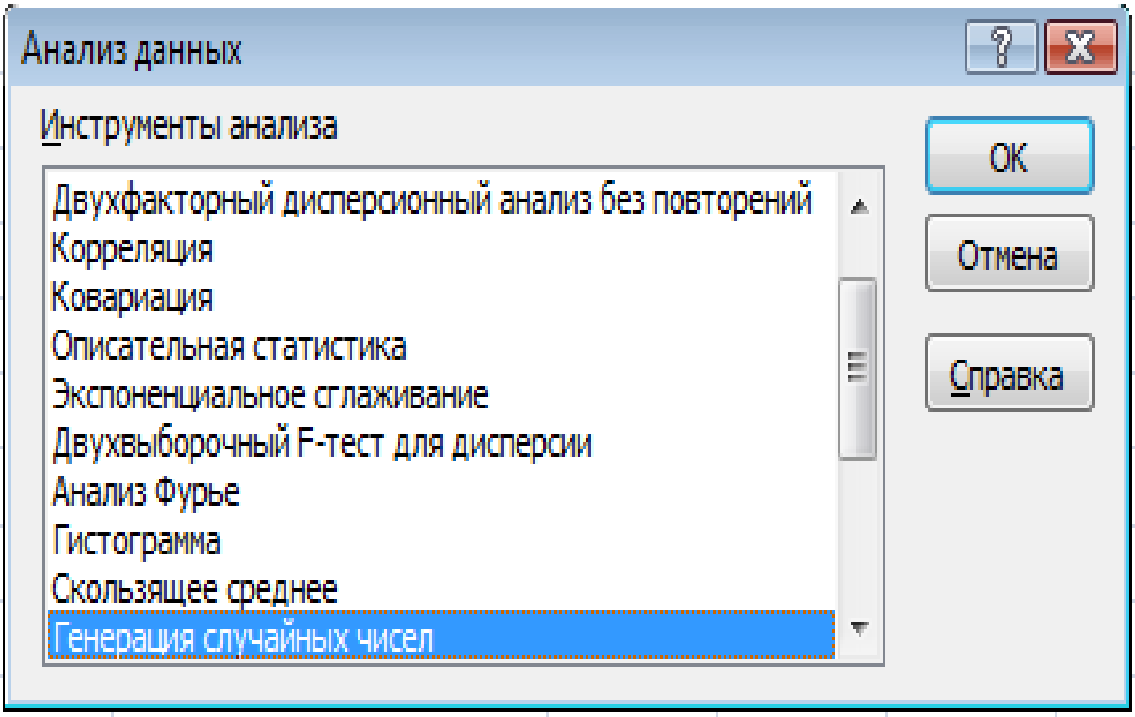


Рисунок 3.9 – Надстройка *Пакет анализа* (*Анализ данных*)

Выберем в *Пакете анализа* инструмент *Описательная статистика* и заполним параметры диалогового окна (рисунки 3.9, 3.10).

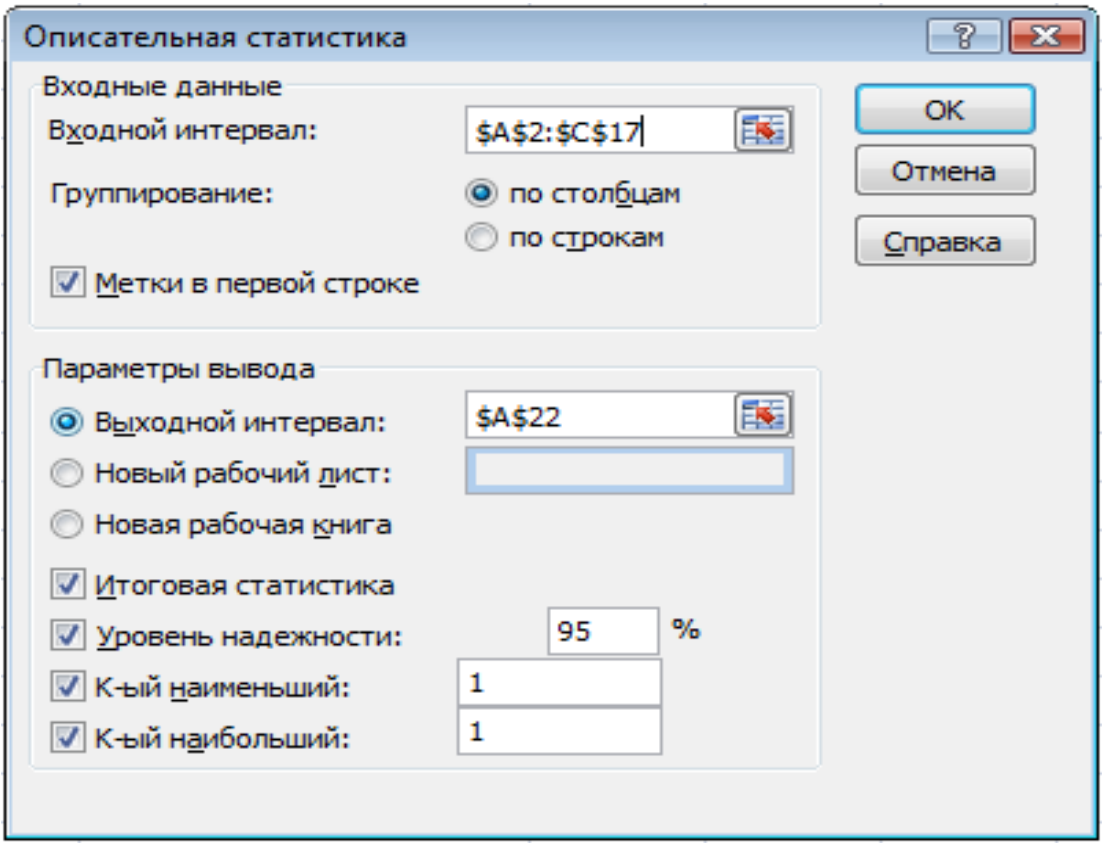


Рисунок 3.9 – Диалоговое окно инструмента Описательная статистика

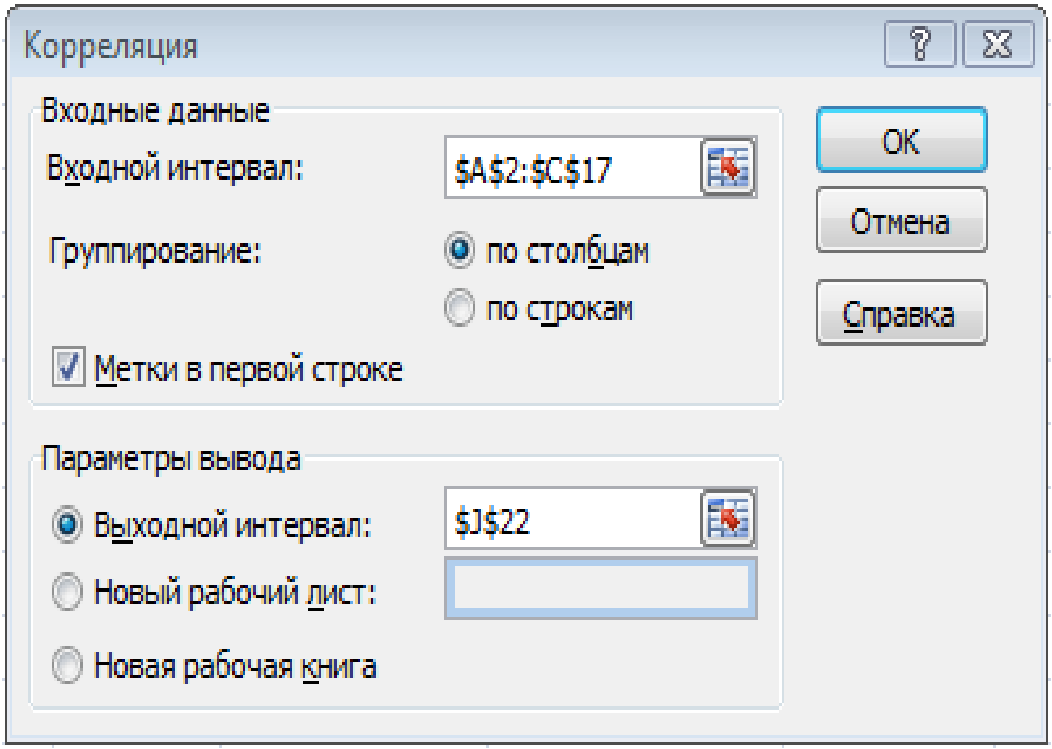


Рисунок 3.10 – Диалоговое окно инструмента Корреляция

Таблица 3.5 – Результаты применения инструмента *Описательная статистика*

	Y	X_1	X_2	Принятые обозначения
Среднее	3949,8	452,333	262,6	$\overline{X} = \sum x_i n_i / n$
Стандартная ошибка	197,0888898	20,900	13,57336	$S_{\overline{X}} = S/\sqrt{n}$
Медиана	4120	467	255	Me
Мода	#Н/Д	#Н/Д	255	Mo
Стандартное отклонение	763,322	80,947	52,56941	S
Дисперсия выборки	582660,457	6552,381	2763,543	$S^2 = \sum (x_i - \overline{X})^2 n_i /(n-1)$
Эксцесс	-0,674	-1,048	-1,43373	$Ex = \sum ((x_i - \overline{X})/S)^4 n_i / n - 3$
Асимметричность	0,0436	-0,187	0,026622	$Sk = \sum ((x_i - \overline{X})/S)^3 n_i / n$
Интервал	2649	266	163	$R=x_{max} - x_{min}$
Минимум	2715	315	180	x_{min}
Максимум	5364	581	343	x_{max}
Сумма	59247	6785	3939	Σx_i
Счет	15	15	15	$n=\Sigma n_i$
Наибольший (1)	5364	581	343	-
Наименьший (1)	2715	315	180	-
Уровень надежности (95,0%)	422,714	44,827	29,112	$\Delta = t_{\alpha;n-1} S_{\overline{X}}$

Для нахождения парных коэффициентов корреляции применим инструмент пакета анализа Корреляция, для этого заполним параметры диалогового окна как на рисунке 3.10.

Таблица 3.6 – Парные коэффициенты корреляции между признаками

	Y	X_1	X_2
Y	1		
X_1	0,773088239	1	
X_2	0,679291686	0,543201832	1

$$r_{yx_1} = 0{,}7731; r_{yx_2} = 0{,}6793; \quad r_{x_1x_2} = 0{,}5432 .$$

Парные коэффициенты корреляции свидетельствуют о тесной связи между факторными и результативным признаками, что дает основание включить данные факторы в уравнение регрессии.

Линейное уравнение множественной регрессии в натуральной форме имеет вид $y=b_0+b_1x_1+b_2x_2+\varepsilon$ найдем параметры этого уравнения, используя инструмент *Пакета анализа – Регрессия*. Заполним параметры диалогового окна (рис. 3.11).

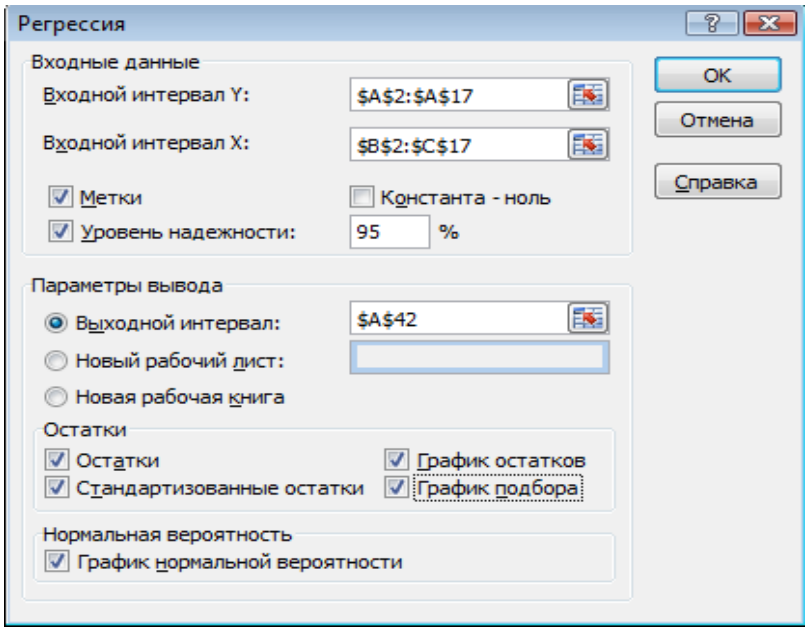


Рисунок 3.11 – Диалоговое окно инструмента Регрессия

Регрессионная статистика	
Множественный *R*	0,8325
R-квадрат	0,6931
Нормированный *R*-квадрат	0,6419
Стандартная ошибка	456,765
Наблюдения	15

Дисперсионный анализ

	df	*SS*	*MS*	*F*	Значимость *F*
Регрессия	2	5653635,943	2826817,971	13,549	0,0008
Остаток	12	2503610,457	208634,2048		
Итого	14	8157246,4			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	101,8239	748,636	0,136	0,894	-1529,314	1732,962
X_1	5,4056	1,796	3,009	0,011	1,492	9,319
X_2	5,3421	2,7658	1,931	0,077	-0,6843	11,368

Рисунок 3.12 – Вывод итогов регрессионного анализа

Получим линейное уравнение множественной регрессии

$$y = 101,8239 + 5,4056x_1 + 5,3421x_2.$$

Коэффициенты множественной регрессии показывают, что при увеличении фондовооруженности на 1 тыс. руб. стоимость валовой продукции в среднем повысится на 5,406 тыс. руб., а при увеличении энергообеспеченности на 1 л.с. в расчете на 100 га с/х угодий стоимость производства валовой продукции увеличится на 5, 342 тыс. руб.

В стандартизованной форме уравнение регрессии имеет вид:

$$t_y = \beta_1 \cdot t_{x_1} + \beta_2 t_{x_2},$$

где

$$t_y = \frac{y - \overline{y}}{\sigma_y}; t_{x_1} = \frac{x_1 - \overline{x}_1}{\sigma_{x_1}}; t_{x_2} = \frac{x_2 - \overline{x}_1}{\sigma_{x_1}}.$$

Найдем β – коэффициенты, используя их связь с коэффициентами b, уравнения регрессии в нормальной форме:

$$\beta_i = b_i \cdot \frac{\sigma_{x_i}}{\sigma_y}.$$

Имеем:

$$\beta_1 = b_1 \cdot \frac{\sigma_{x_1}}{\sigma_y} = 5,4056 \cdot \frac{80,9468}{763,3220} = 0,5732;$$

$$\beta_2 = b_2 \cdot \frac{\sigma_{x_2}}{\sigma_y} = 5,3421 \cdot \frac{52,5694}{763,3220} = 0,3679.$$

β – коэффициенты, можно также найти с помощью парных коэффициентов корреляции по формулам:

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0,7731 - 0,6793 \cdot 0,5432}{1 - 0,5432^2} = 0,5732;$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0,6793 - 0,7731 \cdot 0,5432}{1 - 0,5432^2} = 0,3679.$$

Линейное уравнение множественной регрессии в стандартизованном масштабе имеет вид:

$$t_y = 0,5732t_{x_1} + 0,3679t_{x_2}.$$

По абсолютной величине β – коэффициентов можно сделать вывод об относительной силе влияния факторов на изменение результативного признака. На стоимость производства валовой продукции более сильное влияние оказывает фондовооруженность одного работника, а влияние энергообеспеченности на результаты производства несколько ниже.

Средние коэффициенты эластичности находятся по формуле:

$$Э_{yx_i} = b_i \cdot \frac{\overline{x}_i}{\overline{y}};$$

$$Э_{yx_1} = b_1 \cdot \frac{\overline{x}_1}{\overline{y}} = 5,4056 \cdot \frac{452,333}{3949,8} = 0,619;$$

$$Э_{yx_2} = b_2 \cdot \frac{\overline{x}_2}{\overline{y}} = 5,3421 \cdot \frac{262,6}{3949,8} = 0,355.$$

Значит, при увеличении фондовооруженности 1 работника на 1 % выход валовой продукции на 100 га сельхозугодий увеличивается в среднем на 0, 619% при неизменности влияния второго фактора. Изменение энергообеспеченности предприятия на 1 % приводит к росту стоимости валовой продукции на 0,355%, при исключении влияния первого фактора

Коэффициенты частной корреляции определяются через парные коэффициенты корреляции по формулам:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)\cdot(1-r_{x_1x_2}^2)}} = \frac{0,7731 - 0,6793 \cdot 0,5432}{\sqrt{(1-0,6793^2)(1-0,5432^2)}} = 0,6558$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)\cdot(1-r_{x_1x_2}^2)}} = \frac{0,6793 - 0,7731 \cdot 0,5432}{\sqrt{(1-0,7731^2)(1-0,5432^2)}} = 0,4871$$

$$r_{x_1x_2 \cdot y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)\cdot(1-r_{yx_2}^2)}} = \frac{0,5432 - 0,7731 \cdot 0,6793}{\sqrt{(1-0,7731^2)(1-0,6793^2)}} = 0,0470.$$

Коэффициенты частной корреляции характеризуют тесноту связи между двумя переменными, исключив влияние третьей переменной.

Значит связь между стоимостью валовой продукции и фондовооруженностью прямая и достаточно тесная, между стоимостью валовой продукции и энергообеспеченностью при исключении влияния фондовооруженности средняя. Связь между собой факторных признаков очень слабая.

Коэффициент множественной корреляции находится по формуле:

$$R_{yx_1x_2} = \sqrt{\beta_1 \cdot r_{yx_1} + \beta_2 r_{yx_2}} = \sqrt{0,6732 \cdot 0,7731 + 0,3679 \cdot 0,6793} =$$
$$= \sqrt{0,5205 + 0,2499} = \sqrt{0,7704} = 0,8777.$$

Более точный результат (без округлений) можно получить по формуле

$$R = \sqrt{\frac{\sum(\bar{y}-\hat{y})^2}{\sum(y_i-\hat{y})^2}} = \sqrt{\frac{5653635,943}{8157246,4}} = \sqrt{0,6931} = 0,8325.$$

Величина коэффициента показывает, что связь между Y, X_1 и X_2 довольно тесная, причем 69,3 % вариации стоимости валовой продукции объясняется вариацией фондовооруженности и энергообеспеченоети.

Оценим значимость уравнения регрессии и коэффициента R^2 с помощью критерия F – Фишера. Фактически рассматривается нулевая гипотеза H_0: $R^2=0$($b_1=b_2=0$) и альтернативная гипотеза H_1: $R^2\neq0$ $b_1\neq0$, $b_2\neq0$).

Наблюдаемое или фактическое значение критерия находится по формуле:

$$F_н = \frac{R^2_{yx_1x_2}}{1-R^2_{yx_1x_2}} : \frac{m}{n-m-1},$$

где m – число факторов в линейном уравнении регрессии;
n – число единиц наблюдения.

$$F_н = \frac{0,6931}{1-0,6931} : \frac{2}{15-2-1} = 13,12.$$

При уровне значимости α=0,05 и числе степеней свободы $k_1=m=2$, $k_2=n-m-1=15-2-1=12$ по таблице значений критерия F – Фишера критическое значения составляет 3,80, т.е. $F_{кр}=3,80$. Сравниваем $F_н$ с $F_{кр}$. Так как $F_н > F_{кр}$, то нулевую гипотезу о незначимости величины R^2 отклоним, т.е. уравнение множественной регрессии и R^2 статистически значимы.

В уравнении множественной регрессии не все факторы могут оказывать статистически существенное влияние на изменение результативного признака. Оценка значимости факторов в уравнении регрессии может быть дана с помощью частного F – критерия или критерия t – Стьюдента.

$$F_{нx_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1-R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0,6931-0,6793^2}{1-0,6931} \cdot \frac{15-2-1}{1} = 8,772.$$

При α=0,05, $k_1=1$, $k_2=12$, $F_{кр}=4,75$.

Так как $F_{нx1} > F_{кр}$, то в уравнение регрессии целесообразно включение фактора X_1 после X_2. Фактор X_1 оказывает статистически значимое влияние на Y.

$$F_{нx_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0,6931 - 0,7731^2}{1 - 0,6931} \cdot \frac{15-2-1}{1} = 3,6131.$$

$F_{нx2} < F_{кр}$ - это свидетельствует о статистической незначимости влияния фактора X_2 и нецелесообразности включения его в уравнение множественной регрессии. В данной задаче на стоимость валовой продукции статистически значимое влияние оказывает фондовооруженность одного работника.

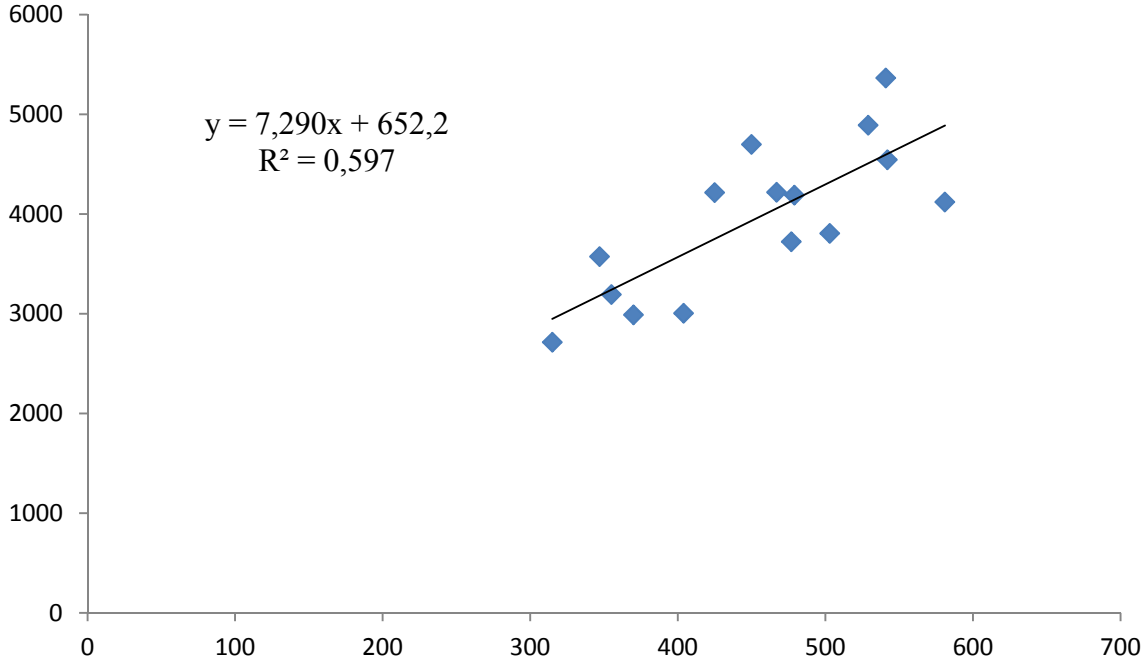


Рисунок 3.13 – Парное уравнение регрессии

Уравнение парной регрессии между X_1 и Y:

$$Y=652,22 + 7,2902x_1 .$$

Значит, при увеличении фондовооруженности одного работника на 1 тыс. руб. производство валовой продукции увеличивается на 7, 29 тыс. руб.

Полученное уравнение объясняется 59,77 % различий в стоимости валовой продукции (что на 9,54 % меньше, чем уравнение с двумя факторами).

Выборка. Создаёт выборку случайным или периодическим (серийным) способом в предположении, что заданный диапазон является генеральной совокупностью.

Парный двухвыборочный тест для средних.

Двухвыборочный *t* - тест с одинаковыми дисперсиями**.**

Двухвыборочный *t* - тест с различными дисперсиями.

Двухвыборочный *z*- тест для средних.

Инструменты Пакета анализа 16-19 позволяют проверить существенность различий между двумя средними для совокупностей, заданных в двух диапазонах.

Задания

По данным, предложенным преподавателем, определить:

- параметры множественного уравнения регрессии в натуральной и стандартизованной форме;
- средние коэффициенты эластичности для каждого фактора;
- коэффициенты частной и множественной корреляции;
- общий и частные критерии F – Фишера.

Вопросы для самоконтроля

- Опишите возможности ввода информации.
- Дисперсия, ее свойства и способы расчета.
- Как проводится дисперсионный анализ.
- Опишите возможности пакета анализа данных.

Практическое занятие № 4

Анализ временных рядов

Цель работы: ознакомиться с возможностями методик анализа временных рядов, получить навыки анализа данных с использованием *Excel* 2007.

Теоретические сведения

Одной из важнейших задач статистики является изучение изменения экономических явлений во времени путем построения и анализа рядов динамики. **Ряд динамики** представляет собой численные значения статистического показателя в последовательные моменты или периоды времени.

В ряду динамики выделяют два элемента: периоды или моменты времени (t) и соответствующие им количественные значения показателя, называемые уровнями ряда (y_t). Если уровни ряда характеризуют значения показателя за определенный период времени, то ряд называется интервальным, а если на определенный момент времени – моментным. Уровни выражаются абсолютными, относительными и средними величинами.

Пример 4.1 Рассмотрим задачу прогнозирования урожайности подсолнечника, при наличии данных с1975г по 2009г (за 35 лет), для этого рассмотрим ряд моделей и выберем ту из них, которая даёт наименьшую ошибку в 2010 году, хотя можно в качестве критерия отбора выбрать минимальную сумму модулей ошибок точек или минимальную дисперсию для последних k-точек и т. д. - в конечном счете, вид лучшего критерия мы узнаем только в 2010 году.

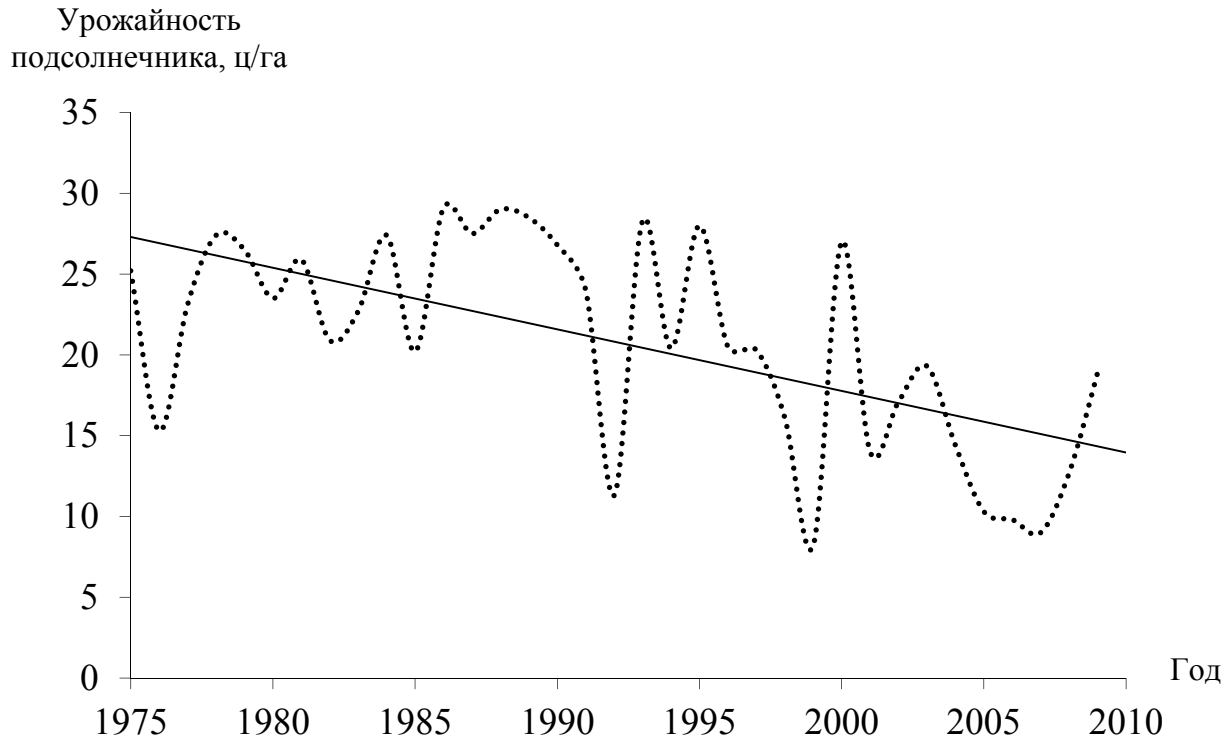


Рисунок 4.1 – Динамика урожайности подсолнечника

Предполагая, что тенденция изменения урожайности не изменится в ближайшие годы, для прогноза экстраполируем выбранную модель для первых 3-5 лет после 2009 года. Используя мастер диаграмм, получим график ряда динамики урожайности (рис. 4.1).

Для аналитического выравнивания и прогноза по уравнению прямой $Y_t = a_0 + bt$ можно использовать в категории Статистические следующие функции (причём для отсчёта времени перейдём к условным годам t: 1,2,…,35) (рис. 4.2):

а) ПРЕДСКАЗ (выделим диапазон D2:D40 и в ячейку D2 введём формулу массива: {=ПРЕДСКАЗ (B2:B40;C2:C36;B2:B36)});

б) ТЕНДЕНЦИЯ (выделим диапазон E36:E40 и в ячейку E36 введём формулу массива: {=ТЕНДЕНЦИЯ(C2:C36;B2:B36;B36:B40;1)});

в) ЛИНЕЙН - позволяет получить коэффициенты уравнения регрессии помощью МНК, которые можно использовать в формуле для выравнивания и прогноза.

Для аналитического выравнивания по уравнению экспоненты $y = b*m^x$:

а) РОСТ (используется для предсказания или выравнивания по экспоненциальной кривой, выделим диапазон F2:F40 и введём формулу массива {=РОСТ(C2:C36;B2:B36;B2:B40;1)});

б) ЛГРФПРИБЛ может использоваться аналогично ЛИНЕЙН.

Самый простой способ прогнозирования на основе линейного или экспоненциального тренда заключается в использовании контекстного меню. Необходимо: 1) выделить диапазон данных; 2) при нажатой правой клавише мыши, протащить маркер заполнения на необходимый период прогнозирования; 3) в открывшемся контекстном меню выбрать вид приближения: Линейное, Экспоненциальное, Прогрессия (рис. 4.3).

D2 f_x {=ПРЕДСКАЗ(B2:B40;C2:C36;B2:B36)}

	A	B	C	D	E	F
1	ГОД	Порядковый номер года, t	Урожайность подсолнечника, ц/га	ПРЕДСКАЗ	ТЕНДЕНЦИЯ	РОСТ
2	1975	1	25,2	27,29571429		28,3618191
3	1976	2	15,3	26,91462185		27,75293051
4	1977	3	23,1	26,53352941		27,15711391
5	1978	4	27,4	26,15243697		26,57408865
6	1979	5	26,5	25,77134454		26,00358013
7	1980	6	23,5	25,3902521		25,44531963
8	1981	7	25,9	25,00915966		24,8990442
9	1982	8	20,9	24,62806723		24,36449655
10	1983	9	22,7	24,24697479		23,84142488
32	2005	31	10,3	15,86294118		14,79030229
33	2006	32	9,8	15,48184874		14,47277518
34	2007	33	9	15,1007563		14,16206494
35	2008	34	12,7	14,71966387		13,85802521
36	2009	35	18,9	14,33857143	14,33857143	13,56051279
37	2010	36		13,95747899	13,95747899	13,26938754
38	2011	37		13,57638655	13,57638655	12,98451235
39	2012	38		13,19529412	13,19529412	12,70575303
40	2013	39		12,81420168	12,81420168	12,43297828

Рисунок 4.2 - Результаты линейной и экспоненциальной экстраполяции

Замечание. Все приведённые выше формулы можно (даже нужно) вводить не в ручную, а используя, *МАСТЕР ФУНКЦИЙ*, категорию *Статистические*; так как рассматриваемые выше формулы обрабатывают массивы данных, то после их введения необходимо нажать *Ctrl + Shift + Enter*.

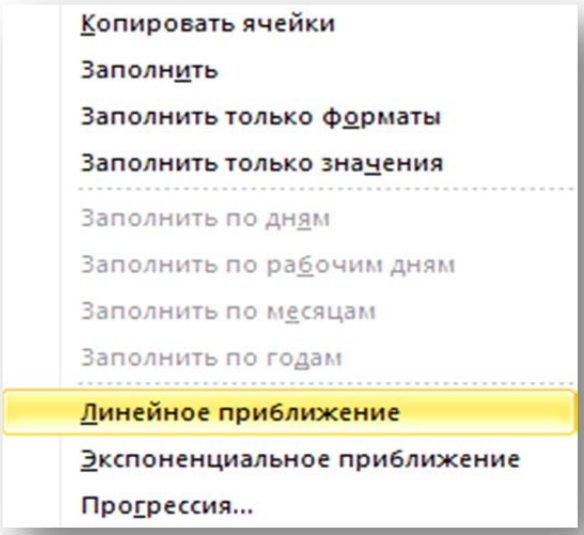


Рисунок 4.3 - Контекстное меню для экстраполяции

Важным методом анализа временных рядов в *Excel* являются диаграммы. Выделим на (рис.4.4) щелчком левой клавиши мыши маркеры наблюдений урожайности подсолнечника по годам; с помощью правой клавиши откроем контекстное меню (рис.4.5) и выберем одну из перечисленных линий трендов (рис.4.5):

Линейная,
Логарифмическая,
Полиномиальная,
Степенная,
Экспоненциальная,
Линейная фильтрация (*Скользящая средняя*).

После выбора одного из трендов, например, линейного - отметим «показывать уравнение на диаграмме» и «поместить на диаграмму коэффициент достоверности аппроксимации (R^2) (рис. 4.6).

Можно выбрать название (назвать тренд самостоятельно) или оставить автоматически предлагаемое Excel; для прогноза согласно выбранной линии тренда на 5 лет вперёд выберем соответствующее значение в диалоговом окне;

Для отображения на диаграмме уравнения тренда и коэффициента детерминации отметим соответствующие элементы вкладки Параметры (рис. 4.5). Далее выберем ОК.

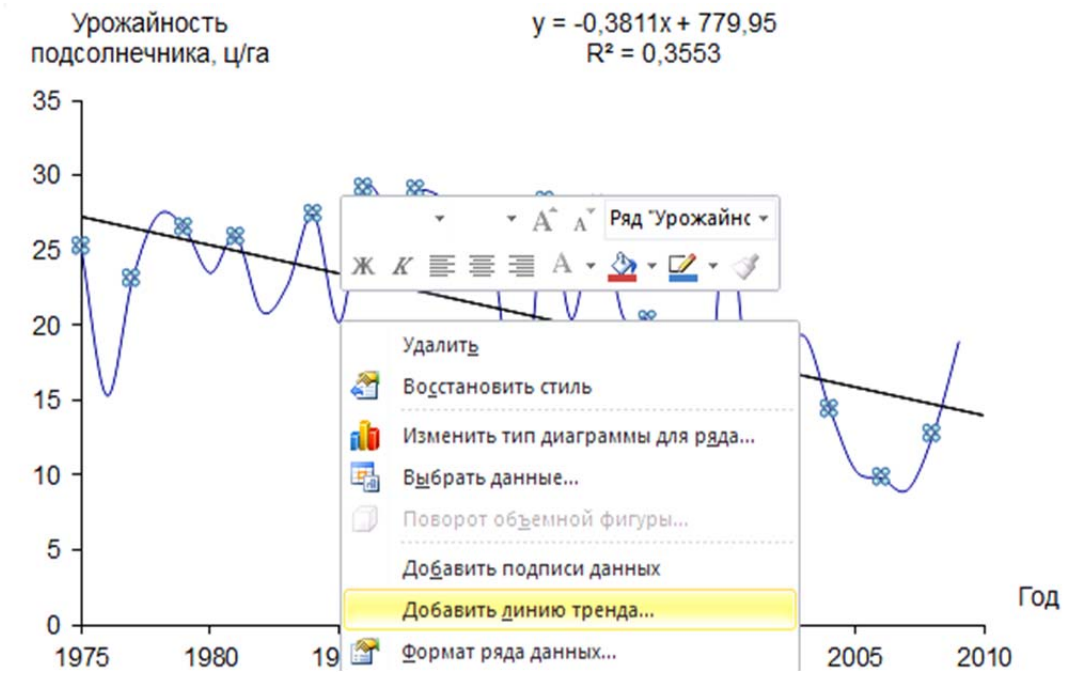


Рисунок 4.4 - Контекстное меню выделенных точек наблюдений

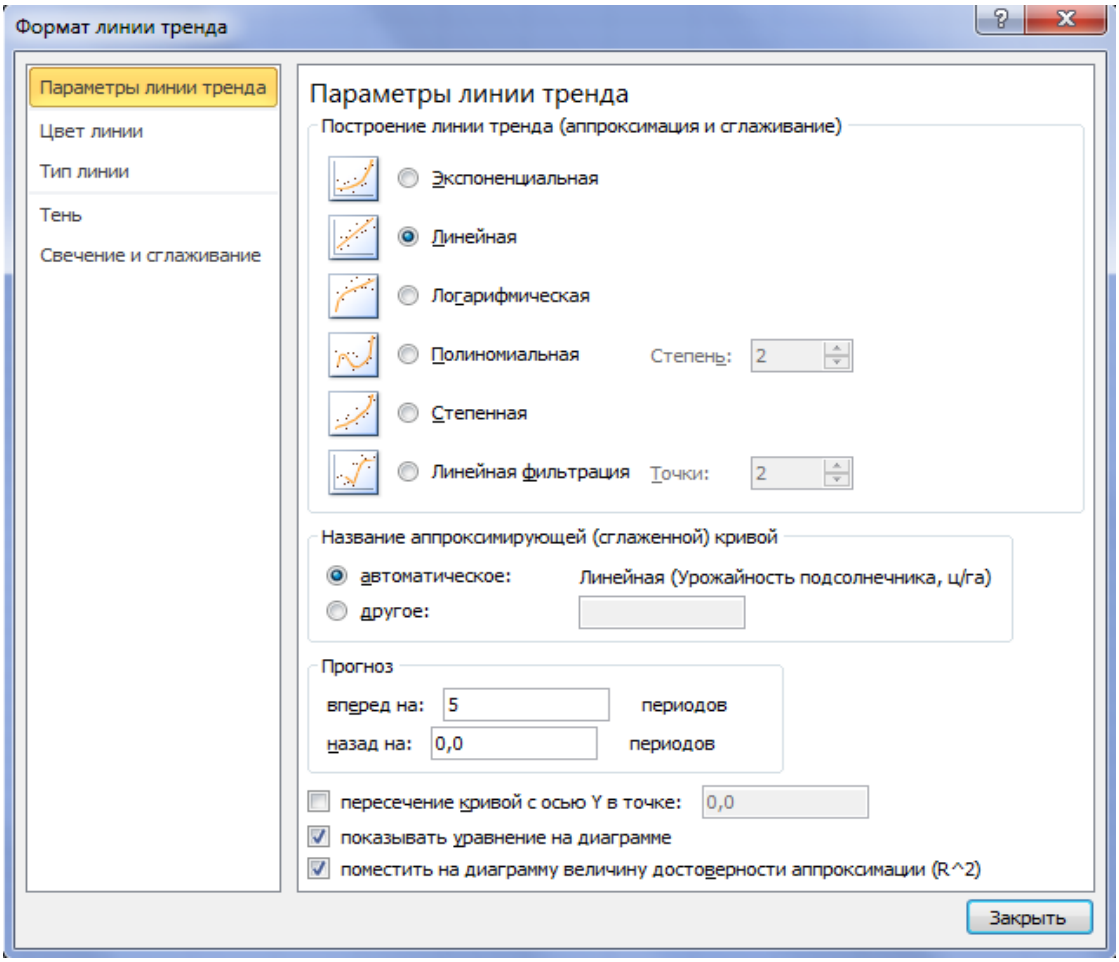


Рисунок 4.5 - Диалоговое окно выбора линии тренда

При выборе других типов линии тренда получим рис.4.6.

Рассмотрим другие типы моделей (рис.4.7).

Прогноз по уравнению третьей степени можно получить, используя рисунок расположенный ранее: Y= 0,0015*T^3 - 8,7017*T^2 + +17292*T - 1E+07, где T год. Но лучше (так как погрешность расчётов меньше) Y=0,0015*t^3 - 0,102*t^2 + 1,6043*t + 18,838, где t - порядковый номер года (для прогноза на 2010 - 2013 годы, t=36,37,38,39).

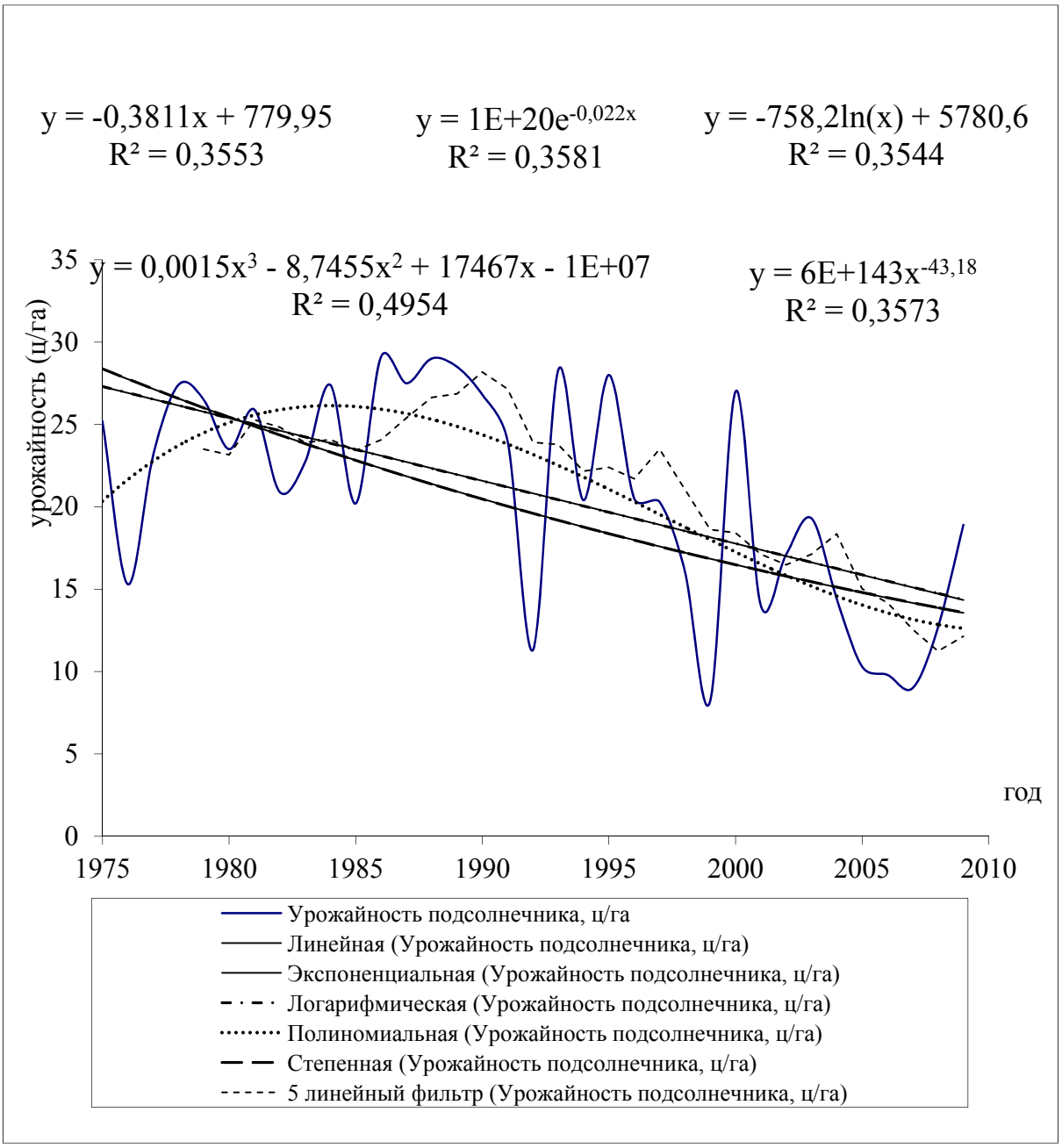


Рисунок 4.6 - Линии тренда

Выравнивание и прогноз по СС - скользящим средним проведём с использованием пакета анализа (инструмент Скользящее среднее).

Взвешенную скользящую среднюю (ВСС) для 5 точек определим для полинома 3 степени (модуль 4) по формуле:

$$z_0 = \frac{1}{35}(-3z_{-2} + 12z_{-1} + 17z_0 + 12z_1 - 3z_2)$$

Для прогнозирования и определения последних 2-х точек, которые нельзя получить с использованием ВСС, с помощью контекстного меню можно определить уравнение полинома третьей степени по последним 5 точкам (предварительно изобразив их на отдельной диаграмме): $y = 0{,}2333k^3 - 0{,}8214k^2 - 0{,}1548k + 11{,}14$, где k - порядковый номер года (для прогноза на 2000-2003 годы k=6,7,8,9).

	A	B	C	D	E	F	G	H	I
1	Год	Подсолнечник	Прогноз по линейному уравнению	Прогноз по уравнению полинома 3степени	СС 3точки	СС 4точки	СС 5точек	СС 11точек	ВСС 5точек
2	1975	25,2	23,5	20,3	-	-	-	-	-
3	1976	15,3	23,1	21,7	-	-	-	-	-
4	1977	23,1	22,7	22,8	21,2	-	-	-	21,4
5	1978	27,4	22,3	23,7	21,9	22,8	-	-	27,0
6	1979	26,5	21,9	24,5	25,7	23,1	23,5	-	26,1
7	1980	23,5	21,6	25,1	25,8	25,1	23,2	-	25,2
8	1981	25,9	21,2	25,6	25,3	25,8	25,3	-	23,6
9	1982	20,9	20,8	25,9	23,4	24,2	24,8	-	22,5
10	1983	22,7	20,4	26,1	23,2	23,3	23,9	-	23,6
11	1984	27,4	20,0	26,1	23,7	24,2	24,1	-	23,7
12	1985	20,2	19,7	26,1	23,4	22,8	23,4	23,5	24,9
13	1986	29,1	19,3	25,9	25,6	24,9	24,1	23,8	25,7
14	1987	27,5	18,9	25,7	25,6	26,1	25,4	24,9	29,1
15	1988	29,0	18,5	25,3	28,5	26,5	26,6	25,5	28,5
16	1989	28,5	18,1	24,9	28,3	28,5	26,9	25,6	28,6
17	1990	26,8	17,8	24,4	28,1	28,0	28,2	25,6	27,6
18	1991	24,0	17,4	23,8	26,4	27,1	27,2	25,6	19,9
19	1992	11,3	17,0	23,2	20,7	22,7	23,9	24,3	19,4
20	1993	28,3	16,6	22,5	21,2	22,6	23,8	25,0	20,2
21	1994	20,4	16,2	21,8	20,0	21,0	22,2	24,8	26,5
22	1995	28,0	15,8	21,1	25,6	22,0	22,4	24,8	23,5
23	1996	20,5	15,5	20,3	23,0	24,3	21,7	24,9	23,4
24	1997	20,3	15,1	19,5	22,9	22,3	23,5	24,1	19,3
25	1998	16,1	14,7	18,8	19,0	21,2	21,1	23,0	13,5
26	1999	8,2	14,3	18,0	14,9	16,3	18,6	21,1	15,8
27	2000	27,0	13,9	17,3	17,1	17,9	18,4	21,0	17,9
28	2001	14,0	13,6	16,5	16,4	16,3	17,1	19,8	19,6
29	2002	17,1	13,2	15,8	19,4	16,6	16,5	19,2	16,2
32	2005	10,3	12,0	14,0	14,7	15,3	15,0	17,7	10,9
33	2006	9,8	11,7	13,6	11,5	13,5	14,2	16,1	9,1
34	2007	9,0	11,3	13,2	9,7	10,9	12,6	15,0	9,6
35	2008	12,7	10,9	12,8	10,5	10,5	11,2	14,4	12,3
36	2009	18,9	10,5	12,6	13,5	12,6	12,1	14,6	19,0
37	2010	13,3	10,1	12,5	15,0	13,5	12,7	15,1	31,0
38	2011	13,0	9,7	12,5	15,1	14,5	13,4	13,8	49,8
39	2012	12,7	9,4	12,6	13,0	14,5	14,1	13,7	76,8
40	2013	12,4	9,0	12,9	12,7	12,8	14,1	13,3	113,3

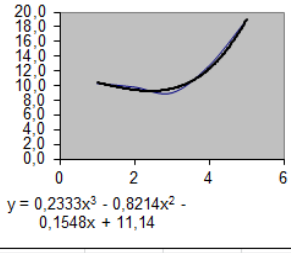


Рисунок 4.7 - Прогнозирование по линиям тренда

В качестве модели прогноза целесообразно выбрать модель, имеющую наименьшую ошибку в последней точке (у нас это модель линейной зависимости и 11- летней скользящей средней), исходя из предположения о сохранении тенденции в дальнейшем.

Проверим автокорреляцию временного ряда (корреляцию между соседними значениями ряда, то есть, влияют ли предыдущие значения на последующие).

Для этого расположим рядом пять столбцов урожайности, сдвинутых каждый относительно предыдущего на единицу (в случае предположения о стационарности ряда смещённые данные каждого столбца можно перенести в его начало).

В результате применения инструмента Корреляция (Пакет анализа) получим корреляционную матрицу (табл. 4.1).

Так как соответствующие коэффициенты автокорреляции значительно отличны от нуля по сравнению с другими. Его легко построить, используя аппарат регрессионного анализа (однако, результат будет не совсем достоверным потому, что не выполняется пятое условие применимости МНК - отсутствие автокорреляции (модуль 4), адекватное решение можно получить, решая систему уравнений Юла - Уокера [6]).

Таблица 4.1 - Корреляционная матрица

	Столбец 1	Столбец 2	Столбец 3	Столбец 4	Столбец 5
Столбец 1	1				
Столбец 2	0,39447398	1			
Столбец 3	0,4662682	0,39447398	1		
Столбец 4	0,3200895	0,4662682	0,394474	1	
Столбец 5	0,19072011	0,3200895	0,4662682	0,394474	1

Очевидно, что уравнение авторегрессии будет иметь вид:

$$Y_t = a_0 + a_1 Y_{t-1} + a_2 Y_{t-2}$$

Рассмотрим получение основных параметров распределения на примере урожайности подсолнечника (ц/га).

Выполним команду Анализ данных - Описательная статистика, заполним параметры диалогового окна (при уровне значимости α=0,05). В результате получим первые два столбца табл. 4.2 , в третьем столбце нами указаны принятые обозначения для статистик.

Таблица 4.2 - Описательная статистика

Урожайность подсолнечника, ц/га		Принятые обозначения
Среднее	20,81714286	$\overline{X} = \sum x_i n_i / n$
Стандартная ошибка	1,10738913	$S_{\overline{X}} = S / \sqrt{n}$
Медиана	20,9	Me
Мода	27,4	Mo
Стандартное отклонение	6,551402441	S
Дисперсия выборки	42,92087395	$S^2 = \sum (x_i - \overline{X})^2 n_i / (n-1)$
Эксцесс	-1,015633106	Ex
Асимметричность	-0,479752816	Sk
Интервал	20,9	W=xmax - xmin
Минимум	8,2	xmin
Максимум	29,1	xmax
Сумма	728,6	Σxi
Счет	35	n=Σni
Наибольший(1)	29,1	-
Наименьший(1)	8,2	-
Уровень надежности(95,0%)	2,250483999	$\Delta = t_{\alpha;n-1} S_{\overline{X}}$

Проведем выравнивание по ряду Фурье для кукурузы (z_t).

Предварительно введя данные и перейдя к условным годам, введём формулы (рис. 4.8):

C2:=(ATAN(1)*8/35)*(A2-1);

D2:=COS(C2)*B2; E2:=COS(2*C2)*B2;

F2:=COS(3*C2)*B2; G2:=COS(4*C2)*B2;

H2:=SIN(C2)*B2; I2:=SIN(2*C2)*B2;

J2:=SIN(3*C2)*B2; K2:=SIN(4*C2)*B2.

Формулу в ячейке C2 скопируем для диапазона C3:C42 (последние пять ячеек соответствуют пяти годам (2000-2004) на которые мы будем давать прогноз); формулы диапазона D2:K2 можно выделить и протащив мышью маркер заполнения скопировать для диапазона C3:K36.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**n**	**Кукуруза**	t	cost	cos2t	cos3t	cos4t	sint	sin2t	sin3t	sin4t	y
2	1	70,3	0,000	70,300	70,300	70,300	70,300	0,000	0,000	0,000	0,000	47,428
3	2	44,2	0,180	43,490	41,382	37,943	33,286	7,892	15,531	22,670	29,081	53,175
4	3	48,4	0,359	45,314	36,449	22,935	6,497	17,007	31,844	42,621	47,962	52,411
5	4	50,9	0,539	43,695	24,120	-2,284	-28,041	26,107	44,822	50,849	42,480	45,434
6	5	40,3	0,718	30,349	5,410	-22,201	-38,848	26,515	39,935	33,633	10,721	35,153
7	6	22,5	0,898	14,029	-5,007	-20,272	-20,272	17,591	21,936	9,762	-9,762	25,855
8	7	20	1,077	9,477	-11,018	-19,919	-7,861	17,612	16,691	-1,793	-18,391	21,356
9	8	16,3	1,257	5,037	-13,187	-13,187	5,037	15,502	9,581	-9,581	-15,502	23,374
10	9	34,6	1,436	4,644	-33,353	-13,599	29,702	34,287	9,205	-31,816	-17,746	30,869
11	10	56,9	1,616	-2,553	-56,671	7,638	55,986	56,843	-5,100	-56,385	10,160	40,631
12	11	20,5	1,795	-4,562	-18,470	12,782	12,782	19,986	-8,895	-16,028	16,028	48,822
13	12	65,9	1,975	-25,900	-45,541	61,698	-2,957	60,597	-47,632	-23,156	65,834	52,732
30	29	52,7	5,027	16,285	-42,635	-42,635	16,285	-50,121	-30,976	30,976	50,121	39,832
31	30	11,9	5,206	5,639	-6,556	-11,852	-4,677	-10,479	-9,931	1,067	10,942	32,228
32	31	30,1	5,386	18,767	-6,698	-27,119	-27,119	-23,533	-29,345	-13,060	13,060	24,702
33	32	32,9	5,565	24,776	4,416	-18,125	-31,714	-21,646	-32,602	-27,457	-8,753	20,201
34	33	35,7	5,745	30,647	16,917	-1,602	-19,667	-18,311	-31,437	-35,664	-29,794	21,013
35	34	6,2	5,924	5,805	4,669	2,938	0,832	-2,179	-4,079	-5,460	-6,144	27,482
36	35	24,6	6,104	24,205	23,031	21,118	18,526	-4,392	-8,644	-12,617	-16,185	37,547
37	36	1558	6,283	-232,344	93,608	122,577	67,146	-55,895	106,688	28,242	144,008	47,428
38	37	44,51428571	6,463	-13,277	5,349	7,004	3,837	-3,194	6,096	1,614	8,229	53,175
39	38	a0	6,642	a1	a2	a3	a4	b1	b2	b3	b4	52,411
40	39		6,822									45,434
41	40		7,001									35,153
42	41		7,181									25,855

Рисунок 4.8 - Экстраполяция урожайности кукурузы (по ряду Фурье)

Просуммируем элементы диапазонов B2:B36 и D2:K36 по столбцам (для этого достаточно выделить диапазон и нажать Σ - автосумма).

Введём в ячейке B38 "=B37/35", а в D38 "=(2/35)*D37" и скопируем последнюю формулу для диапазона E38:K38, в результате получим коэффициенты разложения Фурье:

$$z_t = a_0 + \sum_{j=1}^{k} (a_j \cos jt + b_j \sin jt)$$

(в настоящем примере рассматривается $k = 4$ – четыре гармоники), где

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} z_t, \quad a_j = \frac{2}{n}\sum_{t=1}^{n} z_t \cos jt, \quad b_j = \frac{2}{n}\sum_{t=1}^{n} z_t \sin jt.$$

В ячейку L2 введем формулу (*), которая у нас будет иметь вид:

"=$B$38+$D$38*COS(C2)+$E$38*COS(C2*2)+$F$38*COS(C2*3)+$G$38*COS(C2*4)+$H$38*SIN(C2)+$I$38*SIN(2*C2)+$J$38*SIN(3*C2)+$K$38*SIN(4*C2)"

.

Скопируем формулу из L2 для диапазона L3:L42, в результате в диапазоне L2:L36 получим аппроксимацию с помощью ряда Фурье (1972-2006 гг.), а в диапазоне L37:L42 - прогноз (экстраполяцию) (2008-2012 гг.) (рис. 4.9).

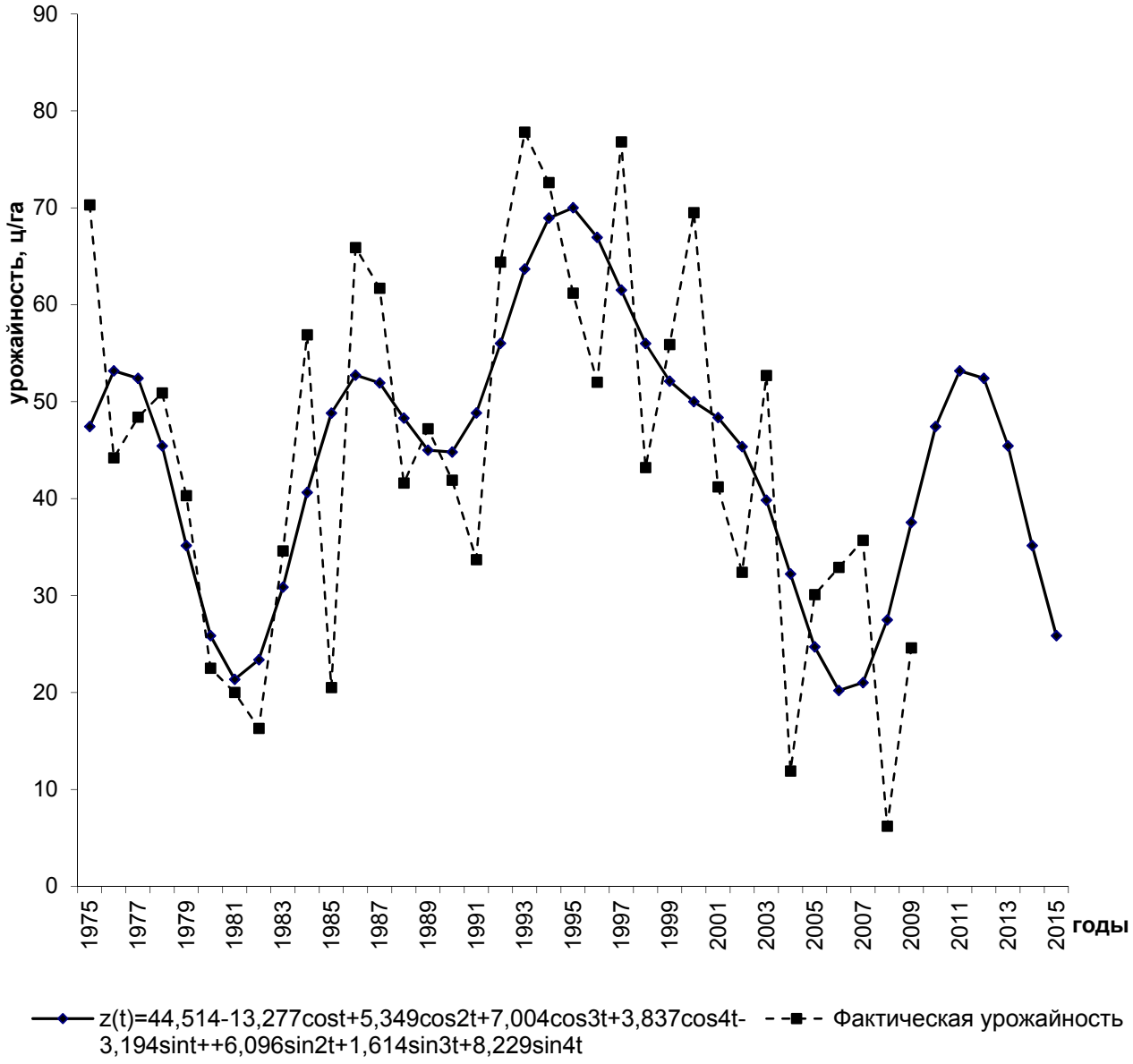


Рисунок 4.9 – Динамика урожайности кукурузы
на зерно в ОАО «Заря»

Построим графики наблюдений (урожайности кукурузы) и её аппроксимации с помощью ряда Фурье. Для этого используем кнопку Мастер диаграмм, расположенную на панели инструментов Стандартная. Выберем График с маркерами, введём диапазоны соответствующих значений (чтобы подписи оси X были годами необходимо ввести ссылку ("Подписи оси X") на диапазон, содержащий соответствующие года.

Задание

По данным приложения провести анализ и прогнозирование временных рядов с использованием различных моделей. Сделать вывод.

Вопросы для самоконтроля

- С какой целью проводится анализ временных рядов.
- Как может быть выявлена основная тенденция в изменениях уровней временных рядов.
- Как выполнить прогноз на будущее с помощью Excel 2007.
- Что понимается под интерполяцией и экстраполяцией.

Практическое занятие№ 5

Финансовые вычисления

Цель работы: Дать целостную концепцию количественного финансового анализа условий и результатов финансово-кредитных и коммерческих сделок, связанных с предоставлением денег в долг.

Теоретические сведения

Финансово-экономические расчеты - это область знаний, в которой излагается методология количественного финансового анализа условий и результатов финансово-кредитных и коммерческих сделок. Они представляют собой совокупность методов и приемов определения изменения стоимости денег, происходящего вследствие их возвратного движения.

Процент рассматривается как плата за пользование заемными средствами, так и показатель доходности любого вложения капитала.

Существует два метода начисления процентов: декурсивный и антисипативный.

Пример 5.1Расчет наибольшей суммы кредита, который, может Вам выдать банк, проводится по формуле

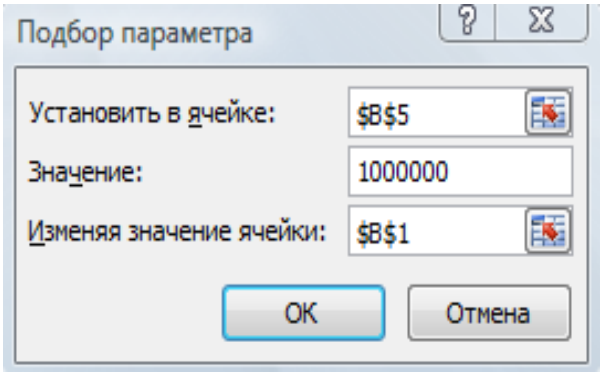
$$\text{Максимальный кредит} = \frac{\text{заработная плата} \times \text{срок кредита (мес)} \times 0{,}6}{1 + \frac{(\text{срок кредита (мес)}+1)\times\text{процент.ст.}}{2\times12\times100}}$$

1) У Вас заработная плата 25000 руб. Найдите наибольшую сумму кредита. Заполним лист Excel

	A	B
1	заработная плата, руб.	25000
2	срок кредита, мес.	60
3	процентная ставка, %	19
4		
5	Максимальная сумма кредита	=(B1*B2*0,6)/(1+(B2+1)*B3/(2*12*100))

В результате получим, что наибольшая сумма кредита на пять лет 606912 руб.

2) Найдем необходимую ежемесячную заработную плату для получения кредита 1000000 руб. на три года. Для этого изменим срок кредита на 36 мес., на вкладке Данные в группе Работа с данными выберем команду Анализ «что-если», а затем – в списке пункт Подбор параметра и заполним параметры диалогового окна.



В результате получим необходимую заработную плату в размере 41192,12 руб.

Пример 5.2 01.03.2010г. Вы хотите взять в банке «АЛЬФА» кредит 100 000 руб. на неотложные нужды под 17 % годовых на 5 лет. Составьте примерный график платежей по кредиту и дайте на его основе оценку эффективной кредитной ставки.

Заполните на лист Excel согласно рисунку 5.1. Обратите внимание:

- одинаковые по внешнему виду формулы вводятся один раз, а затем копируются с помощью маркера заполнения;
- поле дата платежа содержит даты в формате ##.##.####, которые представлены в виде чисел, после включения режима отображения формул (Ctrl+`(левая кавычка, расположенная на одной клавише со знаком ~ «тильда»));
- в формуле процентов нужно учесть, что в високосном году 366 дней;
- $C$2 – абсолютная ссылка на ячейку С2, которая при копировании не изменяется и получается нажатием клавиши F4;
- с помощью справки Excel опишите функцию **ЧИСТВНДОХ( )**.

D2 | 01.03.2010

	A	B	C	D	E	F	G
1	Кредит, руб.	Срок кредита, мес.	Процентная ставка кредита, %	Дата выдачи кредита			
2	100000	60	0,17	40238			
3							
4		Примерный график платежей					
5	Дата платежа	Платеж за расчётный период, ед. валюты				Остаток задолженности по ссуде, ед. валюты	Денежный поток (расходы) получателя ссуды, ед. валюты
6		Сумма платежа	в том числе				
7			Проценты	Погашение основной суммы ссуды	Комиссии и другие платежи		
8	40238	=-A2+E8			1000	=A2	=B8
9	40269	=C9+D9+E9	=(A9-A8)/365*F8*$C$2	=$A$2/$B$2	0	=F8-D9	=B9
10	40299	=C10+D10+E10	=(A10-A9)/365*F9*$C$2	=$A$2/$B$2	0	=F9-D10	=B10
67	42036	=C67+D67+E67	=(A67-A66)/365*F66*$C$2	=$A$2/$B$2	0	=F66-D67	=B67
68	42064	=C68+D68+E68	=(A68-A67)/365*F67*$C$2	=$A$2/$B$2	0	=F67-D68	=B68
69	ИТОГО:		=СУММ(C9:C68)	=СУММ(D9:D6	=СУММ(E		=СУММ(G9:G68)
70	Эффективная процентная ставка:						=ЧИСТВНДОХ(G8:G68;A8:A68)
71						Переплата	=G69/A2-1

Рисунок 5.1 – Примерный график платежей по кредиту в режиме отображения форму

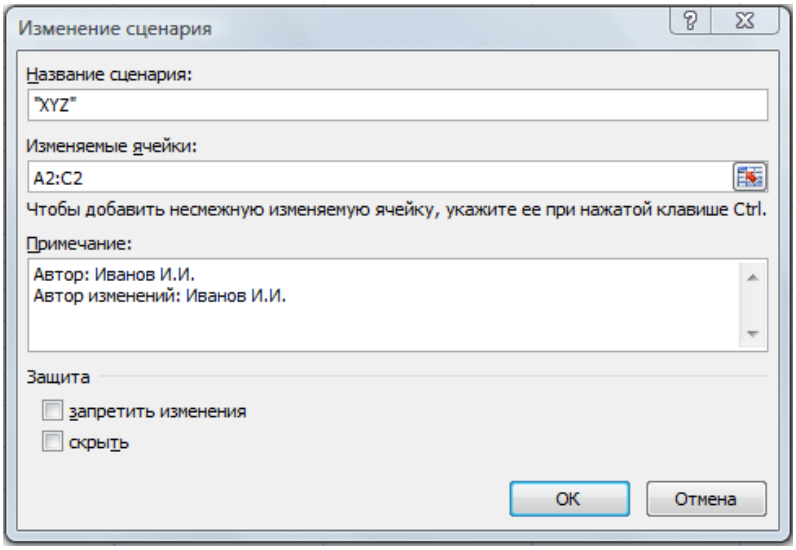
	A	B	C	D	E	F	G
1	Кредит, руб.	Срок кредита, мес.	Процентная ставка кредита, %	Дата выдачи кредита			
2	100000	60	17%	01.03.2010			
3							
4		Примерный график платежей					
5	Дата платежа	Платеж за расчётный период, ед. валюты				Остаток задолженности по ссуде, ед. валюты	Денежный поток (расходы) получателя ссуды, ед. валюты
6			в том числе				
7		Сумма платежа	Проценты	Погашение основной суммы ссуды	Комиссии и другие платежи		
8	01.03.2010	- 99 000,00			1 000,00	100 000,00	- 99 000,00
9	01.04.2010	3 110,50	1 443,84	1 666,67	-	98 333,33	3 110,50
10	01.05.2010	3 040,64	1 373,97	1 666,67	-	96 666,67	3 040,64
67	01.02.2015	1 714,79	48,13	1 666,67	-	1 666,67	1 714,79
68	01.03.2015	1 688,40	21,74	1 666,67	-	0,00	1 688,40
69	ИТОГО:		43 260,35	100 000,00	1 000,00		143 260,35
70	Эффективная процентная ставка:						19,0%
71						Переплата	43%

Рисунок 5.2 – Примерный график платежей по кредиту

Пример 5.3 У Вас есть возможность получить больший кредит в других банках на тот же срок, но с другими процентами:

Банк	**Кредит, руб.**	**Срок кредита, мес.**	**Процентная ставка кредита, %**
«XYZ»	130 000	60	23
«GEO»	150 000	60	27
«И.ГРЕКОВ»	200 000	60	29

Сравните эти предложения и выберите для себя более выгодное.



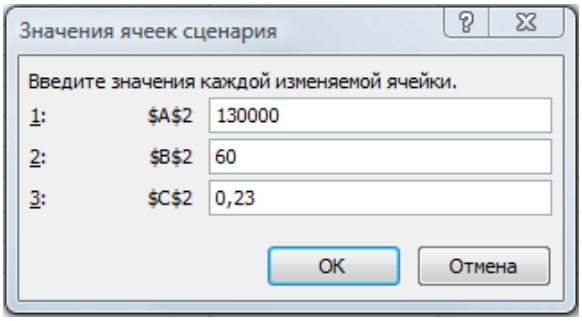


Рисунок 5.3 – Диспетчер сценариев

Для решения этой задачи воспользуемся сценарным подходом.

Сценарий в *Excel* представляет собой некоторое множество исходных значений, предназначенных для подстановки в выбранные зависимости для получения вариантных отчетов.

На вкладке Данные в группе Работа с данными выберите команду Анализ «что-если», а затем выберите в списке пункт Диспетчер сценариев.

Выбрав Добавить введем название сценария «*XYZ*», изменяемые ячейки и затем значения для изменяемых ячеек (рисунок 5.4).

Затем аналогично добавим сценарии «GEO» и «И.ГРЕКОВ» и выберем *Отчет – Тип отчета – Структура* (рисунок 5.4) – ячейки результата G69:G71.

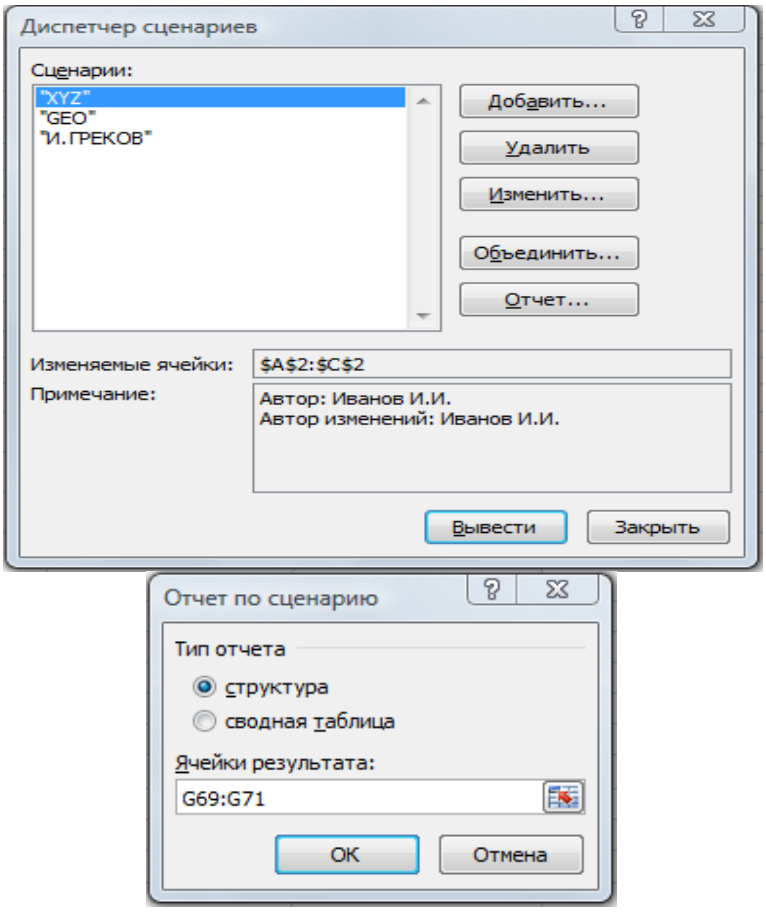


Рисунок 5.4 – Вывод отчета

После выбора OK, получим четыре возможных сценария получения кредита (рисунок 5.5).

Структура сценария					
		Текущие значения:	"XYZ"	"GEO"	"И.ГРЕКОВ"
Изменяемые:					
	$A$2	100000	130000	150000	200000
	$B$2	60	60	60	60
	$C$2	17%	23%	27%	29%
Результат:					
	$G$69	143 260,35	206 087,33	253 061,43	347 594,15
	$G$70	19,0%	26,1%	31,1%	33,6%
	$G$71	43%	59%	69%	74%

Рисунок 5.5 – Сценарии кредитования в разных банках

Переплата при увеличении суммы кредита и процентной ставки естественно растет. Возникает вопрос – какой срок кредитования позволит в других банках переплатить не более 43%?

Задание

1. Вкладка Данные позволяет удалять дубликаты (повторяющиеся строки, например, введите представленную ниже таблицу и выберете Удалить дубликаты. Рассмотрим возможности использования инструмента Консолидация.

1	1
2	2
3	3
4	4
5	5
3	3

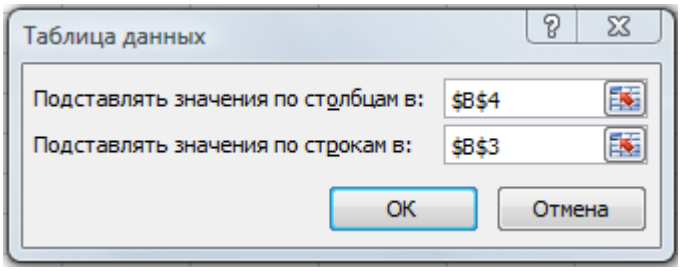
2. По аналогии с примером 12 построить примерный график платежей ипотечного кредита на 20 лет в сумме 1000 000 руб. под 12 % годовых до регистрации ипотеки, и 11 % годовых после регистрации ипотеки (через шесть месяцев). Дайте оценку эффективной кредитной ставки и процент переплаты.

3.(Финансовые функции). Рассмотрим функцию ПЛТ, которая возвращает сумму периодического платежа для аннуитета на основе постоянства сумм платежей и постоянства процентной ставки.

В таблице данных с двумя переменными может быть показано влияние на размер ежемесячных выплат по закладной различных процентных ставок и сроков займа. В следующем примере ячейка C2 содержит формулу вычисления платежа, =ПЛТ(B3/12;B4;-B5), которая ссылается на ячейки ввода B3 и B4.

	A	B	C	D	E
1	Ссуды на недвижимость				
2	Первый взнос	Нет	=ПЛТ(B3/12;B4;-B5)	180	360
3	Процентная ставка	0,095	0,09	=ТАБЛИЦА(B4;B3)	=ТАБЛИЦА(B4;B3)
4	Срок (месяцы)	360	0,0925	=ТАБЛИЦА(B4;B3)	=ТАБЛИЦА(B4;B3)
5	Сумма ссуды	100000	0,095	=ТАБЛИЦА(B4;B3)	=ТАБЛИЦА(B4;B3)

Далее выделим диапазон C2:E5 и выполним команду Данные – Анализ – «что-если» - Таблица данных и заполним параметры диалогового окна:



В результате получим таблицу сумм периодического платежа для аннуитета на основе постоянства сумм платежей и постоянства процентной ставки.

Вопросы для самоконтроля

- Что представляет собой потребительский кредит.
- Сущность ломбардного кредита.
- Практики, используемые при расчете количества дней ссуды.
- Виды финансовых рент

Многомерные статистические методы среди множества возможных вероятностно-статистических моделей позволяют обоснованно выбирать ту, которая наилучшим образом соответствует исходным статистическим данным, характеризующим реальное поведение исследуемой совокупности объектов, оценить надежность и точность выводов, сделанных на основании ограниченного статистического материала.

Дубров А. М., Мхитарян В. С., Трошин Л. И. «Многомерные статистические методы», 1998.

Практическое занятие № 6

Знакомство с системой Statistica 6.0 (краткий обзор пакета)

Цель работы: Ознакомиться с особенностями интерфейса, возможностями настройки и получить навыки ввода данных и вывода результатов анализа. **2**.Ознакомиться с возможностями графического представления данных, получить навыки визуализации данных и редактирования графиков.

Теоретические сведения

Пакет *STATISTICA* 5.5*RU* появился на российском рынке в 1999 г. и с тех пор является одним из лидеров в области визуализации и статистического анализа данных. Имеет полностью русскоязычный интерфейс, контекстную справочную систему, 3000 страниц документации с примерами. Поддерживает все стандарты: импорт из популярных электронных таблиц, публикация результатов в интернете, мастер запросов к *ODBC*-базам данных, язык программирования (*STATISTICA BASIC*) и макрокоманд. Версия *Statistica* 6.1 (русскоязычная) появилась на рынке в 2004году. По сравнению с предыдущей версией вывод результатов упорядочен по структуре и представлен в виде **Рабочих книг** и **Отчетов**, которые содержат в левой части иерархическое оглавление. Все результаты, относящиеся к конкретному виду анализа, помещают в отдельную папку. Может одновременно происходить структуризация информации в рабочей книге и создаваться отчет (который удобно описывать и редактировать). В шестой версии появился язык *Visual Basic*, интегрированный в *STATISTICA* 6, который сохраняет все возможности *STATISTICA BASIC* и *SCL*. *Statistica Visual Basic* предоставляет профессиональную среду для написания собственного приложения. В 6 версии улучшены процедуры импорта данных из файлов различных форматов и баз данных, графические возможности, сняты ограничения на размер текста в названиях, повысилась работоспособность, усовершенствован интерфейс (диалоговые окна разделены на вкладки). Statistica 6 не является модульной по сравнению с 5 версией, где каждый модуль был отдельным *Windows* приложением, что создавало неудобства. Все виды анализа доступны из единой команды меню *Statistics* – Анализы, описание которых можно найти в электронном учебнике или в справке. Для большинства видов анализа снято ограничение на размер файла исходных данных (5000×20000), существовавшее в версии 5.5. *Statsoft* предлагает русскоязычные версии *Statistica* 5.5 и 6.1 различной комплектации. Для версии 6.1:

I.*STATISTICABase for Windows* – базовая версия *STATISTICA* может дополняться следующими видами анализа (которые могут рассматриваться и как отдельные приложения): Углубленные методы анализа, Многомерный разведочный анализ, Промышленная статистика и Шесть сигма, Анализ мощности, Нейронные сети.

После установки *Statistica*, выполнив команду *пуск-программы-Statistica* мы получим пустую электронную таблицу – *Spreadsheet*, либо файл с которым работали в последний раз.

Statistica – система для визуализации и статистического анализа данных, управления базами данных и разработки пользовательских приложений. Пакет содержит широкий набор современных средств анализа данных, объединенных в

свыше 60 модулей (причем многие методы анализа могут быть доступны из разных модулей):

– классических (корреляционно-регрессионный анализ, дисперсионный анализ, кластерный анализ, дискриминантный анализ, факторный анализ, компонентный анализ и др.);

– специальных методов, которые в современной литературе традиционно относят к методам добычи данных – *Data Mining* (нейронные сети, деревья решений и др., а также специализированные процедуры добычи данных в базах данных и текстовой добычи в *Web* и из файлов).

Всего насчитывается более 10 000 вычислительных процедур, сгруппированных по основным направлениям анализа в отдельные модули.

В настоящее время систему Statistica 6.1 можно представить в виде ряда модулей и блоков вычислительных процедур, которые раскрываются из заголовка строки меню **Анализ (*Statistics*)**:

Базовые модули *Statistica* (основные статистики и таблицы, множественная регрессия, дисперсионный анализ, непараметрическая статистика, подгонка распределений) используются для первичного анализа данных.

Углубленные методы анализа содержат различные современные обобщения регрессионного и дисперсионного анализа, логит и пробит регрессию, анализ выживаемости, логлинейный анализ, анализ временных рядов и прогнозироание, моделирование структурными уравнениями.

Многомерный разведочный анализ содержит классические методы многомерного статистического анализа: главных компонент, факторный анализ, дискриминантный и кластерный анализ, деревья классификации, многомерное шкалирование.

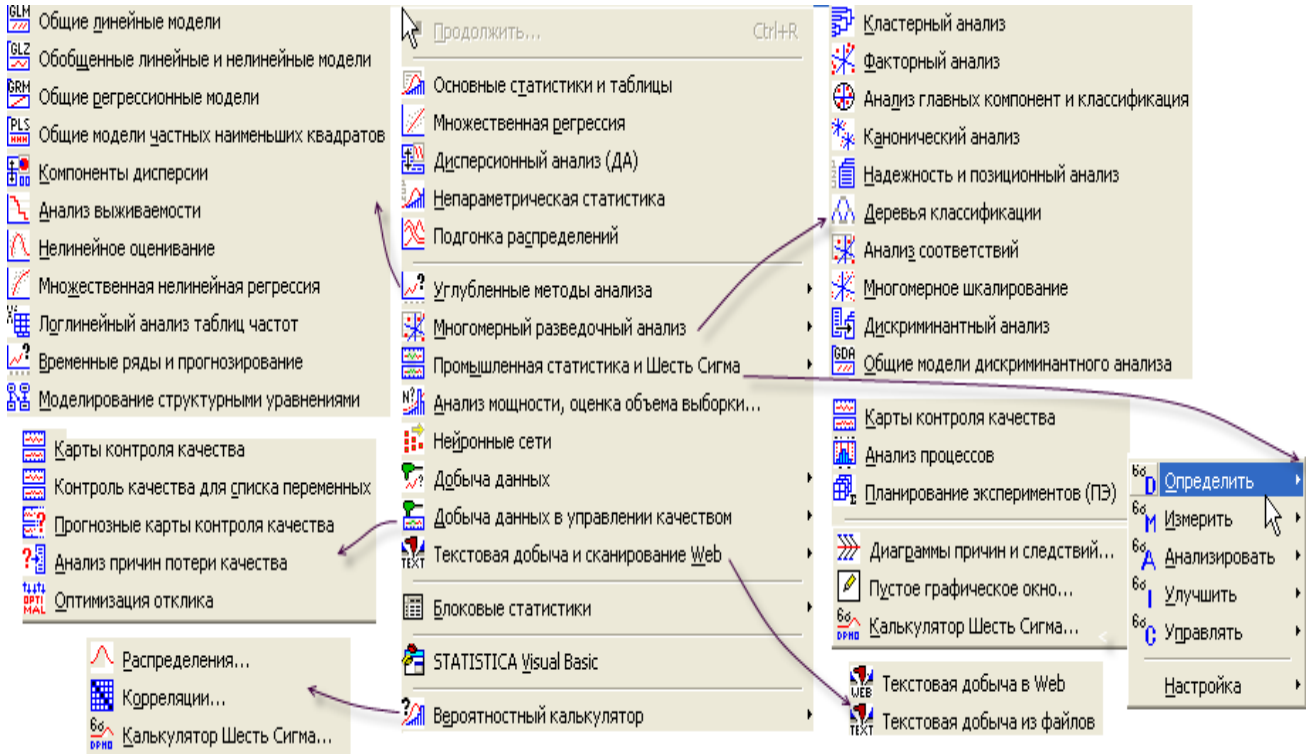


Рисунок 6.1 – Виды анализа данных в системе Statistica 6.1, доступные из команды меню *Statistics* – Анализ

Промышленная статистика(карты контроля качества, анализ процессов, планирование экспериментов).

Анализ мощности, оценка объема выборки в различных схемах выборочного метода и проверки гипотез.

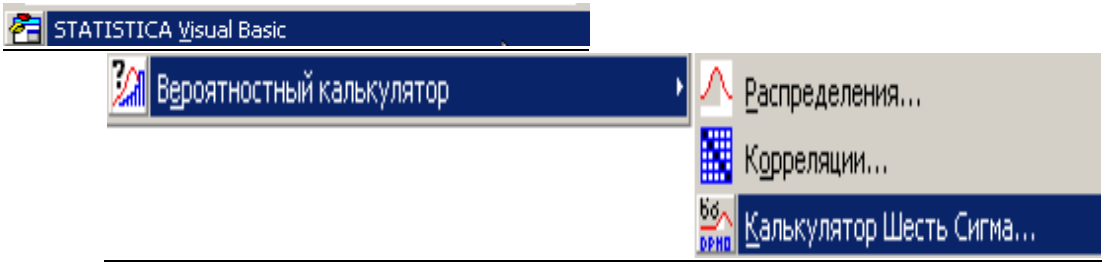
Нейронные сети.

Добыча данных.

Добыча данных в управлении качеством.

Добыча текстов и сканирование *Web*

Блоковые статистики



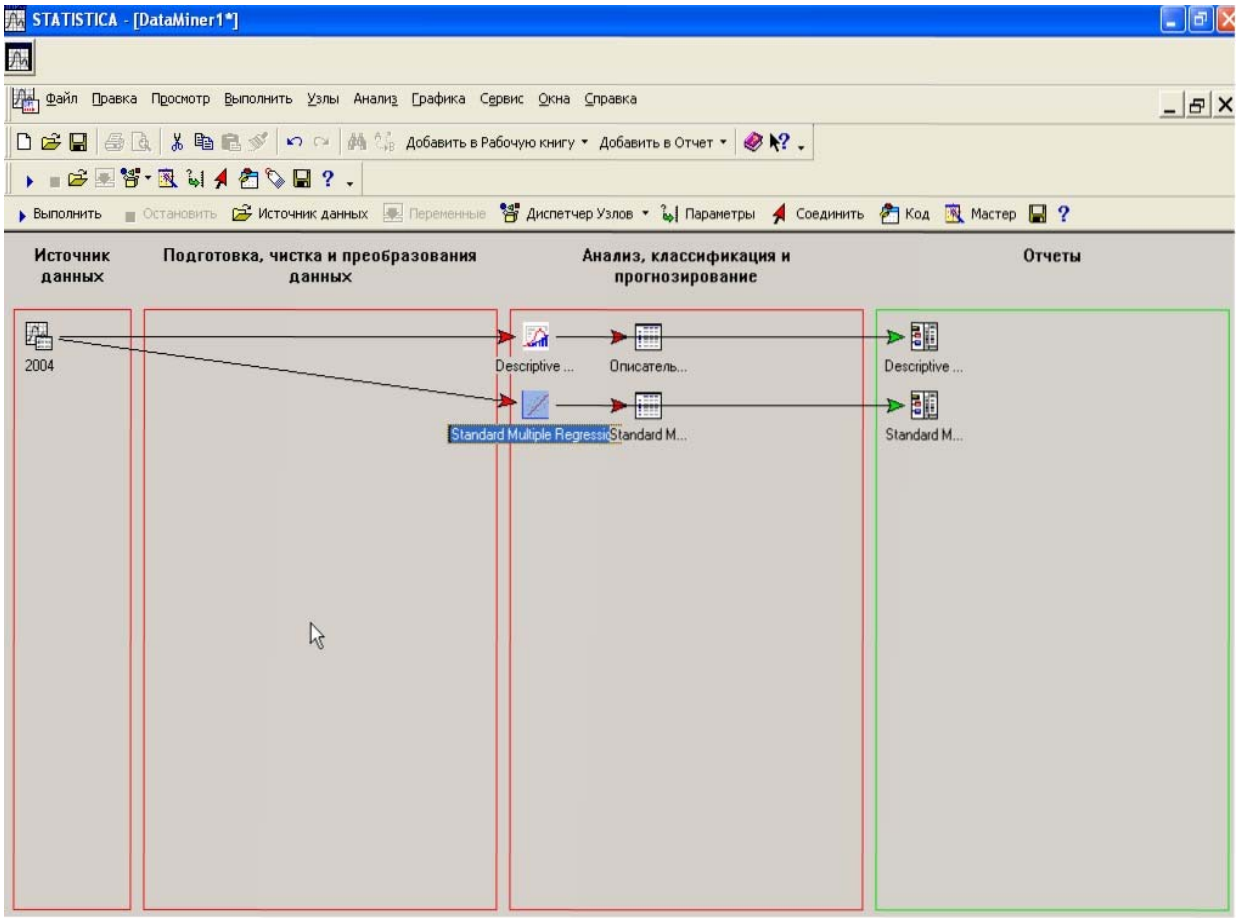


Рисунок 6.2 –Диалоговое окно *STATISTICA Data Miner*

Каждый модуль анализа имеет первые две вкладки, позволяющие проводить первоначальный и более подробный анализ: **Быстрый (*Quick*)** (анализ) и **Допол-**

нительно (***Advanced***). Щёлкнув по кнопке **Переменные** (***Variables***), следует отобрать переменные для анализа (например, в модуле **Множественная регрессия (*Statistics – Multiple Regression*)**: **Зависимые (*Dependent*), Независимые (*Independent*) переменные**. При необходимости выбираются специальные методы обработки и варианты вывода результатов (описание для каждого модуля есть в электронном учебнике и в справке).

II. Корпоративные системы *Statistica*– системы добычи данных, получения зависимостей, прогнозов и классификаций анализа включая современные средства бурения и расслоения, ассоциативные правила обычно используются для работы с большими базами данных. *STATISTICA Enterprise Wide Data Mining System* (*SDM*) – универсальное средство взаимодействия с различными базами данных и создания готовых отчетов, реализующее так называемый графически-ориентированный подход (рис. 6.3) (содержит более 300 основных процедур, специально оптимизированных под задачи *Data Mining*, средства логической связи между ними, управляет потоками данных, позволяет конструировать собственные аналитические методы – рисунок 6.2). *Statistica Enterpri*ce – *Wide Data Analysis System* (*SEDAS*) – многопользовательские системы для решения задач анализа для бизнес приложений в области финансов, маркетинга, а так же для интеграции системы с внешними источниками данных (системами мониторинга, сбора данных в режиме реального времени, измерительными приборами и т.д.). Statistica Enterprice –*Wide SPC System* (*SEWSS*) – локальные и глобальные корпоративные приложения по контролю и улучшению качества, включая методику Шесть Сигма.

Англоязычная *Statistica* 7, предлагаемая сейчас, отличается: внесением дополнений, касающихся возможностей управления данными и их графического изображения; появлением опции «групповой анализ»; дополнительными возможностями экспорта и импорта данных, а также вывода результатов[5].

Продукты Statsoft различают:
по типу (Однопользовательские, Корпоративные, Интернет - технологии);
по применению: Добыча данных (Добытчик данных, Текстовый добытчик, Добытчик качества, нейронные сети, OLAP), Хранилища данных (DataWarehouse), Анализ данных (Углубленные методы, Нейронные сети, Анализ мощности, SEADS, OLAP), Управление документами (Data Warehouse, Document Management System, OLAP), Контроль качества (Карты контроля качества, Анализ процессов, Планирование экспериментов, Добытчик качества, SEWSS).

(Примеры практического применения и другую информацию см. на сайте http://www.statsoft.ru.)

Интерфейс *Statistica.*

Окно системы *Statistica* реализовано согласно стандартам программ, работающих в среде *Windows*.

Пакет *Statistica* 6.0 позволяет выполнять следующие основные функции:

1) загрузить данные представленные в виде файла формата *.sta и просмотреть их в табличном виде (Файл-Открыть).

[5] (http://www.statsoft.ru/_rainbow/documents/NewFeatures_6(7).pdf.)

2) создать файл данных. Задать двойной формат для текстовых переменных.

3) загрузить внешние данные из базы данных, одного из допустимых поставщиков *OLE DB* или из электронных таблиц (*Excel*).

Ввод данных подготовленных в другом приложении можно осуществить следующим образом:

– скопировать и вставить, пользуясь буфером обмена;

– импортировать из файла *Excel* все (выбранные) листы *File-Open-Import all sheets to a Workbook (Import selected sheet to a Spreadsheet);*

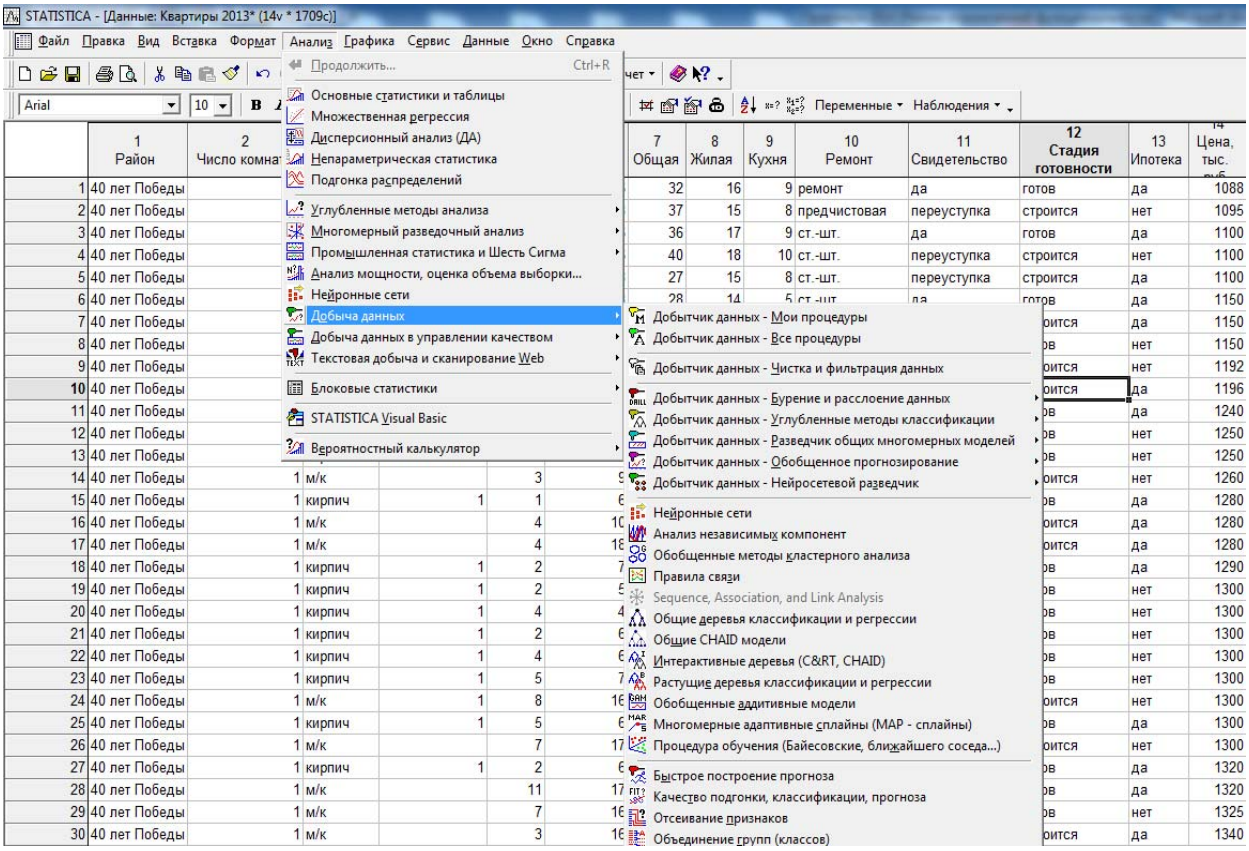


Рисунок 6.3 – Группа методов анализа, объединенных как Добыча данных (*DataMining*) в *Statistica* 6.1

– импортировать с возможностью динамического обмена данными - *Edit-Pastespecial-PasteLink* (либо используя *Edit-DDELinks*);

– импорт файлов из наиболее распространенных баз данных (*Oracle, MSSQLServer, Sybase, MSAccess, FoxPro* и др.) для доступа к данным используется технология *OLE DB*, которая предлагает доступ к большему числу типов данных, чем старая технология ODBC (*Open Data Base Connectivity*).

Доступ к средствам анализа можно осуществить с помощью:

– меню,

– горячих клавиш,

– панелей инструментов,

– пользовательских панелей,

– контекстных меню (вызываемых правой кнопкой мыши).

Программа поддерживает многозадачный режим – есть возможность одновременно работать с несколькими копиями *Statistica*, в каждой из которых можно открыть нескольких анализов одновременно, как над одними, так и над разными данными. В нижней части окна приложения одновременно может представляться несколько функциональных частей называемых «анализами».

Сервис (*Tools*)– настройка позволяет выносить в меню необходимые панели или команды.

Сервис (*Tools*) – параметры (позволяет настроить общие свойства таблиц, графиков, отчетов, рабочих книг и т.д.

Данные (*Data*) – позволяет автоматизировать работу с переменными и наблюдениями.

В *Statistica* 6 можно загрузить (или создать) файл, а затем проводить анализ данных, пользуясь различными средствами анализа (графическими и аналитическими) в одном окне. При этом результаты анализа представляются в виде иерархического дерева, позволяющего иметь доступ к любым результатам и использовать их для дальнейшего анализа (как в Рабочей книге, так и в Отчете).

В системе предлагается три варианта пользовательского интерфейса:

1. Интерактивный.
2. Язык *SVB*.
3. *Web* интерфейс.

Важным моментом является возможность управления выводом данных (Файл-Диспетчер вывода) по трем основным каналам:

1.**Рабочие книги** (***Workboo*k**), в которых автоматически сохраняются все действия с данными в виде графиков и таблиц результатов (в *Statistica* 5.1-5.5 каждый график надо было сохранять отдельно, в шестой версии процедуры работы с результатами анализа значительно упростились). В рабочих книгах результаты анализа (электронные таблицы, графики) представляются в виде вкладок в Рабочей книге.

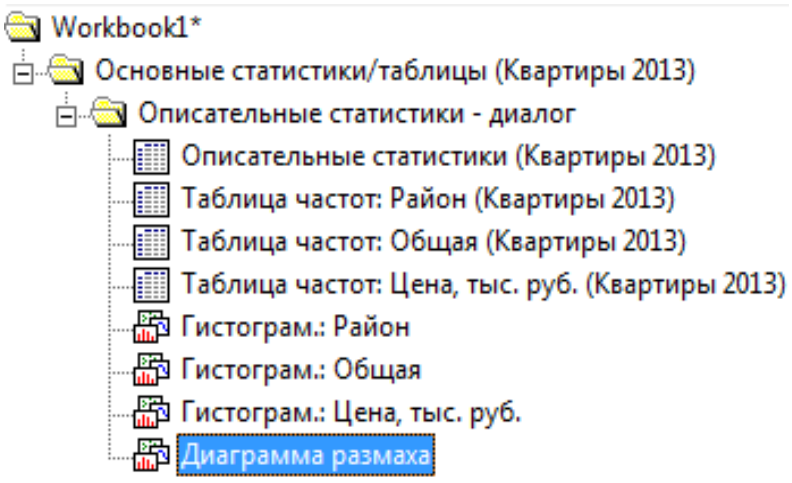


Рисунок 6.4– Дерево просмотра результатов анализа в рабочей книге

Используя дерево просмотра в левой части рабочей книги документы можно организовать в виде иерархии папок или узлов документов (рис.6.4).

Рабочие книги являются оптимизированными *Active-X* контейнерами (документами), что позволяет перенести любую часть дерева в другую Рабочую книгу, Отчет, или в Рабочую область Statistica.

2.**Отчеты** (***Report***) – удобная форма создания отчетов (описания в текстовом режиме таблиц, графиков, моделей анализа, которые импортируются в отчёт – рис.6.5).В предыдущих версиях требовалось для подготовки отчетов копировать таблицы результатов анализа в *Excel*, графики в *Word*, а затем создавать единый документ для описания анализа.

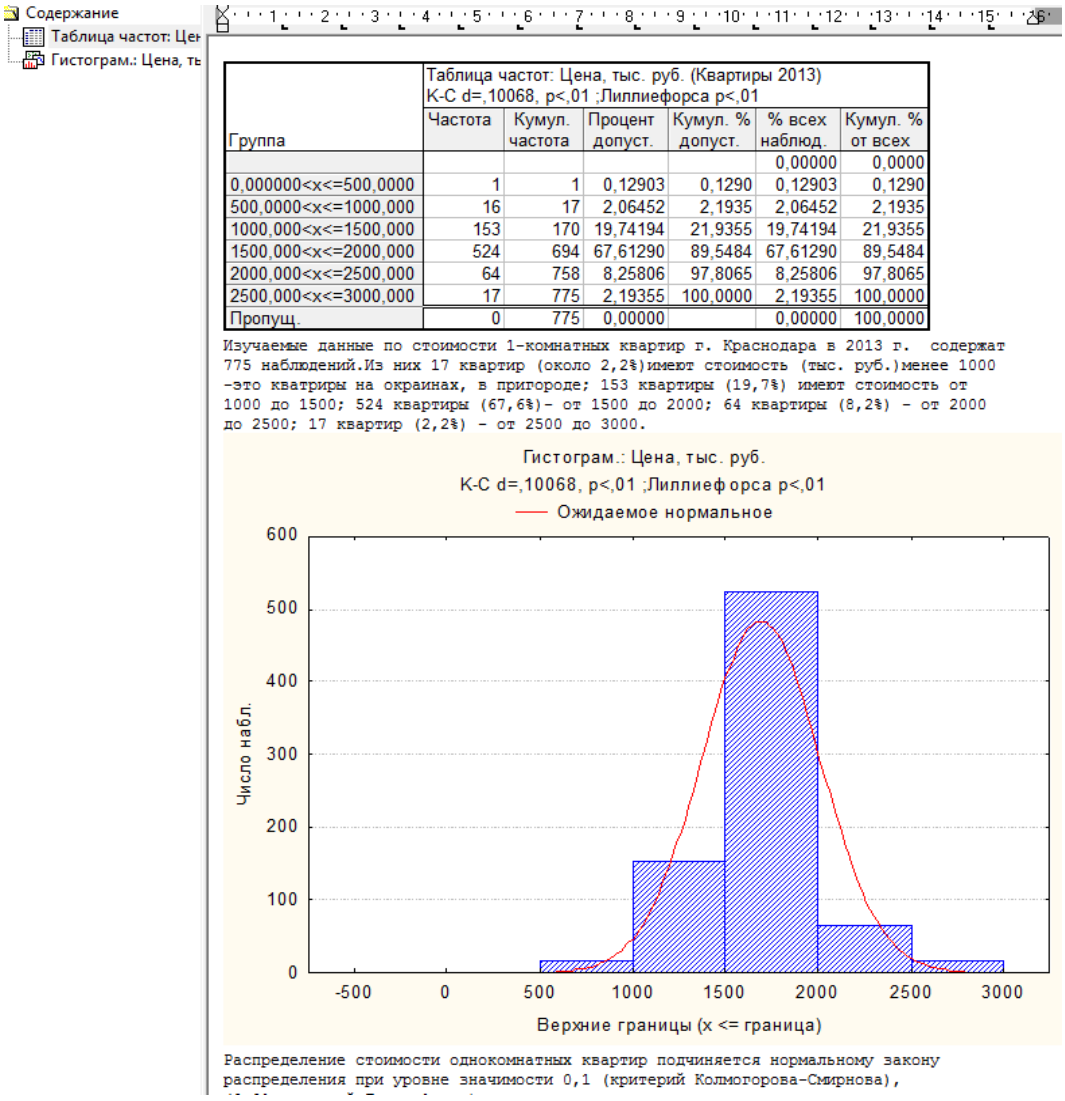


Таблица частот: Цена, тыс. руб. (Квартиры 2013)
K-C d=,10068, p<,01 ;Лиллиефорса p<,01

Группа	Частота	Кумул. частота	Процент допуст.	Кумул. % допуст.	% всех наблюд.	Кумул. % от всех
					0,00000	0,0000
0,000000<x<=500,0000	1	1	0,12903	0,1290	0,12903	0,1290
500,0000<x<=1000,000	16	17	2,06452	2,1935	2,06452	2,1935
1000,000<x<=1500,000	153	170	19,74194	21,9355	19,74194	21,9355
1500,000<x<=2000,000	524	694	67,61290	89,5484	67,61290	89,5484
2000,000<x<=2500,000	64	758	8,25806	97,8065	8,25806	97,8065
2500,000<x<=3000,000	17	775	2,19355	100,0000	2,19355	100,0000
Пропущ.	0	775	0,00000		0,00000	100,0000

Изучаемые данные по стоимости 1-комнатных квартир г. Краснодара в 2013 г. содержат 775 наблюдений.Из них 17 квартир (около 2,2%)имеют стоимость (тыс. руб.)менее 1000 -это кватриры на окраинах, в пригороде; 153 квартиры (19,7%) имеют стоимость от 1000 до 1500; 524 квартиры (67,6%)- от 1500 до 2000; 64 квартиры (8,2%) - от 2000 до 2500; 17 квартир (2,2%) - от 2500 до 3000.

Распределение стоимости однокомнатных квартир подчиняется нормальному закону распределения при уровне значимости 0,1 (критерий Колмогорова-Смирнова),

Рисунок 6.5– Пример отчёта (*Report*) в системе *Statistica*

3.**Автономные окна (*Individual windows*)** – позволяют размещать результаты анализа (таблицы, графики) в отдельных окнах.

Документы системы *Statistica*

1.Рабочие книги.

2.Электронные таблицы (мультимедийные таблицы).

3.Отчеты.

4.Графики.

5.Макросы (программы на языке *SVB*).

Графический анализ данных (визуализация) – это основа разведочного (исследовательского) анализа данных, предложенного Дж.Тьюки в 1962г (смотри стр. 25-27). В Staistica все графики имеют контекстное меню, позволяющее изменять его параметры в том числе и вид подгоняемого распределения. Ряд постоянных настроек графиков доступны в Tools (Сервис) – Options (Параметры).

Первый тип графиков – это **Гистограммы** (**H*istograms***) (термин введен К. Пирсоном в 1895г.), позволяющие увидеть, как распределены значения переменных по интервалам группировки и какому закону распределения они могут соответствовать.

Диаграммы рассеяния (***Scatterplots***) используются для визуального исследования зависимости между двумя переменными, они позволяют находить «выбросы».

Выбросы (нетипичные данные, артефакты– смотри стр. 25-27) искусственным образом занижают или завышают коэффициент корреляции между переменными. При открытой второй вкладке меню диаграммы рассеяния доступны опции вывода на график коэффициента корреляции и уравнения регрессии. Выполнив команду **Сервис - Настройка - Панель инструментов - Графические инструменты** (***Tools – customize – toolbars – Graphtools***), мы можем использовать средство визуального анализа данных Кисть (Brushing)  , которое позволяет интерактивно удалять выбросы и непосредственно наблюдать за изменением аппроксимирующей функции или линии регрессии. Важность графического изучения данных можно проиллюстрировать следующим примером.

Пример 6.1 Зимой 1893-1894гг. Рэлей исследовал плотность азота, полученного различными способами. Измерения Рэлея приведены ниже.

Дата	Исходное вещество	Идентификатор	Очищающий реагент	Вес
29.11.1893	NO	1	Раскаленное железо	2,30143
5.12	NO	1	Раскаленное железо	2,29816
6.12	NO	1	Раскаленное железо	2,30182
8.12	NO	1	Раскаленное железо	2,29890
12.12	Воздух	0	Раскаленное железо	2,31017
14.12	Воздух	0	Раскаленное железо	2,30986
19.12	Воздух	0	Раскаленное железо	2,31010
22.12	Воздух	0	Раскаленное железо	2,31001
26.12	N20	1	Раскаленное железо	2,29889
28.12	N20	1	Раскаленное железо	2,29940
9.01.1894	NH4NO2	1	Раскаленное железо	2,29849
13.01	NH4NO3	1	Раскаленное железо	2,29889
27.01	Воздух	0	Гидроокись железа	2,31024
30.01	Воздух	0	Гидроокись железа	2,31030
1.02	Воздух	0	Гидроокись железа	2,31028

Различия в значениях плотности азота побудило его провести дальнейшие исследования состава воздуха химически очищенного от кислорода. Это привело к открытию Рэлеем нового газообразного элемента аргона и получению за это Нобелевской премии.

Ниже приведены две диаграммы размаха (ящик с усами) для приведенных данных. Главный факт здесь на первой диаграмме – «усы» коротки по сравнению с «ящиком».

Поэтому возникает необходимость более подробного рассмотрения данных. Например, на втором рисунке аналогичные диаграммы для тех же данных классифицированных по признаку «воздух» – 0 и «другие источники» – 1 дают ясную картину о различии в данных.

Схематические диаграммы типа «ящика с усами» не позволяют увидеть, что происходит около середины выборки, для этого используются точечные диаграммы. Рассмотрите оба варианта для предлагаемых данных. Выполните команду[6] **Графика** (***Graphs***) –**2*M* Графики**(***2D Gaphs***): 1) **Диаграммы размаха** (***Box Plots***) – **Переменные** (***Variables***) –**Зависимая переменная** (***Dependent variable***) –Вес – **OK**– получится первый график на рис.6.6.

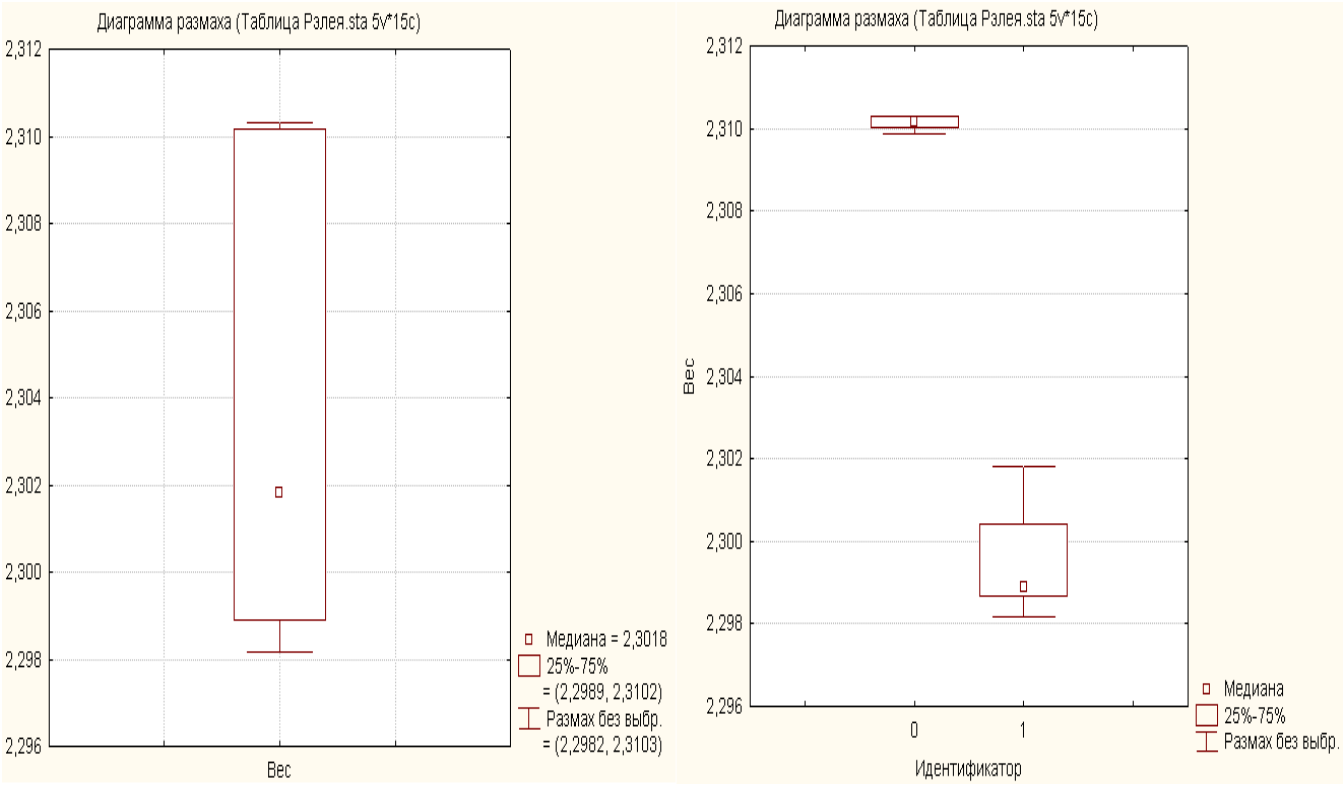


Рисунок 6.6–Диаграммы размаха – «ящик с усами»

[6]Любой пункт любого меню обычно называют *командой*, так как он влечёт за собой выполнение некоторого действия. Словосочетание *Выполнить команду* (*выбрать команду*) означает, что необходимо установить на неё указатель и щёлкнуть левой кнопкой мыши.

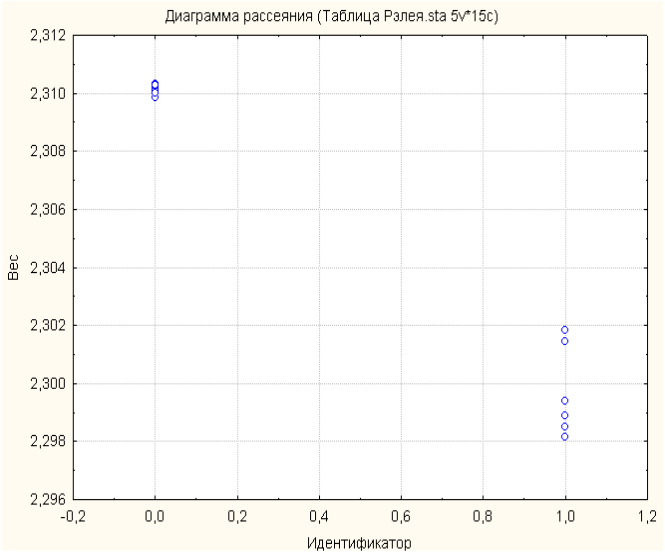


Рисунок 6.7– Диаграмма рассеяния

Затем в качестве **Группирующей переменной** (***Grouping variable***) добавьте-Идентификатор – получится второй график на рис.6.6. 2) **Диаграммы рассеяния** (***Scatterplots***) –**Дополнительно** (***Advanced***) – **Тип графика** (***Graphtype***) – **Простой** (***Regular***) – **Подгонка** (***Fit***) – **выкл** (***off***); **Переменная** (***Variables***) *X* – Идентификатор (либо другая переменная, принимающая постоянное значение), **Переменная** (***Variables***)*Y* –Вес – **OK** – получится график, изображённый на рис.6.7.

Задание

1.Создать файлы, с соответствующими наименованиями переменных начиная с первой переменной (дважды щёлкнув по имени переменной и открыв её свойства): Reklama.sta, (для нахождения площади рекламы, введите в нижнем поле свойств переменной «площадь» – Длинная метка или формула с Функциями формулу: =v1*v2; в этом же поле так же можно описывать переменные); аналогично пределите цену, учитывая, что единица площади стоит 0,9 у.е.

Сохранить файл в своей папке.

A. Reklama1.sta

Ширина	Длина	Площадь	Цена
47	35		
47	73		
47	111		
47	149		
47	187		
47	225		
47	263		
47	301		

2. Описать документы системы Statistica.

3. Изучить калькулятор вероятностных распределений и описать на примерах его работу для *хи*-квадрат Пирсона, экспоненциального, Стьюдента, Фишера и нормального распределений.

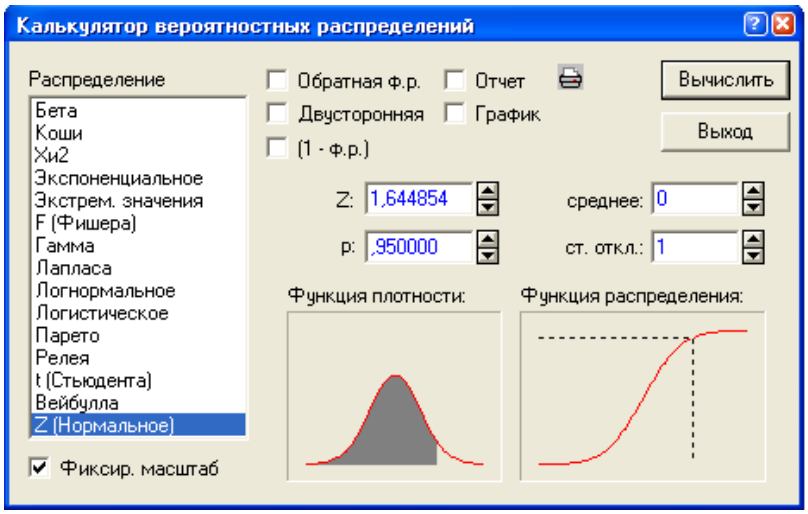


Рисунок 6.8– Вероятностный калькулятор

Найти для нормально распределенной случайной величины при $M(X)=2$, $\sigma=1$ вероятность попадания в интервал: $P(1<Z<5)$.

4.Для вывода результатов анализа в отчёт выполните команду **Сервис – Параметры – Диспетчер вывода – Общий отчет** (***Tools– Options – Output Manager – Single Report***).

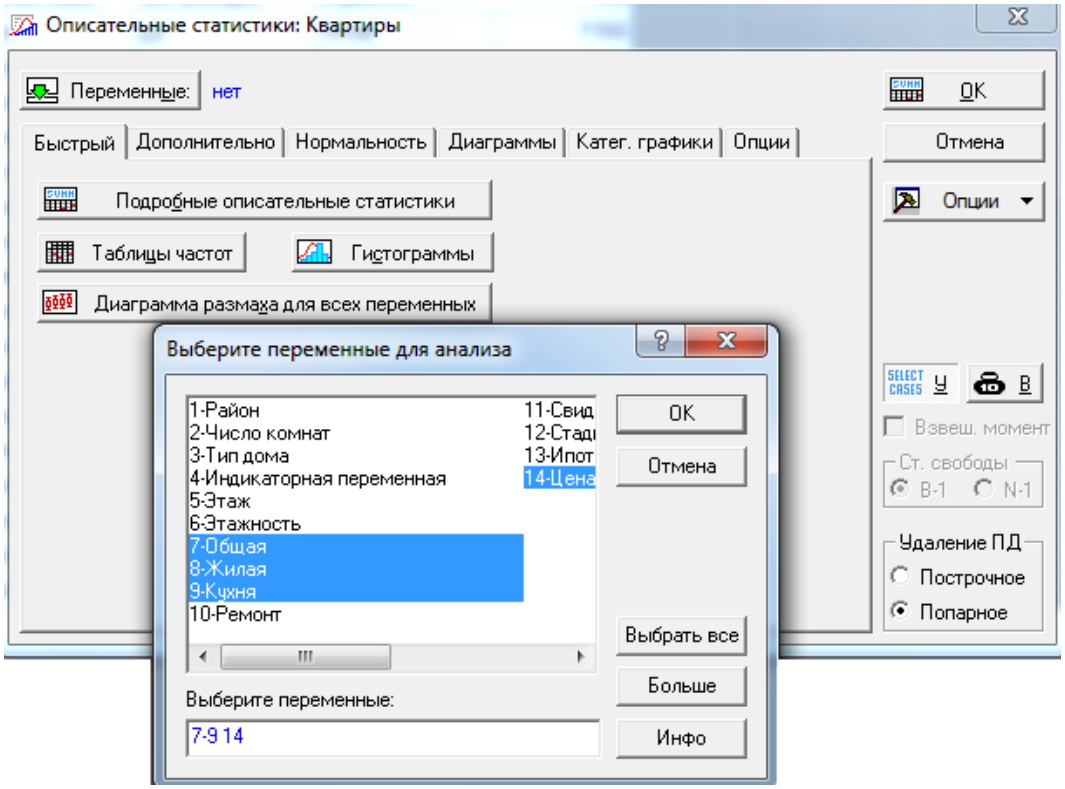


Рисунок 6.9–Модуль Описательные статистики

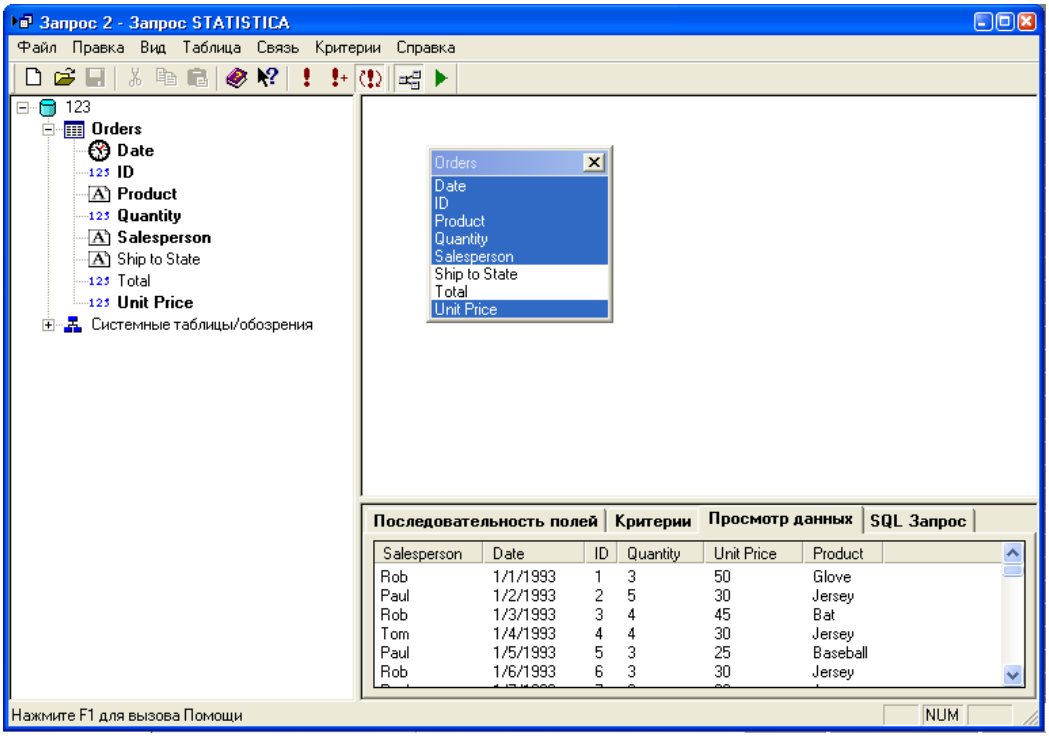


Рисунок 6.10– Окно Запроса Statistica

5. Загрузить файл Квартиры*.xls*, содержащий информацию о квартирах, продававшихся в 2013 году в городе Краснодаре, просмотреть информацию о переменных (предварительно выбрав их с помощью кнопки **Переменные** (***Variables***)), описать с помощью **Основных статистик** (***Basic statistics***) и диаграмм. Результаты вычисления описательных статистик просмотреть в рабочей книге и описать в отчете.

6.Загрузить внешние данные. Для создания связи требуется выполнить команду: *Файл* **– *Внешние данные* – *Создать Запрос* – *Создать***…(новую связь) (***File–Get External Data – Create Query – new***). Далее необходимо выбрать драйвер для организации доступа к базе данных, например, драйвер ***Microsoft Jet* 4.0 OLE DB Provider** (или *Microsoft OLEDB Provider for ODBC Drivers* и имя поставщика базы данных, например, База данных *MS Access*); путь к базе данных, например, *C:\ProgramFiles\Statistica 6.1 \Examples\Examples\Database\baseball*. Выбрать необходимые атрибуты в базе данных (рис. 6.10) и, с помощью кнопки  , импортировать данные в предварительно созданную электронную таблицу.

7. Изучить галерею графиков с помощью электронного учебника (руководства) *Statistica*.

7.1. Изучить самостоятельно, с помощью электронного руководства Statistica, основные типы графиков системы *Statistica* и описать их с условиями применения.

7.2. Загрузить файл *reklama.sta* провести анализ зависимости цены рекламы от длины при фиксированной ширине. Представить данные в виде диаграммы рассеивания с соответствующими заголовками осей и уравнением. Команда: ***Графика – 2М Графики– вкладка Дополнительно – Тип графика Простой – Подгонка Ли-***

нейная – Отметить в группе статистики все элементы –OK (***Graphs – 2D Graphs–Scatterplots–Advanced– Graphtype Regular – FitLinear–Statistics–OK***).

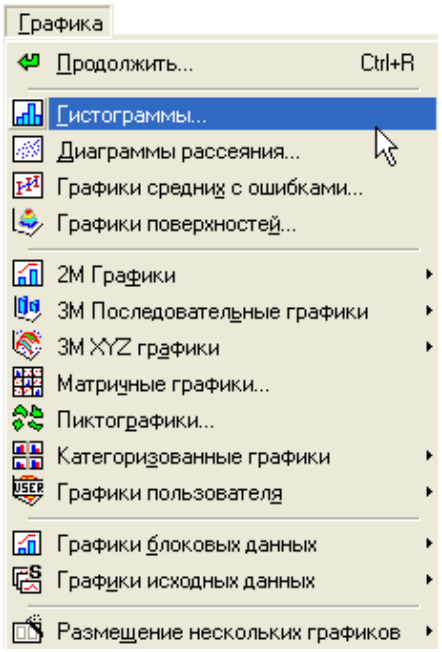


Рисунок 6.11– Галерея графиков (*Graphs*) системы *Statistica*

Вопросы для самоконтроля

- Какие направления анализа данных реализуются в Statistica.
- Опишите возможности ввода и вывода информации.
- Перечислите основные типы графиков и опишите их назначение и особенности.
- Какое назначение инструмента визуального анализа – Кисть.

Практическое занятие №7

Дисперсионный анализ

Цель работы: Ознакомиться с возможностями дисперсионного анализа в системе *Statistica*. Получить навыки дисперсионного анализа данных.

Теоретические сведения

Рассмотрим качественный единичный фактор, который принимает p различных уровней, и предположим, что на каждом уровне сделано n наблюдений, что дает $N=np$ наблюдений. (Ограничимся рассмотрением первой модели дисперсионного анализа – все факторы имеют фиксированные уровни.)

Пусть результаты представлены в виде X_{ij} (i=1, 2…, p; j=1, 2, …, n).

Данные обычно располагают в виде таблицы (табл. 7.1).

Таблица 7.1 – Результаты проведения эксперимента

Номер наблюдения,i	**Уровни фактора**, j			
	A_1	A_2	…	A_p
1	X_{11}	X_{21}	…	X_{p1}
2	X_{12}	X_{22}	…	X_{p2}
3	X_{13}	X_{23}	…	X_{p3}
.	.	.	…	…
.	.	.	…	…
.	.	.	…	…
n	X_{1n}	X_{2n}	…	X_{pn}
ИТОГИ				

Предполагается, что для каждого уровня n наблюдений имеется средняя, которая равна сумме общей средней и ее вариации обусловленной выбранным уровнем:

$$X_{ij} = \mu + A_j + \varepsilon_{ij},$$

где μ - общая средняя;

A_j - эффект, обусловленный j – м уровнем фактора;

ε_{ij} – вариация результатов внутри отдельного уровня фактора. С помощью члена ε_{ij} принимаются в расчет все неконтролируемые факторы.

Пусть наблюдения на фиксированном уровне фактора нормально распределены относительно среднего значения $\mu + A_j$ с общей дисперсией σ^2.

Тогда (точка вместо индекса обозначает усреднения соответствующих наблюдений по этому индексу):

$$X_{ij} - X.. = (X._j - X..) + (X_{ij} - X._j).$$

После возведения в квадрат и суммирования по i и j получим:

$$\sum_{i,j}(X_{ij}-X..)^2=\sum_{i,j}(X._j-X..)^2+\sum_{i,j}(X_{ij}-X._j)^2\ ,$$

так как $\sum_{i,j}(X._j-X..)(X_{ij}-X._j)=\sum_j(X._j-X..)\sum_i(X_{ij}-X._j)$, íî $\sum_i(X_{ij}-X._j)=0.$

Иначе сумму квадратов можно записать: $S = S_1 + S_2$. Величина S_1 вычисляется по отклонениям p средних от общей средней $X..$, поэтому S_1 имеет (p-1) степеней свободы. Величина S_2 вычисляется по отклонениям N наблюдений от p выборочных средних и, следовательно, имеет $N\text{-}p = np - p=p(n\text{-}1)$ степеней свободы. S имеет $(N\text{-}1)$ степеней свободы. По результатам вычислений строится таблица дисперсионного анализа (табл. 7.2).

Таблица 7.2 –Таблица дисперсионного анализа

Источник изменчивости	Суммы квадратов (*SS*)	Степени свободы (*df*)	Средние квадраты (*MS*)
Различия между уровнями	$S_1=n\sum_i(X._j-X_{..})^2$	$p\text{-}1$	$M_1=\frac{S_1}{p-1}$
Различия внутри уровней	$S_2=\sum_{i,j}(X_{ij}-X._j)^2$	$N\text{-}p$	$M_2=\frac{S_2}{N-p}$
Сумма	$S=\sum_{i,j}(X_{ij}-X_{..})^2$	$N\text{-}1$	

Если гипотеза о том, что влияние всех уровней одинаково, справедлива, то обе величины M_1 и M_2 (средние квадраты) будут несмещенными оценками σ^2. Значит, гипотезу можно проверить, вычислив отношение (M_1/M_2) и сравнив его с $F_{кр.}$ с v_1= (p-1) и v_2 = (N-p) степенями свободы.

Если $F_{расч.}>F_{кр.}$, то гипотеза о незначимом влиянии фактора A на результат наблюдений не принимается и требуется определить какие варианты существенно отличаются от остальных, т.е. к каким вариантам относятся существенные разности между уровнями средних. (Если $F_{расч.}< F_{кр,}$ то оценку частных различий не проводят и считают, что разности между парами находятся в пределах ошибки опыта.)

Если гипотеза о равенстве средних отклоняется, то может представлять интерес какие именно группы имеют значимое различие средних. Для этого используются линейные контрасты, которые определяются как линейная комбинация параметров, например, $\beta_1,\beta_2,...,\beta_p$ с весами, сумма которых равна нулю $-Lk=\sum_{j=1}^{p}c_j\beta_j$, где $\sum_{j=1}^{p}c_j=0$. Для пар взаимно независимых оценок $\beta_j(X._j)$ строят доверительные интервалы (при уровне значимости α) для разностей средних и если полученные доверительные интервалы не покрывают нуль, то можно сделать вывод о том, что эти два средних существенно различны.

Оценка линейного контраста $\overline{Lk} = \sum_{j=1}^{p} c_j X_{.j}$, оценка дисперсии $s^2_{\overline{Lk}} = \hat{\sigma}^2 \sum_{j=1}^{p} \frac{c_j^2}{n_j}$.

Границы доверительного интервала для Lk имеют вид:

$$\overline{Lk} \pm \sqrt{(p-1)F_\alpha(p-1, N-p)}\,.$$

Для оценки наименьшей существенности различий в уровне средних (для всех уровней факторов) при $F_{\text{расч.}} > F_{\text{табл.}}$ вычисляют:

а) ошибку опыта $S_{\overline{X}} = \sqrt{\frac{S^2_z}{n}}$,

б) ошибку разности средних $S_d = S_{\overline{X}} \cdot \sqrt{2}$,

в) наименьшую существенную разность $HCP_{\alpha,k} = t_{\alpha,k} \cdot S_d$.

Сравнивая (по вариантам с *HCP*) разности между средними значениями $X_{.j}$ ($j = \overline{2, p}$) и первым (базовым) уровнем, делают вывод о существенности различий в уровне средних. Если фактическая разность больше *HCP*, то она статистически значима, и соответствующий уровень существенно влияет на результат, в противном случае уровень не оказывает статистически существенного влияния на результат.

Замечание. *Практически аналогично рассматривается идеология многофакторного дисперсионного анализа – возрастает лишь сложность вычислений. В дисперсионном анализе предполагается нормальность распределения данных и однородность дисперсий по группам.*

Дисперсионный анализ в *Statistica*

Выполнив команду **Анализ – Дисперсионный анализ** (***Statistics – ANOVA***) мы получим окно дисперсионного анализа. Как видно из рисунка 7.1 модуль дисперсионного анализа позволяет оценивать однофакторные модели (**Однофакторный ДА**), а так же многофакторные без взаимодействия (**Главные эффекты**) и с взаимодействием (**Факторный ДА**), а так же опыты с повторениями (**Повторные измерения ДА**). (Если в модели более четырёх категориальных переменных, то используется модуль ***GLM* – Общих линейных моделей**.)

В пакете *Statistica* факторные признаки задаются отдельными категориальными переменными и их различные сочетания уровней соответствуют результативным (зависимым) переменным (одной или нескольким). При этом уровни факторных признаков могут задаваться как числами (метками) так и категориями – все они перекодируются программой одним из двух способов.

По умолчанию используется сигма-ограниченная модель кодирования переменных, когда сумма уровней равняется нулю. Иначе рассматривается сверхпараметризованная модель – последовательно задающая коды 0, 1 и т.д.

Замечание*. Кроме того, при вычислениях может использоваться один из шести способов образования сумм квадратов*[7]*:*

– сумма квадратов типа I (последовательная) предоставляет разделение предсказанной суммы квадратов для полной модели. Этим свойством не обладает ни один другой тип суммы квадратов. Важ-

[7] Электронное руководство *Statistica* 6.1

ное ограничение для суммы квадратов типа I заключается в том, что сумма квадратов, отнесенная к отдельному эффекту, будет зависеть от порядка включения эффектов в модель;
– сумма квадратов типа II (частная) контролирует влияние других эффектов. В отличие от суммы квадратов типа I сумма квадратов типа II является инвариантной относительно порядка включения эффектов в модель, I и II суммы не рекомендуется использовать для факторных планов с разным числом наблюдений;
– сумма квадратов типа III (ортогональная), отнесенная к некоторому эффекту, вычисляется как сумма квадратов для эффекта, контролируемого для любых эффектов с такой же или меньшей степенью, и ортогональной любым эффектам взаимодействия старшего порядка, которые его содержат. Не рекомендуется использовать для планов с пустыми ячейками;
–сумма квадратов типа IV (оцениваемая) была разработана для проверки "сбалансированных" гипотез для эффектов малого порядка в планах ДА с пропущенными ячейками, однако исследователи не рекомендуют её использовать;
– сумма квадратов типа V (полный ранг) используется в планах ДА с пропущенными ячейками, а так же в дробных факторных планах;
– суммы квадратов типа VI, используют сигма-ограниченное кодирование эффектов категориальных предикторов для получении уникальных оценок эффектов (даже для эффектов малого порядка) гнездовых планов ДА, планов с неоднородными коэффициентами наклона или планов смешанной модели и др. Сумма VI типа обычно задаётся по умолчанию.

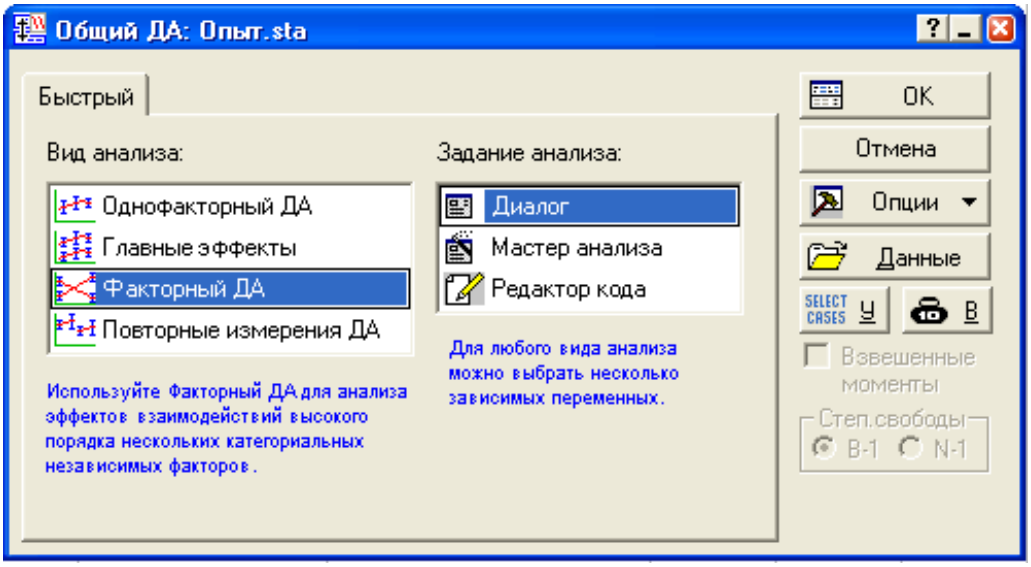


Рисунок 7.1 – Диалоговое окно дисперсионного анализа

Рассмотрим пример трёхфакторного дисперсионного анализа.

Пример 7.1 Создадим файл опыт1.*sta*. Рассматривается учебный пример трёхфакторного опыта 2×2×2 в трёх повторениях (табл. 7.3).

В опыте изучается влияние на урожайность пшеницы следующих факторов:

A – плодородие почвы (0 – исходное плодородие, 2 – 400 т/га навоза + P_{400});

B – система удобрений (0-без применения удобрений, 2 – средняя норма удобрений);

C – система защиты растений от сорняков, вредителей и болезней (0-без применения средств защиты растений, 2-весной в фазе кущения применяли гербицид акрил – *M*).

Выбрав в диалоговом окне (рис.7.1) факторный дисперсионный анализ, получим диалоговое окно (рис.7.2) в котором предлагается выбрать **Зависимые** (***Dependent***) и **Независимые** (***Independent***) переменные.

Таблица 7.3 – Влияние плодородия почвы, системы удобрений, системы защиты растений на урожайность пшеницы

№	A	B	C	Урожайность, ц/га	Тип технологии
1	0	0	0	37,93	экстенсивная
2	0	0	0	42,23	экологически чистая
3	0	0	0	35,63	экстенсивная
4	0	0	2	39,60	экстенсивная
5	0	0	2	41,25	экологически чистая
6	0	0	2	37,31	экстенсивная
7	0	2	0	52,40	экологически чистая
8	0	2	0	57,00	экологически чистая
9	0	2	0	61,28	почвозащитная
10	0	2	2	62,69	почвозащитная
11	0	2	2	63,27	почвозащитная
12	0	2	2	66,34	почвозащитная
13	2	0	0	44,60	экологически чистая
14	2	0	0	50,43	экологически чистая
15	2	0	0	45,73	экологически чистая
16	2	0	2	55,02	экологически чистая
17	2	0	2	53,13	экологически чистая
18	2	0	2	49,05	экологически чистая
19	2	2	0	58,89	экологически чистая
20	2	2	0	60,40	почвозащитная
21	2	2	0	56,82	экологически чистая
22	2	2	2	66,25	почвозащитная
23	2	2	2	71,41	почвозащитная
24	2	2	2	68,74	почвозащитная

В качестве зависимой переменной выберем урожайность, в качестве независимых – факторы *A, B, C*.

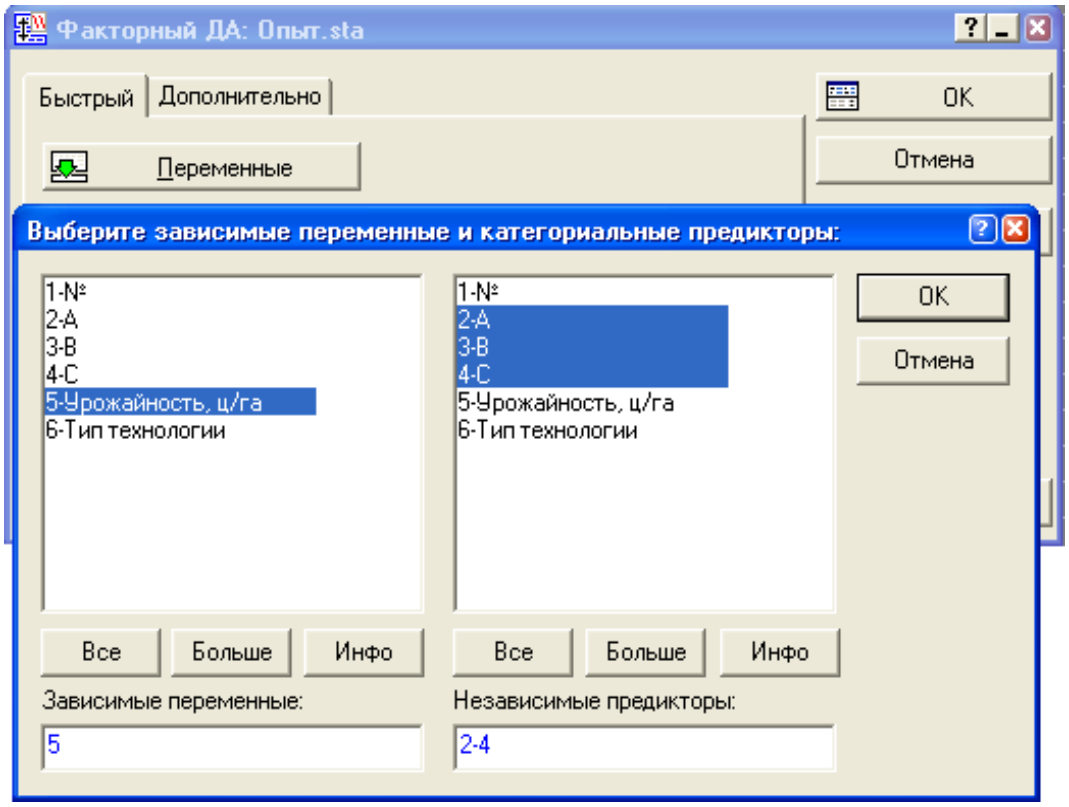


Рисунок 7.2 – Диалоговое окно факторного дисперсионного анализа

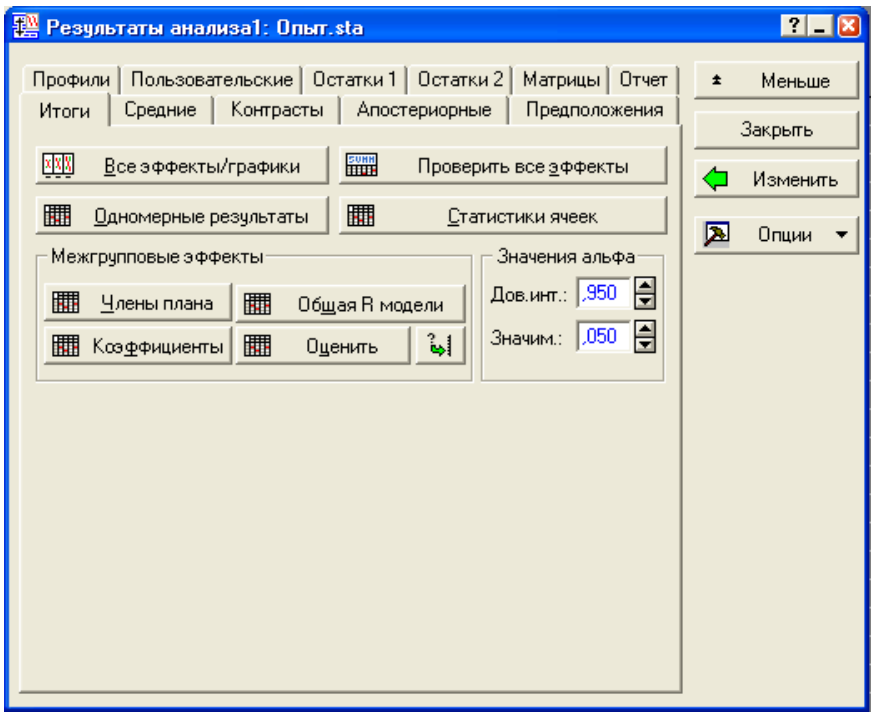


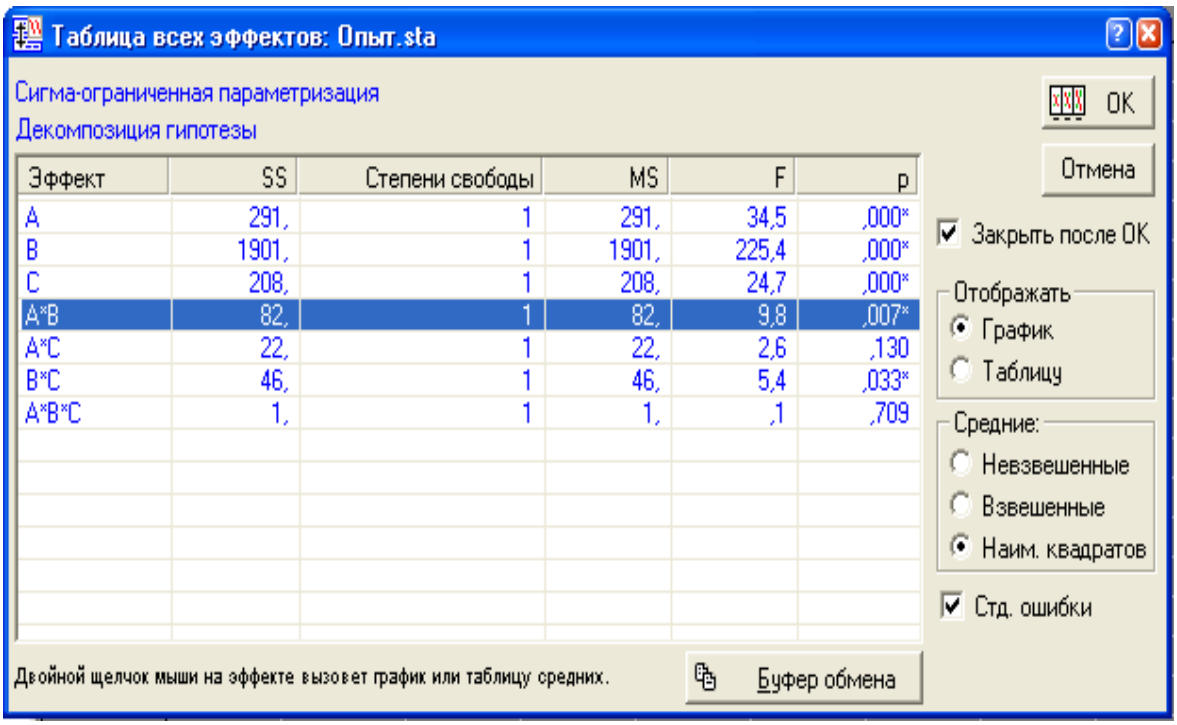
Рисунок 7.3 – Окно анализа результатов: вкладка Итоги

После выбора кнопки **Проверить все эффекты** (***Test all effects***) (рис. 7.3), получим таблицу всех эффектов (рис.7.4), из которой видно, что практически все факторы и их взаимодействия статистически существенно влияют на урожайность (за исключением взаимодействий *AC* и *ABC*) – все значимые эффекты выделяются красным цветом.

Эффект	Одномерный критерий значимости для Урожайность, ц/га (Опыт.sta) Сигма-ограниченная параметризация Декомпозиция гипотезы SS	Степени свободы	MS	F	p
Св. член	67989,62	1	67989,62	8062,050	0,000000
A	290,79	1	290,79	34,481	0,000024
B	1900,68	1	1900,68	225,379	0,000000
C	208,39	1	208,39	24,710	0,000139
A*B	82,44	1	82,44	9,775	0,006508
A*C	21,55	1	21,55	2,555	0,129512
B*C	45,65	1	45,65	5,413	0,033446
A*B*C	1,22	1	1,22	0,144	0,709257
Ошибка	134,93	16	8,43		

Рисунок 7.4 – Таблица всех эффектов

Для визуализации различий урожайности выберем кнопку **Все эффекты/графики** (***All effects /Graphs***), в результате получим диалоговое окно, позволяющее выбирать эффекты и их взаимодействия (рис. 7.5).



Эффект	SS	Степени свободы	MS	F	p
A	291,	1	291,	34,5	,000*
B	1901,	1	1901,	225,4	,000*
C	208,	1	208,	24,7	,000*
A*B	82,	1	82,	9,8	,007*
A*C	22,	1	22,	2,6	,130
B*C	46,	1	46,	5,4	,033*
A*B*C	1,	1	1,	,1	,709

Рисунок 7.5 – Окно Таблица всех эффектов

При выборе взаимодействия факторов *A*B* мы можем получить **Таблицу** (***Spreadsheet***) (рис. 7.6) или **График** (***Graph***) (рис. 7.7), выбрав соответствующие пункты в группе отображать, диалогового окна **Таблица всех эффектов** (***Table of All Effects***). Аналогично получим для взаимодействия факторов *A*C* и *B*C* рисунки 7.5 и 7.6 соответственно.

A*B; МНК средние (Опыт.sta)
Текущ. эффект: F(1, 16)=9,7751, p=,00651
Декомпозиция гипотезы

N ячейки	A	B	Урожайность, ц/га Среднее	Урожайность, ц/га Стд.ош.	Урожайность, ц/га -95,00%	Урожайность, ц/га +95,00%	N
1	0	0	38,99167	1,185558	36,47840	41,50494	6
2	0	2	60,49667	1,185558	57,98340	63,00994	6
3	2	0	49,66000	1,185558	47,14673	52,17327	6
4	2	2	63,75167	1,185558	61,23840	66,26494	6

Рисунок 7.6– Таблица с описательными статистиками по уровням взаимодействия факторов *A* и *B*

На рисунках 7.7-7.9 мы видим средние урожайности: рис.7.7 – по уровням фактора повторных измерений *A* разные и для *B*0, и для *B*2; рис.7.8 – по уровням фактора повторных измерений *A* разные при *A*2 и приблизительно равны при *A*0 для *C*0 и для *C*2; рис.7.9 – по уровням фактора повторных измерений *B* разные при A2 и приблизительно равны при A0 для C0 и для C2.

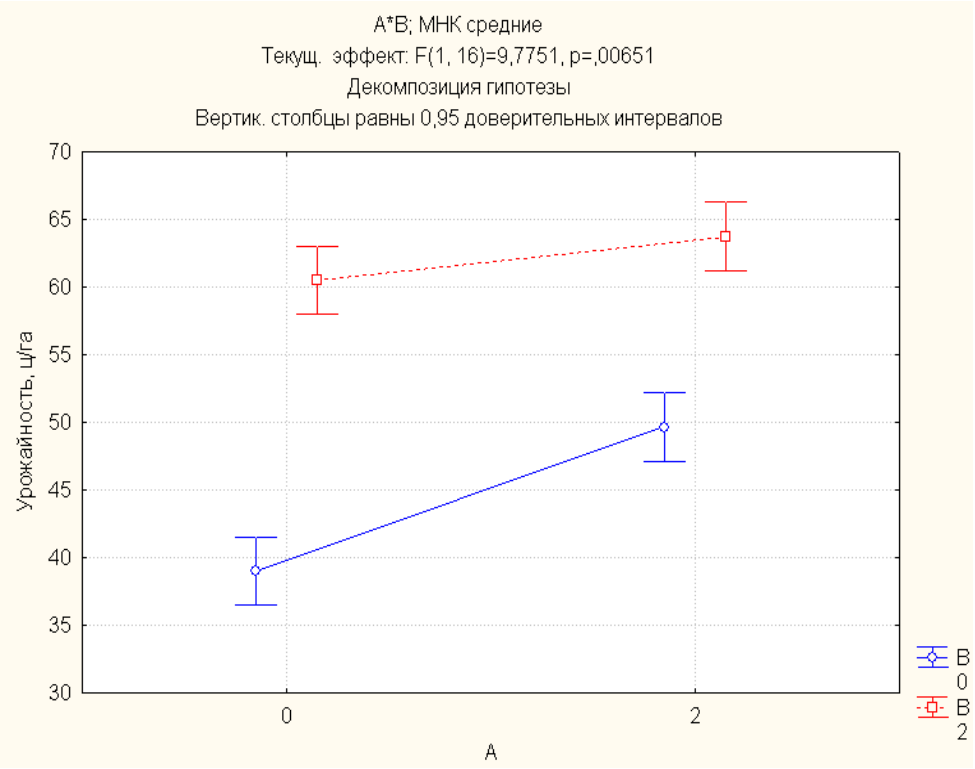


Рисунок 7.7 – График зависимости урожайности от плодородия (*A*) и системы удобрений (*B*)

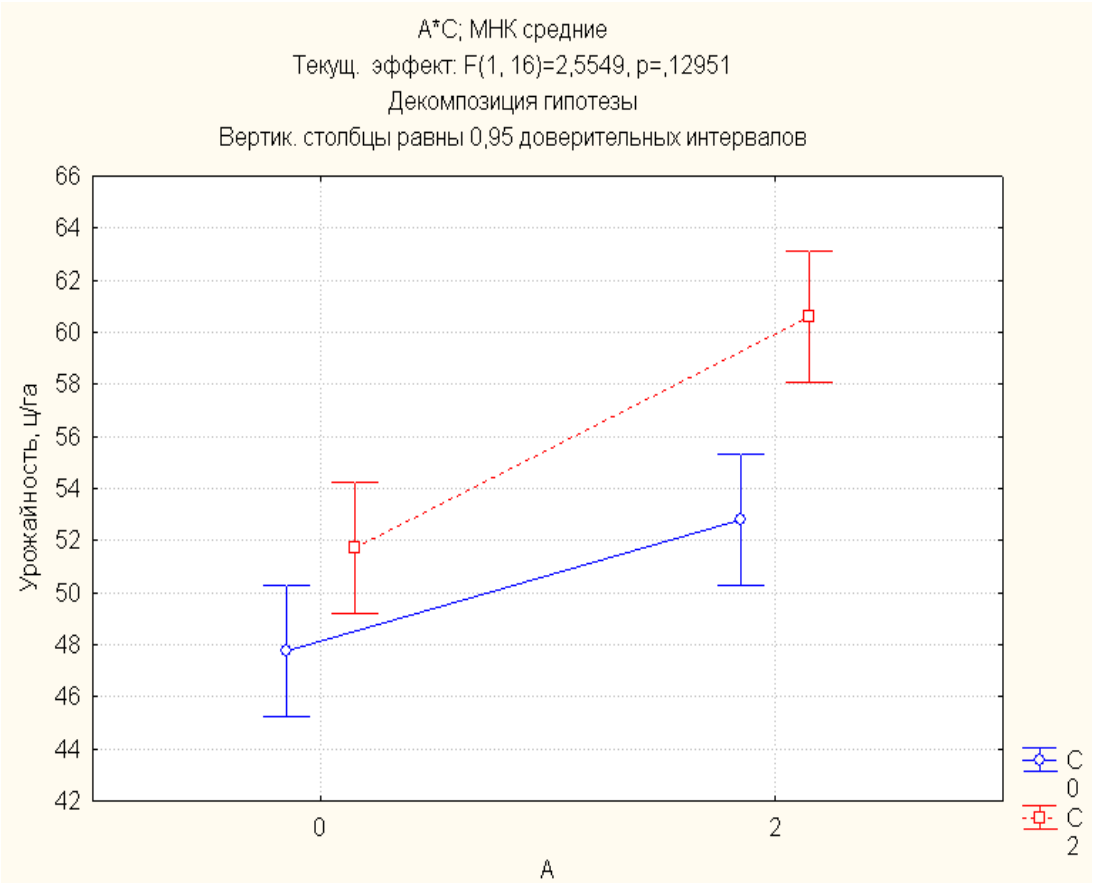


Рисунок 7.8 – График зависимости урожайности от плодородия (*A*) и системы защиты растений (*C*)

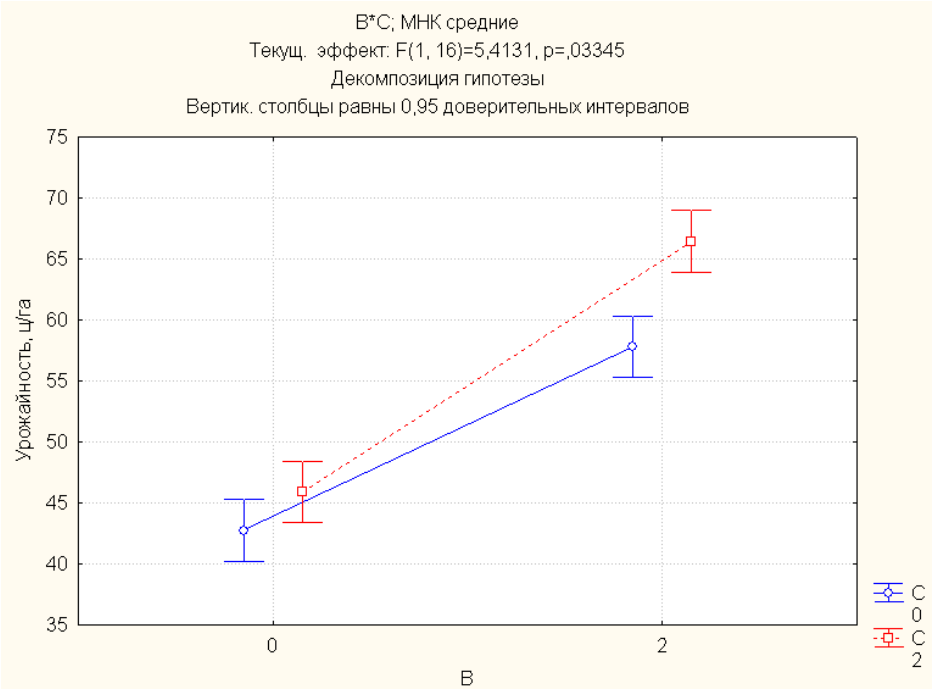


Рисунок 7.9 – График зависимости урожайности от системы удобрений и системы защиты растений (*C*)

Для детального изучения разницы средних урожайностей используем контрасты. Например, сравним *C*0 и *C*2 для A2. В окне **Результаты анализа** выполним команду **Контрасты – Задать контрасты для средних** (**Отдельно для каждого фактора**) (см. рис. 7.10) **– ОК – Вычислить** (***Planed comps – Speciffy contrasts for LS means*** (***Separately for each factor***) **– ОК –** ***Compute***).

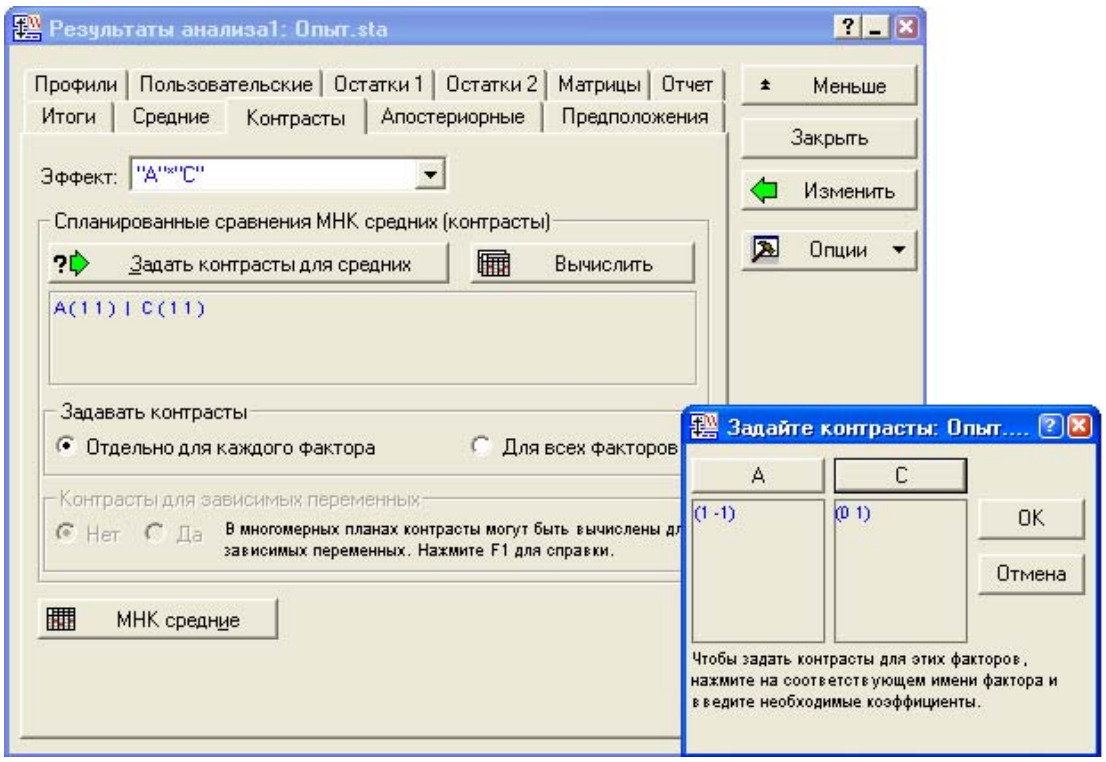


Рисунок 7.10 – Результаты дисперсионного анализа: вкладка Контрасты

В результате получим таблицу одномерного критерия значимости, подтверждающую статистически значимую разность между средними урожайностями (рис. 7.11).

Источник	Одномерный критерий значимости для спланированных сравнений (Опыт.sta) Зависимая перемен.: Урожайность, ц/га							
	Сумма квадрат.	Степени свободы	Среднее квадрат.	F	p			
Эффект	235,3216	1	235,3216	27,90389	0,000074			
Ошибка	134,9327	16	8,4333					

Рисунок 7.11 – Таблица одномерного критерия значимости для средних

Полученные результаты подтверждают, замеченную на графике, значимую разницу между средними урожайностями при уровне плодородия почвы *A*2–без применения средств защиты растений (*C*0) и с применением средств защиты растений (*C*2).С помощью кнопки **Оценить** (***Estimate***) в группе **Межгрупповые эффекты** (***Design terms***) окна **Результаты анализа** (***ANOVA Results***)можно задавать уровни оцениваемых параметров и проверять, таким образом, рабочие гипотезы о существенности влияния определённых уровней факторов. Выбор кнопки **Общая *R*** модели(***Whole model R***) позволяет получить оценку доли изменчивости урожайности, которая объясняется построенной моделью.

Зависим. перемен.	SS модели и SS остатков (Опыт.sta)										
	Множест. R	Множест. R2	Скоррект R2	SS Модель	ст.св. Модель	MS Модель	SS Остаток	ст.св. Остаток	MS Остаток	F	p
Урожайность, ц/га	0,974555	0,949758	0,927777	2550,709	7	364,3870	134,9327	16	8,433292	43,20816	0,000000

Рисунок 7.12 – Таблица *SS* модели и *SS* остатков

В нашем случае (рис.7.12) построенная модель объясняет 94,97% вариации урожайности.

Обычно, после установления различий в среднем значении зависимой переменной для разных категорий требуется установить величину различия для заданных категорий. Для решения подобной задачи исследуют контрасты (см. выше) и *HCP*.

Замечание. При выборе нескольких зависимых переменных, например, урожайность и тип технологии, рассматривается многомерный критерий значимости Уилкса.

Задание

Введите файл исходных данных для анализа и проведите разные виды дисперсионного анализа (Однофакторный ДА, Главные эффекты и т.д.), выбрав в качестве зависимых переменных: урожайность, тип технологии. Исследуйте контрасты.

Вопросы для самоконтроля

- Что изучает дисперсионный анализ?
- В чём сущность построения модели дисперсионного анализа?
- Какая гипотеза проверяется в дисперсионном анализе?
- Если гипотеза о равенстве средних отклоняется, то, как оценить, какие именно группы имеют значимое различие средних?
- Как оценивается наименьшая существенность различий в уровне средних при $F_{расч.}>F_{табл.}$?
- Опишите методы кодирования категориальных переменных.
- Какие типы сумм квадратов могут использоваться в дисперсионном анализе?

Практическое занятие № 8

Регрессионный анализ

Цель работы: Ознакомиться с возможностями корреляционно-регрессионного анализа данных, получить навыки анализа данных с использованием модуля **Множественная регрессия (Statistics – MultipleRegression)**. Провести анализ реальных данных о стоимости жилья с использованием средств модуля MultipleRegression.

Теоретические сведения

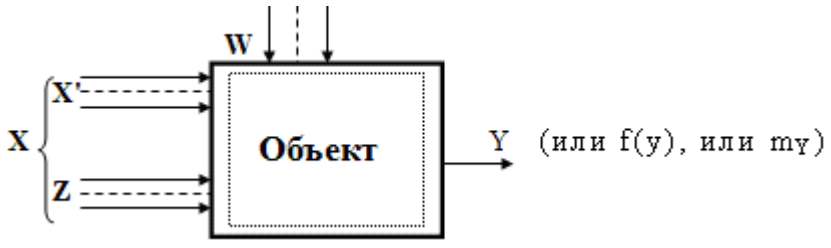
В предисловии к русскому изданию книги [17] ее научные редакторы Ю.П.Адлер и В.Г.Горский написали: “Среди известных методов математической статистики регрессионный анализ занимает исключительное положение. Это связано не столько с тем, что регрессионный анализ является одним из самых распространеннейших методов обработки результатов наблюдений, но и с тем, что он, по существу, служит основой целого ряда других методов математической статистики и, прежде всего – планирования экспериментов, дисперсионного анализа, многомерного статистического анализа… В научной литературе этому методу посвящены тысячи журнальных статей и десятки монографий…”

Объект исследования в регрессионном анализе – экономические, социальные, политические, экологические, технические и др. системы и процессы.

Предмет исследования – математические модели регрессионного анализа.

Цель исследования – установление по результатам статистических наблюдений (пассивных или активных) адекватной аналитической зависимости (уравнения регрессии) между показателями и факторами, которые характеризуют изучаемые системы. Это соответствует одной из наиболее общих задач статистики – оценивания степени и формы связи между величинами.

Обычно для иллюстрации идеи регрессионного анализа используется кибернетический подход, изображающий изучаемую систему в виде чёрного ящика:



где $\boldsymbol{X} = (X', Z)$ – факторы (вектор входных переменных); $\boldsymbol{X'}$ – управляемые, независимые переменные; $\boldsymbol{Z}$ – контролируемые, но неуправляемые факторы;

Y – “отклик” (“показатель качества управления”, “выход”), $f(y)$ – закон распределения , m_Y – математическое ожидание случайной величины Y; $\boldsymbol{W}$ – помехи.

В регрессионном анализе предполагается, что вид зависимости линейный (это наиболее хорошо теоретически разработанный раздел регрессионного анализа).

В матричной форме: $\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{\varepsilon}$, где $\boldsymbol{Y}$– вектор наблюдений; $\boldsymbol{X}$ – матрица значений независимых переменных; $\boldsymbol{\beta}$, $\boldsymbol{B}$ – векторы коэффициентов и их оценоксоответственно; $\boldsymbol{\varepsilon}$– вектор ошибок:

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ y_i \\ . \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1\, x_{11}\, x_{12} \ldots \mathrm{x}_{1\mathrm{j}} \ldots x_{1k} \\ 1\, x_{21}\, x_{22} \ldots \mathrm{x}_{2\mathrm{j}} \ldots x_{2k} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 1\, \mathrm{x}_{\mathrm{i}1}\ \mathrm{x}_{\mathrm{i}2} \ldots \mathrm{x}_{\mathrm{ij}} \ldots \mathrm{x}_{\mathrm{ik}} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 1\, x_{n1}\, x_{n2} \ldots \mathrm{x}_{\mathrm{nj}} \ldots x_{nk} \end{bmatrix}, B=\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ \beta_i \\ . \\ \beta_d \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_i \\ . \\ \varepsilon_n \end{bmatrix}.$$

Решение системы нормальных уравнений (мнк–оценка вектора B):

$$B=(X^TX)^{-1}X^TY$$

(условие разрешимости: детерминант не равен нулю $\det (\mathrm{X^TX})^{-1}\neq 0$.)

Гипотеза H_0: $b_j=0$ о значимости коэффициентов регрессии $y=\beta_0+\sum_{j=1}^{k} b_j x_j$ определяется с использованием t-критерия Стьюдента для двусторонней области при заданном уровне значимости α и $\nu = n\text{-}k\text{-}1$ степенями свободы:

$$\mathrm{t_{расч}}=\frac{\left|\mathrm{b_j}\right|}{\sqrt{\mathrm{S_\varepsilon^2}\cdot \mathrm{c_{jj}}}},$$

где c_{jj}– элемент главной диагонали матрицы $(X^TX)^{-1}$, $t_{\text{кр}}=t_{\text{дв.}\alpha}(n\text{-}k\text{-}1)$, n – число наблюдений, k – число факторов, $S_\varepsilon{}^2=\frac{\sum (y_i-\hat{y}_i)^2}{n-k-1}$ - дисперсия остатков. Если гипотеза принимается для всех j, то на этом регрессионный анализ заканчивается (в противном случае незначимые факторы отбрасываются, однако при коррелированности факторов такое исключение оказывается не надежным и приводит к исключению слишком большого числа факторов).

Для значимых факторов уравнения регрессии рассматривают интервальные оценки коэффициентов и самого уравнения регрессии.

Доверительный интервал для β_j:

$$b_\mathrm{j}-t_{кр.}\sqrt{S_\varepsilon^2\cdot c_{\mathrm{jj}}}\le\beta_\mathrm{j}\le b_\mathrm{j}+t_{кр.}\sqrt{S_\varepsilon^2\cdot c_{\mathrm{jj}}}\ .$$

Адекватность полученной модели обычно определяется с помощью дисперсионного анализа (Таблица 8.1) или критерия значимости множественного коэффициента корреляции.

Таблица 8. 1 - Таблица дисперсионного анализа (основное разложение)

Источник вариации	Число степеней свободы, *df*	Суммы квадратов, *SS*	Средние квадраты, *MS*	$F_{\text{расч.}}$	$F_{\text{кр.}}$
Обусловленный регрессией (SS_1)	*k*	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$MS_R=\frac{SS_1}{k}$	$F=\frac{MS_R}{S^2}$	$F_\alpha(k,\ n\text{-}k\text{-}1)$
Относительно регрессии (остаток) (SS_2)	*n-k-1*	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$S^2=\frac{SS_2}{n-k-1}$		
Общий, скорректированный на среднее *Y* (*SS*)	*n-1*	$\sum_{i=1}^{n}(y_i-\bar{y})^2$			

Если $F_{\text{расч.}}>F_{\text{кр.}}$ при заданном уровне значимости α и соответствующих числах степеней свободы, то гипотеза H_0: $b_j=0$ ($j=\overline{1,k}$)отбрасывается с риском ошибиться не более чем в $\alpha\cdot 100\%$ случаев (и уравнение регрессии считается статистически значимым).

Доля суммы квадратов, объясняемая регрессией называется множественным коэффициентом детерминации (квадратом множественного коэффициента корреляции *R*):

$$R^2=\frac{\textit{объяснённая вариация}}{\textit{общая вариация}}=\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2},\ 0\le \mathrm{R}\le 1.$$

Значимость R^2 для уравнения определяется по F-критерию. $F_{кр.}=F_\alpha(k,n-k-1),\ F_{расч.}=\frac{R^2(n-k-1)}{(1-R^2)k}$. Если $F_{\text{расч.}}>F_{\text{кр.}}$, то гипотезу H_0: $R=0$ отвергают и связь между *X* и *Y* считают статистически значимой.

Если *Y* зависит только от одной переменной *X*, то $R=r$ – парному коэффициенту корреляции.

С учётом потери числа степеней свободы вариации, рассматривается исправленный коэффициент детерминации:

$$\overline{R}^2=1-\frac{(1-R^2)(n-1)}{(n-k-1)}.$$

Частный коэффициент корреляции $r_{lv.t}$ – это корреляция между переменными x_l и x_v при фиксированном влиянии x_t. (Частный коэффициент корреляции может интерпретироваться как способ выявления ложной корреляции, то есть – когда корреляция между x_l и x_v объясняется их взаимодействием с x_t.)

Выбор структуры уравнения наилучшей регрессии (наиболее точно описывающей исследуемый процесс) можно осуществить, используя R^2 – квадрат множественного коэффициента корреляции или дисперсионный анализ. Структура

уравнения регрессии усложняется (например, в полиномиальном случае повышается степень многочлена) до тех пор, пока увеличение соответствующего критерия не станет пренебрежительно малым. (Рекомендуется сначала изучить все возможные парные линейные зависимости).

Однако, вывод о корректности модели по условию $R^2 \approx 1$ не всегда верен. И вот почему. Результата $R^2 \approx 1$ можно добиться, увеличивая число оцениваемых параметров β_j и в случае «насыщенности» $k=N$ значение R^2, будет равно 1, но модель при этом не обязательно корректна. Величина R^2 и суммы квадратов не всегда дают однозначный ответ на вопрос − «адекватна ли модель?».

В экономических исследованиях используют фиктивные переменные, принимающие два значения (0 и 1) для оценки вклада дихотомической переменной в уравнение регрессии, например, при анализе жилья: «дом кирпичный» – 1, иначе – 0.

Замечание.В регрессионном анализе рассматривают:

1) линейную относительно параметров регрессию:

а) парную линейную ($y=b_0+b_1x$),

б) парную криволинейную (например, $y=a+bx+cx^2$, $y=a+bsinx$),

в) множественную линейную ($y = b_0 + \sum_{j=1}^{k} b_j x_j$)

BStatistica: **Анализ** (***Statistics***) – **Множественнаярегрессия** (***Multiple Regression***).

г) множественную нелинейную (например, $y=a_0+a_1x_1+a_2x_2+a_3x_1^2+a_4x_2^2+a_5x_1x_2$)

В Statistica: **Анализ** (***Statistics***) – **Углубленные методы анализа** (***Advanced Linear/Nonlinear Models***) – **Множественная нелинейная регрессия** (***Fixed Nonlinear Regression***).

д) ортогональную полиномиальную регрессию ($y = b_0 + \sum_{j=1}^{k} b_j \varphi_j(x)$, где

φ_j –некоторые функции, например, ортогональные полиномы Чебышева);

2) нелинейную относительно параметров (например, $y=a+be^{cx}$). В Statistica: **Анализ** (***Statistics***) – **Углубленные методы анализа** (***Advanced Linear/Nonlinear Models***) – **Нелинейное оценивание** (***Nonlinear Estimation***).

3) модели бинарных откликов, когда выходная переменная является значением принадлежащим отрезку [0; 1].

Логит-регрессия: $y=exp(b_0 + \Sigma b_jx_j)/ (1+exp(b_0 + \Sigma b_jx_j))$. Пробит-регрессия – линейно связана с независимыми переменными и подчиняется нормальному закону распределения. Модели множественной пробит/логит регрессии являются расширением стандартных логит и пробит регрессионных моделей в случае, когда зависимая переменная имеет более двух категорий (например, не только «Да, Нет», а «Да, Нет, Не знаю»), то есть когда зависимая переменная или переменная отклика, подчиняется мультиномиальному распределению, а не биномиальному.

В Statistica: **Анализ** (***Statistics***) – **Углубленные методы анализа** (***Advanced Linear/Nonlinear Models***) – **Нелинейное оценивание** (***Nonlinear Estimation***).

Практически все перечисленные типы моделей можно построить с помощью той или иной модификации МНК. Адекватность построенных моделей можно про-

верить с помощью дисперсионного анализа (по схемам аналогичным рассмотренным выше) и квадрата множественного коэффициента корреляции R^2.

Вопрос выбора наилучшей регрессии из нескольких построенных моделей регрессии далеко не однозначен. Один из способов, основывающийся на дисперсионном анализе указан выше. Однако он неявно предполагает выполнение классических условий (условий Гаусса – Маркова) применения МНК.

На практике не все выше перечисленные условия применения МНК соблюдаются. Обычно, вид зависимости априори неизвестен, ошибки не подчиняются нормальному закону распределения (так как законы больших чисел для конечных выборок не выполняются). Случайные ошибки имеются не только на выходе, но и на входе и т.д.

В настоящее время существует ряд методов, позволяющих получать эффективные параметры регрессии при нарушении условий нормальности ошибок, использующие функции ошибок (функции потерь): Хубера, Пуанкаре, Винзора, Андрюса, Мешалкина, Рамсея, Гуды; джеккнайф – оценка и т.д. При наличии ошибок, как на выходе, так и на входе, вводится понятие ортогональной регрессии (параметры регрессии находят из условия минимизации суммы квадратов расстояний до поверхности регрессии). Если входные переменные коррелированны (мультиколлинеарность), то используются ридж-оценки, метод главных компонент, метод автоматического отсева переменных и т.д.

В случае множественной регрессии выбор "наилучшей регрессии" осуществляется с помощью пошаговой регрессии последовательно включающей (отбрасывающей) входные переменные, факторного анализа, анализа главных компонент.

Если часть наблюдений, используемых в регрессионном анализе, имеет сильно отличающиеся дисперсии, то используется "взвешенный" МНК.

В последние десятилетия развитие регрессионного анализа характеризуется привлечением топологии, теории групп, функционального анализа. Так в последнем случае – обобщение регрессионных моделей на бесконечномерные гильбертовы пространства носит название коллокационных моделей и позволяет решать ряд задач прогнозирования, например, в финансовой сфере. Существует подход общих линейных моделей (*GLM*), позволяющий оперировать как с непрерывными входными переменными, так и с категориальными; исследовать несколько результативных переменных в одной модели и получать единственное решение для плохо обусловленной матрицы плана ($X^T X$).

Важным моментом при построении искомой зависимости является отбор факторов X_j, существенно влияющих на результативную переменную Y. Известно достаточно много путей отбора, условно их можно разделить на два класса: формальные и содержательные (семантические или смысловые).

Формальные методы основываются на идее перебора различных уравнений (например, различные модификации пошаговой регрессии с последовательным включением или исключением независимых переменных) до момента достижения некоторого критерия, например, F-Фишера (дисперсионный анализ), характеризующего (при заданном уровне значимости α) значимость вклада переменной в регрессию.

Содержательные методы предполагают достижение целей моделирования, при этом, как отмечалось выше, различают:

а) Физические модели, описывающие функциональные особенности изучаемых процессов. Построение функциональной модели – редкий случай, так как принципиально невозможно учесть все причинно – следственные связи и их взаимодействия.

б) Модель для управления процессом. Предполагается возможным для любого y_i найти такие x_{ij} (управляющие воздействия), что, задав их в модели, получим требуемое y_i.

в) Модель для предсказания. Даёт возможность по заданным x_{ij} определить прогнозируемое y_i.

Наблюдения за функционированием сложного объекта можно представить в виде множества точек некоторого фазового пространства. Тогда физические модели, модели управления и предсказания – это возможные проекции объекта на различные плоскости, поэтому на практике эти модели не совпадают.

Очевидно, что практически более необходимы модели, описывающие содержательные стороны изучаемого процесса. А попытка совмещения формальных и содержательных критериев – это типичная многокритериальная задача, решение которой обычно не однозначно.

Альтернативным методом "выбора наилучшей регрессии" являются всевозможные методы машинного обучения, реализуемые с помощью нейронных сетей, методов эволюционного программирования и метода группового учета аргумента (МГУА).

Регрессионный анализ в *Statistica*

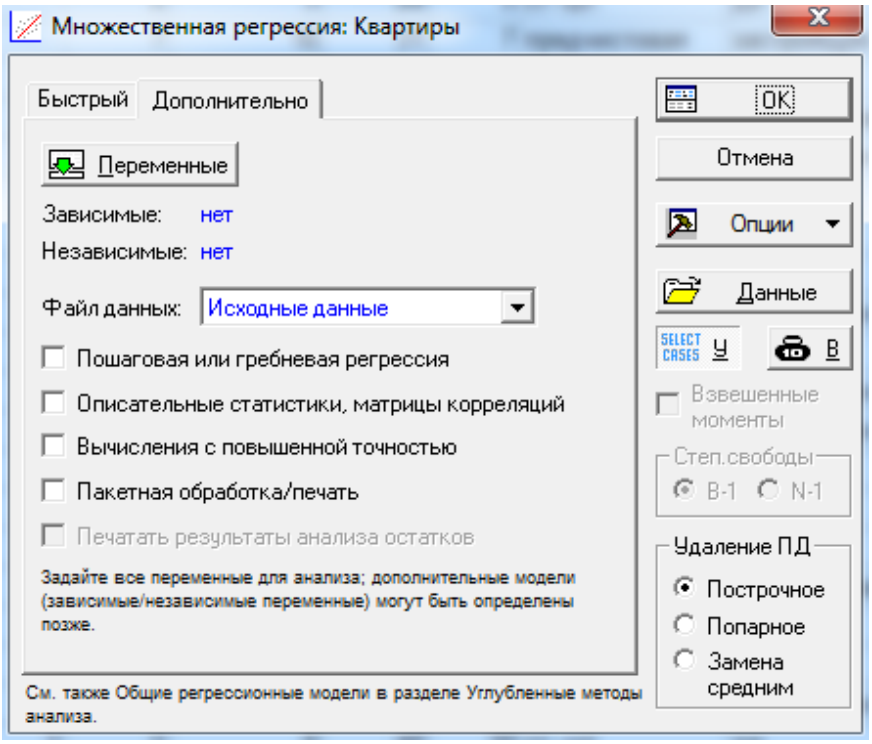


Рисунок 8.1 – Диалоговое окно множественной регрессии в *Statistica* 6.1

Выполните команду: **Анализ – Множественная регрессия** (***Statistics – Multiple Regression***). Вкладка **Переменные** (***Variables***) позволяет выбрать для анализа зависимые (*Dependent*) и независимые (*Independent*) переменные (рис.8.1). Файл исходных данных может представляться как в виде таблицы, так и в виде корреляционной матрицы. Вкладка **Дополнительно** (***Advanced***) позволяет перейти к пошаговой или гребневой регрессии, получать описательные статистики и матицы корреляций, проводить вычисления с повышенной точностью, а так же осуществлять пакетную обработку и печать.

Пример. Рассмотрим построение и анализ модели стоимости жилья в г. Краснодаре на 01 октября 2013г, загрузим файл Квартиры.xls (рис. 8.2).

Рассматриваются переменные:

V1 – порядковый номер; V2 – микрорайон; V3 – расположение; V4 – число комнат; V5 – тип дома; V6 – индикаторная переменная монолитного строительства; V7 – кирпичный дом; V8 – этаж; V9 – этажность; V10 – хороший этаж (2-6); V11 – общая площадь, м^2; V12 – жилая площадь, м^2; V13 – площадь кухни, м^2; V14– наличие ремонта, V15 – индикаторная переменная, характеризующая наличие ремонта, V16 – дом бизнес класса; V17 – евроремонт; V18 – ремонт под ключ; V19 – предчистовая отделка, V20 – отделка стяжка-штукатурка, V21 – требует ремонта; V22 – наличие свидетельства; V23 – стадия готовности дома; V24 – дом сдан; V25 – возможность продажи квартиры в ипотеку; V26 – цена, тыс. руб.

	1 №	2 Район	3 Расположение	4 Число комнат	5 Тип дома	6 Монолит	7 Кирпич	8 Этаж	9 Этажность	10 Хороший этаж	11 Общая	12 Жилая	13 Кухня	14 Ремонт
1	1	40 лет Победы	1	1	кирпич	0	1	4	6	1	32	16	7,04	ремонт
2	2	40 лет Победы	0	1	кирпич	0	1	3	8	1	37	15	8,89	предчистовая
3	3	40 лет Победы	0	1	кирпич	0	1	2	6	1	36	17	7,92	ст.-шт.
4	4	40 лет Победы	0	1	кирпич	0	1	2	6	1	40	18	10,8	ст.-шт.
5	5	40 лет Победы	0	1	м/к	1		7	18	0	27	15	6,19	ст.-шт.
6	6	40 лет Победы	0	1	м/к	1		2	3	1	28	14	4,16	ст.-шт.
7	7	40 лет Победы	0	1	мон	1		7	16	0	31	16	3,07	предчистовая
8	8	40 лет Победы	0	1	кирпич	0	1	3	6	1	34	15	3,98	ст.-шт.
9	9	40 лет Победы	0	1	кирпич	0	1	5	7	1	44	20	9,68	ст.-шт.
10	10	40 лет Победы	0	1	м/к	1		3	12	1	38	15	8,86	ст.-шт.
11	11	40 лет Победы	0	1	кирпич	0	1	2	6	1	37	15	10,89	ст.-шт.
12	12	40 лет Победы	0	1	кирпич	0	1	3	5	1	37	18	6,89	предчистовая
13	13	40 лет Победы	0	1	кирпич	0	1	5	7	1	46	22	10,62	предчистовая
14	14	40 лет Победы	0	1	м/к	1		3	9	1	36	16	8,92	предчистовая
15	15	40 лет Победы	0	1	кирпич	0	1	1	6	1	40	20	6,8	ремонт
16	16	40 лет Победы	0	1	м/к	1		4	10	1	33	16	12,01	предчистовая
17	17	40 лет Победы	0	1	м/к	1		4	18	1	43	18	12,71	предчистовая
18	18	40 лет Победы	0	1	кирпич	0	1	2	7	1	37	20	8,89	предчистовая
19	19	40 лет Победы	0	1	кирпич	0	1	2	5	1	42	20	11,74	ст.-шт.
20	20	40 лет Победы	0	1	кирпич	0	1	4	4	1	29	17	4,13	ремонт
21	21	40 лет Победы	0	1	кирпич	0	1	2	6	1	40	20	5,8	ст.-шт.
22	22	40 лет Победы	0	1	кирпич	0	1	4	6	1	40	20	5,8	предчистовая
23	23	40 лет Победы	0	1	кирпич	0	1	5	7	1	41	19	6,77	предчистовая
24	24	40 лет Победы	0	1	м/к	1	1	8	16	0	36	18	8,92	ст.-шт.
25	25	40 лет Победы	0	1	кирпич	0	1	5	6	1	40	18	7,8	ст.-шт.
26	26	40 лет Победы	0	1	м/к	1		7	17	0	33	16	10,01	предчистовая
27	27	40 лет Победы	0	1	кирпич	0	1	2	6	1	38	18	5,86	ст.-шт.
28	28	40 лет Победы	0	1	м/к	1		11	17	0	32	17	10,04	предчистовая
29	29	40 лет Победы	0	1	м/к	1		7	16	0	31	17	5,07	предчистовая

Рисунок 8.2 – Данные по стоимости жилья в городе Краснодаре на 1 октября 2013 года

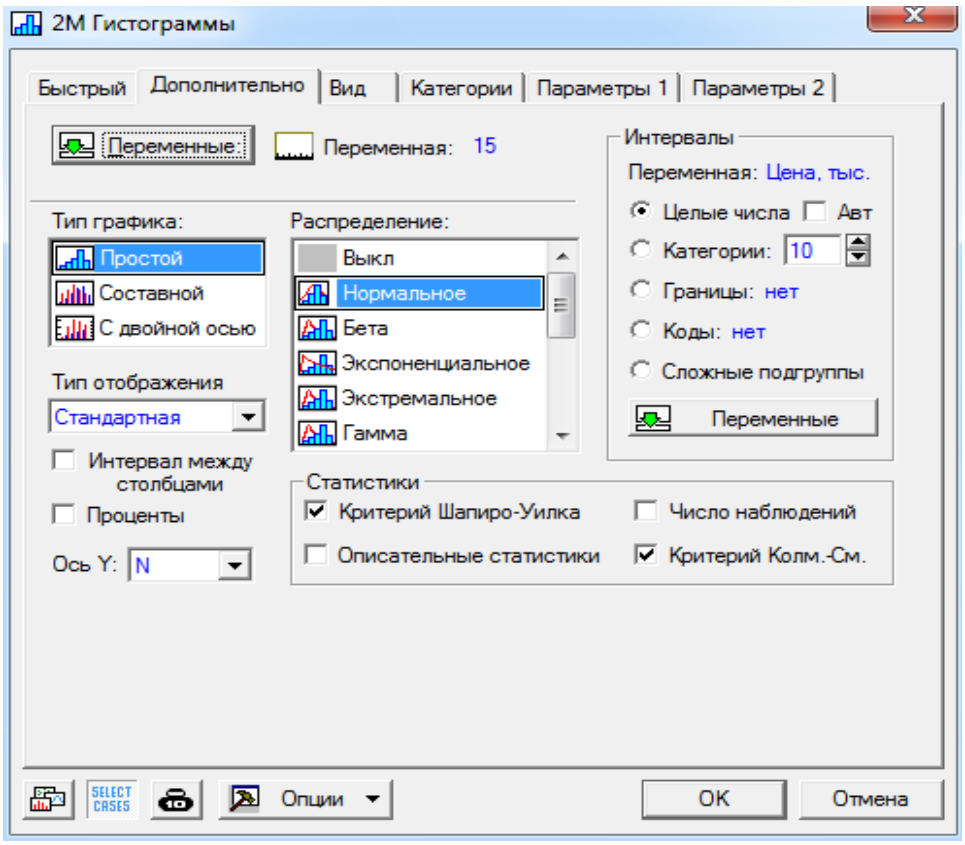


Рисунок 8.3 – Диалоговое окно построения гистограммы

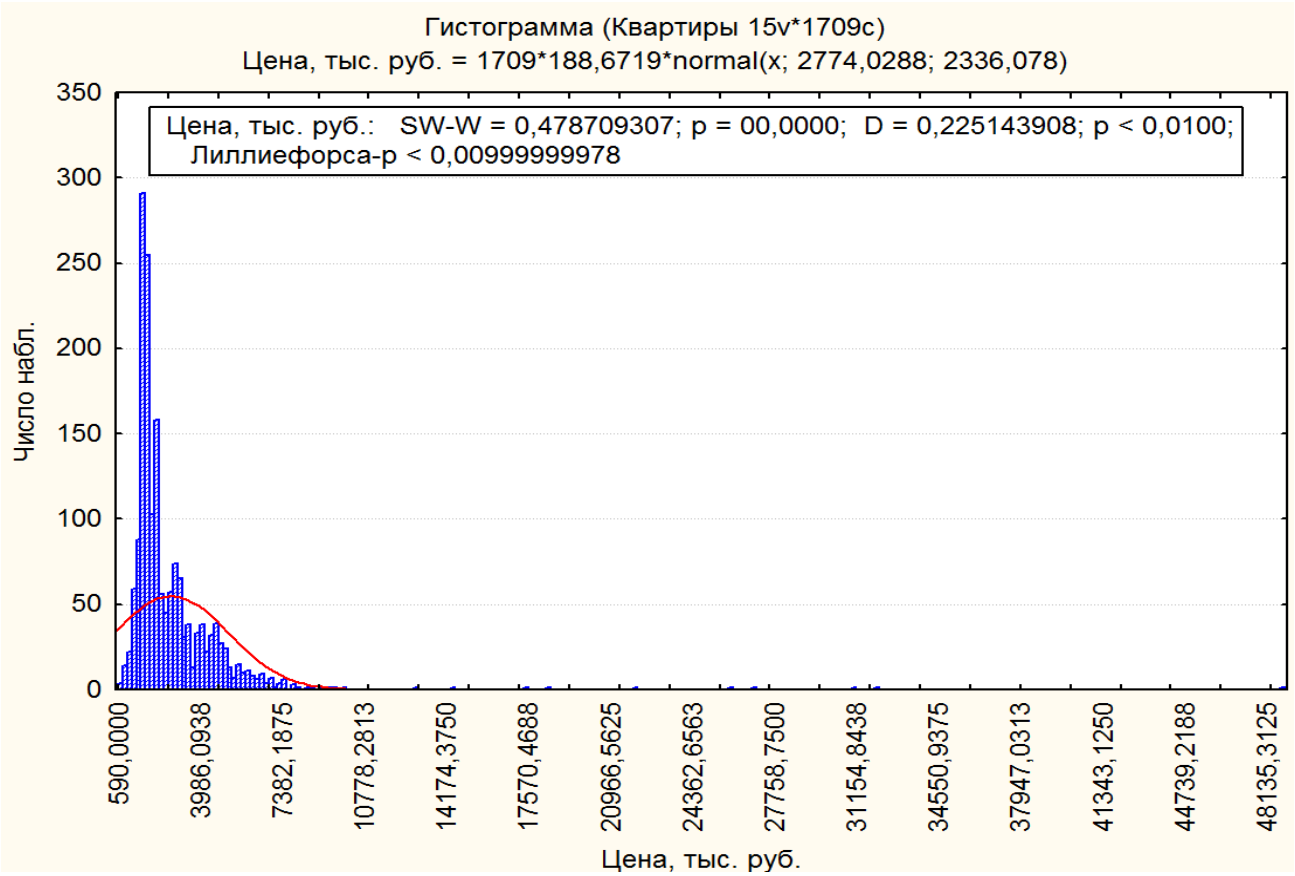


Рисунок 8.4 – Гистограмма стоимости квартир

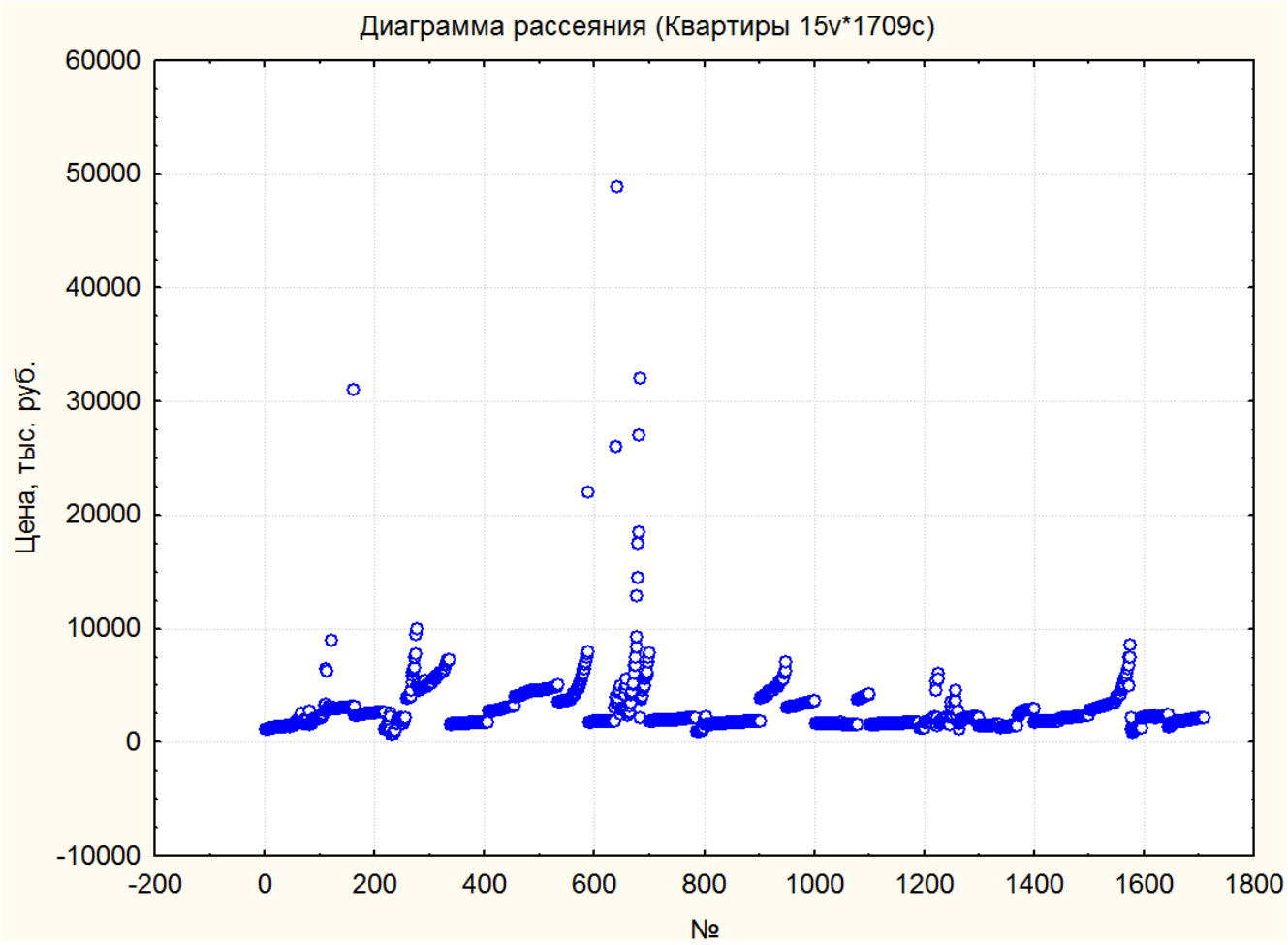


Рисунок 8.5 – Диаграмма рассеяния стоимости квартир

Для анализа переменной V26 выполним команду (рис.8.3): **Графика – 2М Графики – Гистограммы вкладка Дополнительно – Тип графика Простой – Распределение Нормальное – Переменные *V*26– ОК** (***Graphs – 2D Graphs – Histograms – Advanced – Regular – Normal – Variables* V26 – *OK***).

Анализ переменной V26 – цена, показывает, что данные при уровне значимости не более 0,01 подчиняются нормальному закону со средним значением 2774,0288 и средним квадратическим отклонением 2336,078; максимальное значение 48135,3125 и минимальное значение 590 (рис.8.4). Очевидно, что максимальное значение цены не удовлетворяет правилу трёх сигма и, скорее всего, является аномальным или ошибочным значением. Наше предположение подтверждает **Диаграмма рассеяния: Графика - 2М Графики - Диаграмма рассеяния** (***Graphs – 2D Graphs – Scatterplots***) на рисунке 8.5.

Переменная *X* – порядковый номер; *Y* – цена, тыс. руб.

Используя инструмент «кисть» на панели графических инструментов, выделим диапазон квартир стоимостью выше 10000 тыс. руб., т.е. удалим аномальное значение (свыше 10 000 тыс. руб.), в результате получим обновлённую диаграмму рассеивания (рис. 8.6). Очевидно, выделяющиеся наблюдения остались, но они уже значительно меньше отличаются от совокупности, чем-то, которое мы удалили.

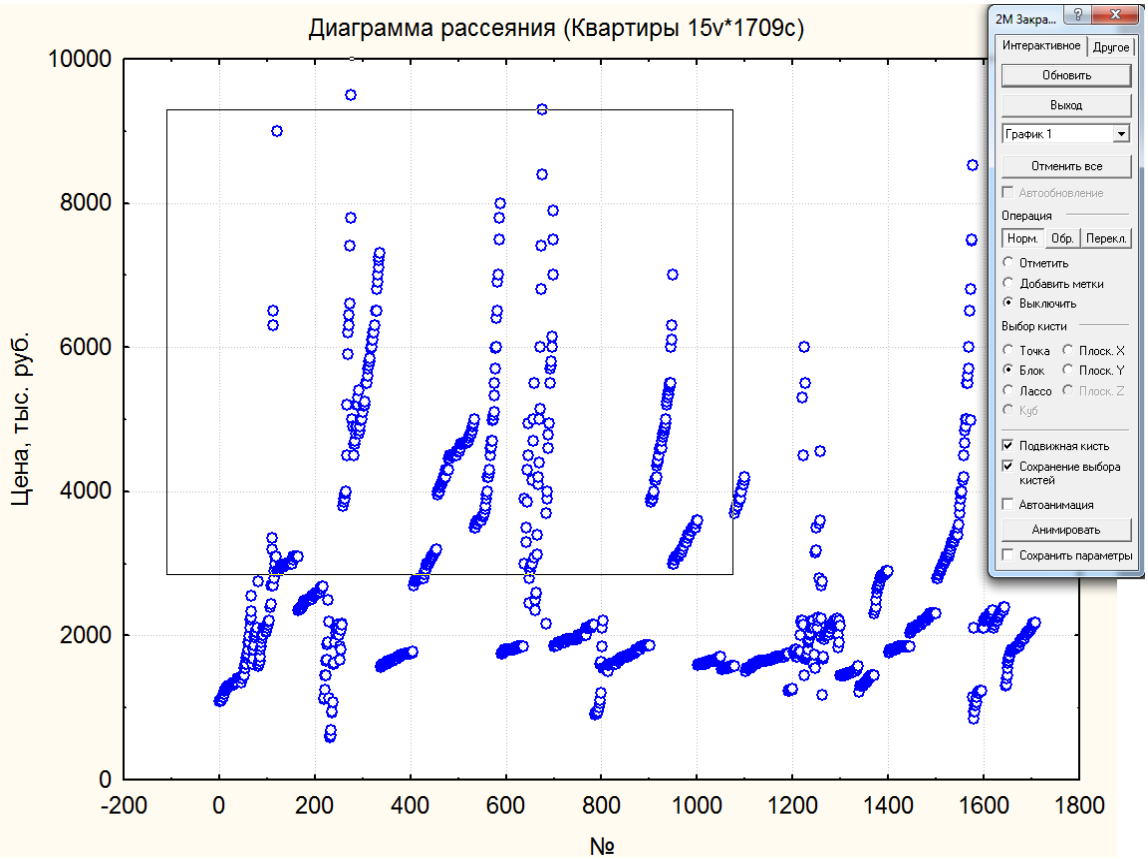


Рисунок 8.6 – Диаграмма рассеяния стоимости квартир после удаления выброса

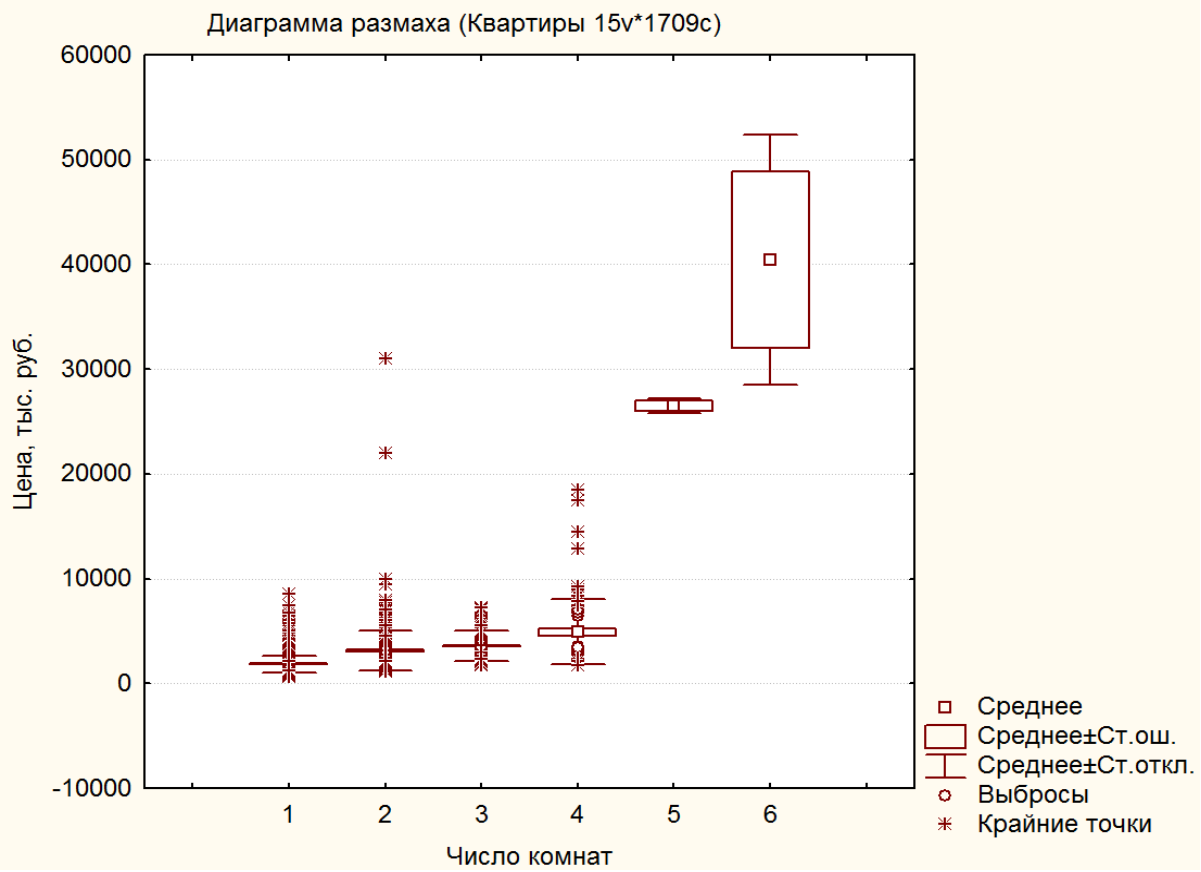


Рисунок 8.7 – Диаграммы размаха стоимости квартир

С помощью диаграмм средних или диаграмм размаха можно увидеть (рис.8.7), что1-5 комнатные квартиры имеют и выбросы, и крайние точки, что естественно указывает на неоднородность данных (команда **Графики** (***Graphs***) **– 2М графики** (***2D Graphs***) **– Диаграммы размаха** (***Box Plots***)). Здесь зависимая переменная будет цена квартиры, независимая число комнат.

Дале мы будем рассматривать только данные по 1 комнатным квартирам общей площадью более 35 м^2, для этого с помощью кнопки SELECT CASES - выбрать случаи в модуле анализа или кнопок панели инструментов Таблица данных , откроем условия выбора переменных (рис. 8.8) и зададим условие, при котором будут отбираться только однокомнатные квартиры с общей площадью равной или более 35 м^2: V4=1 and V*11*>=35 (аномальное значение, как видно из рисунка 8.7 относится к двухкомнатным квартирам).

Проанализируем парные корреляции факторов V11, V12, V13, V26 для этого выполним команду: **Анализ – Основныестатистики и таблицы – Парные и частные корреляции – Матрица парных корреляций** (**Statistics – Basic Statistics – Correlation matrices**). В список переменных введем V11, V12, V13, V26 (переменный выбираются удерживая клавишу *Caps Lock*. В результате получим рисунок 8.8, который показывает, что практически все переменные коррелируют между собой значимо, однако – как известно корреляцию менее 0,6 на практике можно игнорировать.

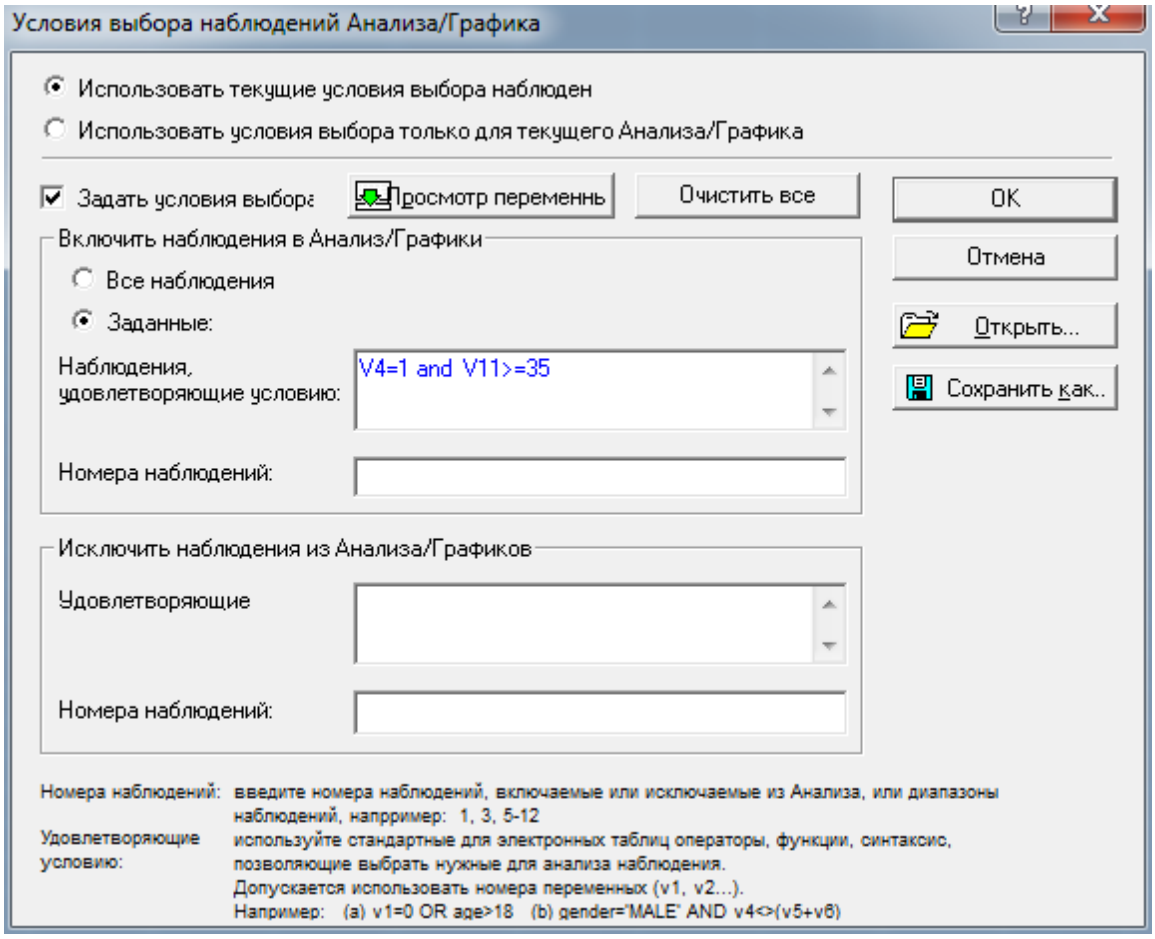


Рисунок 8.8 – Диалоговое окно условий выбора наблюдений

Корреляции (Квартиры)
Отмеченные корреляции значимы на уровне p <,05000
N=617 (Построчное удаление ПД)

Переменная	Общая	Жилая	Кухня	Цена, тыс. руб.
Общая	1,00	0,81	0,46	0,78
Жилая	0,81	1,00	0,23	0,62
Кухня	0,46	0,23	1,00	0,39
Цена, тыс. руб.	0,78	0,62	0,39	1,00

Рисунок 8.9 – Матрица парных корреляций

Поэтому в нашем случае мы оставим в качестве независимых переменных общую площадь и площадь кухни (коэффициент корреляции между жилой и общей площадью 0,81 – поэтому из них нужно выбрать одну. Мы выбираем общую площадь, так как она имеет с ценой наибольший коэффициент корреляции – 0,78).

Далее для множественной регрессии мы будим использовать следующие переменные: зависимые переменные – V26; независимые V3, V11, V13,V17, V18, V25.

Открыв модуль множественной регрессии: **Анализ – Множественная регрессия – вкладка Дополнительно** (***Statistics – Multiple Regression – Advanced***), и выбрав **Описательные статистики, матрицы корреляций** (***Reviev descriptive statistics, correlation matrics***) – **ОК** (рис.8.10), последовательно будем рассматривать результаты регрессионного анализа.

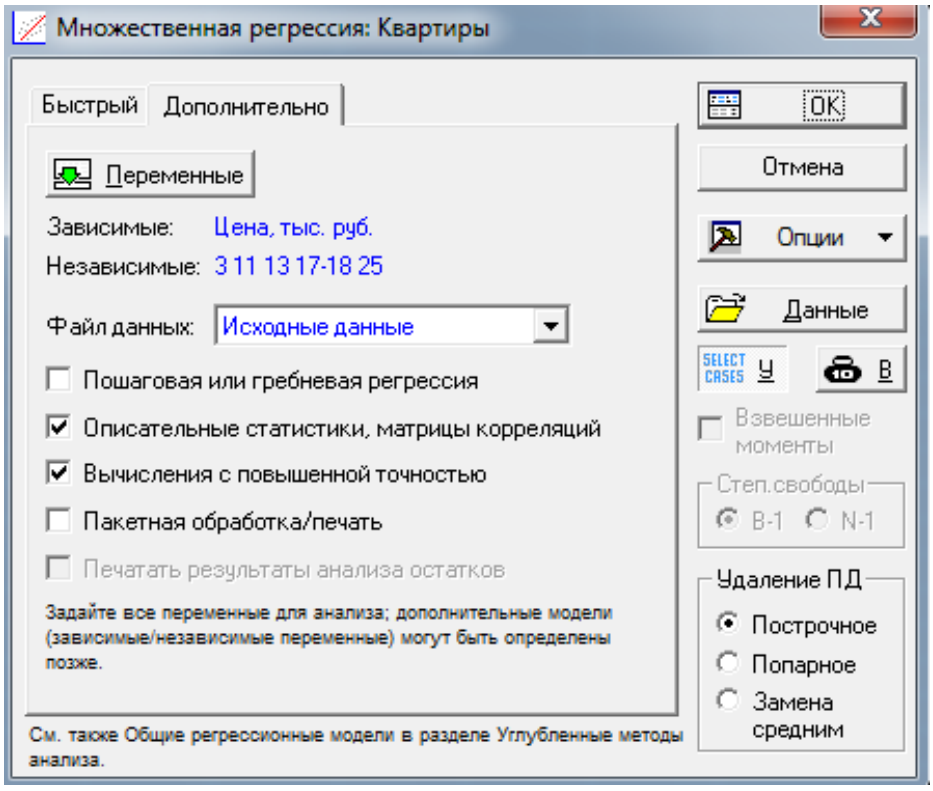


Рисунок 8.10 – Диалоговое окно модуля Множественная регрессия

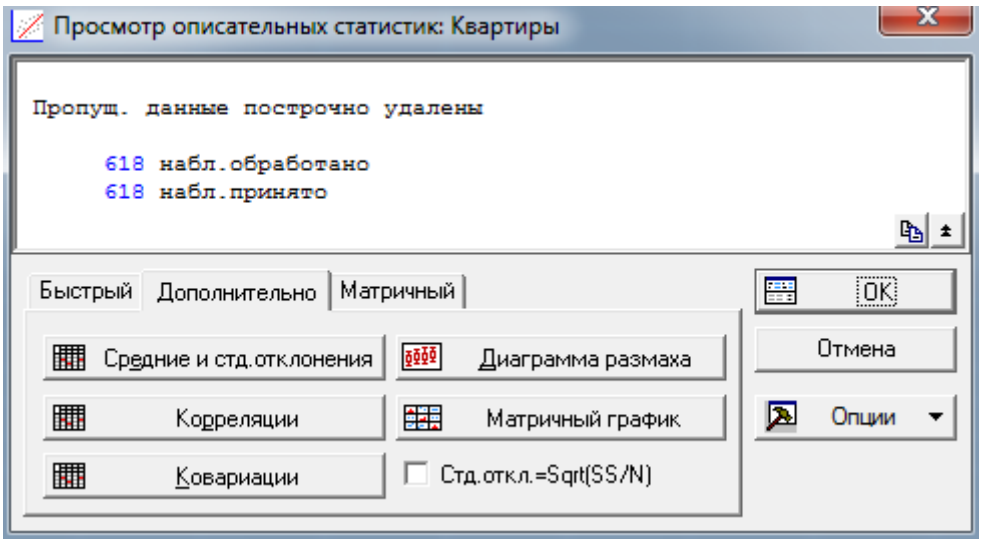


Рисунок 8.11 – Диалоговое окно просмотра описательных статистик

	Средние и станд.отклонения (Квартиры)		
Переменная	**Средние**	Ст.Откл.	N
Расположение	0,519	0,5000	618
Общая	42,722	8,9406	618
Кухня	9,618	2,8250	618
Евроремонт	0,118	0,3230	618
Под ключ	0,068	0,2519	618
Ипотека	0,366	0,4820	618
Цена, тыс. руб.	1974,198	869,8953	618

Рисунок 8.12 – Описательные статистики для однокомнатных квартир

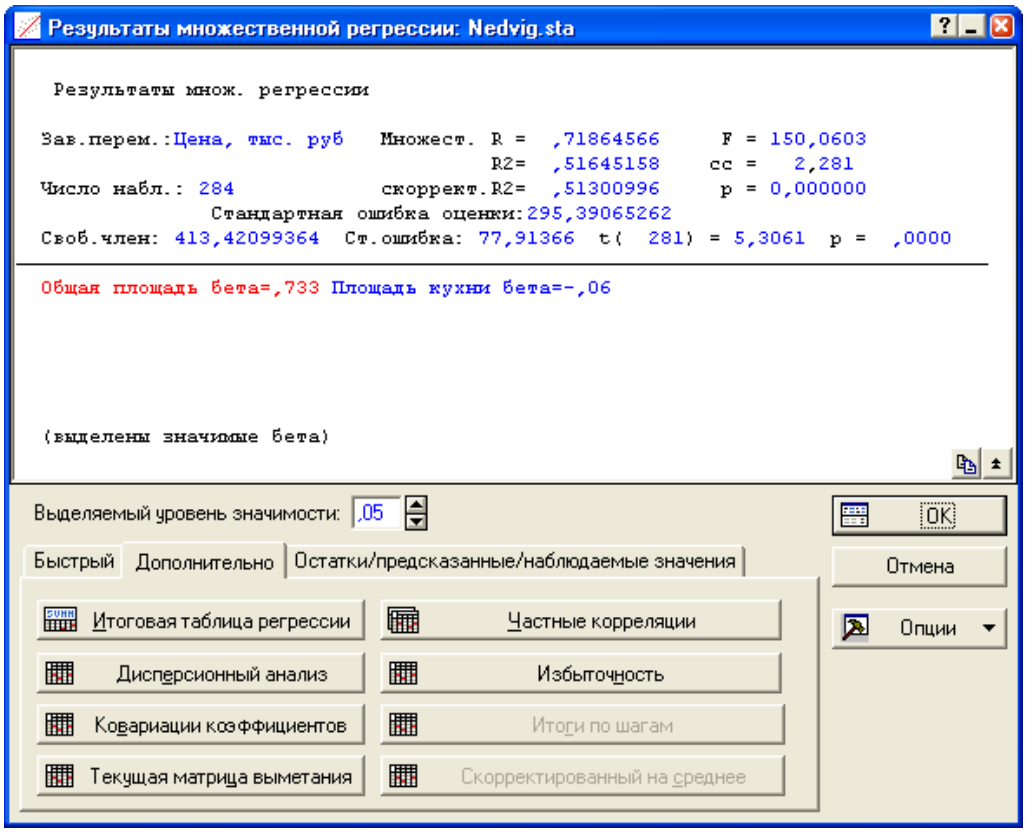


Рисунок 8.13 – Диалоговое окно *Результаты множественной регрессии*

1) Во вкладке **Дополнительно** (***Advanced***) выберем кнопку **Средние и стд. отклонения** (***Means & standard deviations***) (рис.8.13).

Получим, что средняя площадь однокомнатной квартиры – 42,7 м^2 при среднеквадратическом (стандартном) отклонении 8,9; средняя площадь кухни 9,6 м^2 при среднеквадратическом отклонении 8,8; средняя цена 1974,2 тыс. руб. при среднеквадратическом отклонении 869,9 (рис. 8.12).

2) После просмотра описательных статистик выберем **ОК** и перейдём в диалоговое окно **Результаты множественного анализа** (***Multiple Regression Results***) (рис.92).

3) После выбора кнопки **Итоговая таблица регрессии** (***Summary: Regression results***) (рис.8.13), получим таблицу, изображённую на рисунке 8.14. Отметим, что площадь кухни слабо коррелирует с ценой, поэтому статистически значимым в уравнении регрессии является общая площадь жилья (рис.8.14).

Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Квартиры)
R= ,81265853 R2= ,66041389 Скорректир. R2= ,65707917
F(6,611)=198,04 p<0,0000 Станд. ошибка оценки: 509,41

N=618	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(611)	p-уров.
Св.член			-1310,80	109,0942	-12,0153	0,000000
Расположение	0,103669	0,025483	180,35	44,3330	4,0681	0,000054
Общая	0,711926	0,027607	69,27	2,6861	25,7881	0,000000
Кухня	0,064291	0,026659	19,80	8,2091	2,4116	0,016177
Евроремонт	0,174531	0,024035	470,02	64,7275	7,2615	0,000000
Под ключ	0,095114	0,024505	328,48	84,6313	3,8813	0,000115
Ипотека	-0,054915	0,024445	-99,11	44,1157	-2,2465	0,025028

Рисунок 8.14 – Итоги регрессионного анализа

Зависимость имеет вид:

Цена = 180,35V3+69,27V11+19,8V13+470,02V17+328,48V18-99,11V25. Таким образом, увеличение общей площади 1 комнатной квартиры в Краснодаре на 1 м^2 (общей площадью более 35м^2) увеличивает её стоимость на 69,27 тыс. руб., а увеличение площади кухни на 1 м^2 – увеличивает стоимость на 19,8 тыс. руб., евроремонт добавляет к стоимости 470 тыс. руб., хорошее расположение в городе добавляет 180,4 тыс. руб., наличие ипотеки уменьшает стоимость почти на 100 тыс. руб. (соответственно бета-коэффициенты указывают, на сколько стандартных отклонений изменится цена при увеличении площади на 1 стандартное отклонение).

4) Выберем кнопку **Дисперсионный анализ** (***ANOVA***) (рис.8.13). Таблица дисперсионного анализа показывает, что уравнение статистически значимо (рис.8.15).

Проведём анализ остатков,– команда: **ОК – вкладка Вероятностные графики - Нормальный график остатков** (**ОК – *Normal plot of residuals***). Получим график нормальной вероятностной бумаги, который показывает, что остатки отклоняются от нормального закона (рис.8.16) (в идеале они должны находиться на одной линии).

Эффект	Дисперсионный анализ; ЗП: Цена, тыс. руб. (Квартиры) Сумма квадрат	сс	Средн. квадрат	F	р-уров.
Регресс.	308343867	6	51390644	198,0415	0,00
Остатки	158551015	611	259494		
Итого	466894882				

Рисунок 8.15 – Таблица дисперсионного анализа для уравнения регрессии

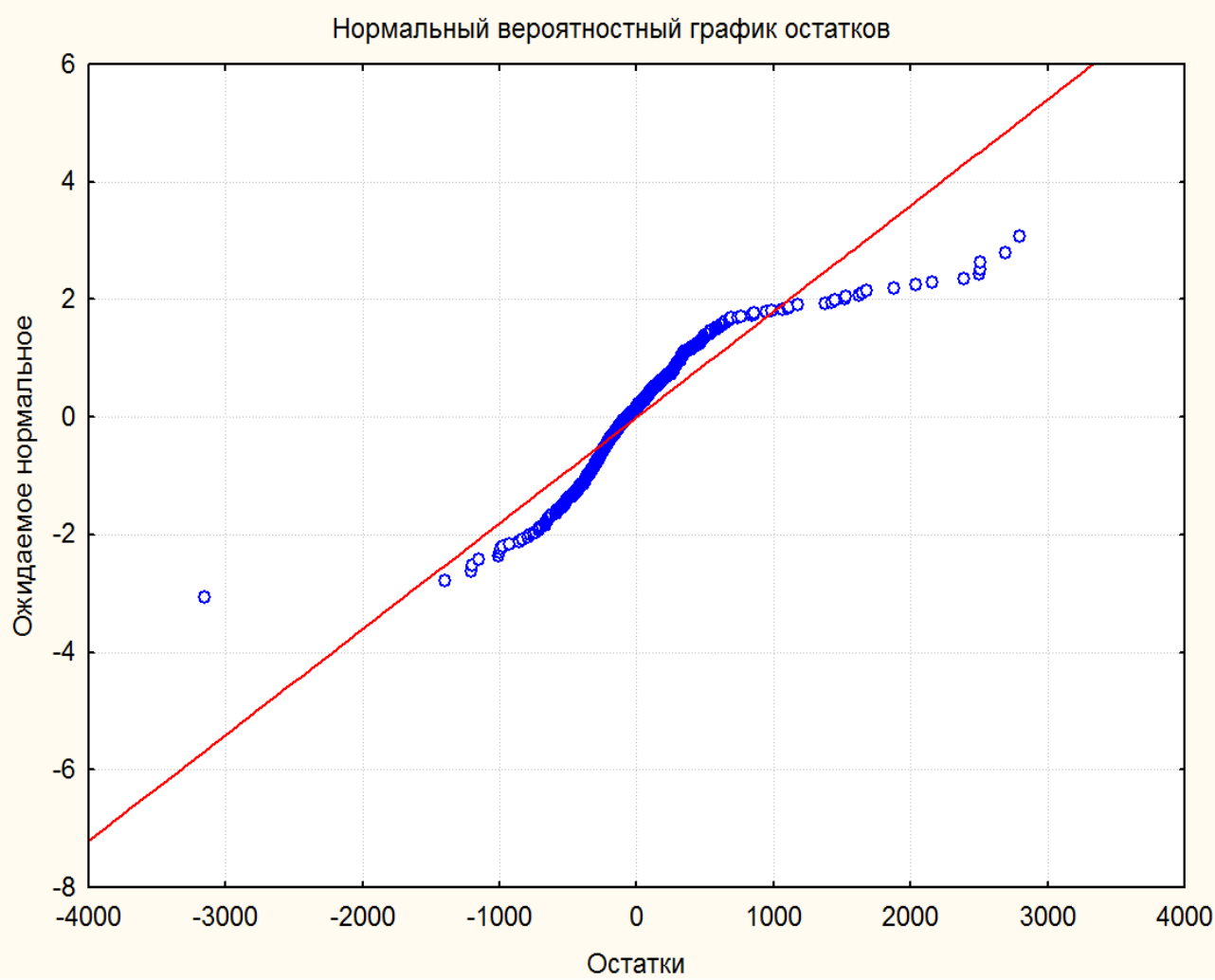


Рисунок 8.16 – График нормальной вероятностной бумаги для остатков

Для построчного анализа выбросов выполним команду вкладка **Выбросы – Стандартные остатки (>2*сигма) – Построчный график выбросов** (***Perform residual analysis – Case wise plot of residuals***) и получим график выбросов, который показывает, какие значения остатков превышают 2 сигма (рис.8.17).

Станд. остатки (Квартиры)
Выбросы

Набл.	Станд. остатки	Наблюд. Значение	Предск. Значение	Остатки	Станд. предск.	Станд. Остатки	Стд.Ош. предск.
123	. . . *	2900,000	4101,790	-1201,79	3,009630	-2,35920	73,6932
794	. . . *. . . .	1000,000	2148,593	-1148,59	0,246694	-2,25477	39,5425
1516	. . . . * . .	3000,000	1471,131	1528,87	-0,711624	3,00128	44,1783
1522	. . . *	3100,000	4307,813	-1207,81	3,301065	-2,37102	79,0137
1539	. . . .* . . .	3350,000	2284,591	1065,41	0,439072	2,09148	39,7505
1542	. . . . .* . .	3400,000	1774,910	1625,09	-0,281908	3,19017	43,5155
1550	. . . .* . . .	3800,000	2620,031	1179,97	0,913577	2,31636	49,1706
1554	. . . . * . . .	4000,000	2620,031	1379,97	0,913577	2,70898	49,1706
1556	. . . . *. . .	4150,000	2726,883	1423,12	1,064726	2,79368	73,7518
1557	. . . .* . . .	4150,000	3040,010	1109,99	1,507667	2,17899	58,0567
1558	. . . . * . .	4200,000	2678,071	1521,93	0,995678	2,98766	57,1269
1559	*	4500,000	7651,530	-3151,53	8,030992	-6,18668	185,9391
1560	. . . *	4675,000	6072,324	-1397,32	5,797092	-2,74305	169,1135
1561	. . . . .* . .	4850,000	3168,010	1681,99	1,688733	3,30186	65,3259
1562	. . . .* . . .	4900,000	3796,595	1103,41	2,577910	2,16606	75,6563
1563	. . . . . * . .	5000,000	3117,782	1882,22	1,617681	3,69493	69,7279
1565	. . . . .* . .	5500,000	3854,756	1645,24	2,660183	3,22973	71,9335
1566	. . . . . .* .	5500,000	3346,370	2153,63	1,941035	4,22773	60,5227
1567	. . . . . . * .	5600,000	3207,670	2392,33	1,744834	4,69632	58,1846
1568	. . . . . . *.	5700,000	3187,873	2512,13	1,716830	4,93149	56,2525
1569	. . . . . . *	6000,000	3306,776	2693,22	1,885027	5,28699	53,3614
1570	. . . . . . *.	6500,000	4001,388	2498,61	2,867605	4,90495	82,9571
1571	. . . . *. . .	6500,000	5052,533	1447,47	4,354525	2,84148	95,7605
1573	. . . . . . *	6800,000	4001,267	2798,73	2,867434	5,49411	81,4633
1574	. . . . . * .	7480,000	5444,783	2035,22	4,909391	3,99528	106,2365
1575	. . . . . . *.	7500,000	4994,559	2505,44	4,272516	4,91836	99,8700
Минимум	*	1000,000	1471,131	-3151,53	-0,711624	-6,18668	39,5425
Максим.	. . . . . . *	7500,000	7651,530	2798,73	8,030992	5,49411	185,9391
Среднее	. . . .* . . .	4771,346	3614,541	1156,81	2,320382	2,27089	74,5875
Медиана	. . . . * . .	4762,500	3257,223	1525,40	1,814930	2,99447	67,5269

Рисунок 8.17 – Построчный график выбросов

5) Исключим выбросы, выходящие за пределы двух сигма от среднего – наблюдения с номерами: 123, 794, 1516, 1522, 1539, 1542, 1550, 1554, 1556-1575 в диалоговом окне изображённом на рисунке 8.18 (в группе **Исключить наблюдения из Анализа/Графиков – Номера наблюдений** (***Excludecases – orcase number***) перечислим соответствующие номера, разделённые пробелом). После этого проведём регрессионный анализ (рис.8.18).

Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Квартиры)
R= ,82442256 R2= ,67967256 Скорректир. R2= ,67693003
F(5,584)=247,83 p<0,0000 Станд. ошибка оценки: 314,75

N=590	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(584)	p-уров.
Св.член			-839,463	82,42921	-10,1840	0,000000
Расположение	0,075289	0,025020	83,314	27,68631	3,0092	0,002732
Общая	0,785089	0,024048	62,843	1,92492	32,6472	0,000000
Евроремонт	0,156276	0,023588	281,953	42,55681	6,6253	0,000000
Под ключ	0,129114	0,024228	277,816	52,13079	5,3292	0,000000
Ипотека	-0,077580	0,024376	-88,930	27,94197	-3,1827	0,001537

Рисунок 8.18 – Итоги регрессии после отбрасывания выбросов

Как видно – возможность продажи однокомнатной квартиры в ипотеку всё также отрицательно влияет на стоимость жилья: стоимость квартиры убывает в среднем на 88,93 тыс. руб., а увеличение общей площади квартиры на 1 м^2 приводит к удорожанию стоимости квартиры на 62,84 тыс. руб., хорошее месторасположение добавляет к цене 83,31 тыс. руб., квартира под ключ и с евроремонтом станет покупателям дороже на 277, 82 и 281, 95 тыс. руб. соответственно. Увеличился коэффициент корреляции (R=0,82442256) и детерминации (R^2=0,67967256). Как показывает дисперсионный анализ – зависимость значима при уровне значимости менее 0,01 (рис. 8.19).

Эффект	Дисперсионный анализ; ЗП: Цена, тыс. руб. (Nedvig.sta)				
	Сумма квадрат	сс	Средн. квадрат	F	p-уров.
Регресс.	22715224	2	11357612	261,3936	0,00
Остатки	11688113	269	43450		
Итого	34403337				

Рисунок 8.19 – Таблица дисперсионного анализа

Рассмотрение в качестве независимых переменных жилой площади и площади кухни (без учёта выбросов) позволило получить зависимость:

Цена = 398,1817+29,3242*V12+33,5215*V13, в которой, несмотря, на меньшее значение множественного коэффициента корреляции значимыми являются оба фактора, причём площадь кухни влияет на стоимость положительно (рис.8.20).

N=284	Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Nedvig.sta) R= ,70693761 R2= ,49976078 Скорректир. R2= ,49620036 F(2,281)=140,37 p<0,0000 Станд. ошибка оценки: 300,45					
	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(281)	p-уров.
Св.член			398,1817	80,55112	4,94322	0,000001
Жилая площадь	0,701591	0,042677	29,3242	1,78376	16,43953	0,000000
Площадь кухни	0,241961	0,042677	33,5215	5,91251	5,66959	0,000000

Рисунок 8.20 – Итоги регрессии с независимыми переменными: жилая площадь и площадь кухни

Недостаток первой модели заключается в том, что площадь кухни входит в общую площадь квартиры, поэтому последняя модель является более приемлемой, несмотря на некоторое уменьшение множественного коэффициента корреляции.

Отбор данных с условием V7=1 – позволяет отобрать для анализа только 1-комнатные квартиры в кирпичных домах.

N=99	Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Nedvig.sta) R= ,63116315 R2= ,39836693 Скорректир. R2= ,38583290 F(2,96)=31,783 p<,00000 Станд. ошибка оценки: 319,31					
	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(96)	p-уров.
Св.член			512,6002	135,2431	3,790212	0,000263
Жилая площадь	0,612441	0,080732	23,8615	3,1454	7,586111	0,000000
Площадь кухни	0,314278	0,080732	34,3728	8,8297	3,892857	0,000183

Рисунок 8.21 – Итоги регрессии для однокомнатных квартир в кирпичных домах

Если дом не кирпичный (V7=0), то получается зависимость стоимости 1-комнатной квартиры от жилой площади и площади кухни, изображённая на рисунке 8.22.

Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Nedvig.sta)
R= ,74919444 R2= ,56129232 Скорректир. R2= ,55647135
F(2,182)=116,43 p<0,0000 Станд. ошибка оценки: 287,37

N=185	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(182)	p-уров.
Св.член			337,5967	101,6748	3,32036	0,001086
Жилая площадь	0,741557	0,049278	32,3648	2,1507	15,04850	0,000000
Площадь кухни	0,187704	0,049278	31,8302	8,3564	3,80910	0,000191

Рисунок 8.22 – Итоги регрессии для однокомнатных квартир в не кирпичных домах

Задание

Загрузить файл с данными о стоимости жилья в Краснодаре Квартиры.xls. Выполнить пример регрессионного анализа. Провести сравнительный корреляционно-регрессионный анализ исходных данных стоимости квартир в г. Краснодаре, а также одного из районов г. Краснодара, указанного в варианте, с учётом ограничений по общей площади и числу комнат.

№ варианта	Микрорайон	Общая площадь, $м^2$	Число комнат
1	МХГ, СМР, СХИ,ФМР	Менее 54	1
2	40 лет Победы, ЗИП	менее40	1
3	ККБ	Менее 45	1
4	КСК, ГМР, ЧМР	38-48	1
5	ККБ	45-70	2
6	ЧМР, ЦМР, Центр	33-75	2
7	КСК,ГМР, ЧМР	37-98	2
8	Центр, ФМР, ЮМР	30-130	3
9	ЦМР, Центр, ЧМР	65 -110	3
10	ФМР, Центр	От 50 до 70	3

Вопросы для самоконтроля

- Опишите задачи регрессионного анализа и основной метод их решения.
- Какие типы регрессионных зависимостей известны?
- Как оцениваются полученные регрессионные модели?
- Какие условия применимости метода наименьших квадратов?

Практическое занятие №9

Ковариационный анализ

Цель работы: Ознакомиться с возможностями модуля *GLM* – общих линейных моделей. Получить навыки ковариационного анализа данных.

Теоретические сведения

Современный этап развития аграрной науки характеризуется поиском наиболее рациональных методов обработки результатов многолетних полевых опытов. Следует отметить, что проблема оценки влияния на урожайность различных факторов, несмотря на почти столетнюю историю (в 1918г Р.Фишер, решая эту задачу, изобрёл дисперсионный анализ) не является завершенной. Основным подходом обычно являются различные модели дисперсионного анализа.

Практически при проведении полевого опыта часто регистрируют целый ряд сопутствующих неконтролируемых переменных, меняющихся при повторении опыта – это элементы погодных условий на разных стадиях развития растений, а так же элементы структуры урожая.

Так в многолетнем многофакторном эксперименте в ст. Ленинградской Краснодарского края фиксировались климатические факторы (I):

.а) содержание влаги в 0-30 см слое почвы:

X_1 – на период посева,

X_2 – на период возобновления весенней вегетации,

X_3 – на период выхода в трубку,

X_4 – на период колошения,

X_5 – на период полной спелости;

б) содержание влаги в 0-100 см слое почвы:

X_6 – на период посева,

X_7 – на период возобновления весенней вегетации,

X_8 – на период выхода в трубку,

X_9 – на период колошения,

X_{10} – на период полной спелости;

в) количество осадков

X_{11} – за с/х год VIII-VII,

X_{12} – за IX-XI – период осенней вегетации,

X_{13} – за IV-VI – период весенне-летней вегетации,

X_{14} – заV-VI – период от колошения до созревания.

Перечисленные выше климатические факторы (I), переменные X_s – ковариаты, наблюдались на фоне двухфакторного иерархического опыта: фактор *B*-доза внесения удобрений «сгруппирован» внутри главного фактора *A*-предшественник. Фактор А наблюдался на 5 уровнях: эспарцет, озимая пшеница, подсолнечник, кукуруза, озимая пшеница.

Фактор В наблюдался на 3 уровнях: без удобрений, средняя доза NPK, органоминеральная система. Опыт проводился с 1979г. по 1998г., результаты опытных

данных были объединены в одну таблицу, годовые данные полевого опыта использовались в качестве повторений (по 20 повторений для каждого сочетания предшественника и дозы внесения удобрений), расчеты проводились с использованием многолетней средней.

Величины X_s, которые часто называют сопутствующими переменными на самом деле могут иметь большее значение для объяснения различий в средней урожайности, чем предшественник или доза внесения удобрений.

Для совместного учета перечисленных количественных и качественных факторов Р.Фишер (1932г) предложил использовать модель ковариационного анализа, которая в нашем случае будет иметь следующий вид:

$$y_{ijk} = \mu + F_i + I_{j(i)} + \sum_{s=1}^{p} b_s (x_{ijsk} - \overline{X_s}) + \varepsilon_{ijk},$$

где μ – многолетняя средняя,

F_i – член соответствующий влиянию главного фактора *A*-предшественник(i=1, 2, 3, 4, 5);

$I_{j(i)}$ – член соответствующий влиянию фактора В-доза внесения удобрений для определённого значения i-главного фактора (j=1, 2, 3);

$\sum_{s=1}^{p} b_s (x_{ijsk} - \overline{X_s})$ – уравнение регрессии Y на $x_1,\ldots,x_p$ коэффициенты b_s которого показывают важность влияния сопутствующих переменных на y_{ij}.

Уравнение регрессии призвано снизить значение остаточной дисперсии для получения более надёжных результатов дисперсионного анализа.

Следует сказать, что ковариационный анализ сводится к дисперсионному, или к регрессионному – так в последнем случае для этого необходимо ввести в уравнение регрессии фиктивные переменные, характеризующие качественные факторные признаки.

Ковариационный анализ в Statistica

Выполнив команду **Анализ – Углубленные методы анализа – Общие линейные модели**(**Statistics –** ***Advanced Linear/Nonlinear Models–GLM (General Linear Models)),*** мы получим соответствующее диалоговое окно (рис.9.1).

Описание всех типов моделей можно найти в справке, следует лишь отметить, что в этом модуле обобщаются практически все модели дисперсионного, регрессионного и ковариационного анализа при разных условиях применения с точки зрения рассмотрения не одной зависимой переменной, а сразу нескольких.

Это позволяет с помощью метода обобщённых обратных матриц получать единственное решение в случае плохой обусловленности матрицы плана, а также использовать многомерные критерии адекватности. (Можно использовать те же суммы квадратов, что и в дисперсионном анализе.)

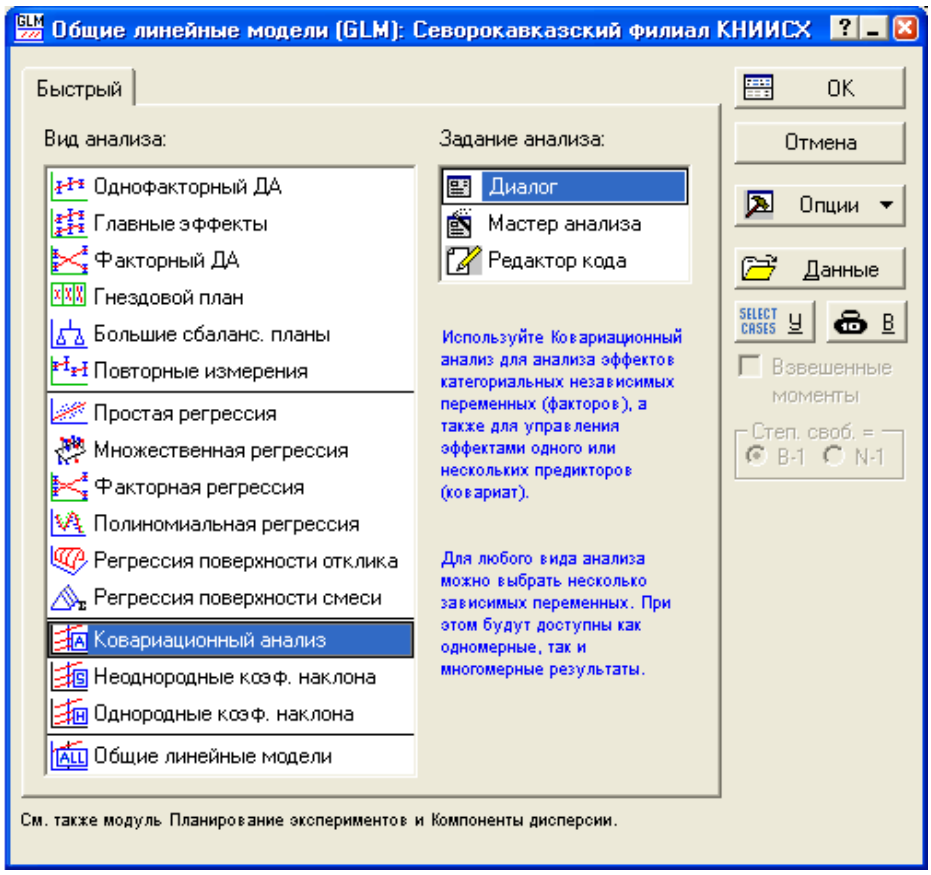


Рисунок 9.1 – Диалоговое окно Общих линейных моделей

Пример. Рассмотрим построение и анализ ковариационной модели урожайности по приведённым выше данным с использованием модуля ***GLM*** – **Общих линейных моделей** (рис.9.1) (исходные данные – файл Северокавказский филиал КНИИСХ.xls).

Выберем **Ковариационный анализ – Диалог**(***Analysis of covariance – Quick specs dialog***) и выберем в диалоговом окне **Зависимые переменные** (***Dependent vars***), **Категориальные**(***Categorical pred***.) и **Непрерывные** (***continuons pred***) **предикторы** (ковариаты) (рис.9.2).

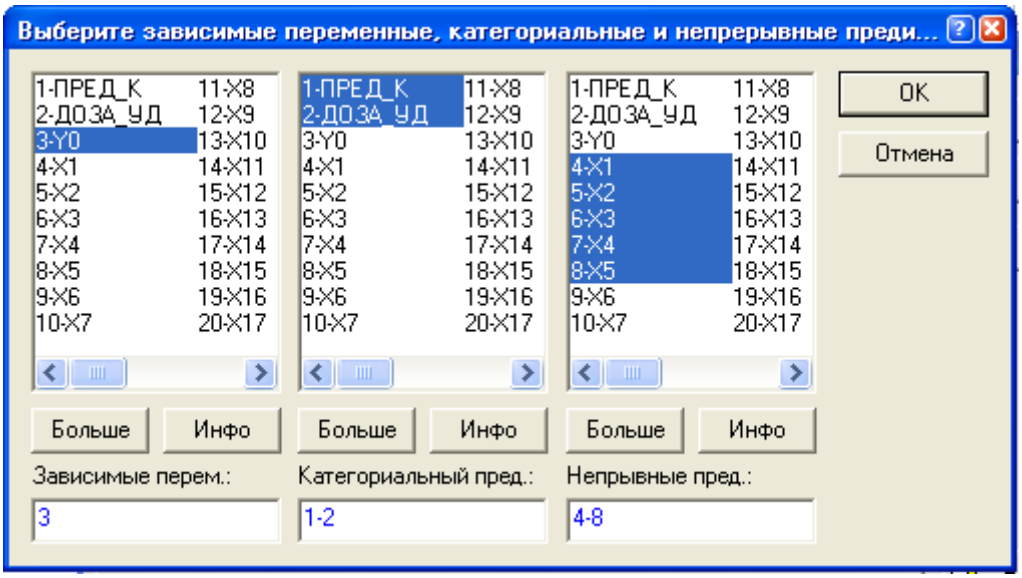


Рисунок 9.2 – Диалоговое окно выбора переменных в модуле *GLM*

После выбора **ОК** получим диалоговое окно вывода результатов, аналогичное окну дисперсионного анализа.

После выбора кнопки **Все эффекты** (***Test all effects***), получим таблицу всех эффектов (рис.9.3) из которой видно, что практически все факторы и их взаимодействия статистически существенно влияют на урожайность (за исключением фактора X_1 – содержание влаги в 0-30 см слое почвы на период посева) – все значимые эффекты выделяются красным цветом.

Эффект	Одномерный критерий значимости для Y0 (Северокавказский филиал КНИИСХ.STA) Сигма-ограниченная параметризация Декомпозиция гипотезы SS	Степени Свободы	MS	F	p
Св. член	4876,74	1	4876,743	48,35132	0,000000
"X1"	287,91	1	287,909	2,85452	0,092219
"X2"	498,21	1	498,214	4,93963	0,027038
"X3"	2293,93	1	2293,928	22,74355	0,000003
"X4"	7477,57	1	7477,571	74,13769	0,000000
"X5"	3987,13	1	3987,133	39,53113	0,000000
ПРЕД_К	4650,04	3	1550,014	15,36788	0,000000
ДОЗА_УД	17238,39	2	8619,196	85,45653	0,000000
ПРЕД_К*ДОЗА_УД	2511,03	6	418,505	4,14934	0,000516
Ошиб.	28543,55	283	100,861		

Рисунок 9.3 – Таблица всех эффектов

Для визуализации различий урожайности выберем кнопку **Все эффекты/графики**(***All effects/Graphs***). В результате получим диалоговое окно – **Таблицы всех эффектов** (***Table of All Effects***), позволяющее выбирать эффекты и их взаимодействия, выберем взаимодействие факторов (рис. 9.4).

Выбрав в группе **Отображать – График** (***Display–Graph***) мы можем получить рисунки изображённые на рисунке 9.5.

Из рисунка 9.5 видно, что если предшественник эспарцет, то во всех случаях (разных доз удобрений) урожайность выше, чем по другим предшественникам. Для средней дозы удобрений наибольшая урожайность по предшественнику эспарцет, затем по подсолнечнику, кукурузе и озимой пшенице. То же соотношение остаётся и при органоминеральной системе удобрений.

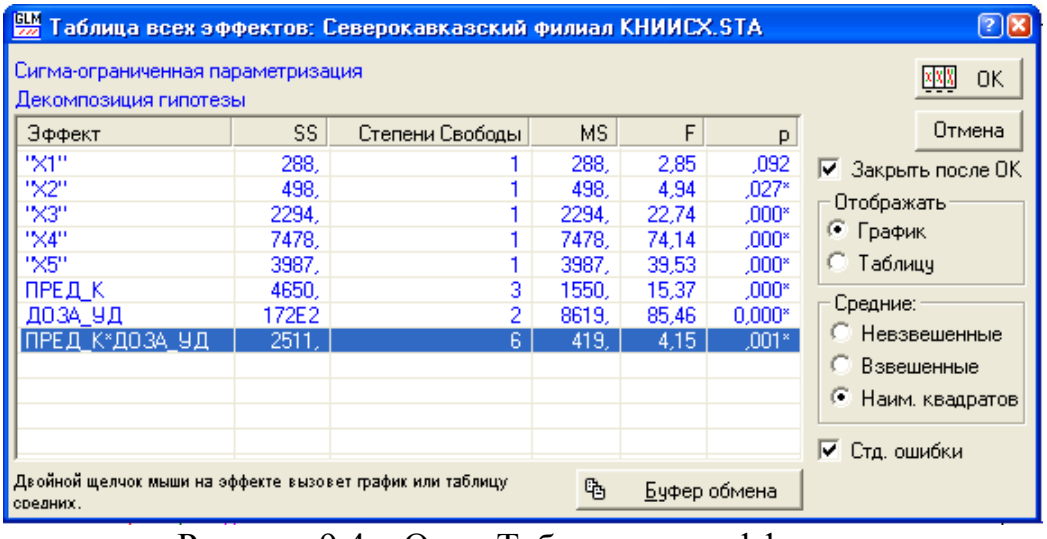


Рисунок 9.4 – Окно Таблица всех эффектов

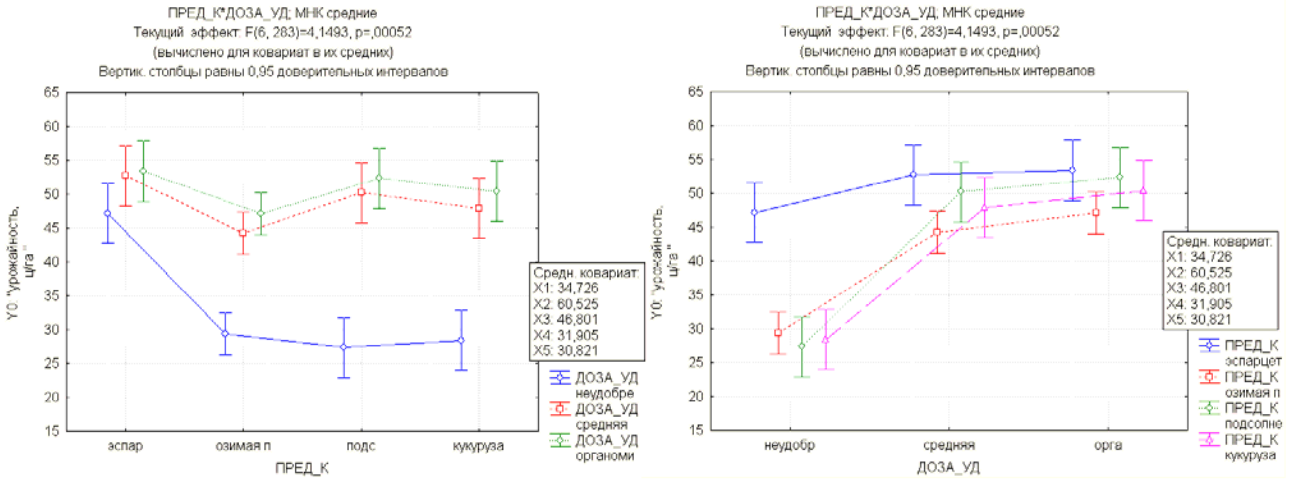


Рисунок 9.5 – Графики зависимости урожайности от взаимодействия Предшественник*Доза удобрений

ПРЕД_К*ДОЗА_УД; МНК средние (Северокавказский филиал КНИИСХ.STA)
Текущий эффект: F(6, 283)=4,1493, p=,00052
(вычислено для ковариат в их средних)

N ячейки	ПРЕД_К	ДОЗА_УД	Y0 Среднее	Y0 Ст.Ош.	Y0 -95,00%	Y0 +95,00%	N
1	эспарцет	неудобре	47,15611	2,258107	42,71130	51,60093	20
2	эспарцет	средняя	52,71111	2,258107	48,26630	57,15593	20
3	эспарцет	органоми	53,37611	2,258107	48,93130	57,82093	20
4	озимая п	неудобре	29,34334	1,590495	26,21264	32,47405	40
5	озимая п	средняя	44,22334	1,590495	41,09264	47,35405	40
6	озимая п	органоми	47,11084	1,590495	43,98014	50,24155	40
7	подсолне	неудобре	27,36094	2,251826	22,92849	31,79339	20
8	подсолне	средняя	50,20094	2,251826	45,76849	54,63339	20
9	подсолне	органоми	52,35094	2,251826	47,91849	56,78339	20
10	кукуруза	неудобре	28,40126	2,249278	23,97382	32,82870	20
11	кукуруза	средняя	47,94126	2,249278	43,51382	52,36870	20
12	кукуруза	органоми	50,40126	2,249278	45,97382	54,82870	20

Рисунок 9.6– Таблица с описательными статистиками по уровням взаимодействия факторов предшественник и доза внесения удобрений

Выбрав в группе **Отображать – Таблицу** (***Display – Spreadsheet***) (кнопка **Все эффекты/графики** (***All effects/Graphs***)), получим таблицу, изображённую на рисунке 9.6. Из неё видно, что при уровне значимости не менее 0,0005 – средние урожайности согласно эффектам взаимодействия и их доверительные интервалы (рис. 9.6).

С помощью кнопки **Оценить** (***Estimate***) в группе **Межгрупповые эффекты** (***Design terms***) окна **Результаты анализа** (***GLM Results***) можно задавать уровни оцениваемых параметров и проверять, таким образом, рабочие гипотезы о существенности влияния определённых уровней факторов. Выбор кнопки **Общая *R* модели** (***Whole model R***) позволяет получить оценку доли изменчивости урожайности, которая объясняется построенной моделью.

Зависим. Перемен.	SS модели и SS остатков (Северокавказский филиал КНИИСХ.STA)										
	Множеств R	Множеств R2	Скоррект R2	SS Модель	cc Модель	MS Модель	SS Остаток	cc Остаток	MS Остаток	F	p
Y0	0,791816	0,626973	0,605883	47975,19	16	2998,450	28543,55	283	100,8606	29,72865	0,00

Рисунок 9.7 – Таблица *SS* модели и *SS* остатков

В нашем случае (рис.9.7) построенная модель объясняет 62,697% вариации урожайности.

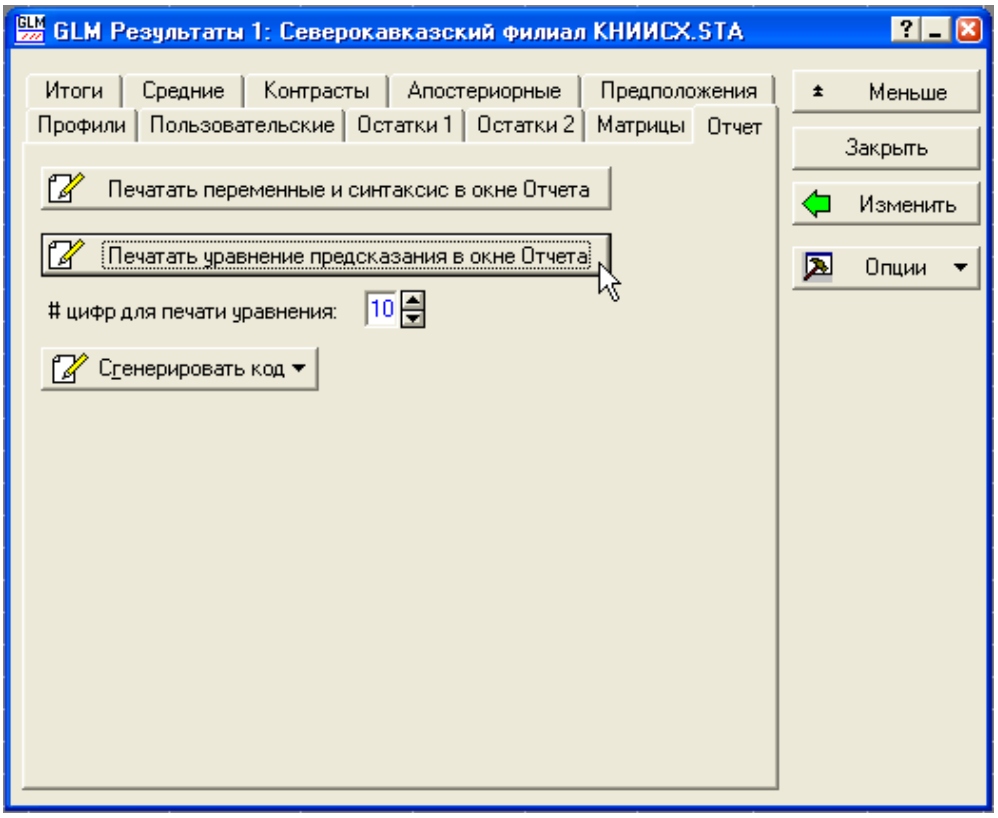


Рисунок 9.8 – Вкладка отчёт окна результатов анализа

С помощью вкладки **Отчёт** (***Report***) мы можем получить уравнение предсказания в виде сигма-ограниченной модели (кодирования категориальных переменных) в которой категориальным переменным ставится в соответствие значения в сумме составляющие 0.

Предск. уравнение для: *Y0*: "урожайность, ц/га (без удобрений)"

Y0=27.0953646+.052946164*"X1"+.133697715*"X2"-.26522900*"X3"
+.408402893*"X4"+.213219871*"X5"+6.86632419*"ПРЕД_К"("эспарцет")
-3.9889446*"ПРЕД_К"("озимая п")-.91051586*"ПРЕД_К"("подсолне")
-11.149375*"ДОЗА_УД"("неудобре")+4.55437500*"ДОЗА_УД"("средняя")
+7.22437500*"ПРЕД_К"*"ДОЗА_УД"(1)-2.9243750*"ПРЕД_К"*"ДОЗА_УД"
(2)+.266875000*"ПРЕД_К"*"ДОЗА_УД"(3)-.55687500*"ПРЕД_К"*"ДОЗА_УД"
(4)-4.7939583*"ПРЕД_К"*"ДОЗА_УД"(5)+2.34229167*"ПРЕД_К"*"ДОЗА_УД"(6)

Обычно, после установления различий в среднем значении зависимой переменной для разных категорий требуется установить величину различия для заданных категорий. Для решения подобной задачи исследуют контрасты (см. занятие № 9).

Замечание. Ковариационный анализ предполагает нормальное распределение опытных данных, однородность дисперсий и равенство коэффициентов регрессии по группам. Для построения адекватной модели в классическом смысле необходимо проверить гипотезу о нормальном распределении и применить многомерные критерии проверки однородности дисперсий и ковариаций Бокса, а также критерий параллельности. Выполнение приведённых выше условий не является критичным, то есть отклонения допускаются.

Задание

Используя файл исходных данных – Северокавказский филиал КНИИСХ.xls, проведите ковариационный анализ. Выберете: в качестве зависимой переменной – урожайность, в качестве категориальных предикторов предшественник и дозу внесения удобрений. В качестве непрерывных предикторов (ковариат) подберите наиболее оптимальный с вашей точки зрения набор переменных X_s (s= 1,…, 19). Сравните результаты анализа с ковариатами и без них.

Вопросы для самоконтроля

- Что изучает ковариационный анализ, в чём сходство и различие с дисперсионным и регрессионным анализом?
- В чём сущность построения модели ковариационного анализа?
- Какие классические условия применения ковариационного анализа?

Практическое занятие №10

Кластерный и дискрминантный анализ

Цель работы: Ознакомиться с возможностями получения и изучения в имеющихся данных однородных (в определённом смысле) групп, с помощью кластерного и дискриминантного анализа.

Теоретические сведения

Кластерный анализ используется для получения в имеющихся данных однородных групп объектов или единиц, которые называются кластерами. Известно четыре прикладных класса задач, сводящихся к применению кластерного анализа:

a. разработка классификации объектов;

b. изучение различных вариантов группировки объектов;

c. получение гипотез на основе анализа данных (т.е. разведочный анализ данных);

d. проверка гипотез о существовании выделенных групп объектов.

Часто указанные задачи решаются параллельно.

Методы кластерного анализа в основном носят эвристический характер, причём разные подходы порождают различные кластеры.

Близость объектов друг к другу характеризуется мерой сходства, которая может быть: коэффициентом корреляции; мерой расстояния; коэффициентом ассоциативности; вероятностной мерой сходства.

Более распространены метрические меры сходства (расстояния), основанные на мерах $\rho_E^2, \rho_E, \rho_C, \rho_\infty, \rho_M$ (см. ниже). Расстояние Махаланобиса не относят к метрическим (так как не выполняются условия метрики (симметрия, неравенство треугольника, различимость нетождественных и неразличимость идентичных объектов), оно связано, с помощью ковариационной матрицы, с корреляциями составляющих векторов наблюдений.

Процент несогласия используется, когда данные являются категориальными. (1 – *r* Пирсона) используется, когда желательно сократить размерность вектора X за счёт выявления групп сильно связанных компонент.

Пусть X_N – некоторое подмножество *k*-мерного пространства R^k: $X_N \subseteq R^k$; $\mathbf{X_i}=(x_{i1}, x_{i2}, \ldots, x_{ik})$, где $i = \overline{1, N}$ – элементы X_N; Σ^{-1} – матрица обратная ковариационной. Тогда можно записать следующие меры близости между произвольными векторами X_i, X_j из X_N:

- Расстояние Махаланобиса – ρ_0 используется, если компоненты вектора X зависимы. Следующие три меры можно считать частными случаями ρ_0:
- Евклидово расстояние – ρ_E используется если компоненты вектора наблюдений X независимы, однородны и подчиняются нормальному закону распределения с постоянной дисперсией (иногда признаки нормируют);

- Квадрат Евклидова расстояния – ρ_E^2 используется, если мы хотим придать большие веса более отдаленным друг от друга объектам;
- Манхэттенское расстояние – ρ_C, приводит практически к тем же результатам, что и евклидово, но влияние больших разностей (выбросов) уменьшается;
- Расстояние Чебышева – ρ_∞ используется, если необходимо различать объекты, отличающиеся хотя бы по одной координате;
- Степенное расстояние Минковского – ρ_M, используется при необходимости изменить (в большую или меньшую сторону) вес размерности по которой объекты отличаются сильно. p – влияет на постепенное взвешивание по координатам, q – влияет на прогрессивное взвешивание больших расстояний между объектами. При p=q=2 – $\rho_M = \rho_E$.

Таблица 10.1 – Основные меры близости, используемые в кластерном анализе

Мера близости (*Distance measure*)	Формула
Расстояние Махаланобиса	$\rho_0(X_i, X_j) = \sqrt{(X_i - X_j)^T \Sigma^{-1} (X_i - X_j)}$
Квадратевклидоварасстояния (*Squared Euclidean distances*)	$\rho_E^2(X_i, X_j) = \sum_{s=1}^{k} (x_{is} - x_{is})^2$
Евклидово расстояние (*Euclidean distances*)	$\rho_E(X_i, X_j) = \sqrt{\sum_{s=1}^{k} (x_{is} - x_{is})^2}$
Манхэттенскоерасстояние (*City-block (Manhatten) distances*)	$\rho_C(X_i, X_j) = \sqrt{\sum_{s-1}^{k} \lvert x_{is} - x_{is} \rvert}$
РасстояниеЧебышева (*Chebychev distance metric*)	$\rho_\infty(X_i, X_j) = \max_{1 \le S \le k} \lvert x_{is} - x_{is} \rvert$
Power: SUM(ABS(x-y)^p)^1/q - Степенное расстояние Минковского	$\rho_M(X_i, X_j) = \sqrt[q]{\sum_{s=1}^{k} (x_{is} - x_{is})^p}$
Процент несогласия (*Percent disagreement*)	$\rho_{PD}(X_i, X_j) = (количество\ x_{is} \ne x_{js})/k$, *где* $s = \overline{1,k}$
1 – *r* Пирсона (1- *Pearson r*) – расстояние Пирсона	$\rho_P(X_i, X_j) = 1 - r$

Основная идея кластерного анализа, использующая понятие близости, состоит в преобразовании пространства описаний таким образом, что все точки одного множества близки друг другу, а точки различных множеств удалены на некоторое расстояние.

Для этого необходимо определить меру расстояния между кластерами, например, если имеется два кластера S_l и S_m с n_l и n_m элементами, которые объединяются в один S_r, кроме того, имеется кластер S_t, ρ_{ij} – расстояние между *i*-м и *j*-м кластерами, то можно установить следующие меры близости между кластерами S_r и S_t, зная расстояния ρ_{lt} *и* ρ_{mt} (первоначально предполагается, что каждая точка – отдельный кластер):

Таблица 10.2 – Процедуры классификации

Правило объединения (*Amalgamation (linkage) rule*)	Формула
Метод одиночной связи (метод «ближайшего соседа») (*Single Linkage*)	$\rho_{\min}(S_r, S_t) = \min(\rho_{lt}, \rho_{mt})$
Метод полной связи (метод «дальнего соседа») (*Complete Linkage*)	$\rho_{\max}(S_r, S_t) = \max(\rho_{lt}, \rho_{mt})$
Невзвешенное попарное среднее (*Unweightedpair-groupaverage*)	$\rho_{UPGMA}(S_r, S_t) = \frac{\rho_{lt} + \rho_{mt}}{2}$
Взвешенное попарное среднее (*Weighted pair-group average*)	$\rho_{WPGMA}(S_r, S_t) = \frac{n_l\rho_{lt} + n_m\rho_{mt}}{n_l + n_m}$
Невзвешенный центроидный метод (*Unweightedpair-groupcentroid*)	$\rho_{UPGMC}(S_r, S_t) = \frac{\rho(\overline{X}_l, \overline{X}_t) + \rho(\overline{X}_m, \overline{X}_t)}{2}$
Взвешенныйцентроидныйметод (*Weighted pair-group centroid (median)*)	$\rho_{WPGMC}(S_r, S_t) = \frac{n_l\rho(\overline{X}_l, \overline{X}_t) + n_m\rho(\overline{X}_m, \overline{X}_t)}{n_l + n_m}$
Метод Варда (Уорда) (*Ward's method*)	$V_r = \sum_{i=1}^{n_r}\sum_{j=1}^{k}(x_{ij} - \overline{x}_{jr})^2$

Меры близости, перечисленные выше, используются в иерархических агломеративных процедурах кластерного анализа, позволяющих из отдельных точек по пречисленным выше правилам получать кластеры, которые обычно изображаются в виде графа – дерева (дендрограммы).

В качестве меры близости используются $\rho_E^2, \rho_E, \rho_C, \rho_\infty, \rho_M$ и др. Для метода «ближайшего соседа» в качестве расстояния между классами выбирается наименьшее расстояние между всеми элементами, принадлежащими разным классам; для метода «дальнего соседа» выбирается наибольшее расстояние между всеми элементами принадлежащими разным классам. Все правила объединения, кроме метода Варда, работают с метриками.

Процесс объединения кластеров в иерархических агломеративных процедурах кластерного анализа происходит последовательно – на основании матрицы расстояний или сходства последовательно объединяются наиболее близкие объекты, пока все данные не объединятся в один кластер. (В иерархических дивизимных процедурах кластерного анализа наоборот из одного кластера получается разложение на ряд дочерних.)

Метод Варда объединяет те объекты (кластеры), которые дают наименьшее приращение величине V_r, тем самым, минимизируя дисперсию внутри кластеров. Метод Варда имеет тенденцию к созданию кластеров примерно равных размеров и имеющих гиперсферическую форму. Следует отметить, что этот метод чаще, чем другие позволяет получать осмысленные результаты.

Общее правило заключается в подборе мер и объединяющих правил для возможности содержательной интерпретации.

Часто рассматривается возможность классификации не только по объектам (наблюдениям), но и по признакам (переменным).

Кроме иерархического кластерного анализа (Joining (treeclustering)), который часто является первым этапом для оценки возможного числа кластеров, используют метод *k*-средних (*K – means clustering*) и двухвходовую процедуру (*Two-wayjoining*).

Метод *K*-средних – позволяет итеративно подобрать *k* – центров для кластеров, для которых расстояния внутри кластеров минимально, а между кластерами максимально.

Двухвходовая процедура объединения используется, если исследователь предполагает, что на образование кластеров одинаково влияют как наблюдения (строки) так и факторы (столбцы).

Оценку адекватности разбиения на кластеры можно получить с помощью дисперсионного анализа.

Дискриминантный анализ. Часто кластерный анализ называют классификацией без учителя, а дискриминантный анализ – классификацией с учителем. В дискриминантном анализе, в отличие от кластерного, известна классификация объектов.

Линейный дискриминантный анализ (Фишера) обычно основывается на предположении, что данные подчиняются многомерному нормальному закону. Рассматривается две задачи:

1) установить правило, согласно которому объект относится к одному из известных классов – обычно это (если выполняется условие линейной разделимости данных) линейная функция от признаков – функция классификации

$S_i = c_i + w_{i1}*x_1 + w_{i2}*x_2 + ... + w_{im}*x_m$.

2) по найденным правилам (функциям классификации) классифицировать новые объекты – объект относится к *i*-му классу, если значение функции классификации S_i – наибольшее.

Значимость различения объектов можно оценивать с помощью дисперсионного анализа. Кроме того, рассматривается специальная характеристика – функция от канонических корней, λ-статистика Уилкса, чем она меньше тем разделение классов лучше.

Кластерный и дискриминантный анализ в Statistica

Выполнив команду **Анализ – Многомерный разведочный анализ – Кластерный анализ** (**Дискриминантный анализ**) ***(Statistics – Multivariate Exploratory Techniques – Cluster Analysis (Discriminant Analysis))*** мы получим соответствующее диалоговое окно.

Замечание. Разберите примеры кластерного (классификации автомобилей различных марок) и дискриминантного анализа (изучения дискриминации для цветов ириса), находящиеся в справке системы *Statistica* 6.0(1).

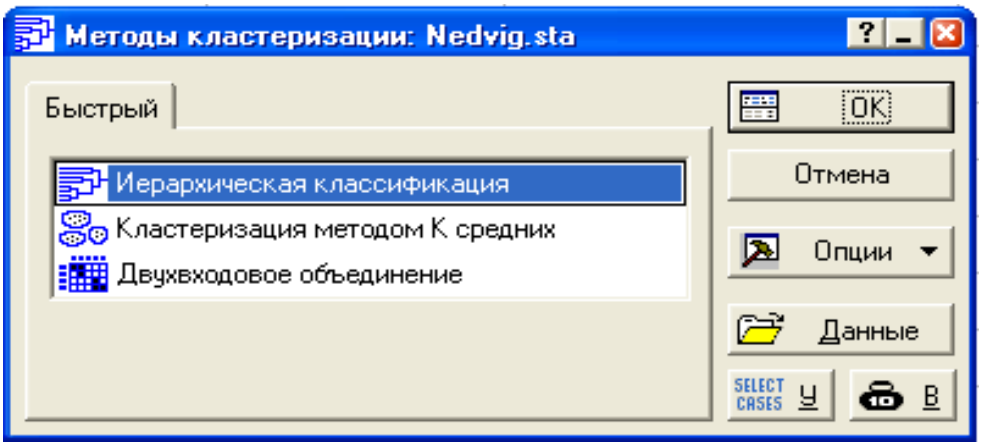


Рисунок 10.1 – Диалоговое окно Методы кластеризации

Пример 1. Кластерный анализ недвижимости Краснодара. По данным практического занятия №8 проведём иерархическую классификацию недвижимости, используя правило объединения (метод) Варда и Евклидову меру близости, в качестве переменных для анализа выберем только количественные переменные для однокомнатных квартир: общая площадь, жилая площадь, площадь кухни, стоимость.

Выполним команду **Анализ** (***Statistics***)– **Многомерный разведочный анализ** (***Multivariate Exploratory Technicues***) – **Кластерный анализ** (***Cluster Analysis***) – **Иерархическая классификация** (***Joining*** (***treeclustering***)). Заполним диалоговое окно в соответствии с рисунком 10.2.

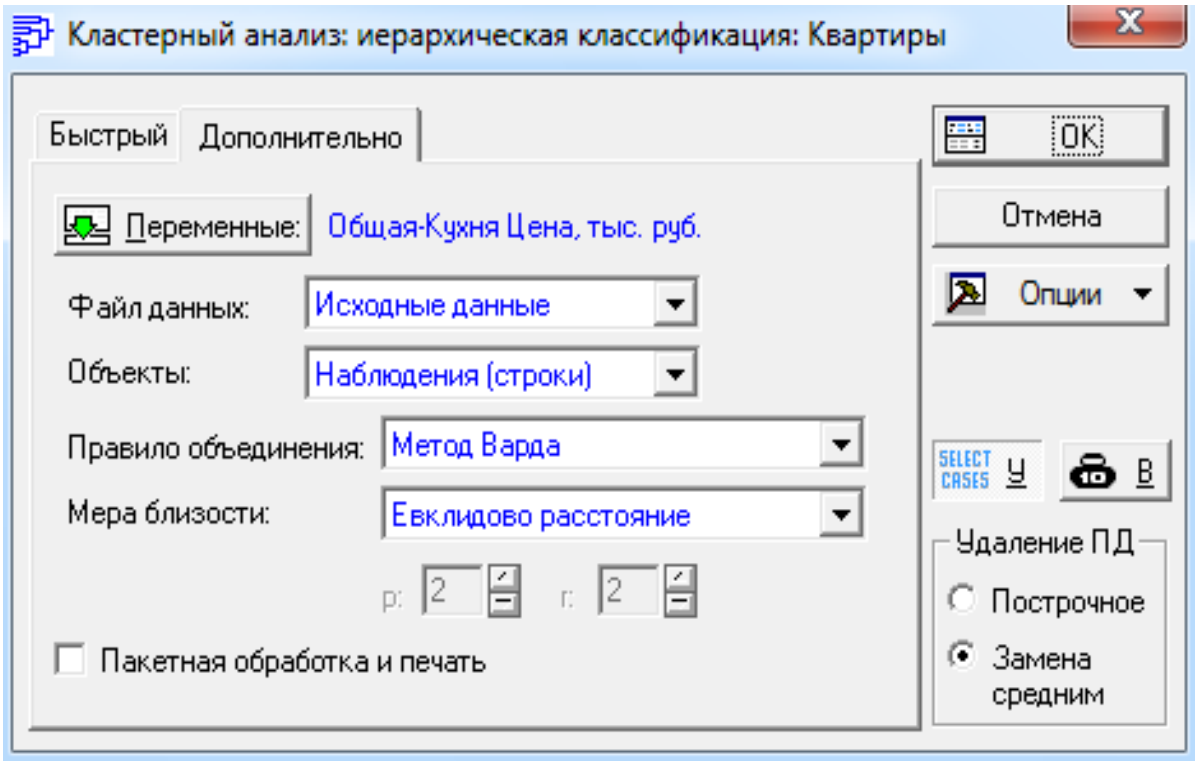


Рисунок 10.2 – Диалоговое окно Кластерный анализ:
иерархическая классификация

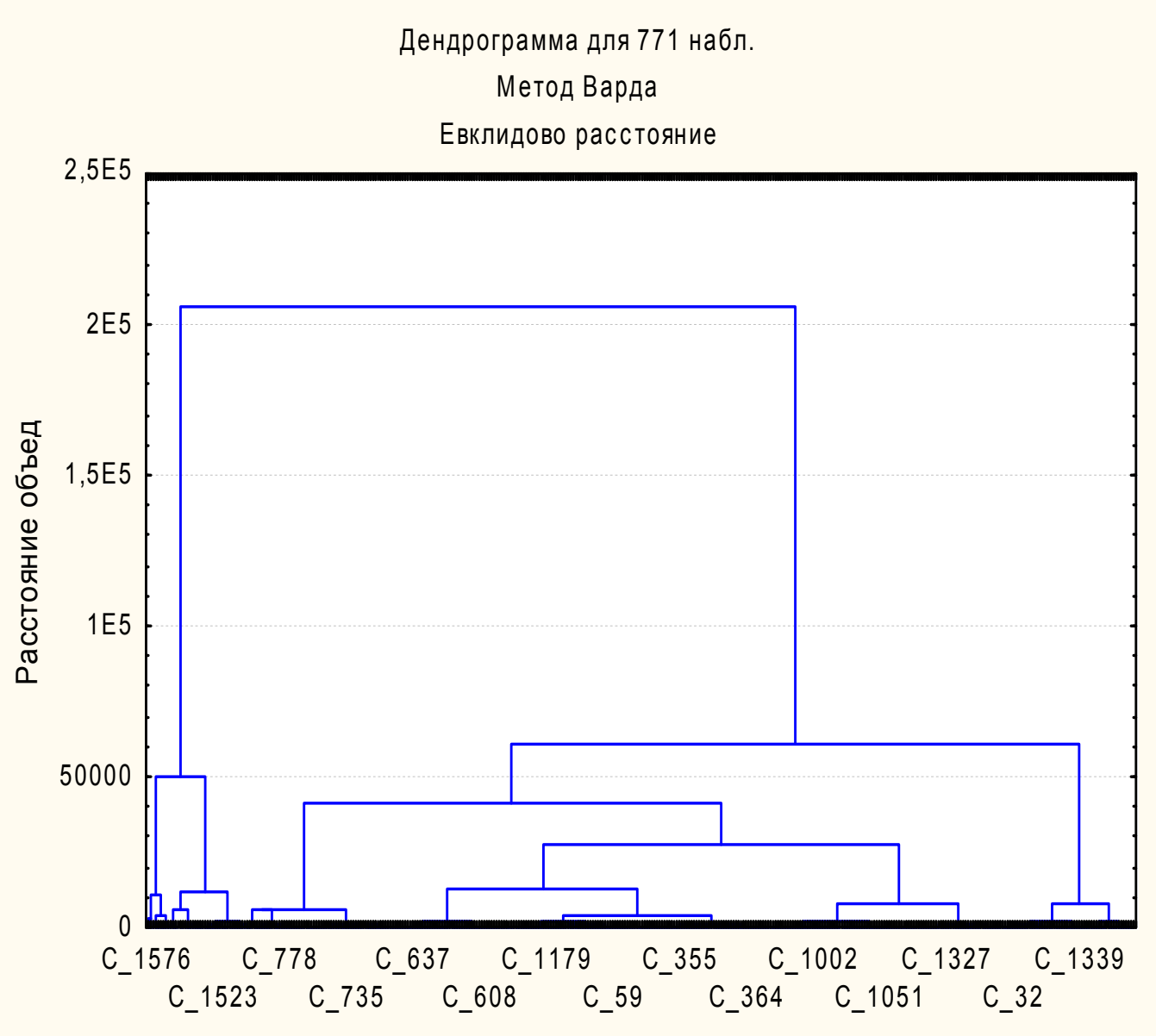


Рисунок 10.3 – Вертикальная дендрограмма иерархической классификации 1 комнатных квартир

Выберем **ОК – Вертикальная диаграмма** (***Verticalicicle plot***) и получим дендрограмму, изображённую на рисунке 10.3.

Визуализация в виде дендрограммы (рис.10.3) позволяет сделать предположение о количестве кластеров (от 2-х до 10) – с нашей точки зрения следует предположить о существовании трёх кластеров (которые соответствуют расстоянию объединения 20 000).

Найдём эти 3 кластера с помощью метода *K* средних (рис.10.1, 10.4). (Для этого вернёмся в предыдущее окно и выберем кластеризацию методом *K* средних: **Отмена** (***Cancel***) – **Кластеризация методом *K* средних** (***K-means clustering***)).

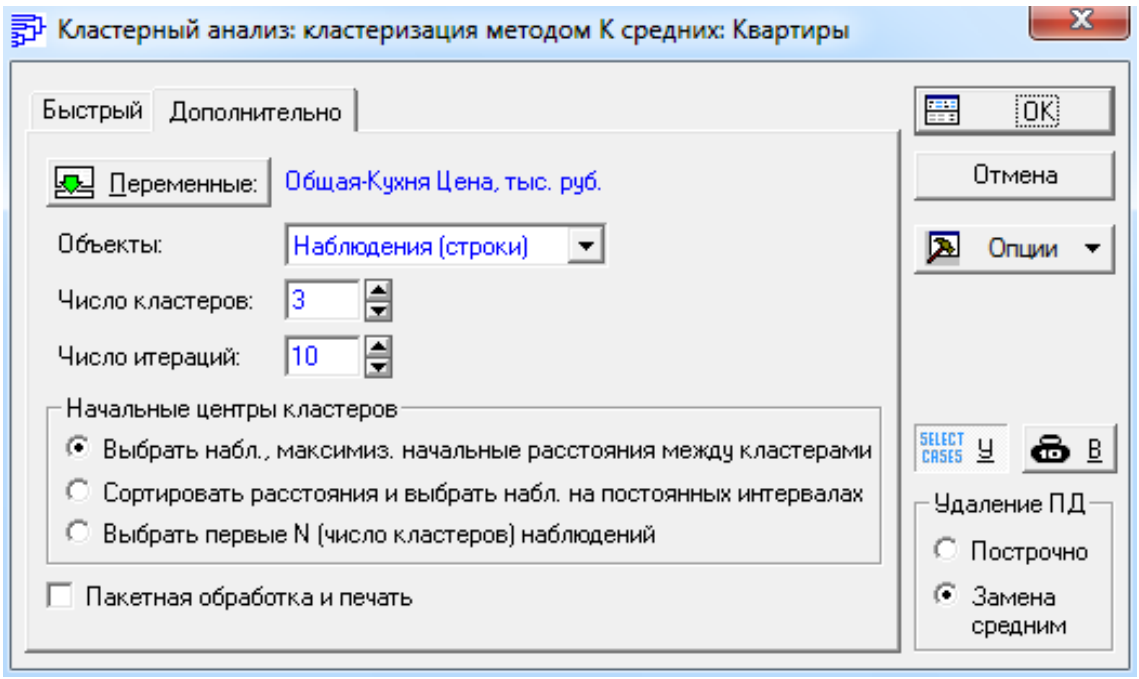


Рисунок 10.4 – Диалоговое окно кластеризация методом *К* средних

Кластер	Евклидовы расст. между кластерами (Квартиры) Расстояния под диагональю Квадраты расстояний над диагональю		
Номер	Но. 1	Но. 2	Но. 3
Но. 1	0,000	4750870	646667
Но. 2	2179,649	0	1891997
Но. 3	804,156	1375	0

Рисунок 10.5 – Матрица расстояний между кластерами и их квадратов

	Средн.класт. (Квартиры)		
перемен.	**Кластер Но. 1**	Кластер Но. 2	Кластер Но. 3
Общая	38,160	72,750	55,212
Жилая	18,461	34,313	25,106
Кухня	8,472	13,380	12,313
Цена, тыс. руб.	1642,433	6001,563	3250,636

Рисунок 10.6 – Среднее для кластеров

Расстояние между первым и вторым кластером 2179,649; между первым и третьим – 804,156; между вторым и третьим – 1375 (рис.10.5). Средние площади и цена для кластеров изображены на рисунке 10.6.

Дисперсионный анализ показывает, что указанные переменные (Общая площадь, площадь кухни, полезная площадь и цена) статистически существенно влияют на результаты классификации (рис.10.7).

	Дисперсионный анализ (Квартиры)					
перемен.	**Между SS**	cc	Внутри SS	cc	F	значим. p
Общая	34680	2	33779	768	394,242	0,000000
Жилая	6313	2	13262	768	182,797	0,000000
Кухня	1216	2	6210	768	75,201	0,000000
Цена, тыс. руб.	434604000	2	76170100	768	2190,990	0,000000

Рисунок 10.7 – Дисперсионный анализ результатов классификации

График средних для кластеров (рис.10.8А) показывает, что найденная классификация достаточно хорошо разбивает данные на классы. Следует отметить, что более наглядный результат можно получить, например, нормировав данные. Для этого скопируйте переменные V11, V12, V13,V26 в соседний диапазон и выполните команду Данные – Стандартизировать (Data - , рис. 10.8В).

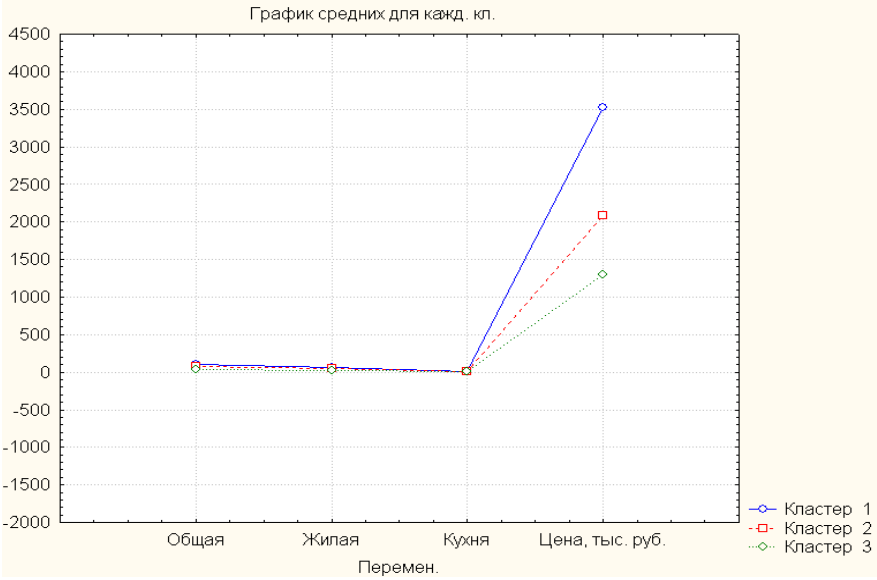


Рисунок 10.8А– График средних для кластеров

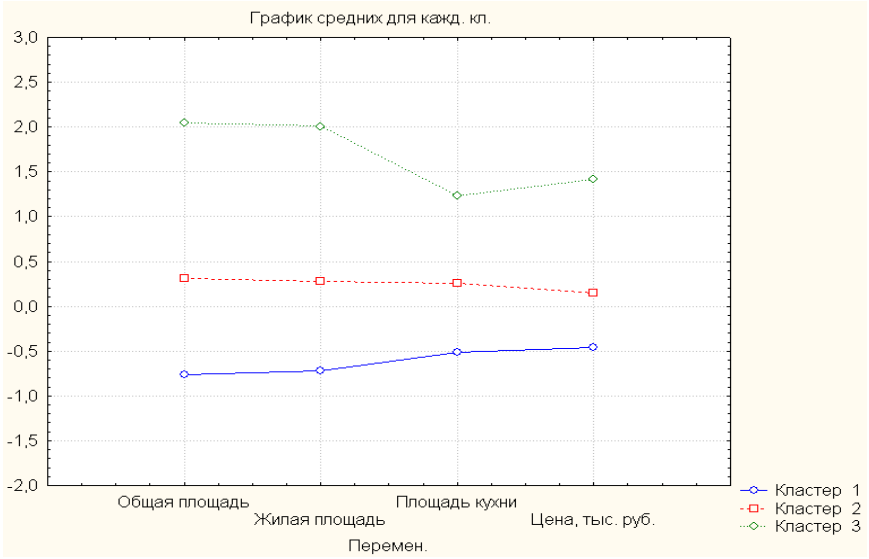


Рисунок 10.8В – График средних для кластеров со стандартизированными значениями переменных

Описательные статистики для найденных кластеров изображены на рисунках 10.9-10.11.

Описат.статистики для кластера 1 (Квартиры)
Кластер содержит 689 набл.

перемен.	Среднее	Стандарт отклон.	Дисперс.
Общая	38,160	5,0013	25,01
Жилая	18,461	2,2805	5,20
Кухня	8,472	2,7888	7,78
Цена, тыс. руб.	1642,433	259,4828	67331,35

Рисунок 10.9 – Описательные статистики для 1 кластера

Рисунок 10.9 показывает, что для первого кластера однокомнатных квартир, средняя площадь квартир выставленных на продажу составила 38 кв. м., жилая площадь – 18 кв. м., площадь кухни – 8 кв. м. Средняя цена, сложившаяся на 1 октября 2013 года однокомнатных составила 1642 тыс. руб. Стандартное отклонение показывает, что площадь квартир находится в пределах от 32 кв. м. до 43 кв. м., жилая площадь и площадь кухни может откланяться в среднем на 2 – 3 кв. м., цена отклоняется на 260 тыс. руб. Значит, минимальная стоимость однокомнатной квартиры составляла 1382 тыс. руб. (1642-260), квартиру за эту цену можно приобрести на этапе строительства, а максимальная стоимость составила 1902 тыс. руб., что соответствует покупки квартиры в готовом сданном доме с отделкой «под ключ».

Описат.статистики для кластера 2 (Квартиры)
Кластер содержит 16 набл.

перемен.	**Среднее**	Стандарт отклон.	Дисперс.
Общая	72,750	22,831	521
Жилая	34,313	18,786	353
Кухня	13,380	3,066	9
Цена, тыс. руб.	6001,563	1124,541	1264592

Рисунок 10.10 – Описательные статистики для 2 кластера

Анализируя 2 кластер (рис. 10.10) видно, что средняя площадь квартиры составила 73 кв. м. и колебалась от 50 кв. м. до 96 кв. м., жилая площадь в среднем составила 34 кв. м., кухня 13 кв. м. Стандартное отклонение стоимости однокомнатных квартир во вотом клатере в 4 раза выше, чем у однокомнатных квартир первого кластера. Так, минимальная цена квартиры составила 4876 тыс. руб., а самая дорогая квартира стоила во втором кластере 7126 тыс. руб.

Описат.статистики для кластера 3 (Квартиры)
Кластер содержит 66 набл.

перемен.	Среднее	Стандарт отклон.	Дисперс.
Общая	55,212	11,6031	134,6
Жилая	25,106	8,2184	67,5
Кухня	12,313	3,3245	11,1
Цена, тыс. руб.	3250,636	409,0750	167342,3

Рисунок 10.11 – Описательные статистики для 3 кластера

На рисунке 10.11 представлен 3 кластер однокомнатных квартир. Так, средняя площадь квартиры третьего кластера составила 55 кв. м. и колебалась от 43 кв. м. до 67 кв. м., площадь кухни колебалась от 9 до 15 кв. м. Средняя цена квартир составила 3251 тыс. руб. Самая дешевая квартира третьего кластера стоит 2842 тыс. руб., а самая дорогая – 3660 тыс. руб. В этом кластере наблюдается относительно небольшая колеблемость факторов, но она выше, чем в первом кластере.

Сохраним, найденную классификацию и расстояния – рисунки 10.12-10.14.

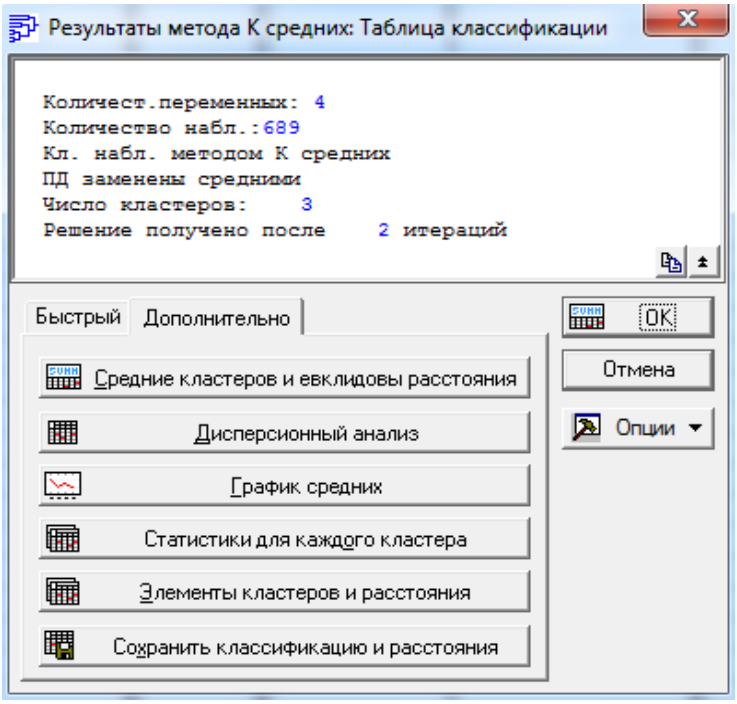


Рисунок 10.12 – Сохранение найденной классификации.

Кроме переменных, использованных при кластеризации, отметим ещё переменные: район, тип дома (рис. 10.13).

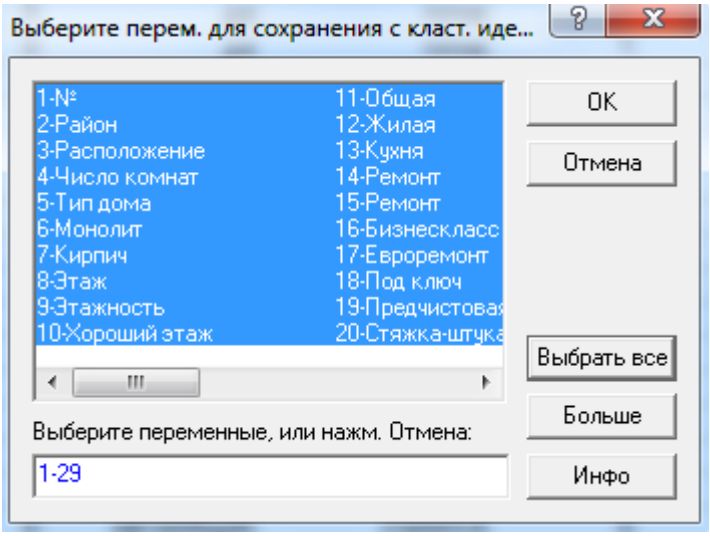


Рисунок 10.13 – Выбор сохраняемых переменных

	Nedvig.sta									
	1 Район	2 Тип дома	3 Индикаторная переменная	4 Общая площадь	5 Жилая площадь	6 Площадь кухни	7 Цена, тыс. руб.	8 НАБЛ_НО	9 КЛАСТЕР	10 РАССТ.
C_1	ЮМР	блочный	0	40	20	10	1596	1	3	149,64
C_40	Центр	блочный	0	37	17	8	1344	40	3	24,03
C_76	ЧМР	блочный	0	38	18	9	1148	76	3	74,47
C_77	ЧМР	блочный	0	27	14	6	686	77	3	305,52
C_78	ЧМР	блочный	0	30	17	6	1008	78	3	144,57
C_79	СМР	кирпичный	1	34	17	12	1176	79	3	60,65
C_101	40-летПобеды	блочный	0	35	18	7	1140	101	3	78,54
C_102	40-летПобеды	блочный	0	40	20	10	1450	102	3	76,66
C_103	40-летПобеды	блочный	0	40	20	10	1400	103	3	51,67
C_104	40-летПобеды	блочный	0	44	22	11	980	104	3	158,38
C_105	Авиагородок	блочный	0	40	20	10	1250	105	3	23,49
C_106	40-летПобеды	кирпичный	1	43	16	10,2	1500	106	3	101,68
C_107	Витаминкомбинат	блочный	0	50	20	10	1600	107	3	151,67
C_108	ГМР	кирпичный	1	49	22	14	1740	108	2	171,09
C_109	ГМР	кирпичный	1	49	22	14	1700	109	2	191,01
C_110	ГМР	монолитный	0	50,5	20	12	1385	110	3	44,31
C_111	ЗИП	монолитный	0	40	20	10	940	111	3	178,39
C_112	ЗИП	блочный	0	44	20	10,8	1000	112	3	148,39
C_113	ККБ	монолитный	0	32	16	6,5	1020	113	3	138,54
C_114	ККБ	кирпичный	1	55	20	19	1550	114	3	126,86
C_115	ККБ	кирпичный	1	52	20	16	1350	115	3	27,21
C_116	КМР	блочный	0	30	15	7	1250	116	3	24,66
C_117	КМР	блочный	0	40	20	10	1500	117	3	101,65
C_118	КМР	блочный	0	34	20	7,5	1300	118	3	5,27
C_119	ФМР	блочный	0	40	20	10	1350	119	3	26,72
C_120	ФМР	кирпичный	1	30	17	6	1200	120	3	48,96
C_121	СМР	кирпичный	1	33	19,5	6,5	1250	121	3	24,05
C_122	СХИ	блочный	0	40	18	10	1480	122	3	91,67
C_123	СХИ	блочный	0	42	22	10	1300	123	3	1,86

Рисунок 10.14 – Таблица сохранённых кластеров

Теперь, используя полученную таблицу (рисунок 10.14), с помощью дискриминантного анализа найдем функции классификации, позволяющие, как известно, отнести новый объект (квартиру) к одному из классов по наибольшему значению соответствующей функции. Для этого закроем окно кластерного анализа и выполним команду Анализ (*Statistics*)– Многомерный разведочный анализ (*Multivariate Exploratory Technicues*) – Дискриминантный анализ (*Discriminant Analysis*) и выберем группирующую и независимые переменные (см. рис.10.15-10.16). На рисунке 10.17 изображены, полученные линейные функции классификации.

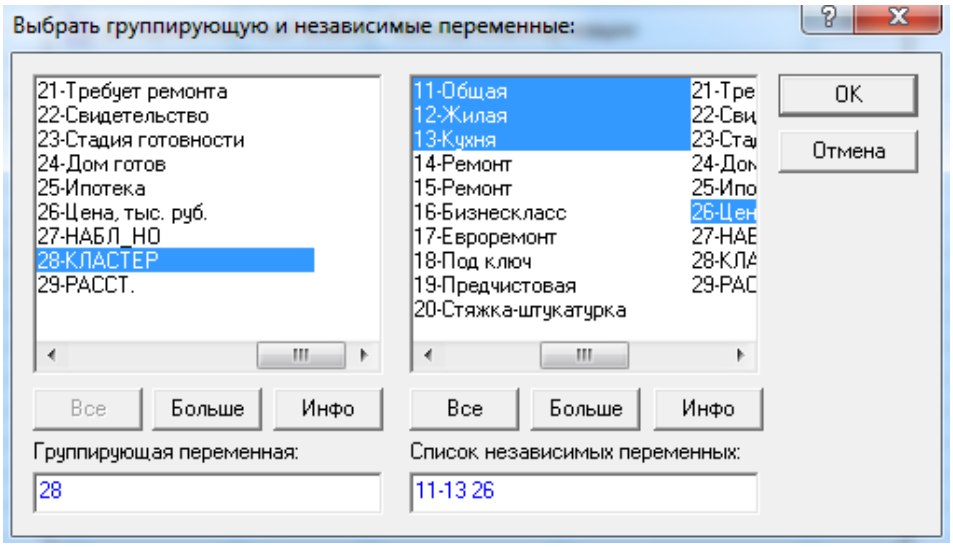


Рисунок 10.15 – Выбор группирующей и независимых переменных

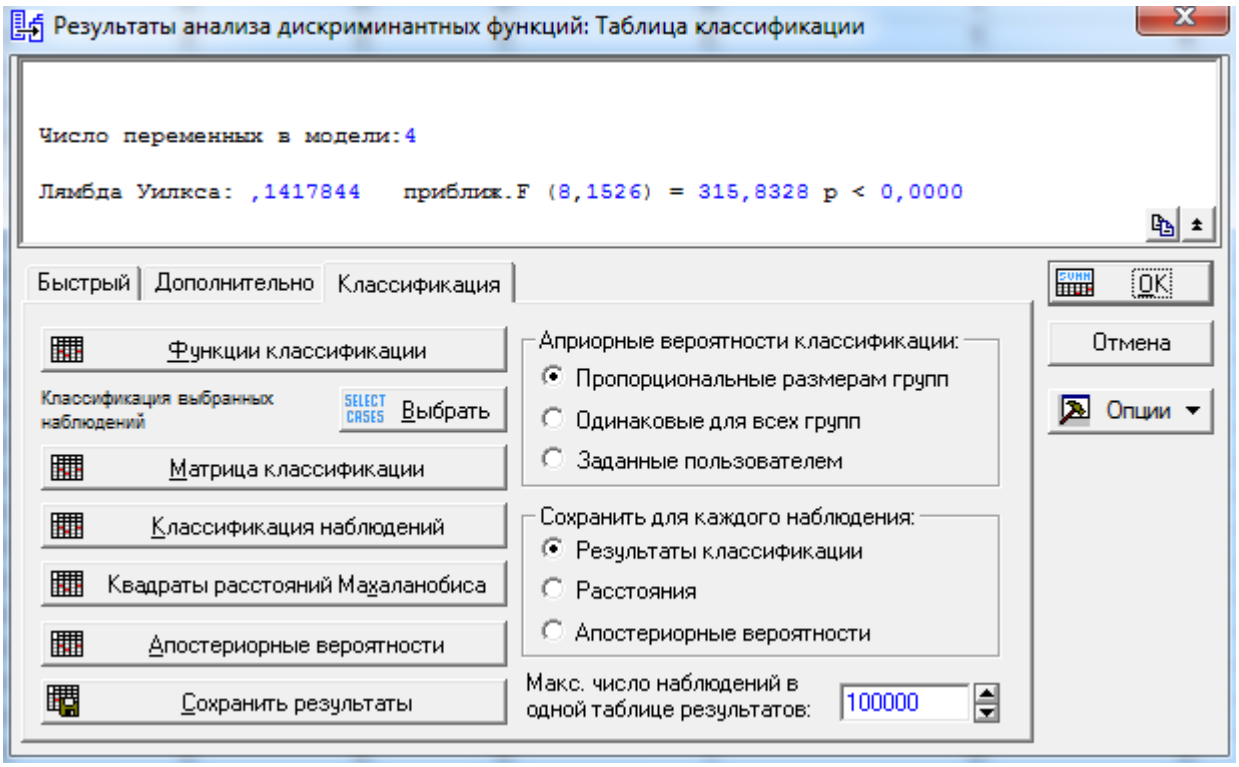


Рисунок 10.16 – Диалоговое окно модуля дискриминантного анализа

	Функции классификации; группировка: КЛАСТЕР (Таблица классификации)		
Переменная	G_1:1 p=,89337	G_2:2 p=,02081	G_3:3 p=,08583
Общая	0,4729	0,371	0,5500
Жилая	0,2913	0,114	0,1733
Кухня	0,2666	-0,260	0,2519
Цена, тыс. руб.	0,0099	0,057	0,0258
Конст-та	-21,0511	-187,683	-63,3616

Рисунок 10.17 – Функции классификации

Матрица классификации показывает, что практически 95% квартир, классифицированы правильно (рис. 10.18).

	Матрица классификации (Таблица классификации) Строки: наблюдаемые классы Столбцы: предсказанные классы			
Группа	**Процент правиль.**	G_1:1 p=,89337	G_2:2 p=,02081	G_3:3 p=,08583
G_1:1	100,0000	687	0	0
G_2:2	93,7500	0	15	1
G_3:3	100,0000	0	0	66
Всего	99,8700	687	15	67

Рисунок 10.18 – Матрица классификации

Графическое изображение классов так же указывает на хорошую классификацию квартир – рисунок 10.19. Первый канонический корень дискриминирует

первый класс и совокупность второго и третьего классов. (Вкладка **дополнительно – Канонический анализ – Диаграмма рассеяния для канонических значений)**.

Матрица факторной структуры (таблица коэффициентов корреляции между параметрами и факторами – каноническими корнями), полученного решения, позволяет оценить вклад переменных в полученную классификацию, посредством связи с каноническими корнями – первый корень наиболее сильно связан с ценой, а второй с жилой площадью (рис. 10.20). Таким образом, можно считать, что имеющиеся данные в основном обусловлены указанными выше факторами. (Вкладка **дополнительно – Канонический анализ – Дополнительно – Факторная структура**).

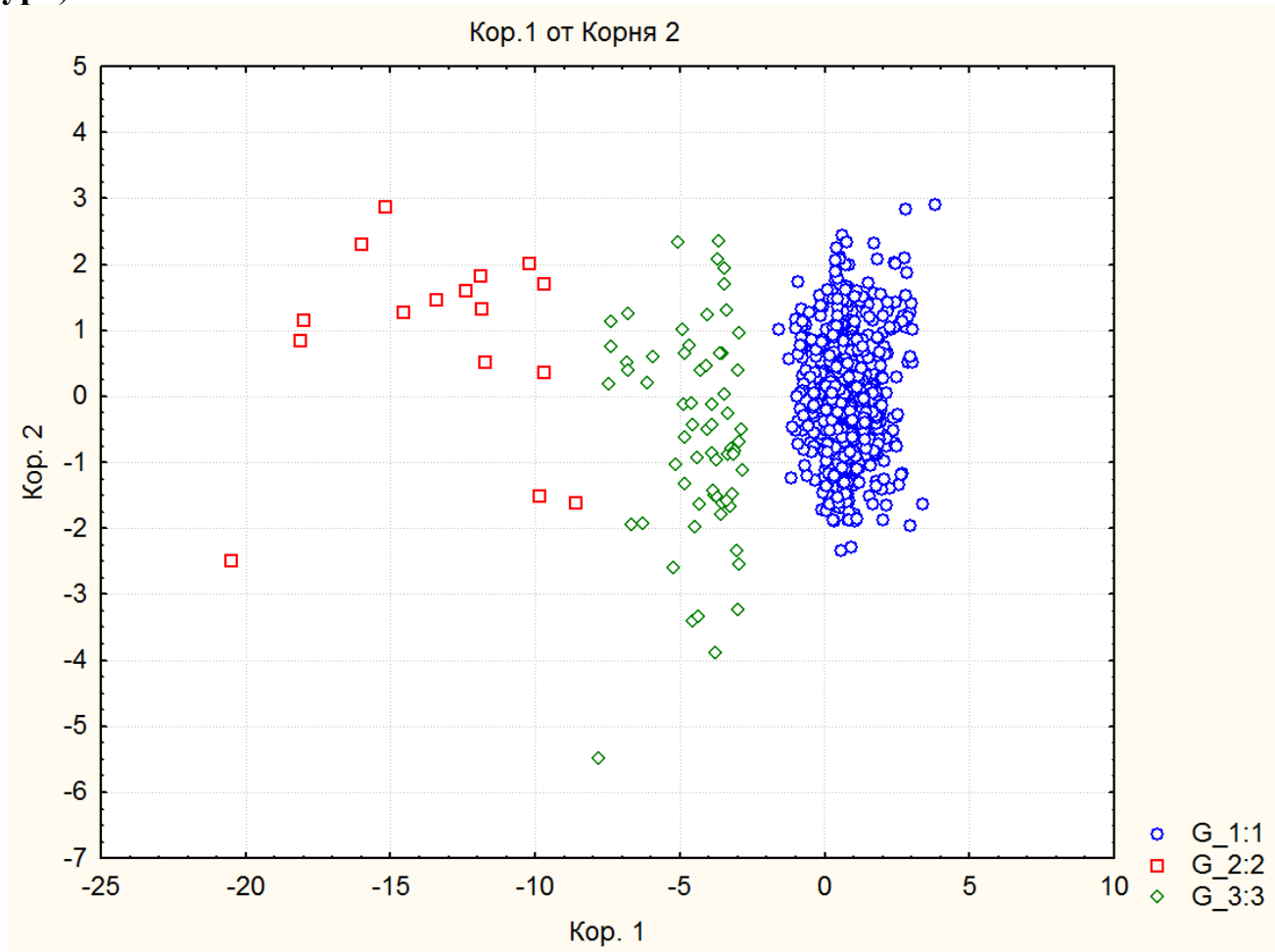


Рисунок 10.19 – Разделение трёх классов квартир

Матрица факторной структуры (Таблица классификации)
Корр. переменных и функции дискрим.
(объединенные внутригруп. корреляции)

Переменная	Кор. 1	Кор. 2
Общая	-0,417240	-0,749151
Жилая	-0,286849	-0,234114
Кухня	-0,172092	-0,811330
Цена, тыс. руб.	-0,995395	-0,072253

Рисунок 10.20 – Корреляционная матрица переменных и канонических корней

Теперь задача поиска аналитической зависимости цены однокомнатной квартиры от других переменных может решаться в каждом из классов. При появлении новой квартиры с заявленной ценой, с помощью функций классификации, её следует отнести к одному из трёх классов, а затем, имея регрессионную зависимость – дать оценку средней ожидаемой стоимости квартиры.

Итоги регрессии для зависимой переменной: Цена, тыс. руб. (Таблица классификации)
R= ,52264728 R2= ,27316018 Скорректир. R2= ,26890966
F(4,684)=64,265 p<0,0000 Станд. ошибка оценки: 221,87

N=689	БЕТА	Стд.Ош. БЕТА	B	Стд.Ош. B	t(684)	p-уров.
Св.член			817,2777	71,30427	11,46183	0,000000
Расположение	-0,154570	0,035122	-80,1684	18,21634	-4,40091	0,000013
Общая	0,429266	0,033919	22,2745	1,76004	12,65572	0,000000
Евроремонт	0,101466	0,032964	90,6363	29,44570	3,07808	0,002167
Под ключ	0,117206	0,033854	129,9597	37,53751	3,46213	0,000569

Рисунок 10.21 – Итоги регрессионного анализа

Замечание. Задача оценки рыночной стоимости жилья с использованием компьютера является актуальной. В практических работах № 8, 10, 14 для решения этой проблемы предлагается использовать методы многомерного статистического анализа (интеллектуального анализа данных).

Общий подход заключается в применении разведочного анализа данных (визуализации данных для изучения совокупности квартир выдвигаемых на продажу; построение регрессионных моделей, как для всей совокупности, так и для классов на которые её можно предварительно разбить). Очевидно, что цены индивидуальных продаж (даже абсолютно одинаковых квартир) могут отличаться от средней рыночной стоимости (из-за личных мотивов, осведомлённости, условий сделки и т.д.).

В практикуме предлагается подход к оценке рыночной стоимости квартир (на примере данных о стоимости жилья в г.Краснодаре на 01 октября 2013 г.) как к среднему значению (математическому ожиданию) этой случайной величины для данной квартиры, зависящий от: числа комнат, района, типа дома, района, жилой площади, общей площади и площади кухни. Разумеется:

1) Число и состав переменных может быть изменён, данные должны обновляться, быть массовыми и т.д.

2) Цены, которые могут быть различны (цена: продавца, покупателя, реализации; предложения, спроса, сделки).

Мы рассматривали цены предложения (так как эта информация наиболее доступна), хотя конечно наибольший интерес представляет средняя цена сделки, наиболее точно отражающая ситуацию на рынке недвижимости.

Следует отметить, что общая идеология оценки средней стоимости квартир на основании имеющейся статистической информации известна и применяется как у нас в стране, так и за рубежом. Кроме того, существует возможность получения законченного аналитического решения для оценки стоимости недвижимости с использованием любого из рассматриваемых в практикуме продуктов (*Statistica, Pol-*

yAnalyst, Deductor) при условии постоянного пополнения базы данных о рынке недвижимости.

Задание

1. Выполнить приведённые выше примеры.
2. Провести классификацию *N* – комнатных (в соответствии с вариантами занятия по регрессионному анализу, без учёта районов) квартир и построить в классах уравнения зависимости цены квартиры от входных переменных.
3. На основе корректных обучающих выборок, полученных в задании 2 и классификационных функций, провести классификацию квартир одного из районов, представленных в вашем варианте практического занятия № 8 и дать оценку возможной средней стоимости.

Вопросы для самоконтроля

- Какие общие задачи решают методы кластерного и дискриминантного анализа?
- Чем отличаются методы классификации с учителем и без учителя.
- Какие известны меры сходства и различия элементов (объектов) и классов?

Практическое занятие №11

Факторный анализ в системе Statistica

Цель работы: Ознакомиться с возможностями анализа данных с помощью метода главных компонент и факторного анализа.

Теоретические сведения

Задача факторного анализа состоит в том, чтобы выразить параметр x_j в терминах скрытых гипотетических факторов. Наиболее распространены два варианта такого описания, соответствующие целям моделирования: 1) выделение максимальной дисперсии, 2) «наилучшая» аппроксимация выборочных корреляций. Это соответствует двум подходам к решению указанной задачи – методу главных компонент и факторному анализу.

Модель компонентного анализа имеет вид:

$$x_j = \sum_{p=1}^{k} a_{jp} F_p \ ,$$

где каждый из наблюдённых параметров x_j ($j = \overline{1,k}$) зависит от k некоррелированных между собой компонентов (факторов) $F_1, F_2, F_3, \ldots , F_k$, причём каждая следующая компонента даёт максимальный вклад в суммарную дисперсию параметров – вся дисперсия описывается полностью (иногда часть незначимых компонент отбрасывается).

Модель факторного анализа имеет вид:

$$x_j = \sum_{p=1}^{s} a_{jp} F_p + d_j U_j \, ,$$

где каждый из наблюдённых параметров x_j ($j = \overline{1,k}$) зависит от sобщих факторов ($F_1, F_2, F_3, \ldots , F_s$) и одного характерного (U_j). Общие факторы учитывают корреляции между параметрами, характерный фактор учитывает оставшуюся (в том числе связанную с различными погрешностями) дисперсию. Коэффициенты при факторах называют нагрузками.

Основная проблема, с которой сталкиваются исследователи – сложность интерпретации полученного факторного решения. Обычно, после того как факторное решение найдено, система факторов подвергается вращению, чтобы полученное решение интерпретировалось специалистами предметной области.Известны различные методы вращения факторов. Целью этих методов является получение понятной (интерпретируемой) матрицы нагрузок, то есть факторов, которые ясно отмечены высокими нагрузками для некоторых переменных и низкими – для других. Типичными методами вращения являются стратегии: варимакс, квартимакс и эквимакс,облимакс, квартимин, облимин и др.[36].

Наибольшее число приложений факторный анализ имеет в области психологии и социальных наук.

Существует много различных методов решающих проблему факторного анализа. Различия часто поверхностны и обуславливаются расхождением в небольшом числе исходных предположений. Более того – все методы с точки зрения практики

приводят к одним решения. Остроумное пояснение этого факта есть у Г.Хармана [36]:

«Факторную теорию можно определить как математически разумную гипотезу. Специалист в области факторного анализа – это субъект, одержимый некой навязчивой идеей о природе умственных способностей или личности. Применяя высшую математику к исследуемому предмету, он доказывает, что его оригинальная точка зрения верна и неизбежна. Обычно он доказывает также, что все другие специалисты в факторном анализе – опасные сумасшедшие и единственное их спасение состоит в том, чтобы принять его теорию; только в этом случае выяснится истина об их болезни. Поскольку противники никогда не поддерживают такое обвинение, то он обзывает их безнадёжными, и устремляется в области математики, наверняка им не известные; тем самым доказывая не только необходимость, но и достаточность неизлечимости оппонентов». Несмотря на указанные проблемы, факторный анализ часто используется для решения задач: сокращения числа переменных (сжатия информации), определения структуры связей между переменными (классификации переменных), подтверждающего факторного анализа – о факторной структуре различных выборок.

Факторный анализ в Statistica

Выполнив команду **Анализ – Многомерный разведочный анализ – Факторный анализ** (***Statistics – Multivariate Exploratory Techniques – Factor Analysis***) мы получим соответствующее диалоговое окно (рис.11.1).

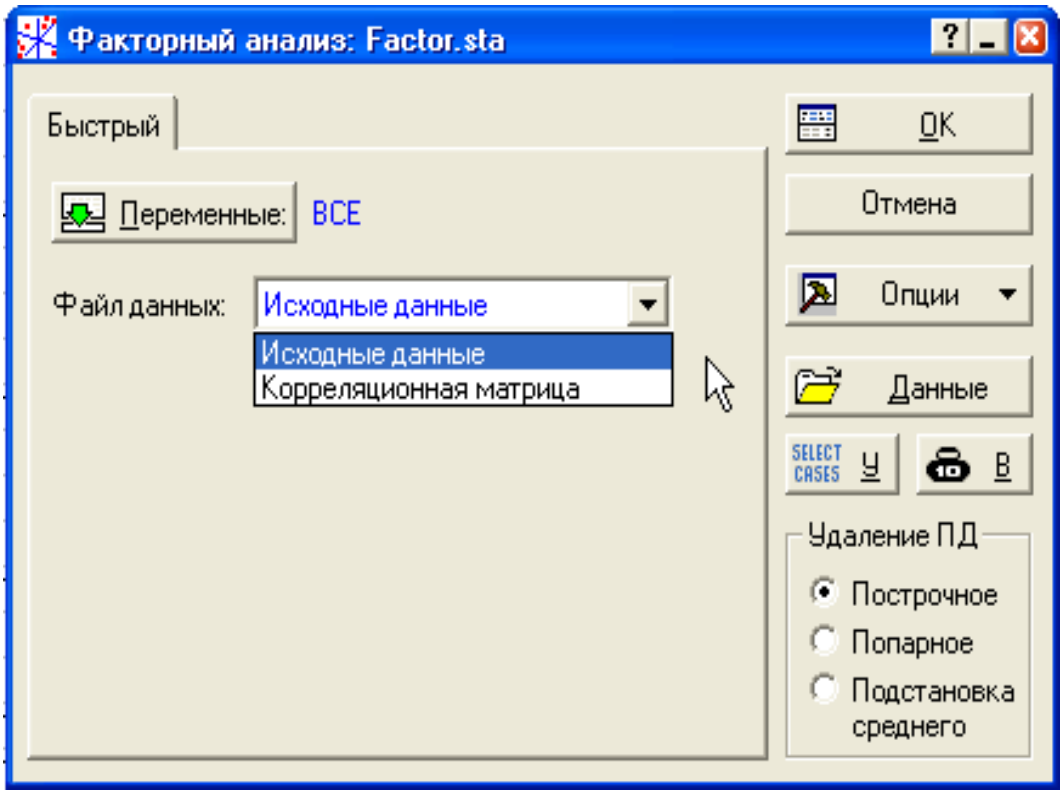


Рисунок 11.1– Диалоговое окно модуля факторный анализ

В модуле возможны два типа исходных данных: **Исходные данные** (***Raw Data***), **Корреляционная матрица**(***Correlation Matrix***).

Группа **Удаление ПД** (***Mising Data deletion***) задает способ обработки пропущенных значений:

CaseWise – Построчное исключение пропущенных случаев,

PairWise– Попарное исключение пропущенных значений,

Mean substitution – Подстановка среднего вместо пропущенных значений.

Пример. Откроем файл *Factor.sta*[8], в котором собраны результаты опросов 100 взрослых людей относительно степени их удовлетворенности жизнью. Выбрав все переменные и воспользовавшись кнопкой Больше (Spread) мы получим расширенное описание переменных (рис.11.2).

Work 1 – удовлетворенность работой, – первая компонента,
Work 1 – удовлетворенность работой, – вторая компонента,
Work 1– удовлетворенность работой, – третья компонента,
Hobby 1 – удовлетворенность свободным временем, – первая компонента,
Hobby 2 – удовлетворенность свободным временем, – первая компонента,
Home 1 – удовлетворенность домашней жизнью, – первая компонента,
Home 2 – удовлетворенность домашней жизнью, – вторая компонента,
Home 3 – удовлетворенность домашней жизнью, – вторая компонента,
Miscel 1 – общая удовлетворенность, – первая компонента,
Miscel 2 – общая удовлетворенность, – вторая компонента.

Щелкнем кнопку **OK** начнем анализ выбранных переменных.

Первый шаг – это вычисление корреляционной матрицы, если она не задана сразу. Третья вкладка позволяет вычислить корреляции, средние, стандартные отклонения, построить диаграммы рассеяния, провести регрессионный анализ.

После выбора **Главные компоненты** (***Principal components***) получим результаты факторного анализа, изображённые на рисунке 11.4.

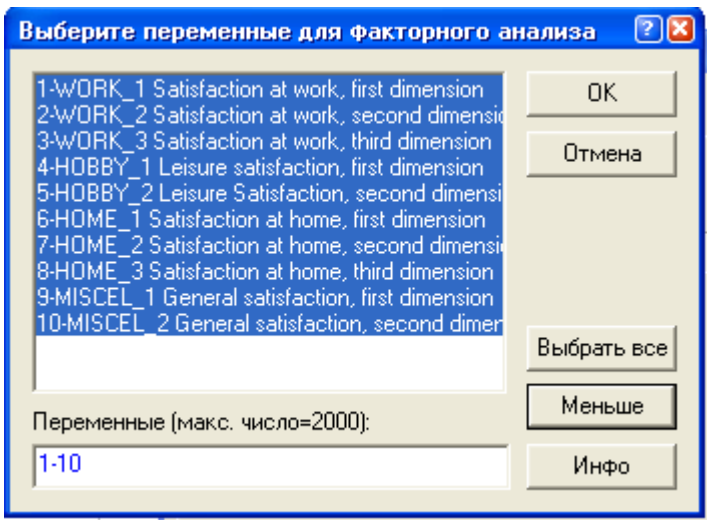


Рисунок 11.2– Окно выбора переменных

[8] Стандартный пример системы *Statistica* 6.0 (1)

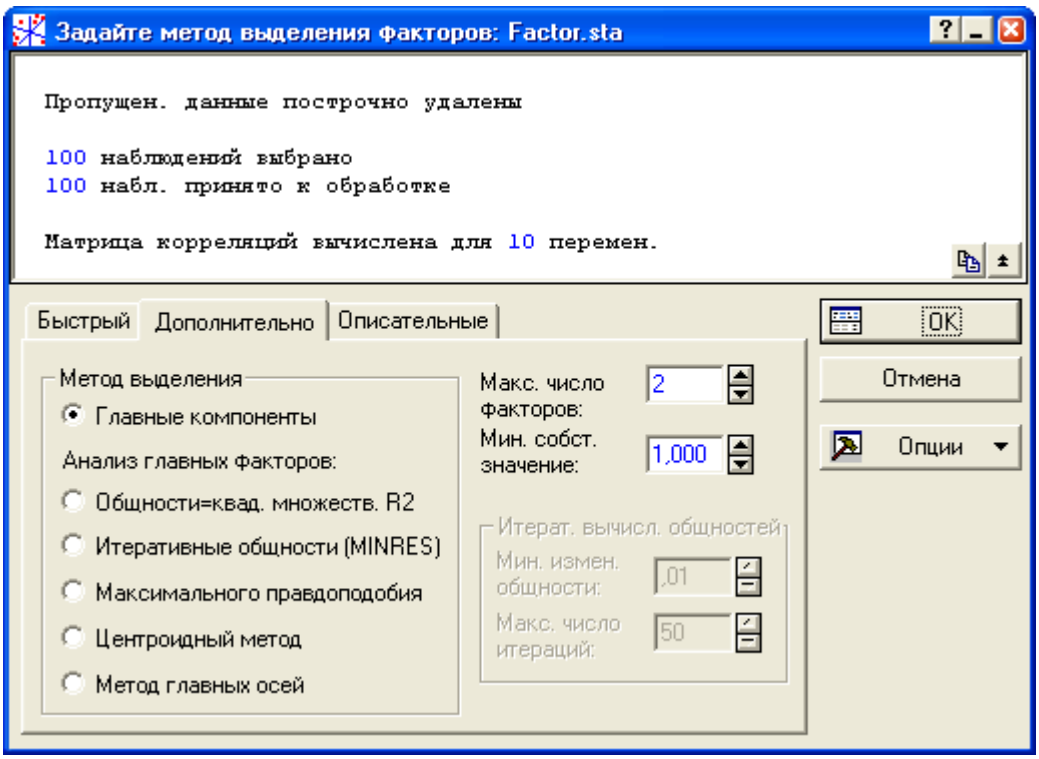


Рисунок 11.3–Диалоговое окно выбора метода факторного анализа

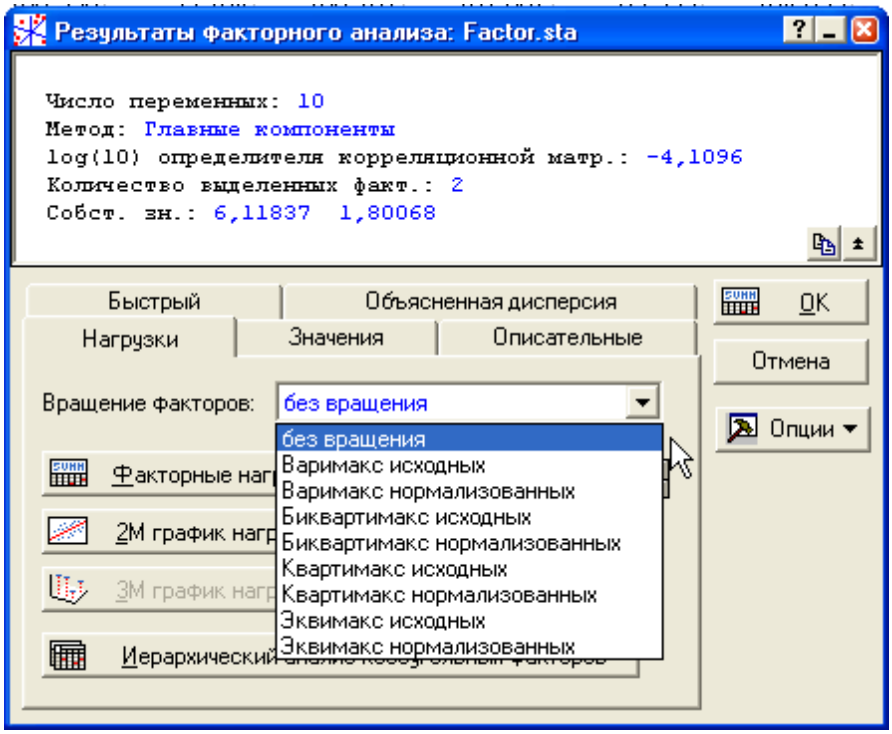


Рисунок 11.4– Методы вращения факторов в системе *Statistica* 6.0(1)

Вкладка **Вращение факторов** (***Factor rotation***) –позволяет выбрать различные варианты поворота осей… Главное – найти интерпретируемое решение.

Выберем первую строку – **без вращения**(***Unrotated***). Шелкнем кнопку **2М график нагрузок** (***Plot of loadings*** **2D**) (рис.11.5).

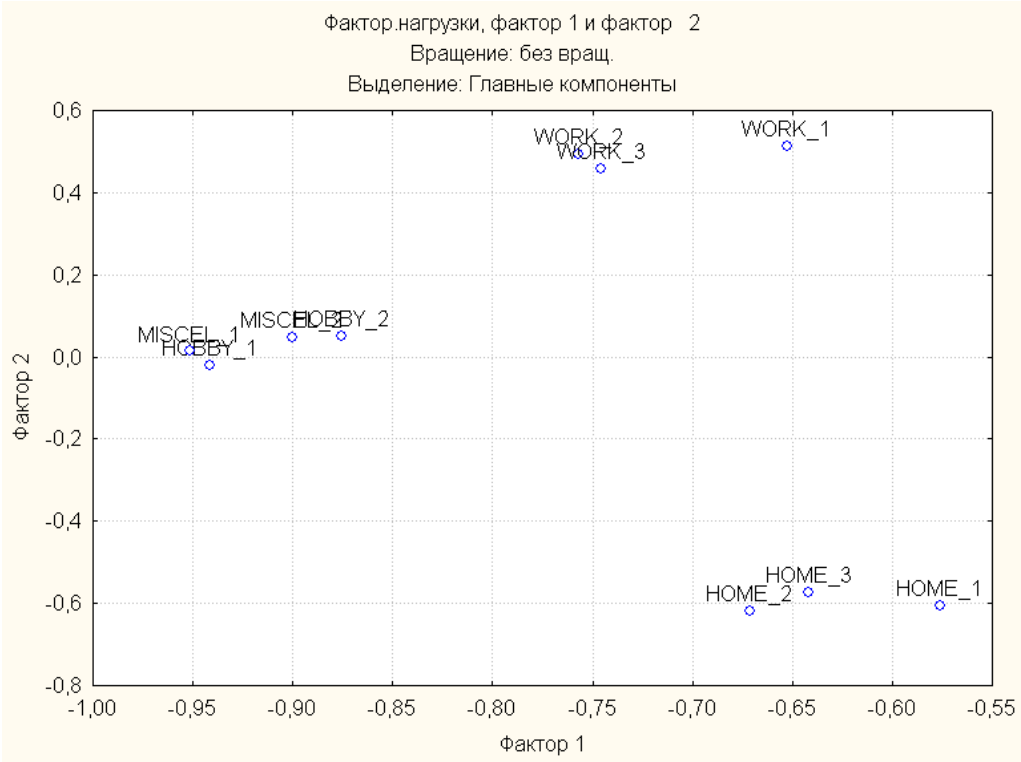


Рисунок 11.5– График факторных нагрузок без вращения

В данном случае не совсем ясно –какой смысл придать двум выделенным факторам и как в этих терминах описывать удовлетворенность жизнью индивидуума. Рассмотрим вращение факторов – **Варимакс нормализованных** (***Varimax-normalized***). Щелкнем кнопку **2М график нагрузок** (***Plot of loadings*** **2D**) (рис.11.6).

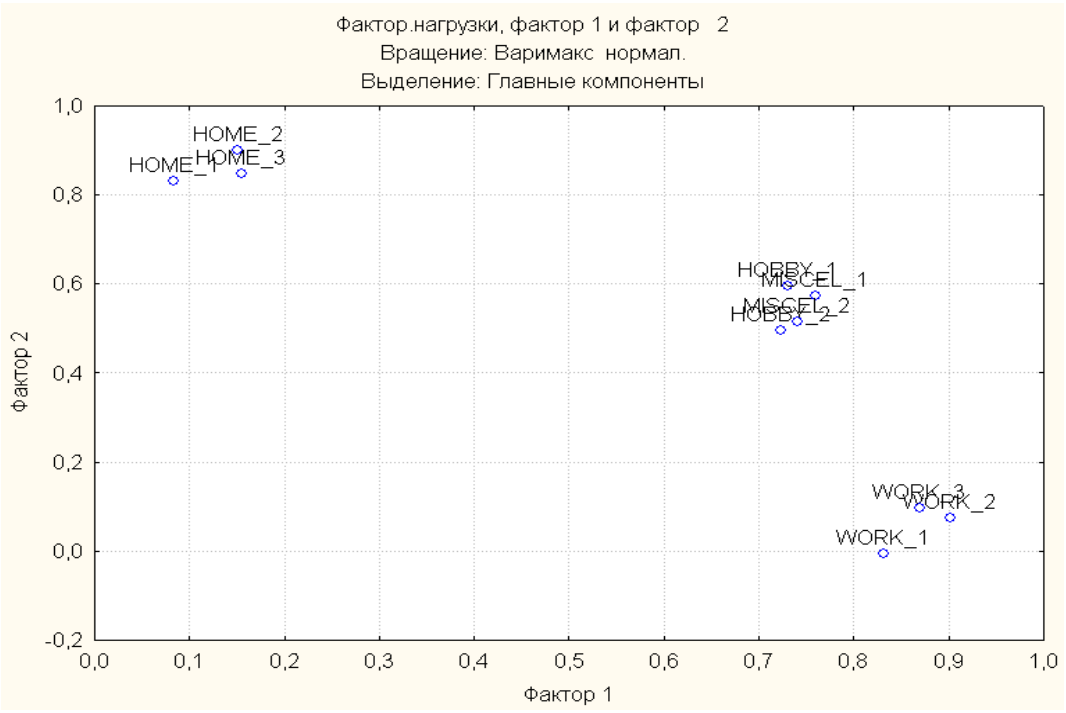


Рисунок 11.6– График факторных нагрузок после вращения методом Варимакс

Перемен.	Фактор.нагрузки (Варимакс нормал.) (Factor.sta) Выделение: Главные компоненты (Отмечены нагрузки >,700000) **Фактор 1**	Фактор 2
WORK_1	0,830827	-0,005746
WORK_2	0,901325	0,073641
WORK_3	0,869058	0,096808
HOBBY_1	0,730235	0,594896
HOBBY_2	0,723177	0,496371
HOME_1	0,083802	0,831157
HOME_2	0,151040	0,899830
HOME_3	0,154555	0,846798
MISCEL_1	0,759727	0,573045
MISCEL_2	0,740557	0,514289
Общ.дис.	4,493483	3,425568
Доля общ	0,449348	0,342557

Рисунок 11.7– Корреляционная матрица переменных и канонических корней

Фактор 1 можно интерпретировать как удовлетворение, получаемое опрошенными на работе, удовлетворенность работой (переменные *Work* 1-3) максимально большие по этой оси и малы по другой.

Фактор 2 измеряет удовлетворенность домашней жизнью.

Общий вывод: общая удовлетворенность человека определяется двумя факторами – удовлетворенностью на работе и дома.

Задание

1. Создайте, используя справку, корреляционную матрицу **Перепись.smx** (рис.11.8) .

Данные: Перепись.smx* (5v * 9c)

	1 **Население**	2 Образование	3 Общая занятость	4 Профессиональная занятость	5 Средняя стоимость жилья
Население	1,00000	0,90816	0,99985	0,95542	0,98659
Образование	0,90816	1,00000	0,91499	0,74708	0,96279
Общая занятость	0,99985	0,91499	1,00000	0,95025	0,98923
Профессиональная занятость	0,95542	0,74708	0,95025	1,00000	0,89444
Средняя стоимость жилья	0,98659	0,96279	0,98923	0,89444	1,00000
Средние	1383,49193	2,27407	510,96620	34,20324	3338,68358
Ст.откл.	2496,59157	4,06708	926,94544	57,27057	6474,61415
Кол-во N	7,00000				
Матрица	1,00000				

Рисунок 11.8 – Корреляционная матрица переменных и канонических корней

Проведите факторный анализ на основании этой матрицы, полученной анкетированием 1000 респондентов.

2. Проанализировать исходные данные о стоимости жилья по вариантам, не учитывая район (см. работу № 8) с использованием факторного анализа.

Вопросы для самоконтроля

- Какие общие задачи решают методы главных компонент и факторного анализа?
- Чем отличаются методы главных компонент и факторного анализа?
- Какие существуют проблемы практического использования метода главных компонент и факторного анализа?

Практическое занятие №12

Анализ временных рядов

Цель работы: Ознакомиться с возможностями методик анализа временных рядов, получить навыки анализа данных с использованием модуля **Временные ряды и прогнозирование (*Statistics – Advanced Linear/Nonlinear Models – Time Series/ Forecasting*)**.

Теоретические сведения

Дискретный временной ряд – это последовательность измерений значений переменной (процесса) за определенный период через одинаковые промежутки времени:

$$Z_1, Z_2, Z_3, \ldots, Z_t, \ldots, Z_n. \qquad (12.1)$$

С детерминистской точки зрения можно представить как:

$$Z_t = f(t) + \varepsilon_t,$$

где t=1, 2, …, n; f – гладкая (непрерывная и дифференцируемая) функция, характеризующая долгосрочное движение в зависимости от времени – тренд; ε_t – случайный ряд возмущений, наложенный на систематическую часть.

Анализ структуры временного ряда, содержащего сезонные или циклические колебания начинается с предположения о том, что $Z_t = T + C + Ц + \varepsilon_t$ – аддитивная модель или $Z_t = T \cdot C \cdot Ц \cdot \varepsilon_t$ – мультипликативная модель, где T – тренд, C – сезонность, $Ц$ – цикличность (цикличность отличается от сезонности большей протяженностью и может не выделяться отдельно). Если амплитуда колебаний увеличивается или уменьшается, то рассматривается мультипликативная модель, в противном случае (колебания практически не изменяются) – аддитивная модель.

Последовательные наблюдения обычно зависимы. При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционная зависимость между последовательными уровнями временного ряда называют автокорреляцией уровней ряда. Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на один или несколько шагов во времени, называемого коэффициентом автокорреляции.

Так как коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции, то по нему можно судить о наличии линейной или близкой к линейной тенденции. Чем ближе коэффициент автокорреляции первого порядка к единице, тем более выражена линейная тенденция. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

Последовательность коэффициентов автокорреляции уровней первого, второго и т.д. порядков называют автокорреляционной функцией временного ряда. Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то ряд содержит циклические или сезон-

ные колебания с периодичностью в τ моментов времени. Если ни один коэффициент не является значимым, можно сделать вывод – либо ряд не содержит тенденции и циклических колебаний, либо содержит сильную не линейную тенденцию.

Число периодов или моментов времени, по которым рассчитывается коэффициент автокорреляции, называют лагом.

Построение аналитической функции для моделирования тенденции (тренда) временного ряда называют аналитическим выравниванием временного ряда. Тенденция во времени может принимать разные формы, для ее формализации используют функции подобранные путём визуализации временного ряда. Такой подход, несмотря на заслуженную критику, используется и в настоящее время.

Второй подход (стохастический) заложил Эдни Юл в 1927 г. Он предложил для его объяснения пример, объясняющий, что будущее состояние системы зависит как от детерминированной составляющей, так и от воздействия случайностей. Эта концепция приводит к теории стохастических процессов, важнейшим разделом которой является теория стохастических временных рядов.

Третий подход к анализу временных рядов – это спектральный анализ в частотной области, который обычно рассматривается, если данные имеют периодичность.

Используются следующие понятия: частота (ν) – это число циклов (периодов) в единицу времени; период ($T=1/\nu$)– это продолжительность по времени полного цикла. Цель спектрального анализа – это выявление циклов различной длины в изучаемых рядах и представление временного ряда в виде суммы различных синусоидальных функций.

В частном случае можно получить выравнивание по ряду Фурье (при этом обычно рассматривается не более 5 гармоник (j= 1,2,3,4,5)):

$$z_t = a_0 + \sum_{j=1}^{k}(a_j \cos jt + b_j \sin jt).$$

Параметры a_j и b_j находятся с помощью МНК, в результате применения, которого получим:

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} z_t, \quad \mathrm{a_j} = \frac{2}{n}\sum_{t=1}^{n} z_t \cos jt, \quad \mathrm{b_j} = \frac{2}{n}\sum_{t=1}^{n} z_t \sin jt.$$

В спектральном анализе для выявления (подтверждения) сезонности или цикличности используется периодограмма, которую можно подсчитать как

$$\mathrm{P_j} = (\mathrm{a_j^2} + \mathrm{b_j^2}) \cdot \frac{\mathrm{n}}{2},$$

где P_j – значение периодограммы на частоте ν_j, n – длина ряда. Интерпретируют P_j как дисперсию (вариацию) данных на j-ой частоте. Периодограмму изображают в зависимости от частот, периодов, логарифмов периодов.

Если ряд не имеет циклов и все значения взаимно независимы, то в случае если мы имеем белый шум (случайный процесс, который является реализацией нормально распределённых случайных величин с постоянной дисперсией и нулевым математическим ожиданием) – значения периодограммы будут иметь экспо-

ненциальное распределение. Перед началом анализа рекомендуется привести ряд к стационарному виду (см. ниже), то есть – вычесть тренд, среднее значение.

Оценка периодичности по периодограмме не всегда однозначна. Например, месячная периодичность порождает эффект появления пиков, которые соответствуют двухмесячным, трёхмесячным, четырёхмесячным и т.д. периодам. Это так называемые эхо-эффекты, соответствующие повторениям спектра на низких частотах.

Отделить тренд и сезонность в общем случае невозможно, так как они взаимно проникают друг в друга. При выделении тренда и сезонности остается колеблющийся ряд. Удаление тренда (сглаживание временного ряда) можно осуществить с помощью скользящей средней (СС). Скользящая средняя, в отличие от простой средней для всей выборки, содержит сведения о тенденциях изменения данных.

Для этого к первым ($2m$+1) точкам ряда (12.1) подбирают полином

$$Q_P(t) = a_p t^p + a_{p-1} t^{p-1} + a_{p-2} t^{p-2} + \ldots + a_1 t + a_0$$

(для определения значения тренда в (m+1) точке) и минимизируют:

$$\sum_{t=-m}^{m} (z_t - a_p t^p - a_{p-1} t^{p-1} - a_{p-2} t^{p-2} - \ldots - a_1 t - a_0)^2 .$$

Затем подбирают полином того же порядка для второго, третьего,…, ($2m$+2) наблюдения. Эта процедура продолжается вдоль всего ряда до последней группы из ($2m$+1) точек. На самом деле нет необходимости подбирать полином каждый раз. Например, для полинома третьей степени (p=3) и пяти точек - значение тренда в какой-либо точке равно средневзвешенному значению пяти точек с данной точкой в качестве центральной и весами $\frac{1}{35}$[-3,12,17,12,-3] . Для пяти точек и p = 1 получаем простую скользящую среднюю:

$$a_0 = \frac{1}{5}(z_{-2} + z_{-1} + z_0 + z_1 + z_2) .$$

Кроме рассмотренного подхода к выводу формул взвешенных СС существуют другие способы определения СС: использование простых СС, формулы Спенсера и т.д.

При рассмотрении СС, в рамках нашего примера (p=3, m=2), следует отметить проблему крайних двух точек – они не оцениваются.

Сглаживание может производиться скользящей медианой, которая более устойчива и является альтернативной оценкой центра распределения.

Рассмотренные выше СС (их называют иногда линейными фильтрами), являются симметрическими (т.е. коэффициенты (веса) симметричны относительно среднего).

Для прогнозирования в статистике используют асимметричные фильтры. Так в *Statistica* при выборе опции *По пред* (*Prior*) скользящая средняя, скользящая медиана заменяют не средний, а последний уровень ряда в промежутке сглаживания, она (СС) используется для расчета значений в прогнозируемом периоде, на основе значения переменной для указанного числа предшествующих периодов, например, для средней, по формуле:

$$F_{t+1} = \frac{1}{N}\sum_{h=0}^{N} Z_{t-h+1},$$

где N – число предшествующих периодов, входящих в СС;

Z_h – фактическое значение в момент времени h;

F_h – прогнозируемое значение в момент времени h.

Асимметричные СС иногда могут учитывать степень "устаревания данных", т.е. каждое новое наблюдение будет иметь вес больше предыдущих, например:

$$F_{t+1} = (1-\alpha)(A_{t+1} + \alpha A_t + \alpha^2 A_{t-1} + \ldots) = (1-\alpha)\sum_{i=0}^{\infty} \alpha^i A_{t-i+1}, \quad 0 < \alpha < 1 \quad .$$

Рассмотренный подход к определению асимметричных СС носит название экспоненциальное сглаживание (ЭС) (или экспоненциальных средних). ЭС предназначается для предсказания значения F_{t+1} на основе прогноза для предыдущего периода F_t, скорректированного с учетом погрешностей в этом прогнозе (A_t-F_t), можно получить, что:

$$F_{t+1}=F_t+\alpha\,(A_t-F_t)= \alpha A_t+(1-\alpha)F_t,$$

где F_{t+1} – прогноз в момент времени (t+1), F_t – прогноз в момент времени (t), A_t – преобразованное значение ряда в момент времени t. Изменяя α ($0<\alpha<1$) можно регулировать влияние текущих или предшествующих наблюдений.

Перечисленные выше фильтры могут комбинироваться, как, например, в 4253H фильтре. Он включает несколько последовательных преобразований:

1) 4-х точечная скользящая медиана, центрированная скользящей медианой 2,
2) 5-ти точечная скользящая медиана,
3) 3-х точечная скользящая медиана,
4) 3-точечное взвешенное скользящее среднее с весами (0,25; 0,5; 0,25) (Hanning weights),
5) вычисляются остатки вычитанием преобразованного ряда из исходного ряда,
6) шаги 1–4 повторяются для остатков,
7) преобразованные остатки добавляются к преобразованному ряду.

На практике этот метод фильтрации дает сглаженный ряд, сохраняя основные характеристики исходного ряда. (Справочная система *Statistica* 6.1.)

Сглаживая временной ряд, мы устраняем случайные колебания и получаем сглаженный (усреднённый) ряд, который проще прогнозировать. Кроме того, выделенный тренд позволяет планировать управляющие воздействия (например, для экономического временного ряда – это может быть план закупок товаров, проведение рекламных акций и т.д.)

Временной интервал, для которого производится сглаживание, называют окном. Выбор ширины скользящего окна не является однозначным: при выборе малой ширины – усреднение шума не происходит, при выборе большой ширины усредняется не только шум, но и регулярная составляющая. По-видимому, следует рассматривать несколько вариантов скользящего окна и отбирать нужный, исходя из содержательной интерпретации (объяснение поведения или получения наилучшего прогноза).

Существует ещё целый ряд методов сглаживания и экстраполяции: модель Хольта-Уинтерса (содержит три параметра α_1 , α_2, α_3 и позволяет учесть сезон-

ность), модель Харрисона является модификацией предыдущей и выражает сезонность через гармоники. Указанные методы были разработаны для анализа экономических процессов. Широко известны также модель Бокса – Дженкинса, фильтры Калмана и Бюсси.

Применение скользящих средних сопряжено с рядом проблем: может искажаться циклическое движение; случайные колебания после сглаживания теряют часть движения и могут приниматься за долгосрочную тенденцию; производный ряд становится более гладким, чем исходный случайный ряд, и в нём могут проявляться систематические колебания, проявляющиеся в появлении ненулевых корреляций (эффект Слуцкого-Юла).

В 60-е годы XX века Шискиным была разработана программа для Бюро переписи США, известная как *Gensus Mark* II. Её цель была разделить сезонные и остаточные колебания.

Известно несколько вариантов этой программы. В системе Statistica имеются соответствующие модули для выделения сезонностей – *Gensus* I (классическая сезонная декомпозиция) и *Gensus* II (производится месячная и квартальная корректировка), описание которых можно найти в справке.

Практически все рассмотренные методы содержат предположения относительно вида генератора (модели) изучаемого временного ряда. Критерием адекватности той или иной модели может служить только практическое достижение первоначальных целей анализа временных рядов (описание поведения ряда, объяснение изменения наблюдений, прогноз и т.д.).

Стационарные временные ряды. Временной ряд, не имеющий тренда (либо с исключённым трендом), называется стационарным или иначе – если его свойства не зависят от начала отсчёта времени (механизм, генерирующий ряд не меняется со временем, хотя и носит вероятностный характер). Поэтому перечисленные ниже параметры для данного ряда являются постоянными:

$M(z_t)=M,\ M(z_t-M)^2=\sigma^2=D(z_t);\ M[(z_t-M)(z_{t+k}-M)]=c_k$ – k-ая автоковариация,

$\rho_k=\rho_{-k}=\frac{c_k}{\sigma^2}$ – соответствующая автокорреляция.

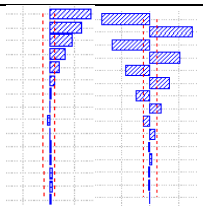
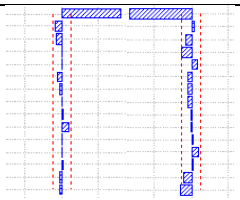
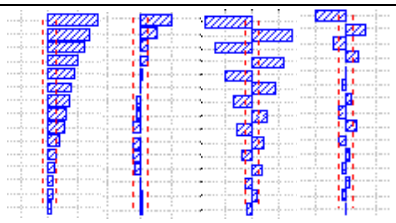
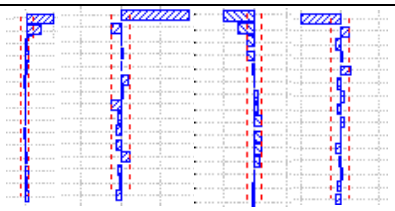
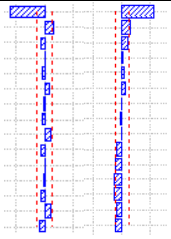
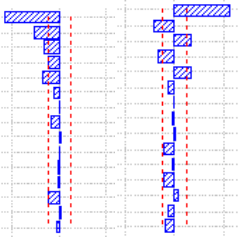
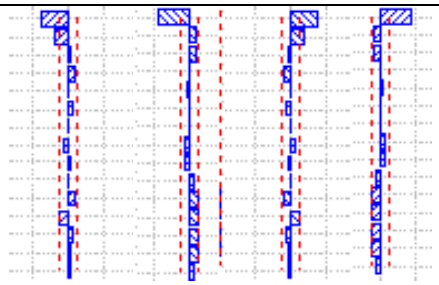
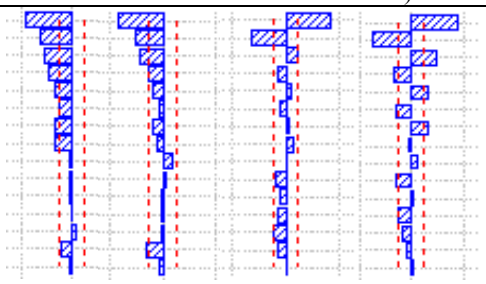
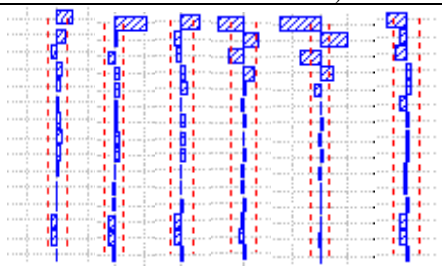
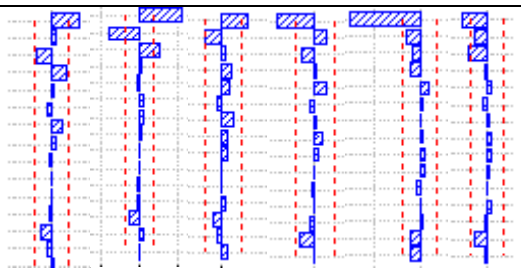
Иногда совокупность значений ρ_k представляется на графике и называется коррелограммой или автокорреляционной функцией (АКФ). Если ρ_k вычисляется с учётом исключения влияния наблюдений с лагом меньше k, то соответственно получается частная автокорреляционная функция (ЧАКФ).

Если процесс не удовлетворяет условию стационарности, то его преобразуют – выделяя тренды, логарифмируя значения ряда.

Один из способов удаления тренда заключается в переходе к разностям ряда – это используется, например, в рассматриваемой ниже модели Бокса и Дженкинса – авторегрессии-проинтегрированного скользящего среднего (АРПСС). При этом следует помнить, что если случайные компоненты ряда были взаимно независимы, то после взятия разностей новые случайные компоненты будут взаимно коррелировать.

Для стационарного процесса рассматривают 3 основных типа моделей (соответствующих определённым типам стационарных стохастических процессов).

Таблица 12.1 – Примерные критерии подбора моделей АРСС (p, q), для которых $p + q \le 2$

	АКФ	ЧАКФ
p=1 , q=0	 $a_1>0$ $a_1<0$ Экспоненциально затухает	 $a_1>0$ $a_1<0$ Выброс на лаге 1 (нет корреляций для других задержек)
p=2 , q=0	 $a_1>0$ $a_1>0$ $a_1<0$ $a_1<0$ $a_2>0$ $a_2<0$ $a_2<0$ $a_2>0$ Затухает синусоидально или экспоненциально	 $a_1>0$ $a_1>0$ $a_1<0$ $a_1<0$ $a_2>0$ $a_2<0$ $a_2<0$ $a_2>0$ Выбросы на сдвигах 1 и 2 (нет корреляций для других задержек)
p=0 , q=1	 $b_1>0$ $b_1<0$ Выброс на сдвиге 1 (остальные значения нулевые)	 $b_1>0$ $b_1<0$ Экспоненциально затухает монотонно или осциллируя (меняя знак)
p=0 q=2	 $b_1>0$ $b_1>0$ $b_1<0$ $b_1<0$ $b_2>0$ $b_2<0$ $b_2<0$ $b_2>0$ Выбросы на сдвигах 1 и 2 (остальные значения нулевые)	 $b_1>0$ $b_1>0$ $b_1<0$ $b_1<0$ $b_2>0$ $b_2<0$ $b_2<0$ $b_2>0$ Синусоидальная волна или экспоненциально затухает
p=1 , q=1	 $0<a_1<1$ $0<a_1<1$ $b_1<a_1<0$ $-1<a_1<0$ $-1<a_1<0$ $a_1<b_1<1$ $0<b_1<a_1$ $-1<b_1<0$ $-1<b_1<0$ $a_1<b_1<0$ $0<b_1<1$ $0<b_1<1$ Экспоненциально затухает (монотонно или колебательно)	 $0<a_1<1$ $0<a_1<1$ $b_1<a_1<0$ $-1<a_1<0$ $-1<a1<0$ $a_1<b_1<1$ $0<b_1<a_1$ $-1<b_1<0$ $-1<b_1<0$ $a_1<b_1<0$ $0<b1<1$ $0<b_1<1$ Экспоненциально затухает монотонно или осциллируя (меняя знак)

1. Авторегрессии (АР) порядка p. В этой модели текущее значение t выражается через линейную комбинацию p предыдущих значений процесса плюс случайный импульс ε_t:

$z_t = a_1 z_{t-1} + a_2 z_{t-2} + a_3 z_{t-3} + \ldots + a_{t-p} z_{t-p} + \varepsilon_t$.

Важными частными случаями являются:

а) при p=1, модель процесса Маркова (процесс с отсутствием последействия - каждое следующее значение зависит только от предыдущего):

$z_t = a_1 z_{t-1} + \varepsilon_t$;

б) при p=2, модель процесса Юла:

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + \varepsilon_t.$$

2. Скользящего среднего (СС):

$$z_t = \varepsilon_t - b_1 \varepsilon_{t-1} - b_2 \varepsilon_{t-2} - b_3 \varepsilon_{t-3} - \ldots - b_{t-q} \varepsilon_{t-q}$$

(термин СС не означает, что сумма весов при ε_i равна 1). Предполагается, что последовательные значения ряда сильно зависимы и генерируются последовательностью независимых импульсов ε_t, которые являются реализацией случайных величин, подчиняющихся нормальному закону распределения с нулевым средним и дисперсией σ_ε^2 (в технике последовательность ε_t, ε_{t-1}, ε_{t-2} , … называется белым шумом).

3. Авторегрессии - скользящего среднего (АРСС):

$$z_t = a_1 z_{t-1} + a_2 z_{t-2} + a_3 z_{t-3} + \ldots + a_{t-p} z_{t-p} - b_1 \varepsilon_{t-1} - b_2 \varepsilon_{t-2} - b_3 \varepsilon_{t-3} - \ldots - b_{t-q} \varepsilon_{t-q} + \varepsilon_t.$$

Обычно, на практике достаточно рассматривать модели АР, СС, АРСС при p и q не превышающих 2.

Для описания нестационарных процессов пользуются экспоненциально взвешенными средними (см. выше). В более общем случае рассматривают модели Бокса и Дженкинса - авторегрессии-проинтегрированного скользящего среднего (АРПСС), [8,17]. В этой модели тренд исключается переходом к разностям ряда ($\nabla Z_t = Z_t - Z_{t-1}$) и допускается коррелированность остатков.

Без учёта сезонных эффектов модель имеет вид:

$$\nabla^d Z_t - \alpha_1 \nabla^d Z_{t-1} - \alpha_2 \nabla^d Z_{t-2} - \ldots - \alpha_p \nabla^d Z_{t-p} = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}.$$

Рассмотрим оператор сдвига назад B, определяемый как $BZ_t = Z_{t-1}$, тогда $\nabla = 1 - B$ и предыдущую формулу можно переписать в виде

$$\alpha(B)(1-B)^d Z_t = \beta(B)\varepsilon_t,$$

где $\alpha(B)$ и $\beta(B)$ – полиномы от B порядка p и q. Необходимо определить три параметра: p, q, d. Считается, что на практике они не превышают 2.

Идентификация модели АРПСС (p, d, q). С помощью графиков (изучаемого ряда и его автокорреляционной и частной автокорреляционной функций) визуально оценивают стационарностьили нестационарность ряда. Если ряд признаётся нестационарным, то вычисляют разности ряда до момента, пока он не станет стацио-

нарным и таким образом дают оценку *d*. Необходимая для достижения стационарности разность *d* (0, 1, 2) предполагает затухание автокорреляционной функции, соответствующей порядку разности.

Параметры *p* и *q* определяют, используя автокорреляционную функцию (АКФ) и частную автокорреляционную функцию (ЧАКФ) [10,15]. Приближённые критерии оценки параметров *p* и *q* модели АРСС (*p,q*) рассмотрены в таблице 12.1.

Для процесса авторегрессии порядка *p* – автокорреляционная функция процесса спадает плавно, а её частная автокорреляционная функция обрывается после задержки (лага) *p*. Для процесса скользящего среднего порядка *q* – автокорреляционная функция обрывается после задержки *q*, а её частная автокорреляционная функция спадает плавно. Для смешанного процесса АРСС (*p, q*) – автокорреляционная функция после (*q-p*) задержек приближённо представляется в виде суммы экспонент и затухающих синусоид, а её частная автокорреляционная функция приближённо представляется в виде суммы экспонент и затухающих синусоид после (*p-q*) задержек. Поведение АКФ процесса авторегрессии похоже на поведение ЧАКФ процесса скользящего среднего.

На практике обычно не достигается полного сходства между выборочной и теоретической АКФ. Необходимо ориентироваться на главные характеристики. Так, например, для реальных процессов АКФ может иметь всплески и тренды.

Следует отметить, что модель АРПСС(*p, d, q*) может быть обобщена и представлена в виде мультипликативной сезонной модели АРПСС (*p, d, q*)×(*Ps, ds, Qs*)$_s$. В этой модели к параметрам *p, d, q* – добавлены сезонный параметр авторегрессии *Ps*, сезонная разность ds и сезонный параметр скользящего среднего *Qs*. Например, модель АРПСС $(0,1,1) \times (0,1,1)_{12}$ содержит один параметр скользящего среднего и один сезонный параметр скользящего среднего, полученные после взятия разности с лагом 1, а затем сезонной разности с лагом *s*=12 (*s* больше единицы).

Сезонность может идентифицироваться с помощью АКФ и ЧАКФ, а также с помощью графика спектральной плотности изучаемого ряда. Адекватность полученных моделей оценивается с помощью АКФ остатков. Если график АКФ не содержит периодических колебаний, систематического смещения и сильных корреляций (более 0,5-0,6), то модель адекватна.

Если в силу внешних причин поведение ряда резко изменяется, то можно проводить анализ моделей АРПСС с интервенцией.

Для оценки запаздывающей зависимости между временными рядами используют Анализ распределённых лагов, позволяющий построить регрессию одного ряда на другой (если одни измерения опережают другие).

Замечание.

1.Задачи анализа и обработки временных рядов в настоящее время являются типичными во многих областях науки и техники: микроэкономике, макроэкономике, демографии, криминалистике, экологии, финансах, сельском хозяйстве, радиоэлектронике, геофизике, астрономии, медицине, радиолокации, проектировании и испытании мостов, самолетов и автомобилей и т.д. Кроме того, анализ временных рядов интенсивно используется на валютном и фондовом рынках.

2. Следует отметить, что приведённые выше (и другие модели) используют априорные предположения о процессе, генерирующем изучаемый временной ряд. Реальные данные (особенно малого объёма – порядка нескольких десятков) часто не укладываются в описанные выше модели.

Именно поэтому созданы десятки методов анализа и прогнозирования временных рядов, предполагающих использование искусственного интеллекта, теории хаоса, оптимизационных моделей, нейронных сетей, вейвлет-преобразований и т.д. Выше описываются классические методы анализа временных рядов. Критерием выбора может являться только достижение практических целей анализа: описание поведения ряда, объяснение изменения наблюдений, прогноз.

Анализ временных рядов в Statistica

Выполнив команду **Анализ – Углубленные методы анализа – Временные ряды и прогнозирование (*Statistics – Advanced Linear/Nonlinear Models – Time Series / Forecasting*)** мы получим соответствующее диалоговое окно (рис.12.1).

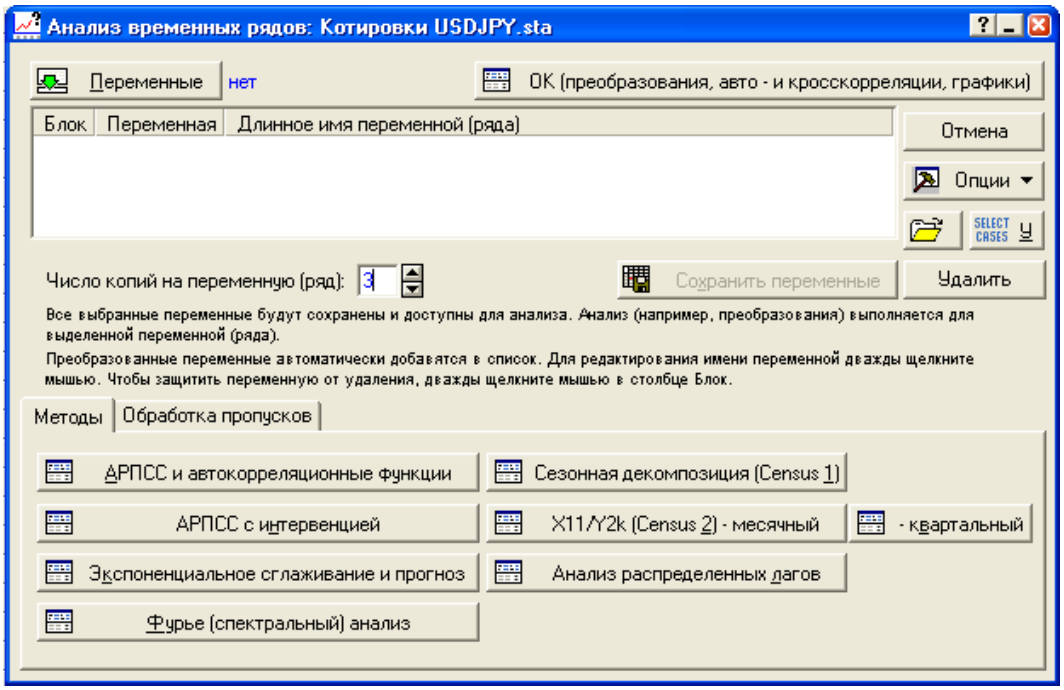


Рисунок 12.1А– Диалоговое окно Анализ временных рядов (***Time Series Analysis***)

Рассмотрим операции доступные в диалоговом окне (рис.12.1А). При выборе кнопки **Переменные (*Variables*)** будут доступны переменные в открытом в настоящее время файле.

В верхней части, в информационном поле, записываются имена анализируемых и преобразованных переменных. Для каждого временного ряда можно оставлять от 3 до 99 копий, например, если выбрано 3 копии, то при четвёртом преобразовании первое преобразование удалится. **(Число копий на переменную (ряд) – *Number of backups per variable* (*series*).)**Если дважды щёлкнуть левой кнопкой мыши в поле преобразованной переменной на колонке Блок, то переменная будет заблокирована (появится, буква *L*). Анализ временных рядов, как и другие методы анализа, предполагает, что данные не имеют пропусков, поэтому предлагаются различные способы заполнения пропусков. Вкладка Обработка пропусков (*Missing data*) позволяет осуществлять замену пропущенных данных внутри ряда: общим средним, рассчитанным по всему ряду; интерполяцией по ближайшим (не пропу-

щенным) точкам; средним *N* ближайших значений; медианой *N* ближайших значений; значения линейной регрессии (рис.12.1В).

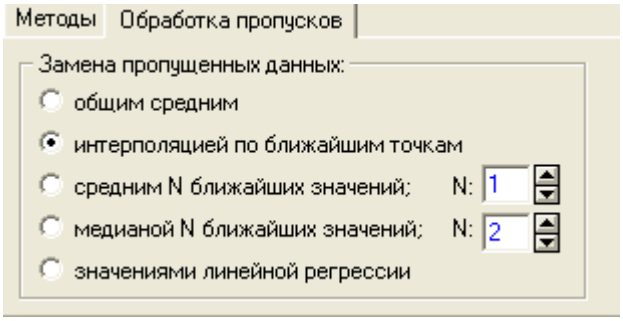


Рисунок 12.1В - Диалоговое окно Анализ временных рядов Вкладка Обработка пропусков (*Missing data*)

Практически все методы анализа временных рядов предполагают наличие в ряде регулярной составляющей и шума. Регулярная составляющая может быть трендом или сезонностью (цикличностью), повторяющейся через определённый шаг (лаг). Для фильтрации шума используются различные методы преобразований и сглаживания, доступные после выбора кнопки **OK (Преобразования, авто – и кросскорреляции, графики – *Transformations, autocorrelations, crosscorrelations, plots*)** (рисунки 12.2 -12.5 соответственно).

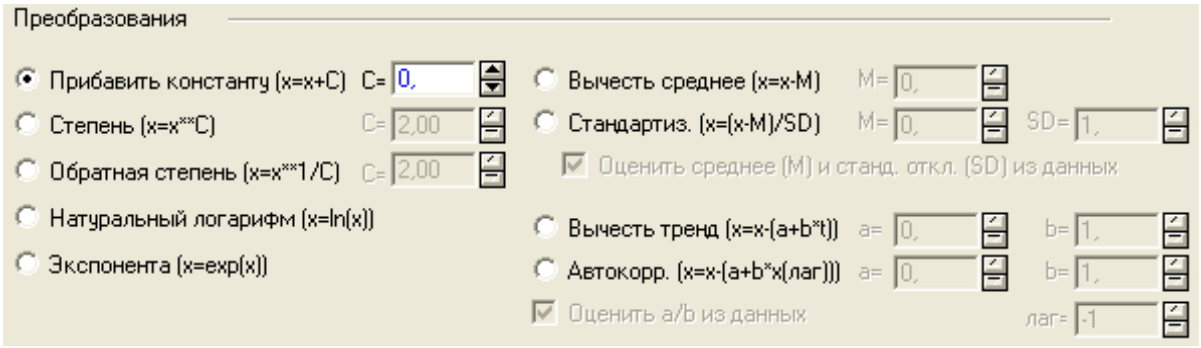


Рисунок 12.2 – Окно OK (Преобразования) – вкладка *x=f(x)*

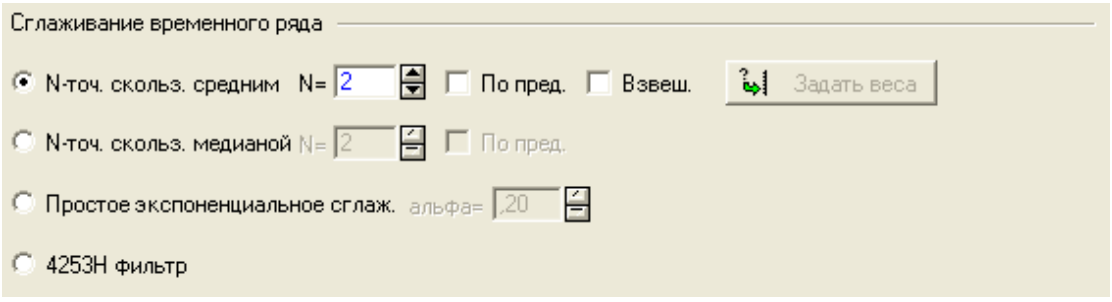


Рисунок 12.3 – Окно OK (Преобразования) – вкладка Сглаживание (*Smoothing*)

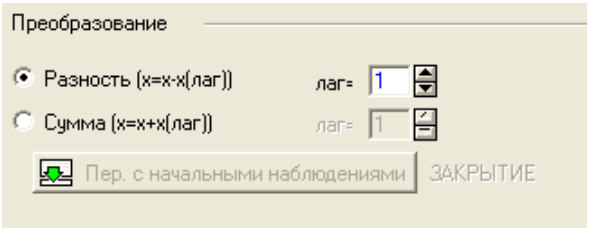


Рисунок 12.4 – Окно ОК (Преобразования) – вкладка Разность, сумма (*Difference, integrate*)

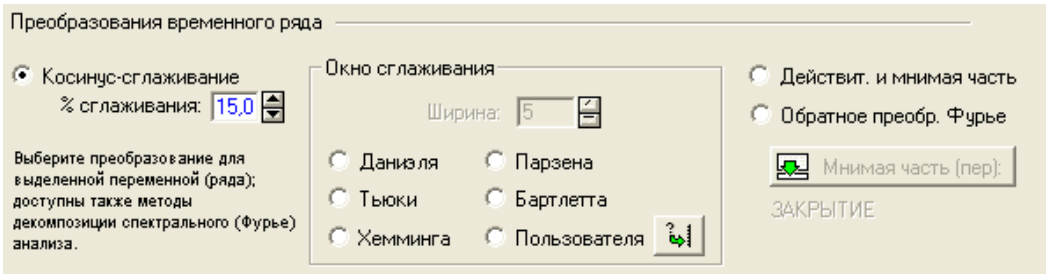


Рисунок 12.5 – Окно ОК (Преобразования) – вкладка Фурье (*Fourier*)

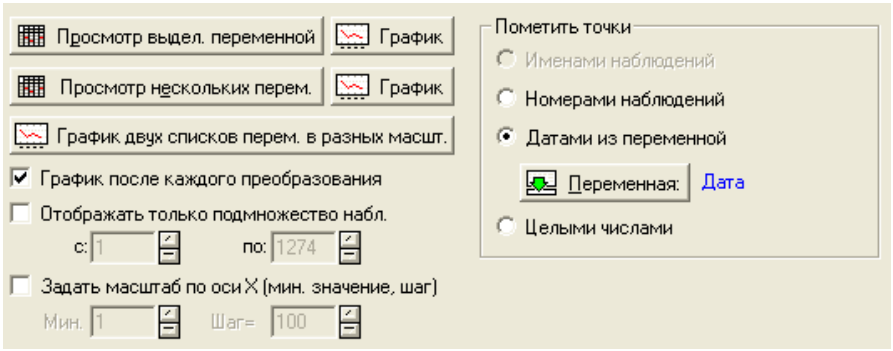


Рисунок 12.6 – Окно ОК (Преобразования) – вкладка Графики (*Review & plot*)

Вкладка Графики (*Review&plot*) позволяет построить графики одной или нескольких переменных (как исходных, так и преобразованных) в том числе и с разными масштабами (рис.12.6). С использованием вкладки Описательные (*Descriptives*) можно найти описательные статистики ряда; построить гистограммы, графики на нормальной вероятностной бумаге разных типов. Вкладка Сдвиг (*Shift*) позволяет осуществить сдвиг ряда на заданный лаг вперёд или назад.

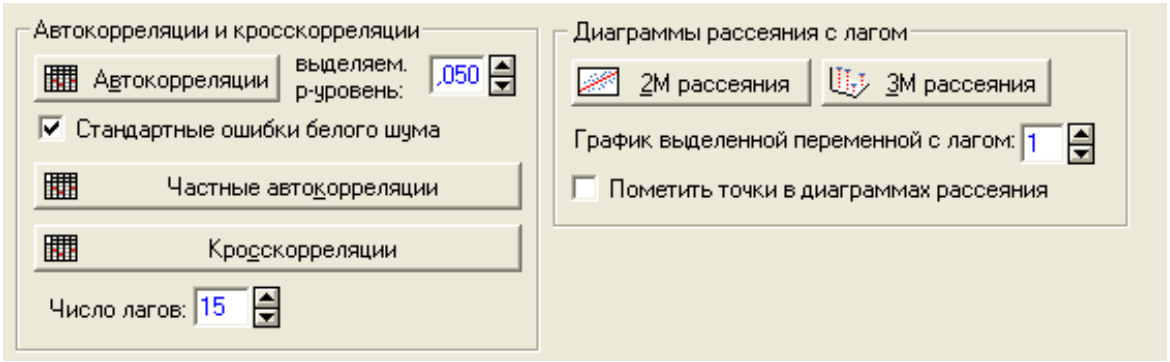


Рисунок 12.7 – Окно ОК (Преобразования) вкладка Автокорреляции (*Autocorrs*)

Сезонности и циклы можно определять, используя автокорреляции и частные автокорреляции, устраняющие влияние автокорреляций меньшего порядка (рис. 12.7). Кроме того, в системе *Statistica* используются специальные методы поиска циклов: *Census* 1 и *Census* 2.

Для анализа нестационарных процессов часто используется модель АРПСС (рис.12.8).

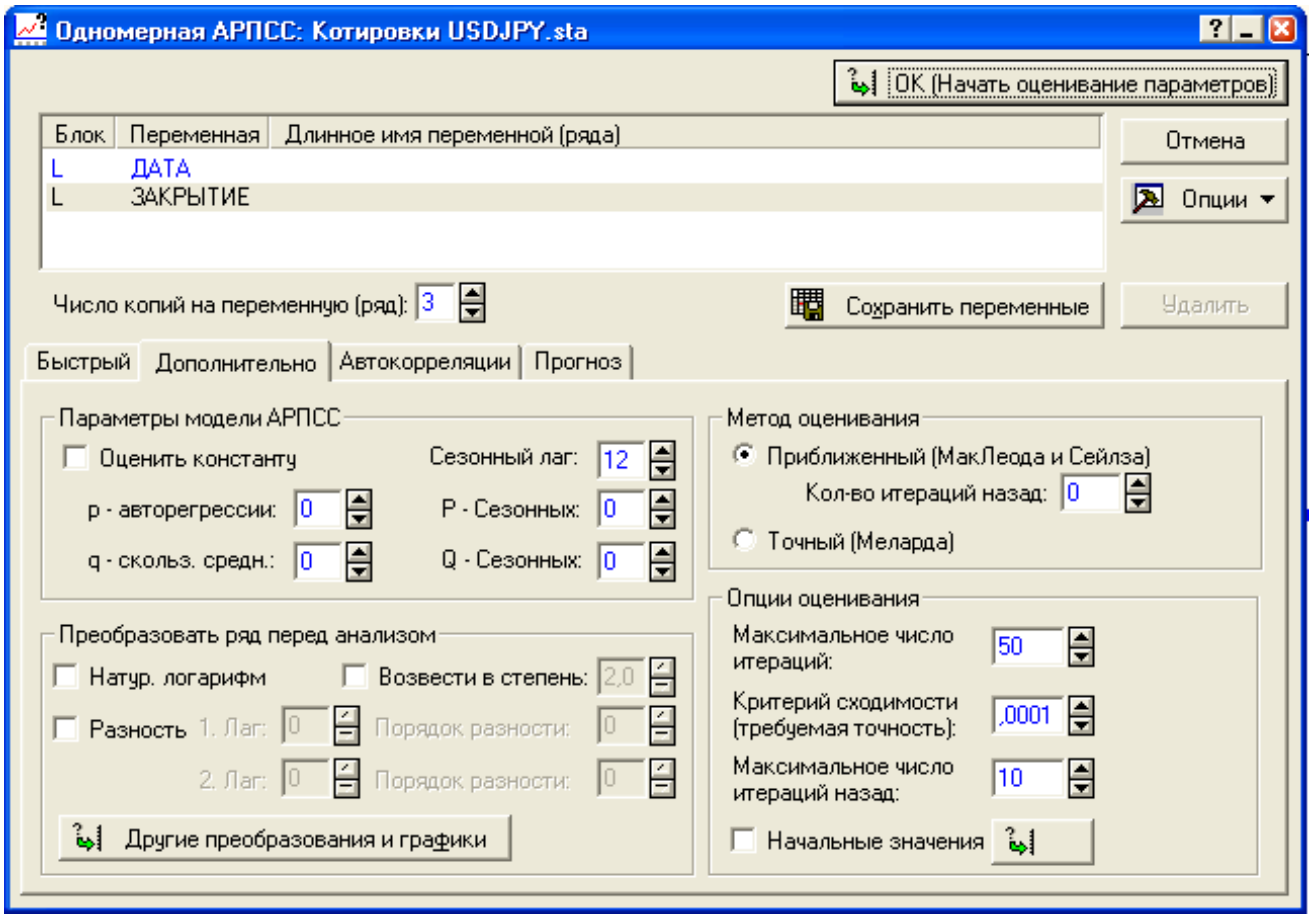


Рисунок 12.8 – Окно АРПСС

Пример 1. (Преобразования переменных.) Рассмотрим возможности по преобразованию переменных модуля анализа временных рядов на примере котировок валютных пар доллар США – японская Йена (доллар/йена - *USDJPY*) с 1 января 2002г по 17 ноября 2006г (файл котировки.xls).

По мнению многих аналитиков, цена закрытия позволяет лучше оценить основные тенденции, поэтому в качестве анализируемого выберем именно этот временной ряд, для этого выделим переменную Закрытие.

Выполнив команду **OK (выполнить преобразования, авто - и кросскорреляции, графики *-Transformations, autocorrelations, crosscorrelations, plots*)** – выделим переменную Закрытие – **График** (***Review & plot***) **– Пометить точки данными из переменной** (***Dates from a variable***) Дата. График указывает на наличие тренда и нестационарность ряда (рис. 12.9).

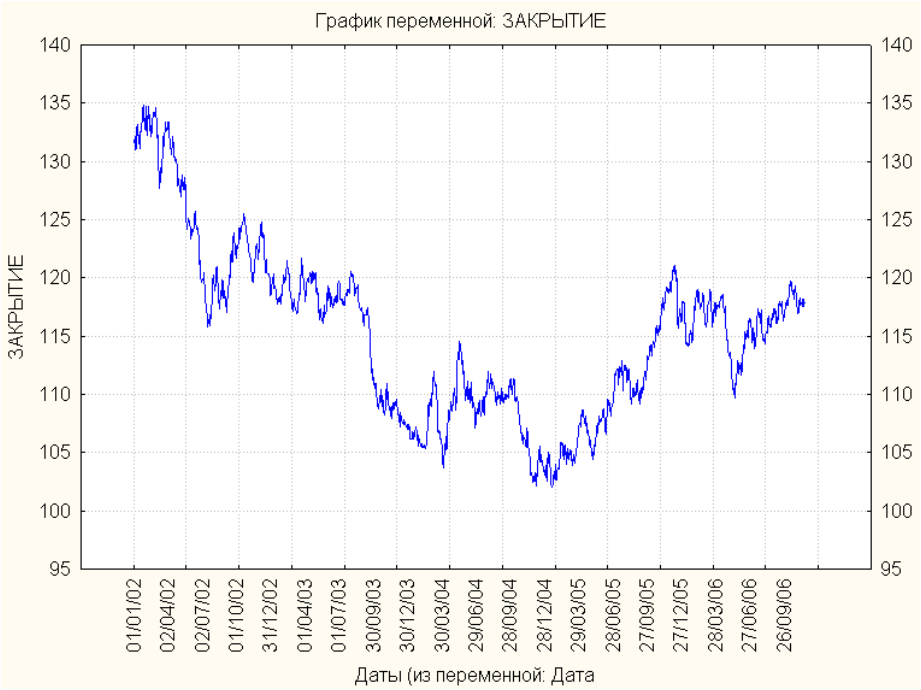


Рисунок 12.9 – График переменной Закрытие

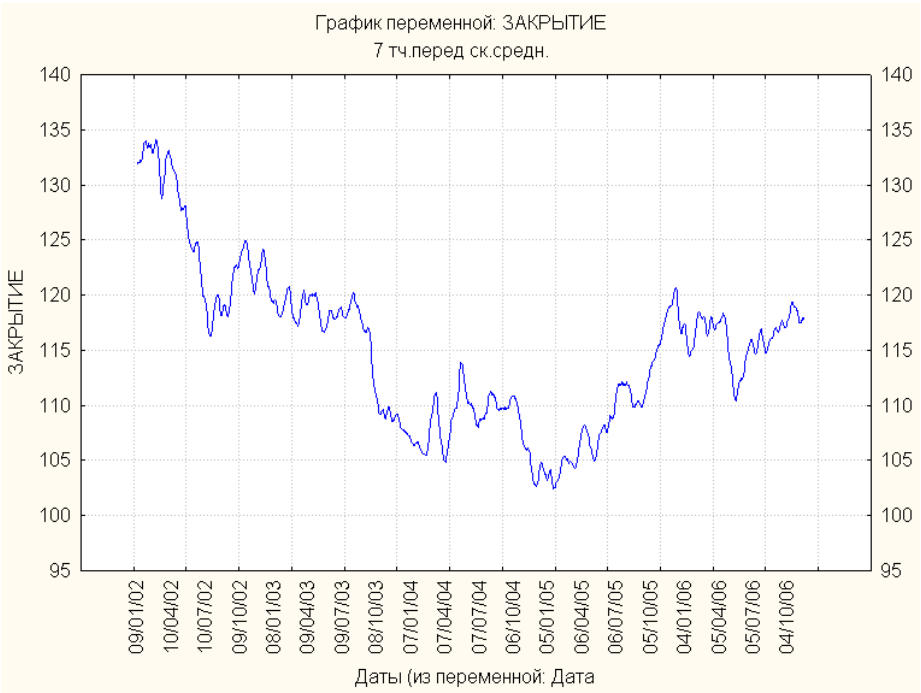


Рисунок 12.10 – График переменной Закрытие преобразованной с помощью 7 дневной скользящей средней

Выполнив команду**OK (выполнить преобразования, авто – и кросскорреляции, графики** – ***Transformations, autocorrelations, crosscorrelations, plots***) – выделим переменную Закрытие – откроем вкладку **Сглаживание (*Smoothing*)** 7 точечным скользящим средним (*N*=7) – откроем вкладку График **(*Review & plot*) – Пометить точки данными из переменной** (***Dates from a variable***) Дата, – мы по-

лучим сглаженный временной ряд, который имеет меньше «зубцов» и лучше видна тенденция на убывание. Эта тенденция смещает оценки авторорреляционной функции, поэтому следует удалить тренд.

Удалим тренд переменной Закрытие (команда x=f(x) – **вычесть тренд (*trends ubtract*)** (*x=x*-(*a*+*b***t*)–**OK (преобразовать выделенную переменную – *transform selected series*,** рис.12.11).

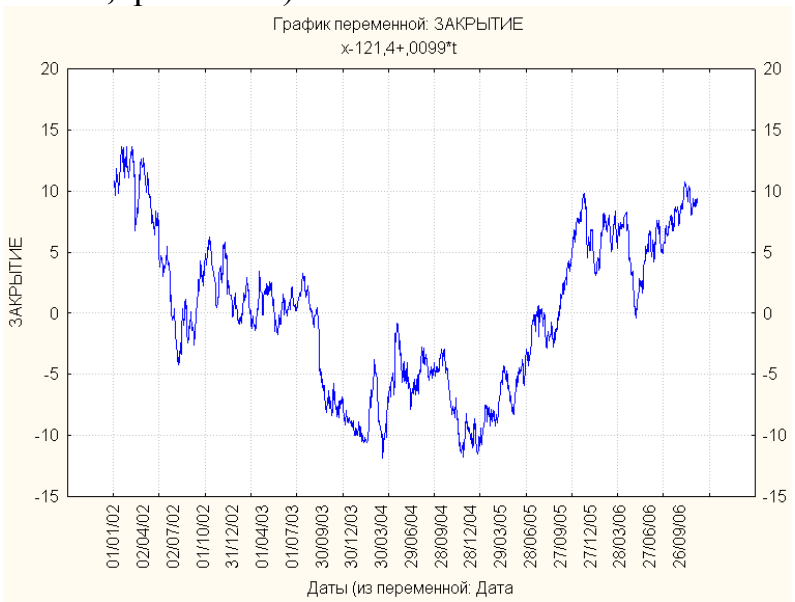


Рисунок 12.11 – График переменной Закрытие после удаления линейного тренда

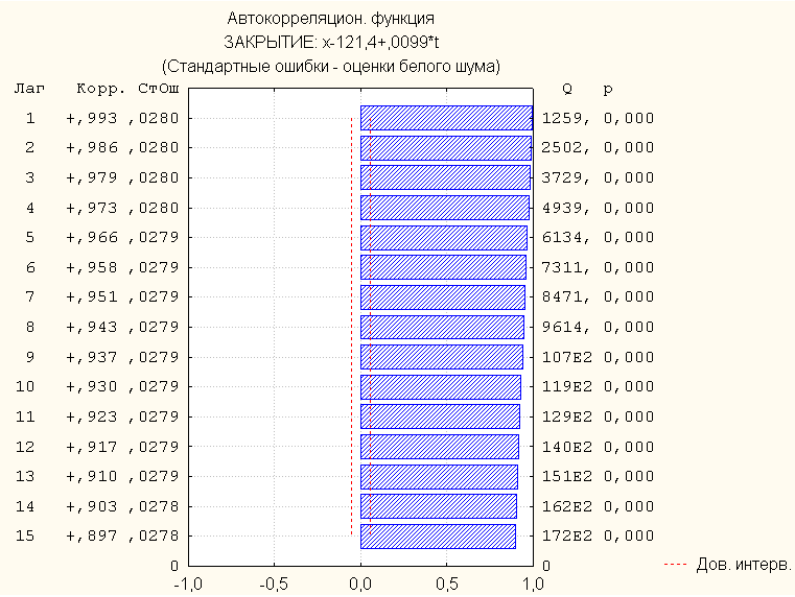


Рисунок 12.12 – График автокорреляционной функции для переменной Закрытие после удаления линейного тренда

Выделим преобразованный ряд, выберем кнопку **Автокорреляции (*Autocorr.*)** (рис.12.7) и получим, что для всех значений лага от 1 до 15 наблюдается сильная автокорреляция (рис. 12.12). Для проверки этого факта используем кнопку **Частная автокорреляция (*Partial autocorrelations*)**, которая позволяет искать автокорреляции для всех уровней лага без учёта влияния автокорреляций с меньшими лагами (рис. 12.13).

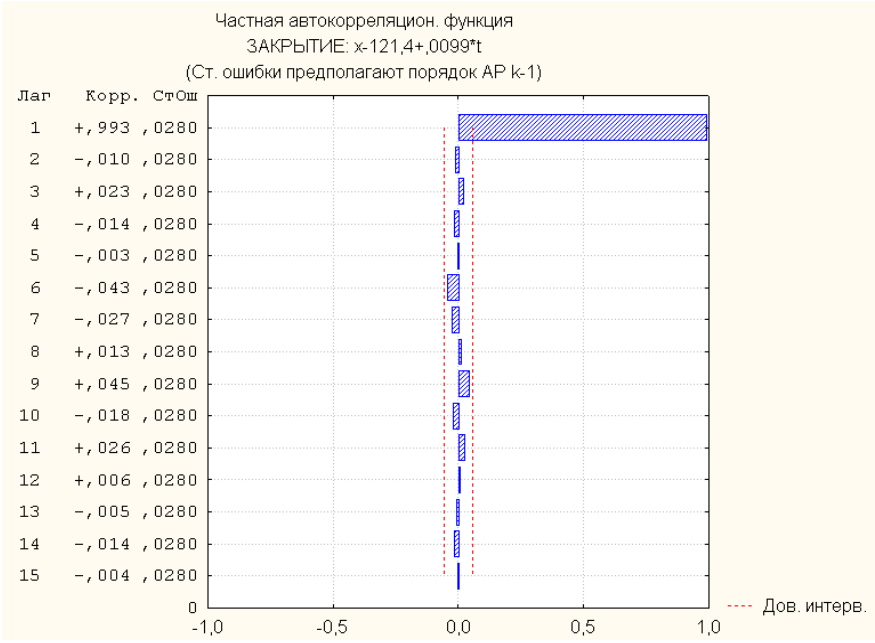


Рисунок 12.13 – График частной автокорреляционной функции для переменной

Мы видим, что автокорреляции с лагом выше 1 обуславливались автокорреляциями с меньшими лагами. Таким образом, на каждое последующее значение ряда влияет только значение предыдущего.

Удалим автокорреляцию в окне преобразования переменных команда ***x=f(x)*** **– Автокорр.** ***(x=x-(a+b*x(лаг=1)*** **– OK (преобразовать выделенную переменную –** ***transform selected series*****)**. Или рассмотрим разность первого порядка для переменной Закрытие (вкладка **Разность, сумма (*Difference, integrate*) – OK**) и получим рисунок, указывающий на возможную стационарность процесса (рис.12.14).

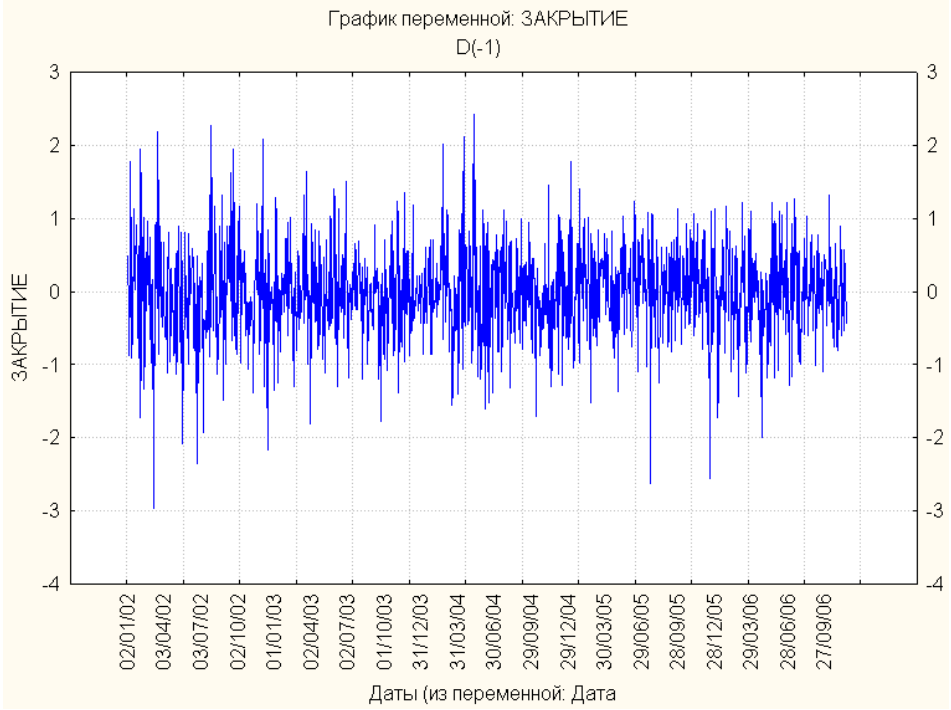
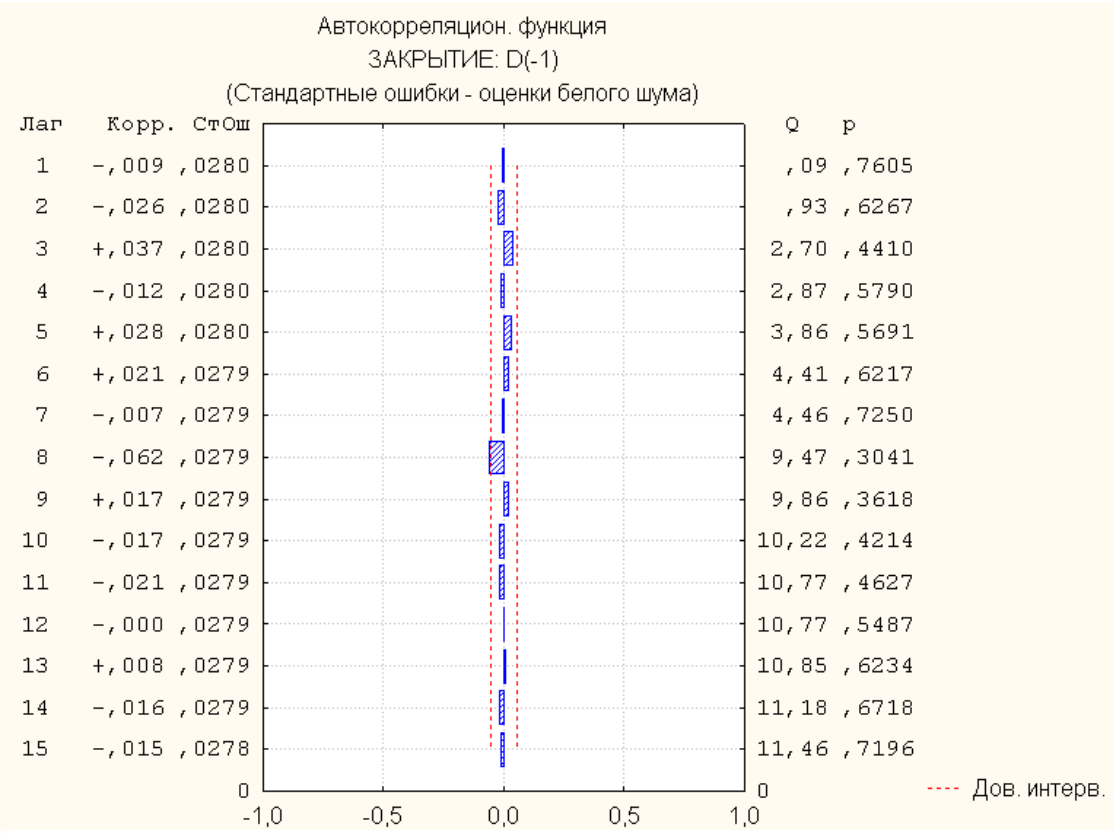


Рисунок 12.14 – График переменной Закрытие после взятия разности 1 порядка

Рассмотрим АКФ м ЧАКФ переменной Закрытие после взятия разности 1 порядка.



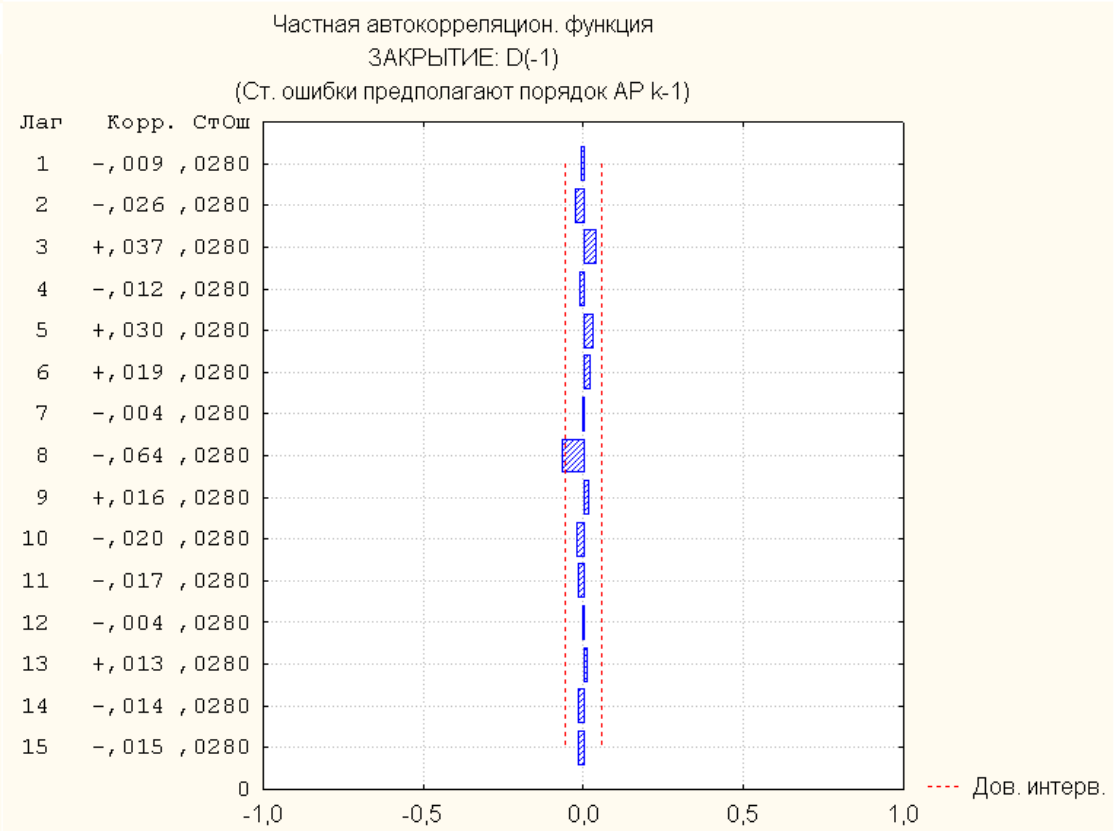


Рисунок 12.15 – Графики автокорреляционной и частной автокорреляционной функции для переменной Закрытие после взятия разности 1 порядка

Так как после взятия разности 1 порядка нет значимых значений АКФ и ЧАКФ (рис. 12.15), то модель не идентифицируется как модель АРСС. Однако, если взять разности второго порядка (рис. 12.16), то получаются значимые автокорреляции (рис. 12.17). Это объясняется наличием тренда второго порядка, заметного на рисунке 12.9.

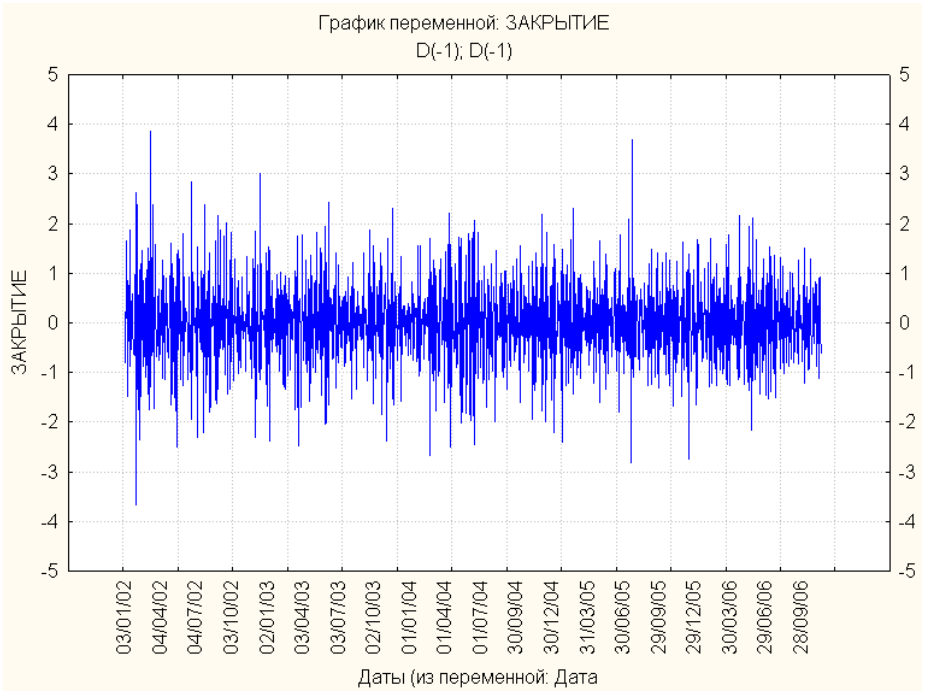


Рисунок 12.16 – График переменной Закрытие после взятия разности второго порядка

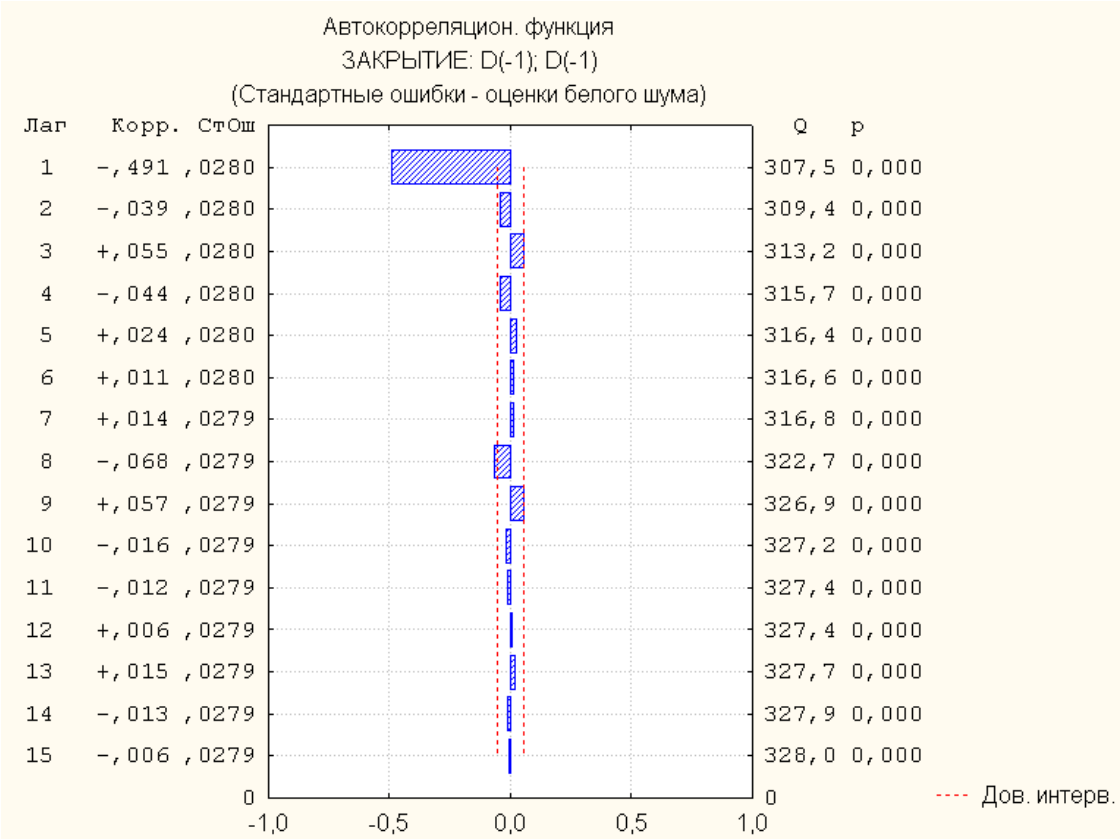


Рис. 12.17 –График автокорреляционной функции для переменной

Закрытие после взятия разности второго порядка

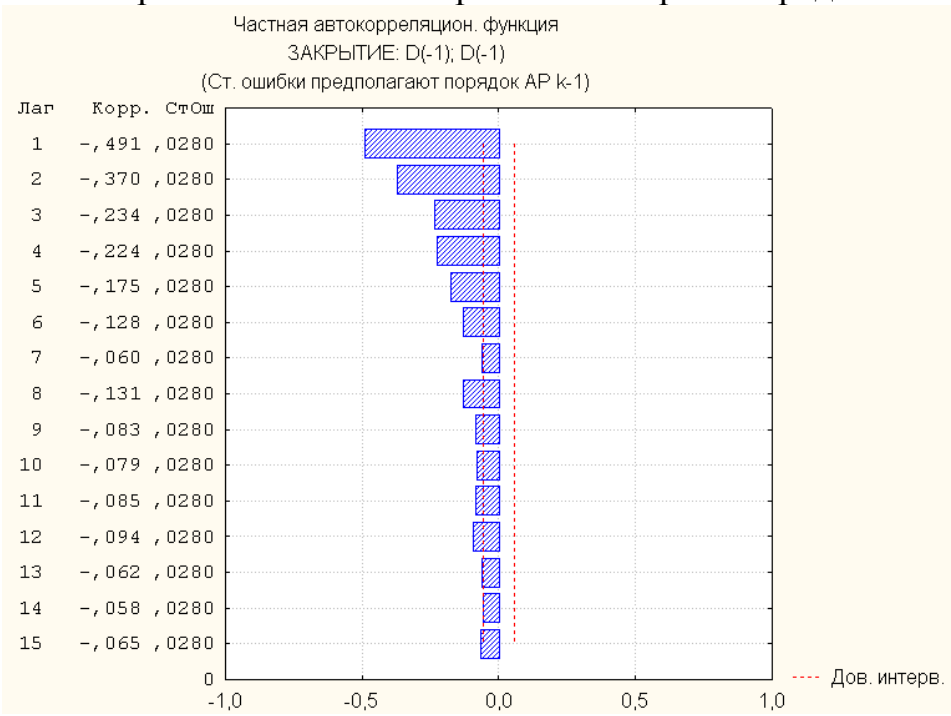


Рис. 12.18 –График частной автокорреляционной функции для переменной Закрытие после взятия разности 2 порядка

Из рисунков 12.17-12.18 следует, что согласно, приведённых выше критериев идентификации (см. табл.12.1), процесс, получившийся после взятия разности 2 порядка следует отнести к процессу СС первого порядка. Это можно объяснить наличием нелинейной составляющей 2 порядка, которая заметна на графике исходного ряда (рис.12.4). Итак, мы идентифицировали процесс котировок валютных пар доллар/йена по переменной Закрытие – как процесс АРПСС (0, 2, 1).

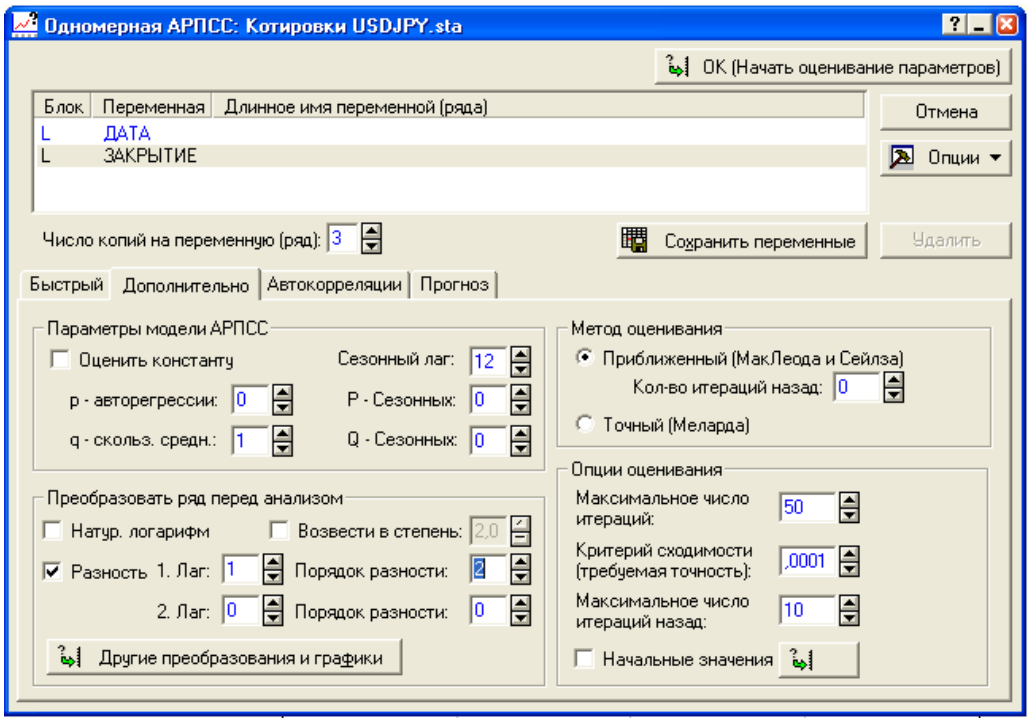


Рисунок 12.19 – Диалоговое окно АРПСС (*ARIMA*)

Найдём параметры модели. Выйдем из окна преобразований (***Cancel***), выполним команду **Методы – АРПСС и автокорреляционные функции (*Quick – ARIMA & autocorrelation functions*) – ОК**, предварительно заполнив диалоговое окно в соответствии с рисунком 12.19.

В диалоговом окне **Результаты АРПСС (*Single Series ARIMA Results*)** выберем кнопку **Оценки параметров (*Summary: Parameter estimates*)** (рис. 12.20) и получим таблицу, изображённую на рисунке 12.21.

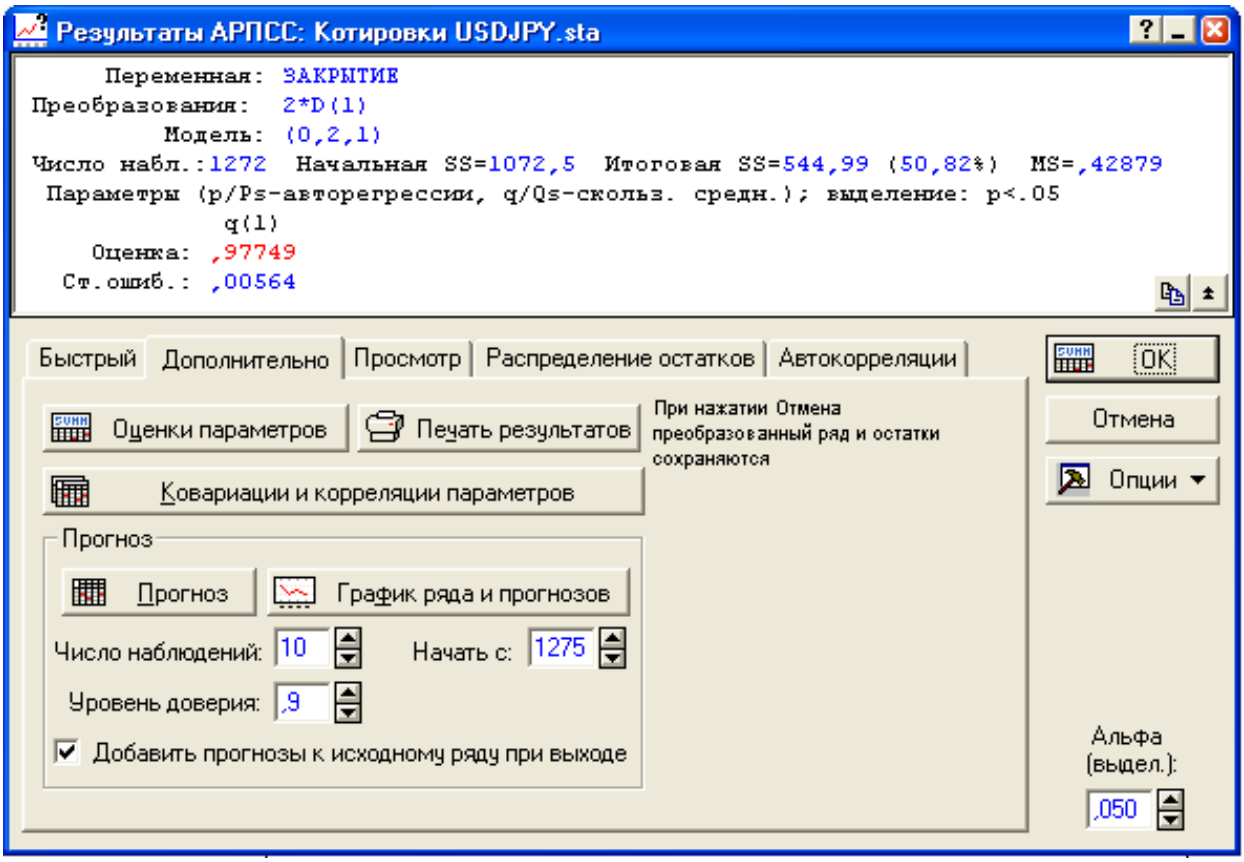


Рисунок 12.20 – Диалоговое окно результаты АРПСС

Параметр	Исход.:ЗАКРЫТИЕ (Котировки USDJPY.sta) Преобразования: 2*D(1) Модель(0,2,1) MS Остаток= ,42879					
	Парам.	Асимпт. Ст.ошиб.	Асимпт. t(1271)	p	Нижняя 95% дов.	Верхняя 95% дов.
q(1)	0,977495	0,005645	173,1658	0,00	0,966420	0,988569

Рисунок 12.21 – Оценка параметра b_1

Параметр модели СС(1) b_1=0,977495 доверительным интервалом (0,966420; 0,988569) при уровне значимости 0,05 (доверительной вероятностью 0,95). Параметр b_1 значим при уровне значимости не более 0,01.

Далее выберем вкладку **Распределение остатков – Гистограмма и Нормальный (*Distribution of residuals – Histogram and Normal Probability Plot*)**. Мы

получим гистограмму остатков и график нормальной вероятностной бумаги для остатков (рис.12.22), которые не противоречат гипотезе о нормальности распределения остатков.

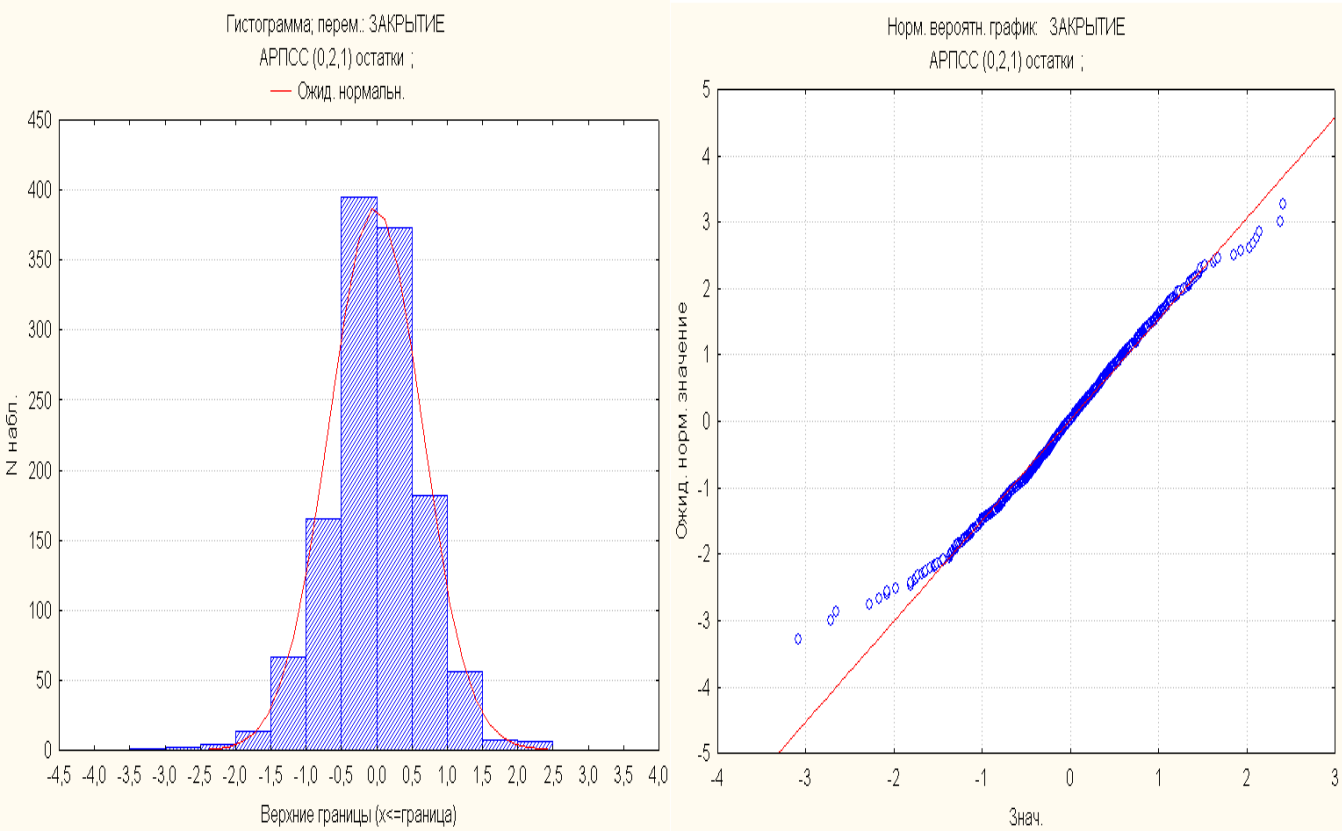


Рисунок 12.22 – Гистограмма и график вероятностной бумаги для остатков

Следующий шаг – это проверка коррелированности остатков. В окне результатов откроем вкладку **Автокорреляции (*Autocorrelations*)** и выберем последовательно Автокорреляции и Частные автокорреляции. В результате получим графики, изображённые на рисунке 12.23. Графики указывают на то, что остатки не коррелированны.

Выбрав число наблюдений 24 и кнопки **Прогноз (*Forecastcases*)**, **График ряда и прогнозов (*Plot & forecasts*)** – мы получим прогноз на заданное число дней.

Модель хорошо описывает процесс Закрытие, однако, к прогнозу, полученному с её помощью, следует относиться осторожно. Этому есть несколько причин: параметр СС почти равный 1 говорит о том, что процесс близок к нестационарному (в этом можно убедиться выбрав точный способ оценки параметров – см. рис.12.19); границы доверительной области расходятся (рис.12.24), то есть, несмотря на то, что нам далось подобрать модель – ряд очень близок к нестационарному и не поддаётся удовлетворительному прогнозу.

Фактически мы получили классический результат – имеет место модель случайного блуждания для исходных данных.

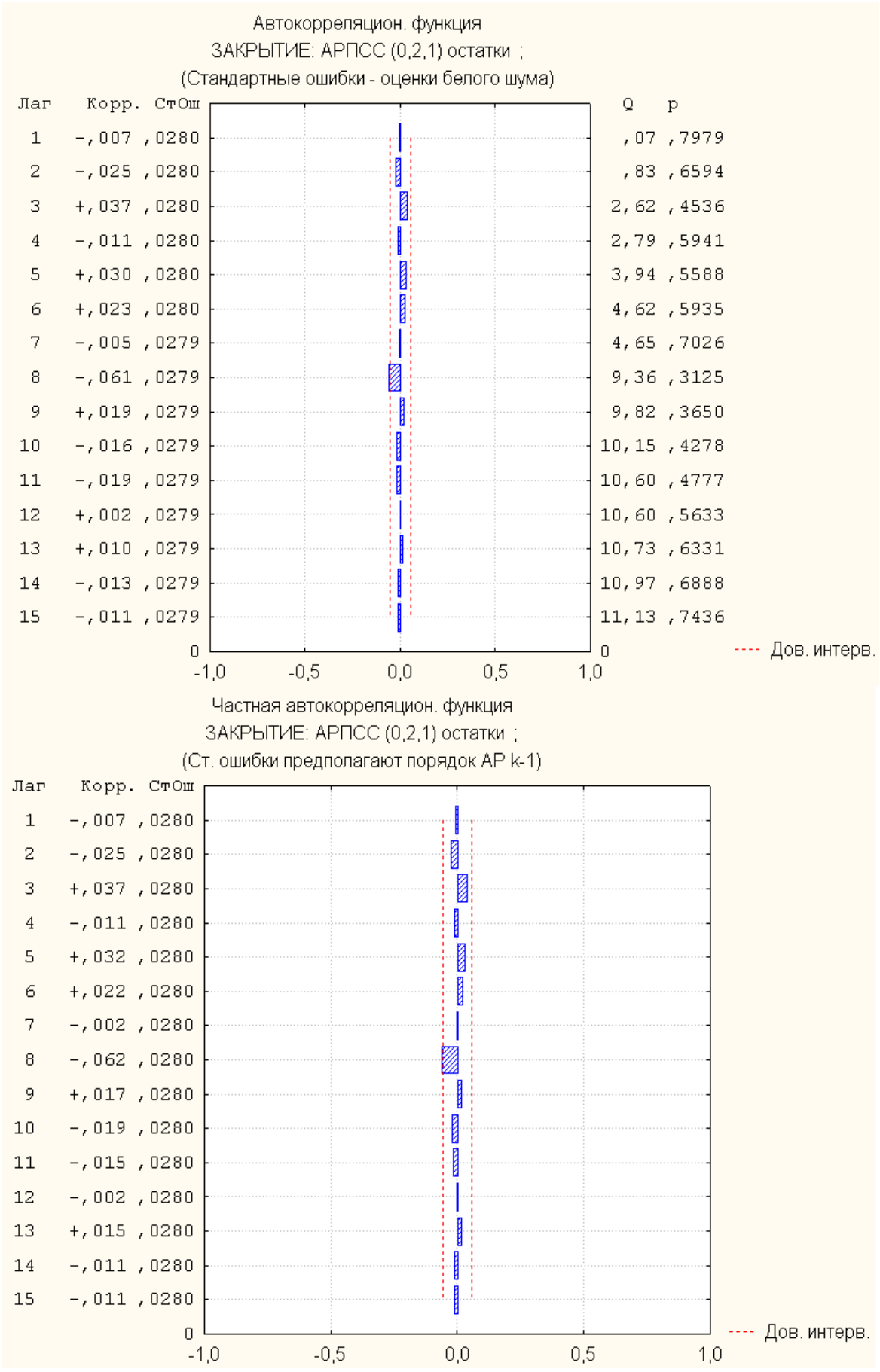


Рисунок 12.23 – АКФ и ЧКФ остатков

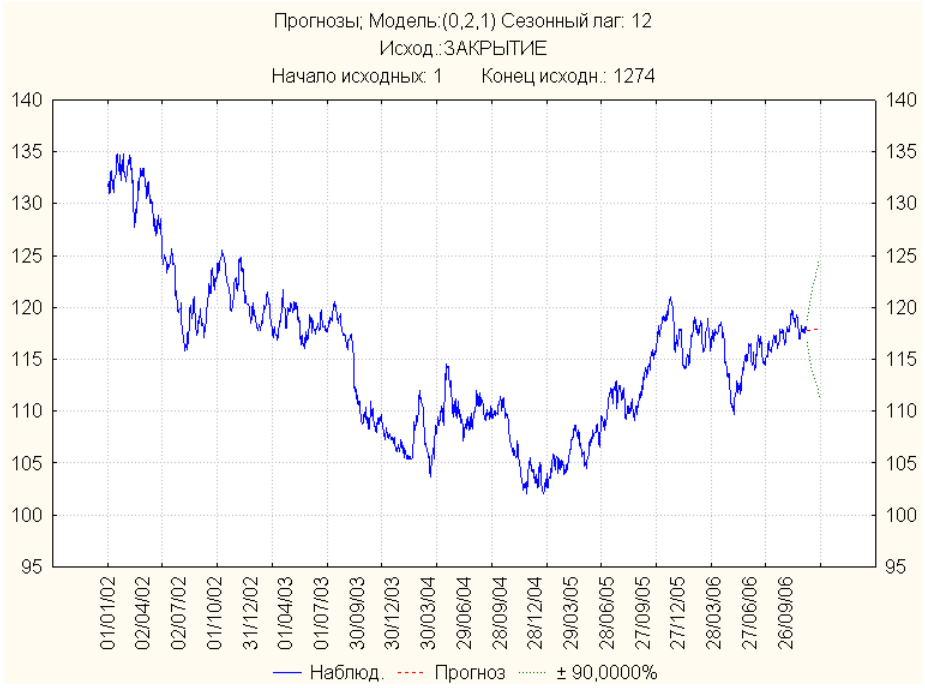


Рисунок 12.24 – Графики прогноза ряда Закрытие

Качество полученной модели можно оценить визуально. Вернёмся в исходное окно Анализа временных рядов и выполним команду **ОК (выполнить преобразования, авто – и кросскорреляции, графики** – ***Transformations, autocorrelations, crosscorrelations, plots***) Далее, открыв вкладку **Графики (*Review&plot*)**, выберем для просмотра нескольких переменных кнопку **График (*Plot*)** во второй строке, а затем отметим переменные: **ЗАКРЫТИЕ** и **ЗАКРЫТИЕ: +прогнозы, Модель: (0,2,1)**. Выберем **ОК** и получим рисунок 12.25, иллюстрирующий хорошее соответствие модели исходному временному ряду.

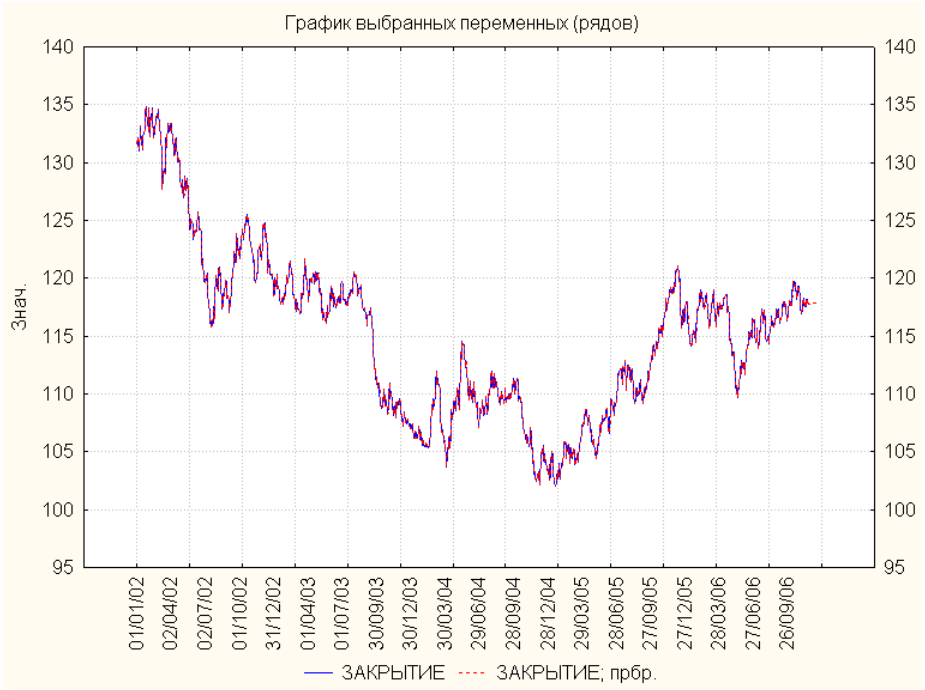


Рисунок 12.25 – Графики исходного ряда Закрытие и аппроксимирующей модели

Замечание. В настоящее время в России ещё нет исторически сравнимых длинных временных рядов. Поэтому для иллюстрации аппарата анализа временных рядов часто обращаются к котировкам акций и валютным парам на биржах. Следует отметить, что, несмотря на универсальность аппарата анализа временных рядов, каждая предметная область привносит свои особенности. Так при анализе ценных бумаг и валютных пар стал традиционным технический анализ, позволяющий по графикам, специальным индикаторам, внешним факторам и т.д. прогнозировать движение стоимости акций. (Заинтересованным лицам рекомендуем книгу классика анализа валютного и фондовых рынков Томаса Р. Демарка: «Технический анализ – Новая наука», М.: Диаграмма, 2001, 280с.)

Пример 2. Рассмотрим классический пример модели АРПСС Дж. Бокса и Г.Дженкинса. Рассматривается ряд пассажирских перевозок на международных авиалиниях – месячные итоги (в тысячах пассажиров) с января 1949 по декабрь 1960 [Бокс, "Ряд *G*"]. Откроем стандартный пример *Statistica – Series_G.sta* из папки *Examples – Datasets* и изобразим его на графике (рис. 12.26).

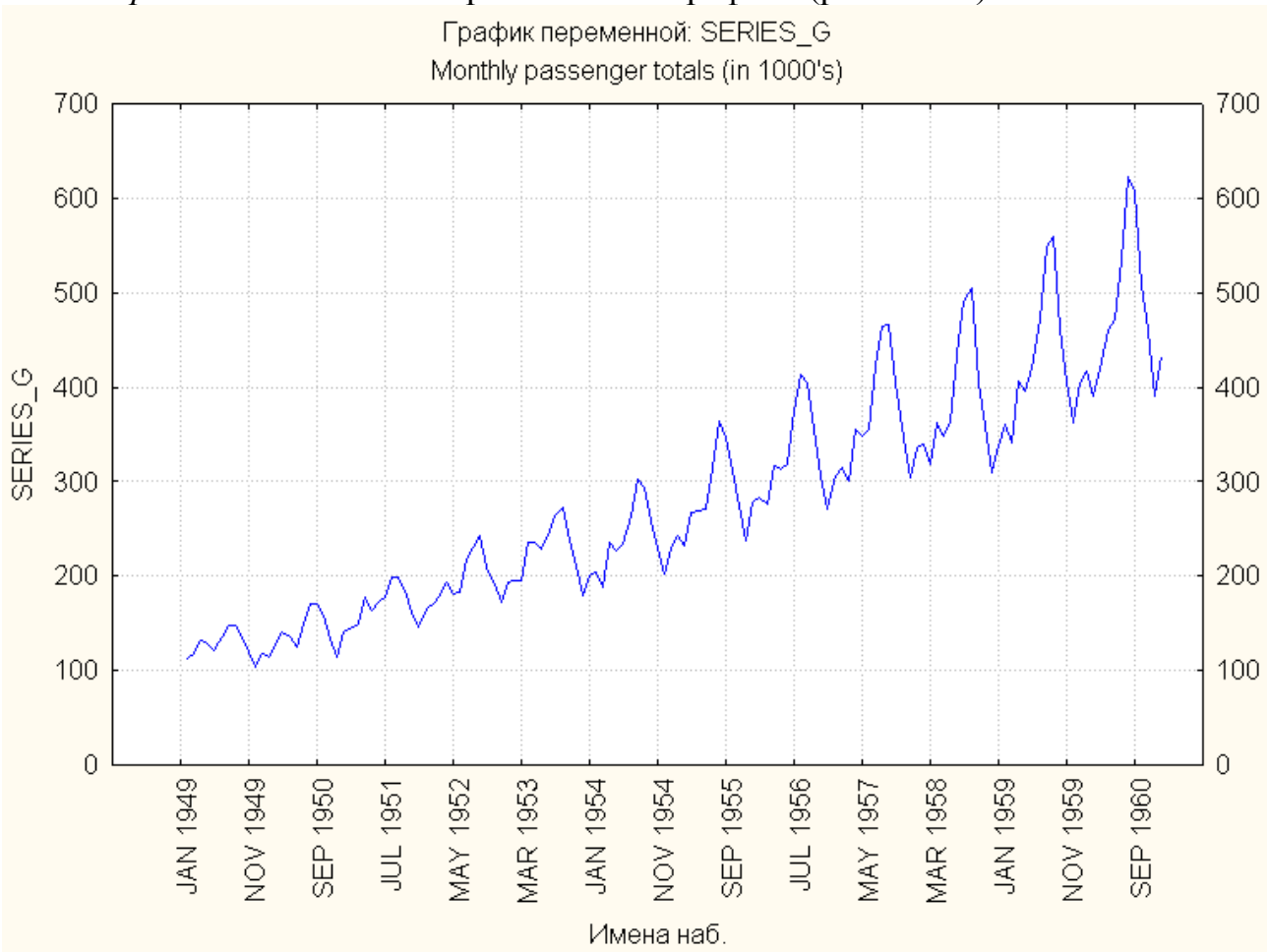


Рисунок 12.26 – График ряда G – пассажирских перевозок на международных авиалиниях

Из рисунка видно, что ряд имеет линейный тренд и сезонность в силу того, что со временем сезонность увеличивается – можно предположить мультипликативность модели. Кроме того, амплитуда со временем возрастает. Таким образом, мы можем для устранения эффекта мультипликативности прологарифмировать

ряд. Выполним команду **OK** (**Преобразования, авто – и кросскорреляции, графики – *Transformations, autocorrelations, crosscorrelations, plots***) – вкладка x=f(x) – Натуральный логарифм (***Natural log***) *(x=ln(x))*. На рисунке 12.22 можно увидеть, что амплитуда и сезонность стали стабильными.

Из рисунка 12.27 следует наличие линейного тренда, для его устранения рассмотрим разности первого порядка (рис. 12.28).

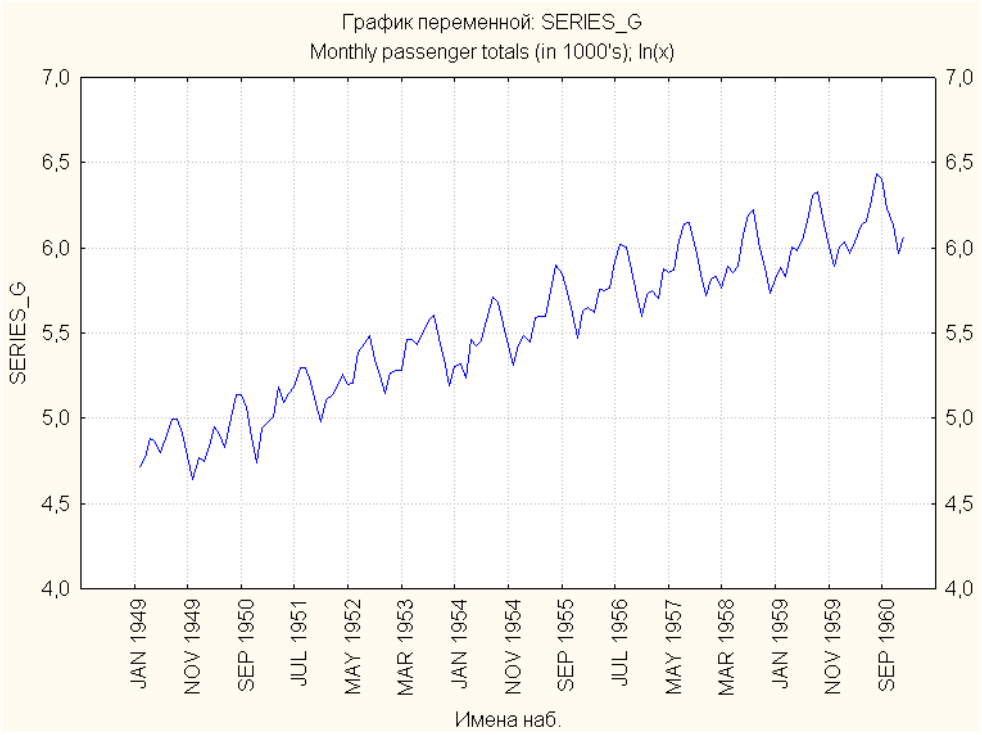


Рисунок 12.27 – График прологарифмированных значений ряда G

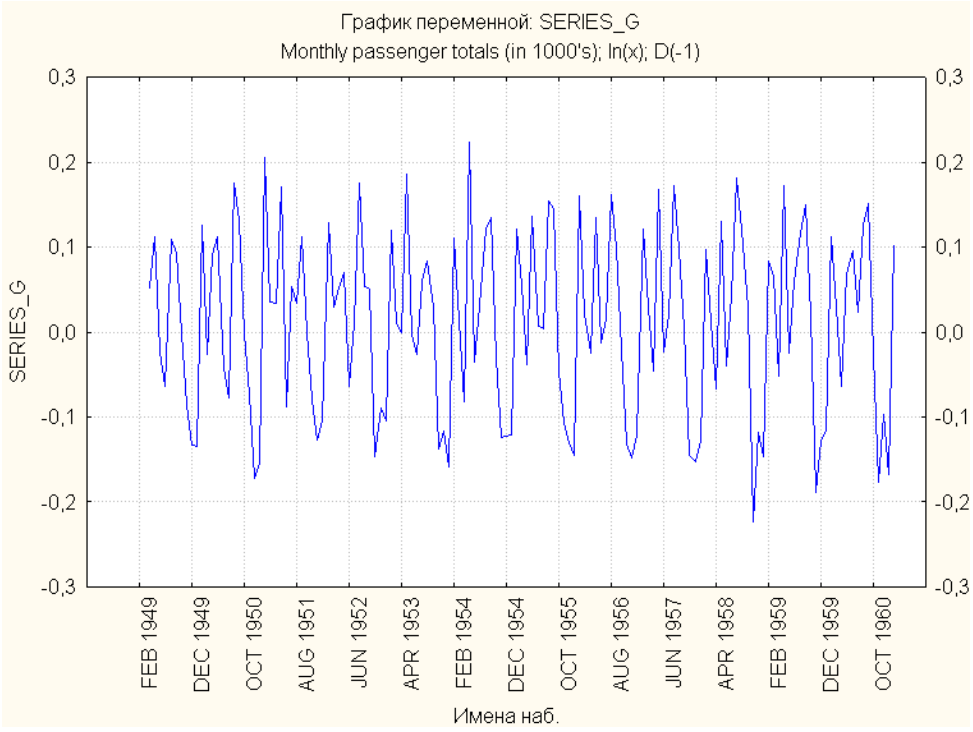


Рисунок 12.28 – График прологарифмированных значений ряда G после взятия разностей 1 порядка

Теперь ряд внешне похож на стационарный, для проверки этого предположения рассмотрим автокорреляции и частные автокорреляции с лагом 20 (рис. 12.29-12.30).

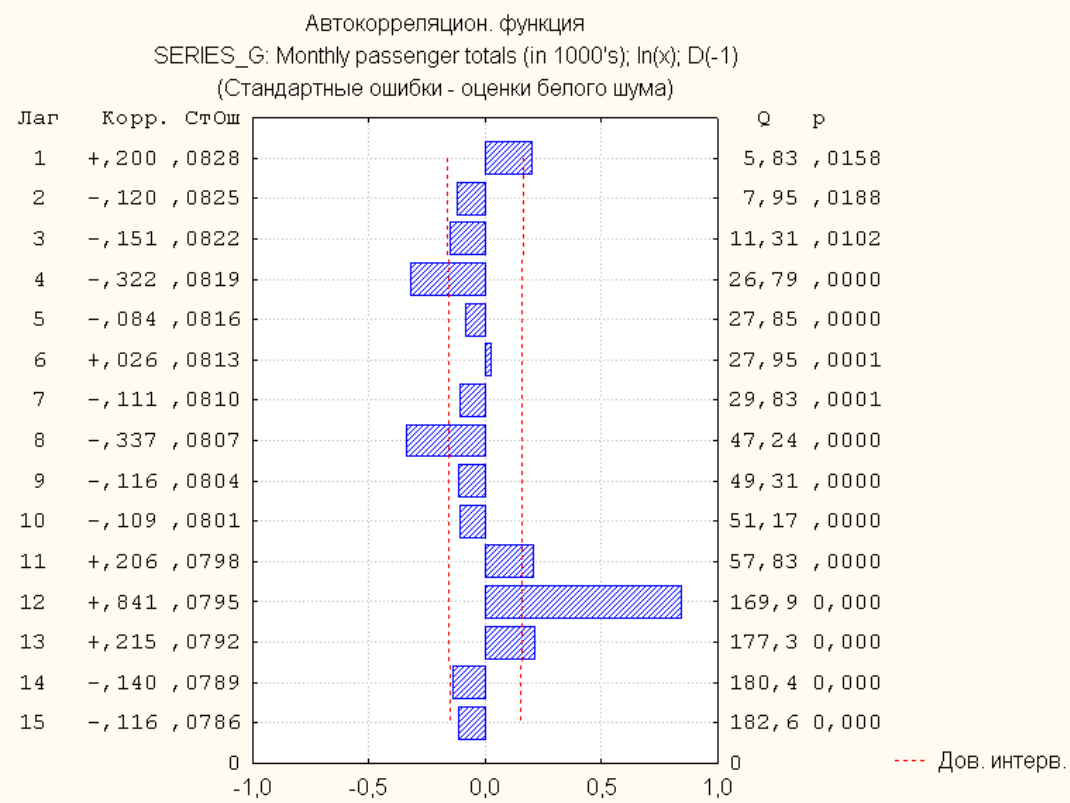


Рисунок 12.29 – АКФ ряда G после взятия первой разности

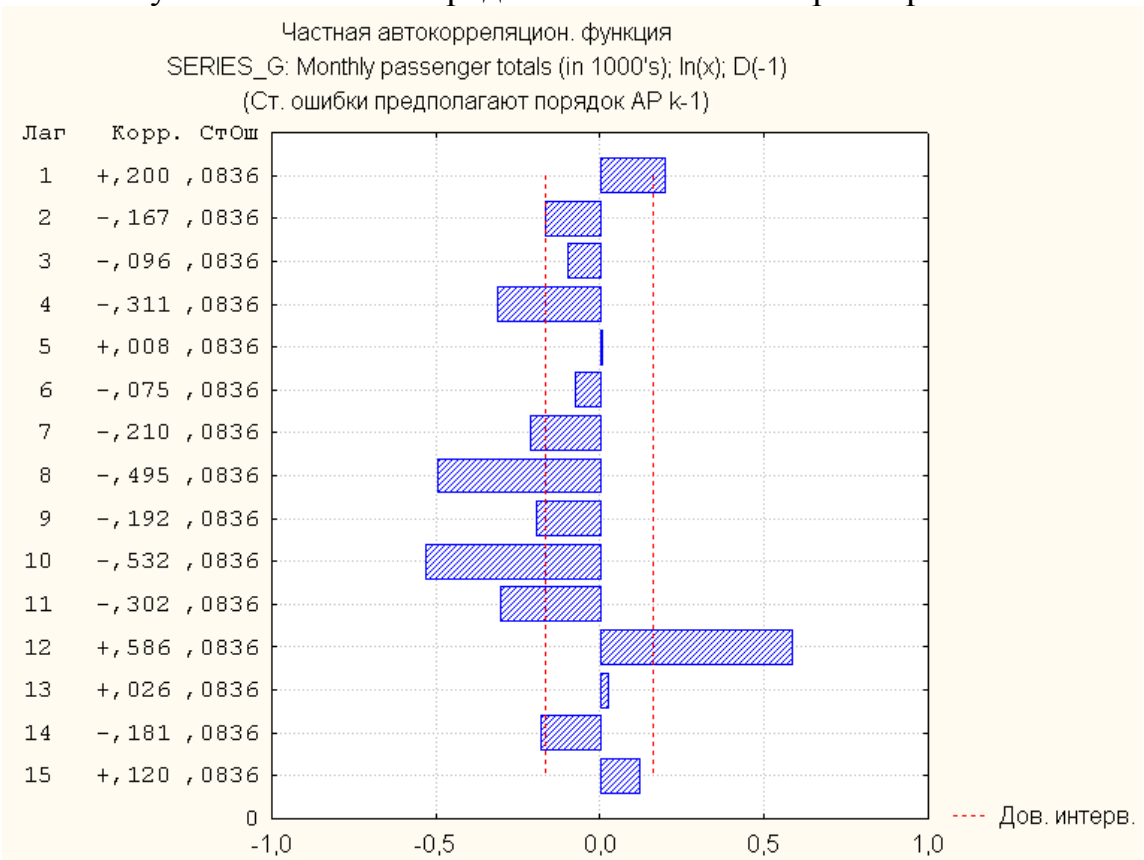


Рисунок 12.30 – ЧАКФ ряда *G* после взятия первой разности

Рисунки 12.29 и 12.30 показывают наличие сезонности в изучаемых данных с лагом 12 месяцев. Исключим сезонность, взятием разностей с лагом 12, в результате получим рисунки 12.31-12.32.

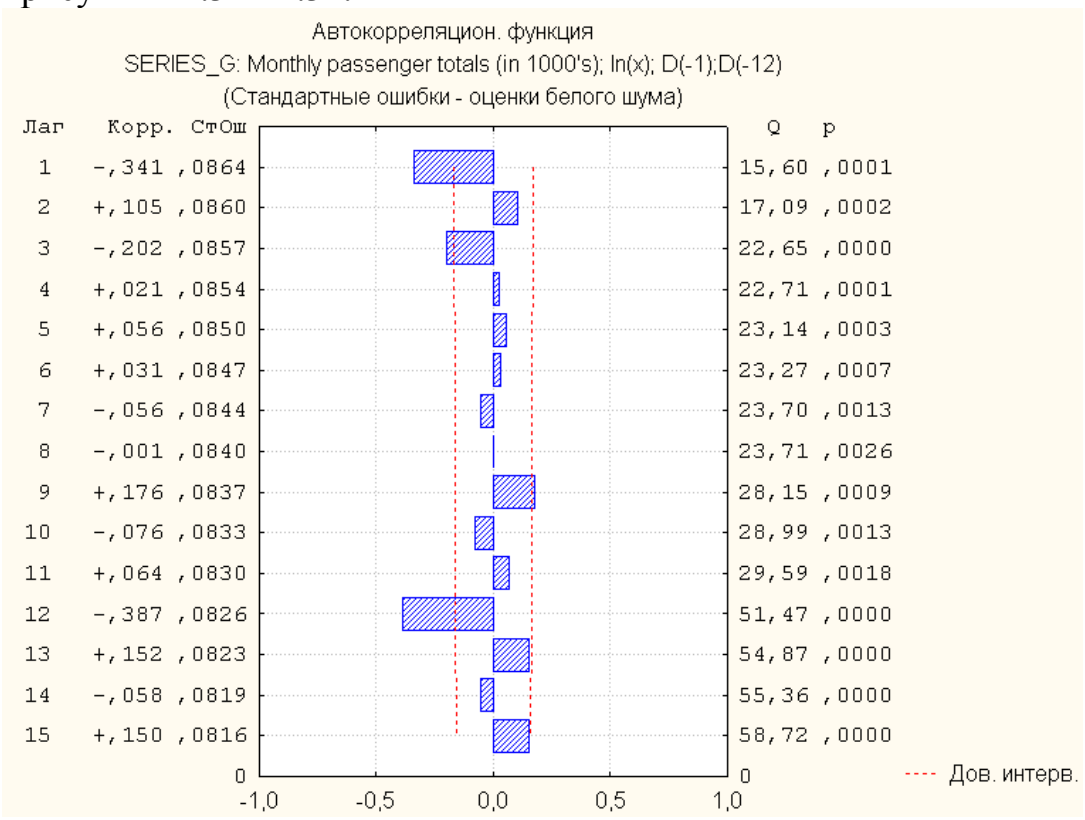


Рисунок 12.31– АКФ ряда *G* после взятия второй разности с лагом 12

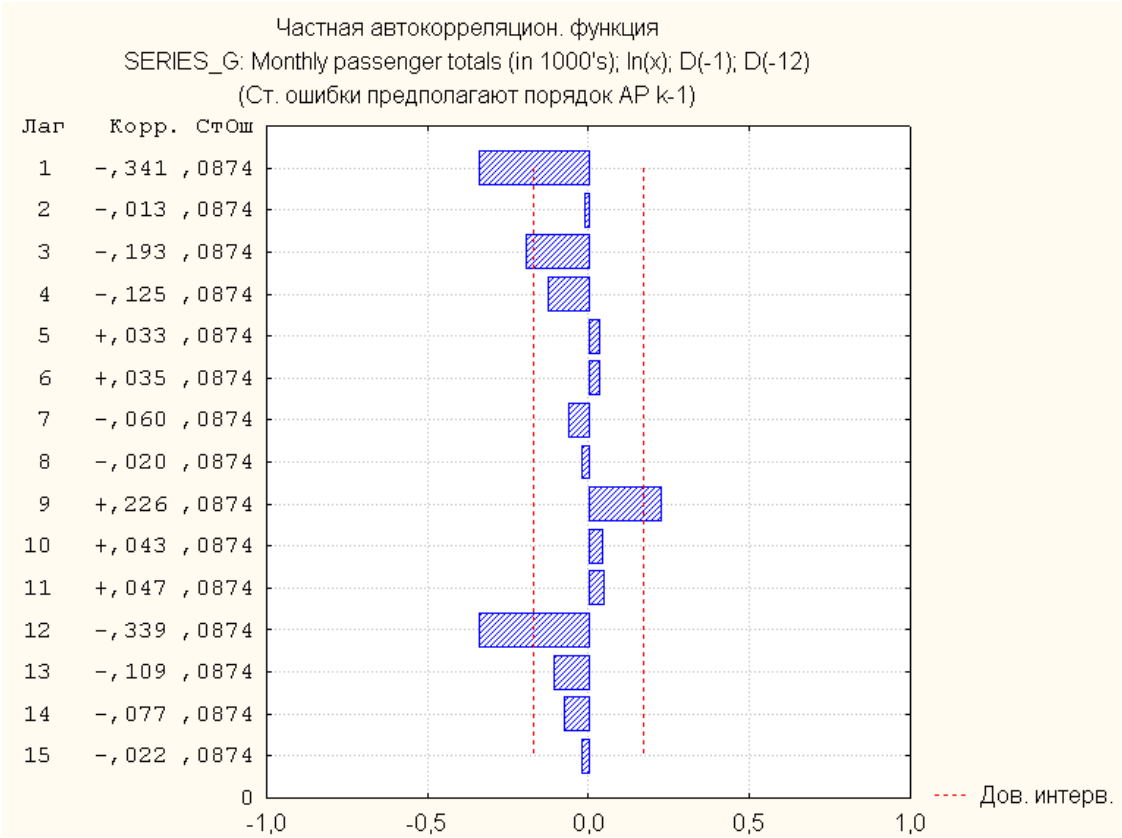


Рисунок 12.32 – ЧАКФ ряда *G* после взятия второй разности с лагом 12

Исходя из правил идентификации стационарных рядов (табл.12.1) мы видим (рис. 12.31-12.32), что автокорреляционная функция обрывается после задержки q=1, а её частная автокорреляционная функция спадает плавно – это соответствует процессу скользящего среднего порядка q=1. Автокорреляциям с задержкой 12 можно поставить в соответствие сезонный лаг Q=1. Итак, выбрав вкладку **АРПСС** (***ARIMA***) и заполнив её в соответствии с рисунком 12.33, мы получим окно результатов анализа (рис.12.34).

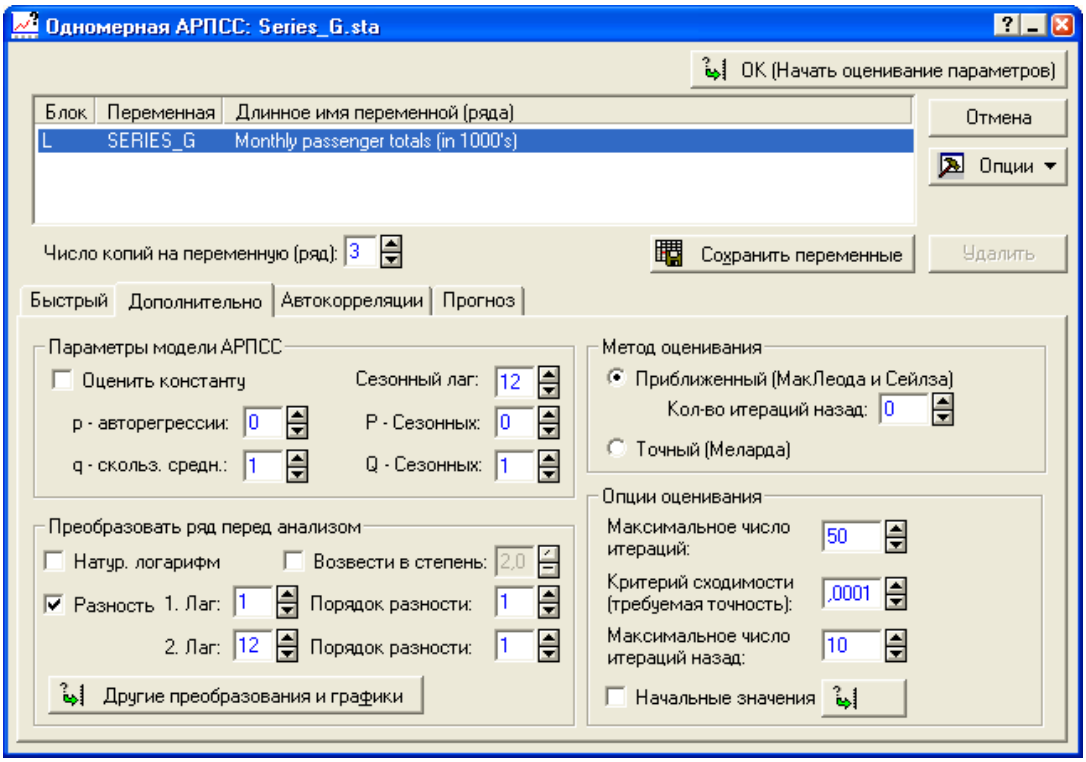


Рисунок 12.33 – Диалоговое окно модуля АРПСС (ARIMA)

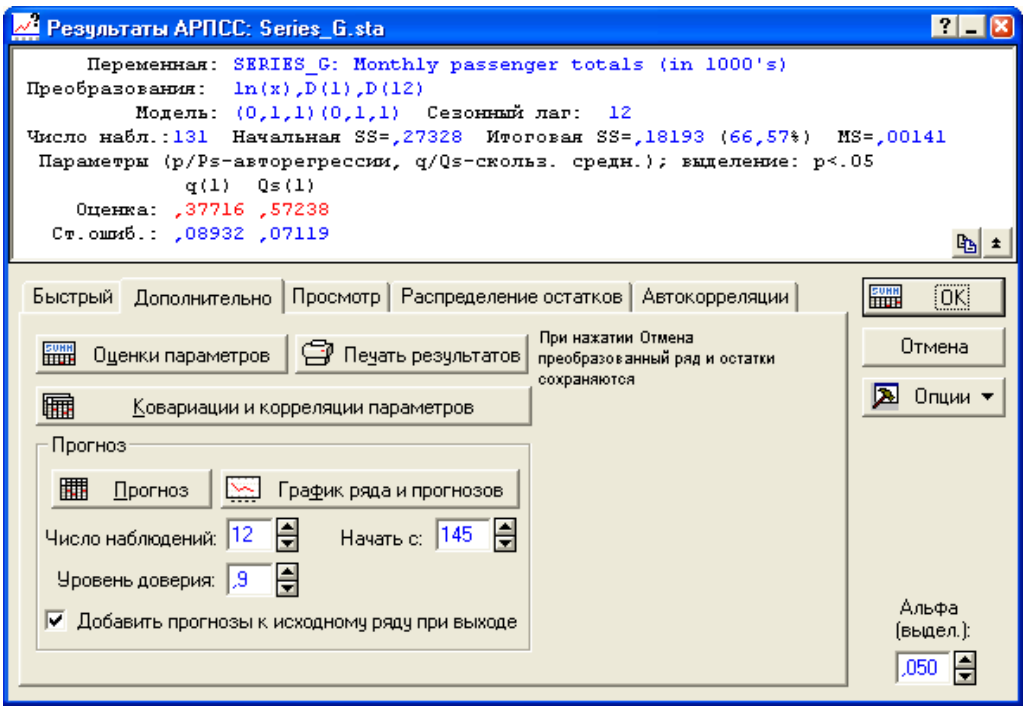


Рисунок 12.34 – Окно результатов АРПСС (*ARIMA*)

Параметр	Исход.:SERIES_G: Monthly passenger totals (in 1000's) (Series_G.sta) Преобразования: ln(x),D(1),D(12) Модель(0,1,1)(0,1,1) Сезонный лаг: 12 MS Остаток= ,00141					
	Парам.	Асимпт. Ст.ошиб.	Асимпт. t(129)	p	Нижняя 95% дов.	Верхняя 95% дов.
q(1)	0,377162	0,089318	4,222697	0,000045	0,200445	0,553880
Qs(1)	0,572379	0,071189	8,040233	0,000000	0,431529	0,713229

Рисунок 12.35 – Оценки параметров мультипликативной модели АРПСС (*ARIMA*) $(0,1,1)\times(0,1,1)_{12}$

Выбрав кнопку **Оценки параметров** (***Summary: Parameter estimates***), мы получим при уровне значимости 0,05 (или иначе с доверительной вероятностью 0,95) оценки параметров мультипликативной модели АРПСС (*ARIMA*) $(0,1,1)\times(0,1,1)_{12}$: b_1= 0,377162 с доверительным интервалом (0,200445; 0,553880); *Qs*(1)=0,572379 с доверительным интервалом (0,431529; 0,713229) и т.д. (рис.12.35).

Явно, полученную модель можно записать, например, в виде разностного уравнения:

$$z_t - z_{t-1} + z_{t-13} - z_{t-12} = \varepsilon_t - 0{,}377162\ \varepsilon_{t-1} - 0{,}572379\ \varepsilon_{t-12} + 0{,}377162\times 0{,}572379\ \varepsilon_{t-13}.$$

Или для прогноза на *k* шагов вперёд:

$$z_{t+k} = z_{t+k-1} - z_{t+k-13} + z_{t+k-12} + \varepsilon_{t+k} - 0{,}377162\ \varepsilon_{t+k-1} - 0{,}572379\ \varepsilon_{t+k-12} + 0{,}215879608\varepsilon_{t+k-13}.$$

Обычно для получения прогнозов: неизвестные значения z_t заменяются прогнозами, а неизвестные ε_t – нулями; известные ε_t – это уже вычисленные ошибки на шаг вперёд

$$\varepsilon_t = z_t - \hat{z}_{t-1}(1).$$

Выбрав последовательно вкладки **Распределение остатков – Гистограмма и Нормальный** (***Distribution of residuals – Histogram*** *and* ***Normal Probability Plot***) **и Автокорреляции** (***Autocorrelations***) – **Автокорреляции и Частные автокорреляции** мы получим рисунки 12.36 и 12.37.

Рисунок 12.36 иллюстрирует, что остатки удовлетворяют предположению о нормальности распределения. Автокорреляционная и частная автокорреляционная функции, изображённые на рисунке 12.37 показывают, что остатки не коррелированны.

Таким образом, все исходные предположения выполняются, и полученная модель может считаться адекватной.

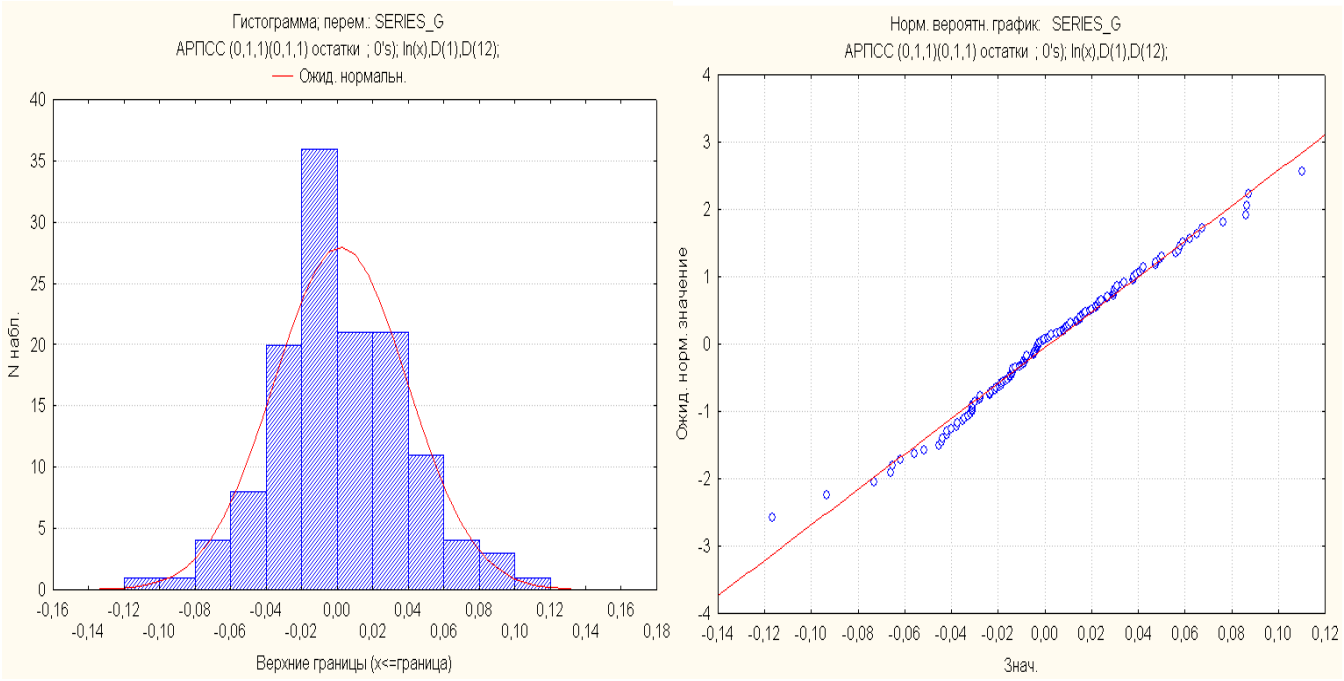


Рисунок 12.36 – Гистограмма и график вероятностной бумаги для остатков

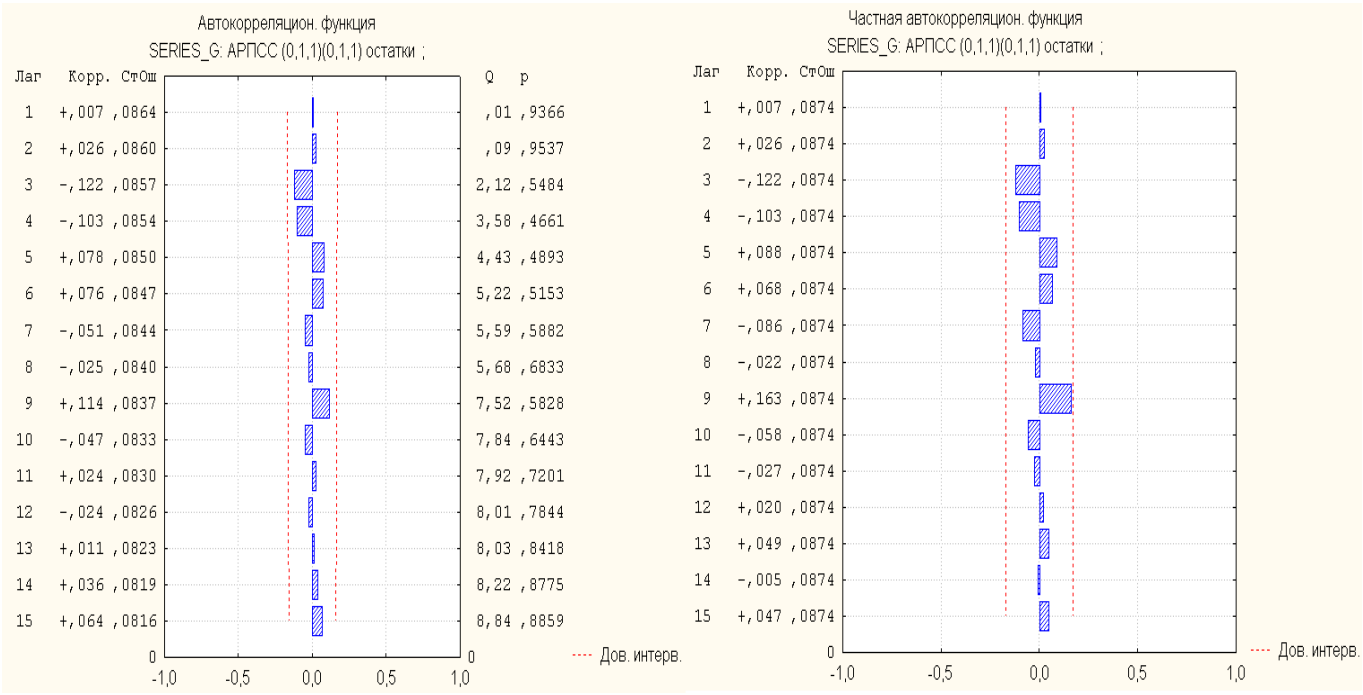


Рисунок 12.37 – АКФ и ЧАКФ для остатков ряда *G*

Выберем группу **Прогноз (*Forecasting*)** и установим: **Число наблюдений (*Number of cases*)** 36, **Начать с (*Start ad case*)** 133, **Уровень доверия (*Confidence level*)** 0,9, **Добавить прогнозы к исходному ряду при выходе (*Append forecasts to original series on Exit*)**.

В результате выбора кнопки **График ряда и прогнозы (*Plot series & forecasts*)** мы получим график прогноза для ряда *G* и соответствующие доверительные интервалы (рис.12.38).

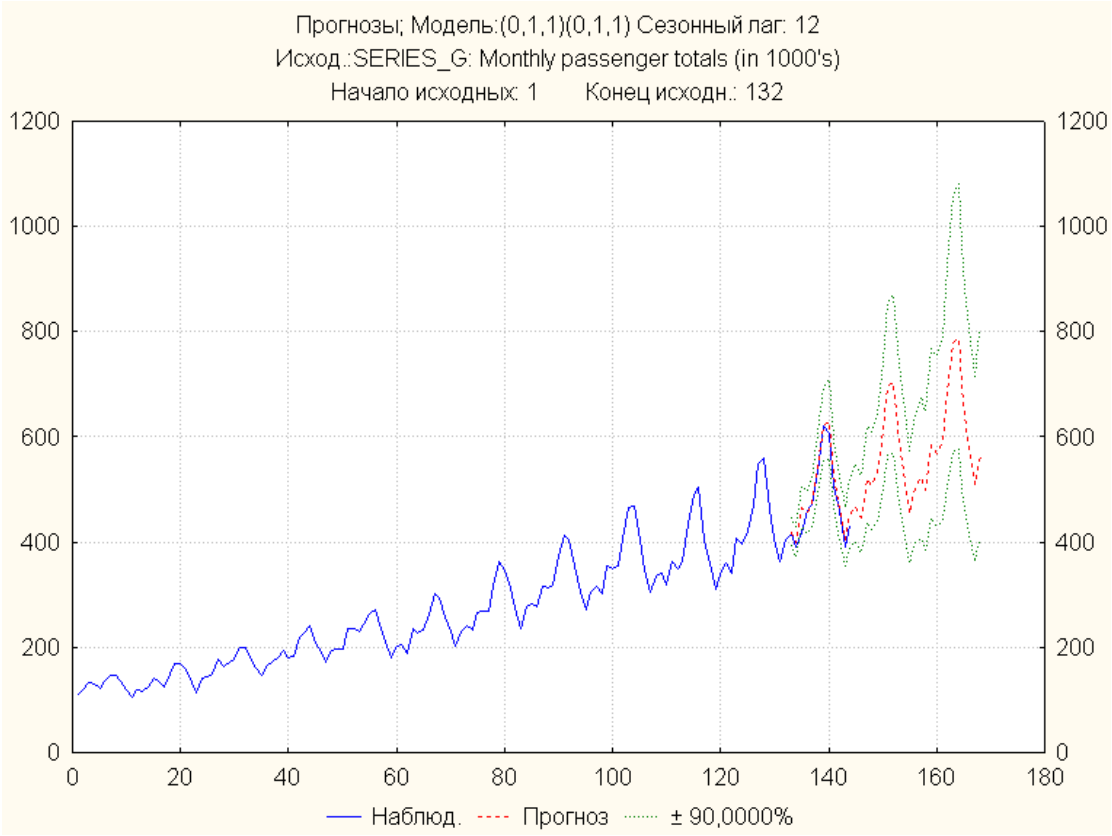


Рисунок 12.38 – График прогноза для ряда G

Рассмотрим применение к исходному ряду G процедуры экспоненциального сглаживания. Для этого в диалоговом окне Анализ временных рядов выберем кнопку **Экспоненциальное сглаживание и прогноз**(***Exponential Smoothing & forecasting***). Исходя из предположения, что ряд имеет мультипликативную сезонную компоненту с лагом 12 и мы желаем получить прогноз на 24 дня – заполним диалоговое окно в соответствии с рисунком 12.39.

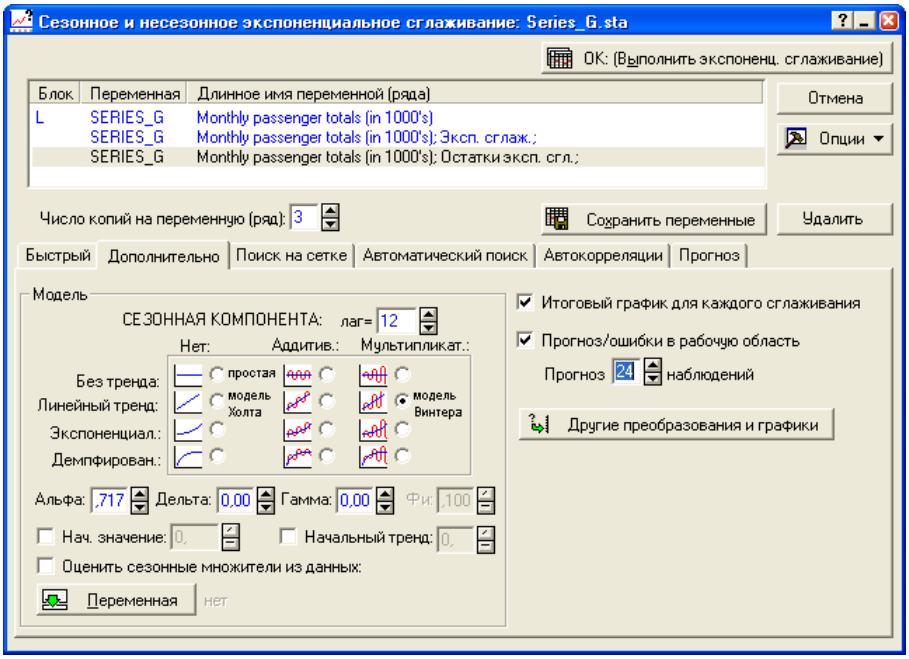


Рисунок 12.39 – Диалоговое окно экспоненциального сглаживания для ряда *G*

Для поиска параметров процедуры экспоненциального сглаживания воспользуемся вкладкой **Автоматический поиск** (***Automatic search – Automatic estimation***).

В результате получим сглаженный ряд и прогноз на 24 дня (рис.12.40), найденное значение Альфа (Alpha) модели равно 0,717.

Набл.	Эксп. сглажив.: Мульт. сезон. (12) S0=110,8 T0=2,648 (Series_G.sta) Лин.тренд,мульт.сезон.; Альфа= ,717 Дельт=0,00 Гамма=0,00 SERIES_G: Monthly passenger totals (in 1000's)			
	SERIES_G	Сглажен. ряд	Остатки	Сезонные составл.
1	112,0000	103,4301	8,5699	91,1856
2	118,0000	108,3489	9,6511	88,2217
3	132,0000	134,3813	-2,3813	100,8068
4	129,0000	130,6282	-1,6282	97,2951
5	121,0000	133,1656	-12,1656	98,1268
6	135,0000	144,1407	-9,1407	111,3347
7	148,0000	155,4299	-7,4299	123,1350
8	148,0000	151,2839	-3,2839	121,4657
9	136,0000	132,5109	3,4891	105,7905
10	119,0000	120,0690	-1,0690	92,1691
11	104,0000	106,0500	-2,0500	80,2883
12	118,0000	119,8533	-1,8533	90,1807
157		471,6331		
158		458,6386		
159		526,7340		
160		510,9609		
161		517,9269		
162		590,5879		
163		656,4444		
164		650,7613		
165		569,5809		
166		498,6833		
167		436,5276		
168		492,7004		

Рисунок 12.40 – Таблица итогов Экспоненциального сглаживания ряда *G* и прогноз на 24 дня

Изобразим, получившийся сглаженный ряд *G* и прогноз, а также исходный ряд на рисунке 12.41 (откроем вкладку ***Forecasting*** (**Прогноз**), во второй строке выберем график, отметим нужные ряды и щёлкнем по кнопке ОК).

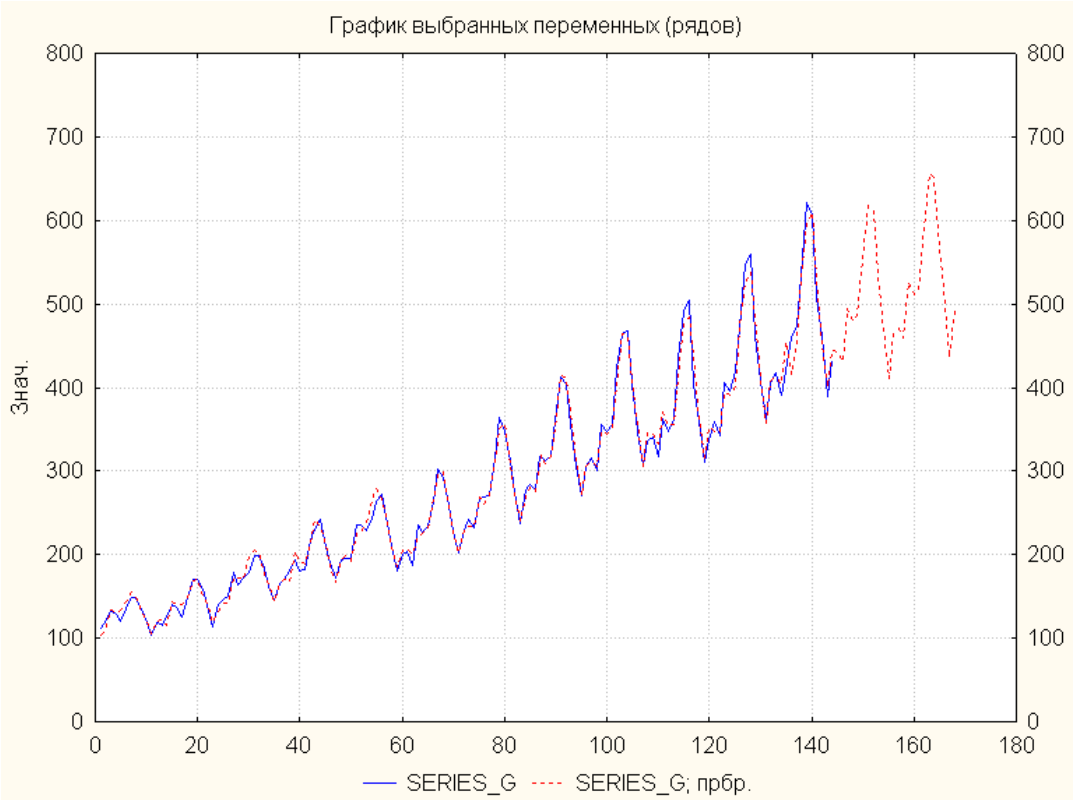


Рисунок 12.41– График ряда G, сглаженного ряда G и прогноз на 24 дня

Даже визуально видно, что экспоненциальное сглаживание, также как и рассмотренная выше модель АРПСС, даёт хороший результат и по воспроизведению ряда G, и по прогнозу. Недостатком модели считается отсутствие вероятностной интерпретации модели – нет возможности построения доверительных интервалов для прогноза.

Модель экспоненциального сглаживания считается простой моделью – более широкий класс процессов описывается в виде АРПСС (ARIMA). Следует отметить, что модель АРПСС позволяет описывать как стационарные процессы, так и достаточно широкий класс нестационарных процессов. А именно процессов со стационарными приращениями порядка d (обычно d =0,1,2), причём d-ая разность является стационарным процессом с определёнными свойствами [Бокс]. Кроме того, имеется возможность рассмотрения «сезонностей». Однако, при изучении процессов, происходящих в природе и обществе вероятностные модели не всегда применимы. Хотя без сомнения их необходимо рассматривать в качестве одной из альтернатив. Рассмотрим в качестве примера временной ряд месячных продаж вина в Австралии: январь 1980 года – июнь 1994.

Пример 3. Загрузим исходные данные из файла (продажи вина. sta). В качестве анализируемой переменной выберем общую величину продаж (Total). График переменной Total изображён на рисунке 180. мы видим, что ряд имеет сезонности и слабо выраженные тенденции – роста продаж примерно до июля 1987г, затем спада до января 1990г. Эти тенденции можно заметить по амплитудам, хотя в среднее значение продаж практически не изменяется. Рассмотрение автокорреляционной и частной автокорреляционной функций к данному ряду указывает на наличие сезонности с лагом 12 месяцев. После удаления сезонности путём взятия соответ-

ствующей разности – АКФ и ЧАКФ показывают отсутствие значимых автокорреляций, следовательно, модель не идентифицируется как модель АРСС (согласно таблице 12.1).

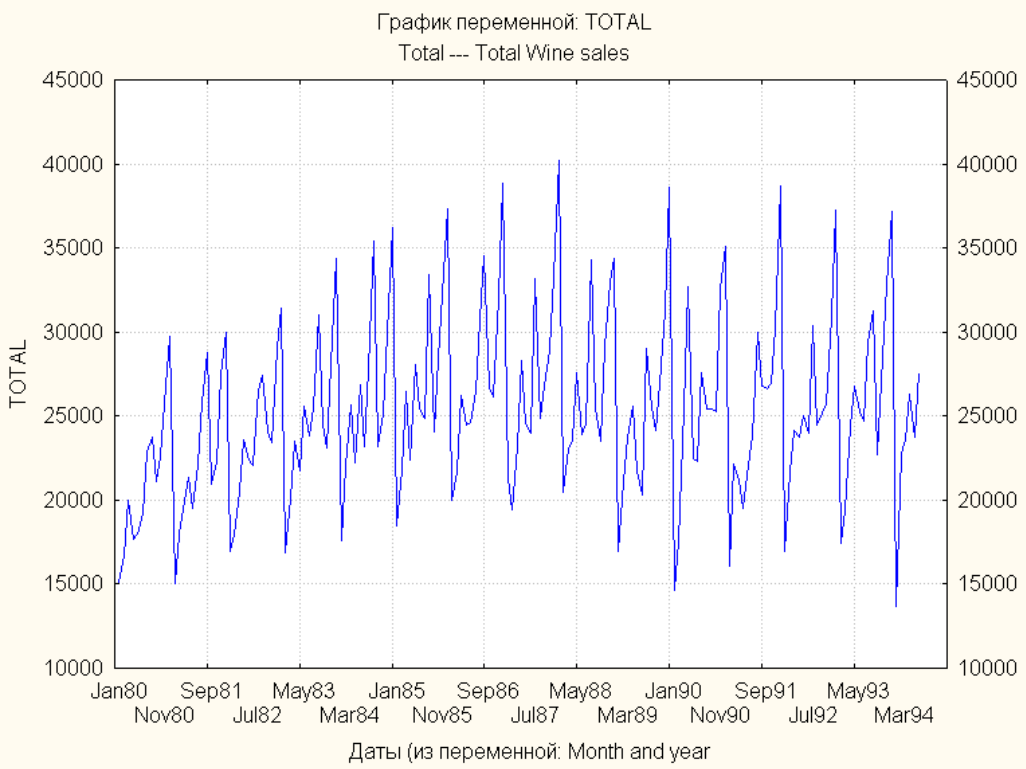


Рисунок 12.42 – График общих ежемесячных продаж вина в Австралии (Jan 1980– Jun 1994)

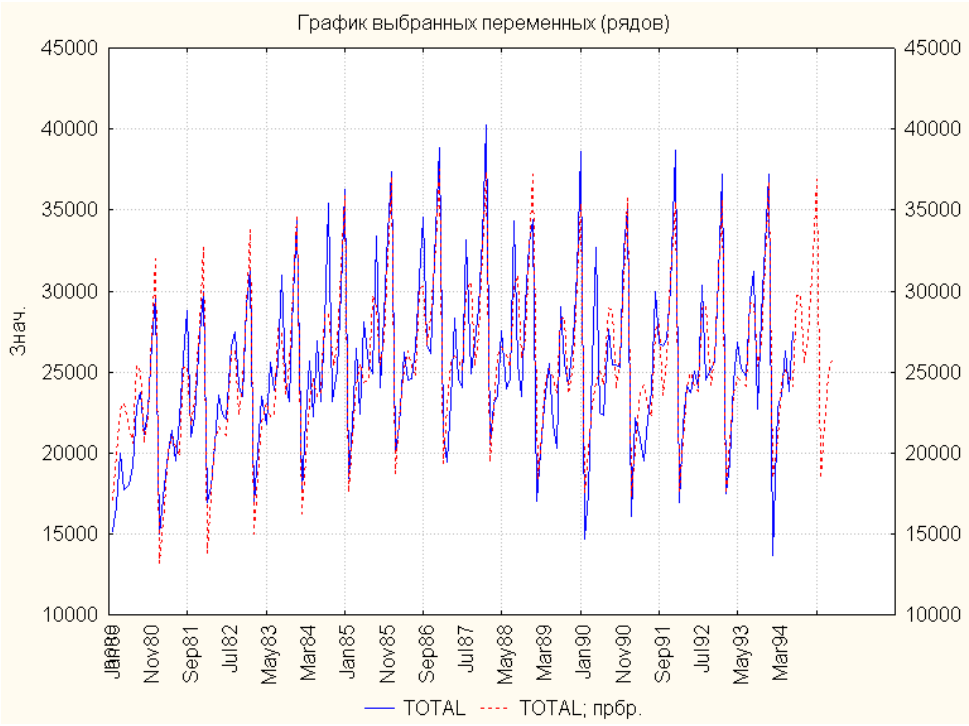


Рисунок 12.43 – График общих ежемесячных продаж вина в Австралии (Jan 1980– Jun 1994) и соответствующий ряд, сглаженный процедурой экспоненциального сглаживания

Экспоненциальное сглаживание даёт приемлемый результат и по воспроизведению ряда, и по прогнозу. (Открыв вкладку **Дополнительно (*Advanced*)** модуля экспоненциального сглаживания, выберем: сезонную компоненту с лагом 12 (***Seasonal component: lag* 12)**, аддитивную модель без тренда (***Notrend – Additive***). С использованием процедуры автоматического поиска (см. пример 2), найдём значение параметра Альфа равное 0,163. Построим график ряда *Total* и соответствующую сглаженную модель, изображённые на рисунке 12.43.)

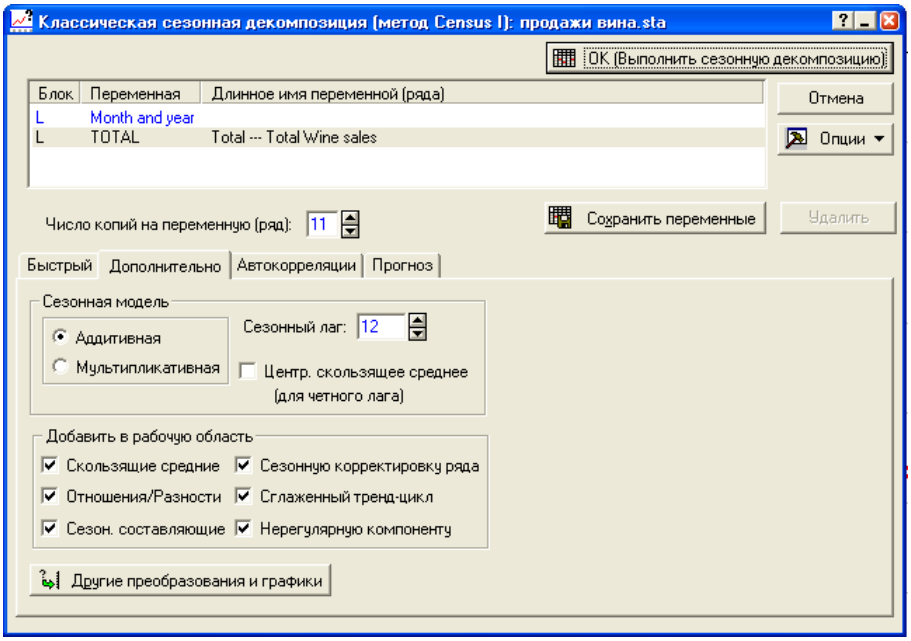


Рисунок 12.44 – Диалоговое окно модуля Сезонная декомпозиция (Census_1)

Однако при более подробном рассмотрении мы заметим, что остатки колеблются от нескольких сотен до нескольких тысяч, поэтому вряд ли эту модель стоит использовать для прогноза (хотя общую тенденцию она описывает).

Выделим все предполагаемые составляющие ряда с использованием модуля Сезонная декомпозиция (Census1). Заполним диалоговое окно в соответствии с рисунком 12.44.

Будет рассматривать **Аддитивную (*Additive*)** сезонную модель с лагом (***Seasonal lag***) 12. На первом шаге вычисляется скользящая средняя с шириной окна равной сезонному периоду (при чётном лаге для того, чтобы первое и последнее наблюдение в окне имели неравные веса необходимо выбрать параметр **Цент. скольз. среднее – *Centred moving averages* (*vor even Seasonal lag only*)**). В группе **Добавить в рабочую область (*On OK append components to active work area*)** можно выбрать для аддитивной модели компоненты, которые будут добавлены в активную рабочую область (предварительно согласовав **Число копий на ряд – *Number of backups per variable*(*series*)**):

– **Скользящее среднее (*Moving averages*)** (скользящие средние с шириной окна равной сезонному периоду);
– **Разности (*Ratios/Differences*)** (из наблюдаемого ряда вычитаются значения скользящих средних);

– **Сезон. составляющие (*Seasonal factors*)**(среднее всех значений соответствующих сезону);
– **Сезонная корректировка ряда (*Seasonally adj. Series*)** (разность между значениями исходного ряда и сезонной составляющей);
– **Сглаженный тренд-цикл (*Smoothed trend cucle*)** (выявление путём сглаживания циклов больших, чем сезонность);
– **Нерегулярная компонента** или погрешность **(*Irregular component*)** (разность между рядом с сезонной корректировкой и рядом с тренд-циклической компонентой).

После выбора ***OK*** мы получим таблицу со всеми выделенными компонентами.

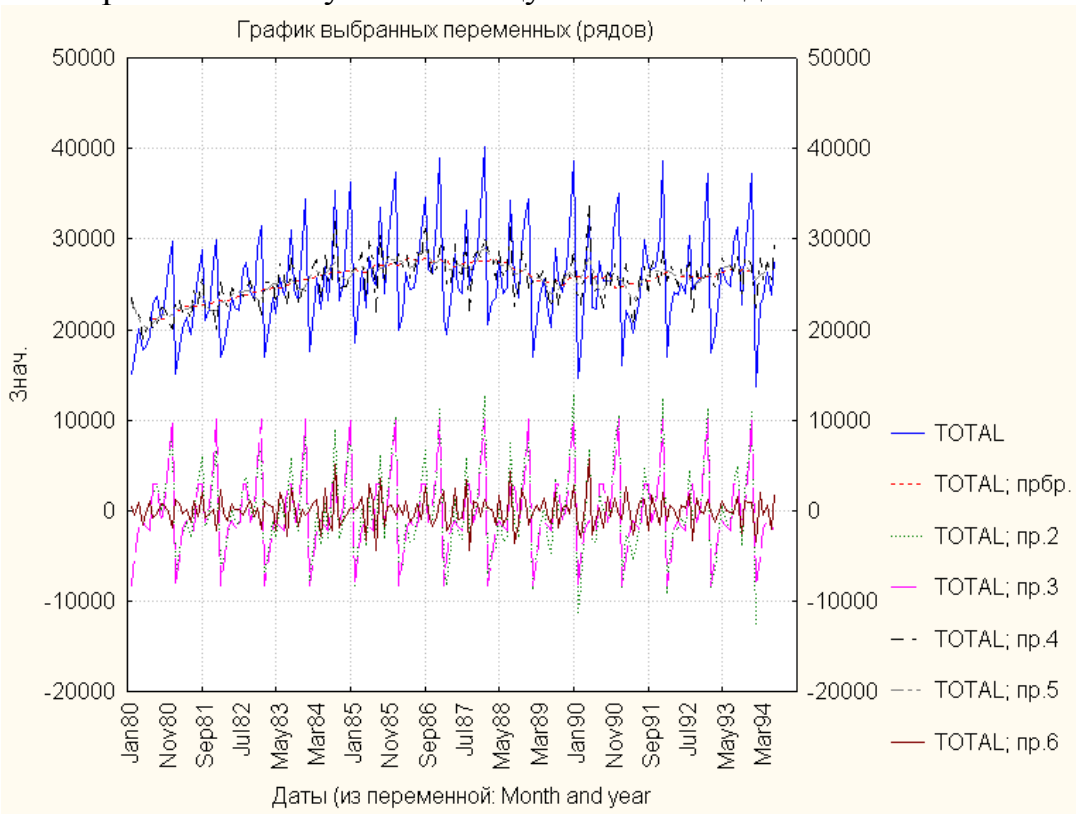


Рисунок 12.45 – Ряд Total и все выделенные компоненты его аддитивной модели

Используя вкладку Прогноз (*Review series*) – График (*Plot*) можно получить как отдельные графики компонент, так и все графики вместе (рис. 12.45). Порядок описания компонент на графике соответствует приведённому выше описанию компонент.

Получившиеся в активном окне компоненты ряда можно изучать по отдельности и использовать для построения прогноза.

Следует отметить, что для выявления сезонности и периодичности, да и вообще для анализа и прогноза явных периодических зависимостей, лучше использовать спектральный анализ. Для выявления сезонностей и циклов переменной Total откроем в диалоговом окне анализа временных рядов **Фурье (спектральный) анализ (*Spectral (Fourier) analysis*)**. Выполним команду: Одномерный анализ Фурье – Просмотр и графики – График – Период – Периодограмма (***Single series Fourier analysis – Review & plot – Plotby Period – Periodogram***). В результате полу-

чим рисунок периодограммы (рис.12.46). Используя кнопку – увеличить, выделим прямоугольник с тремя максимальными значениями и получим рисунок 184. Мы видим, что исходный ряд имеет 4-месячную сезонность, менее выраженную 6-месячную и ещё менее выраженную 12-месячную.Таким образом, визуальный и содержательный анализ, предполагающий годовую сезонность, нас подвёл – на самом деле сильнее выражены 4-месячная и 6-месячная сезонности.

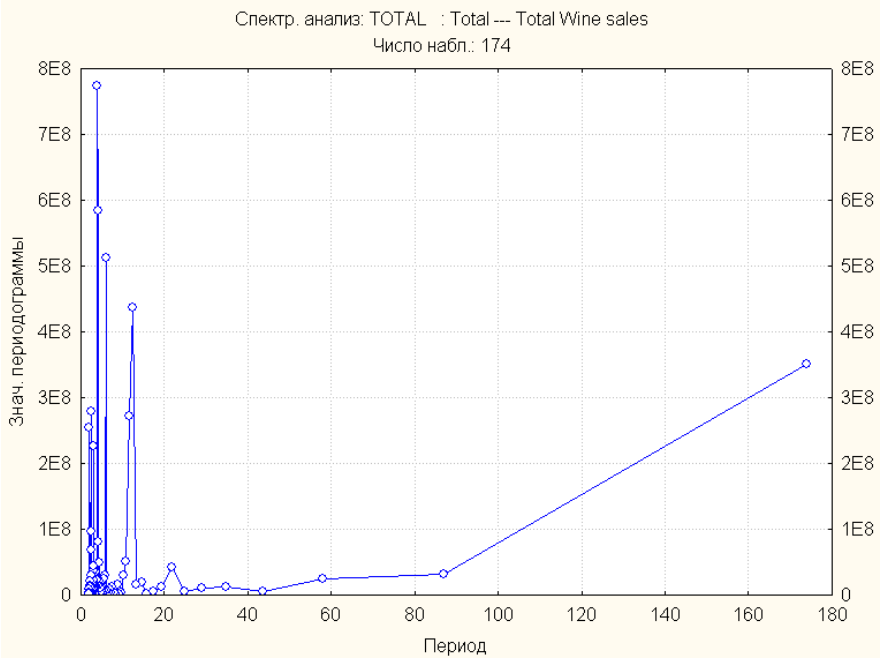


Рисунок 12.46 – Периодограмма для ряда Total в стандартном виде

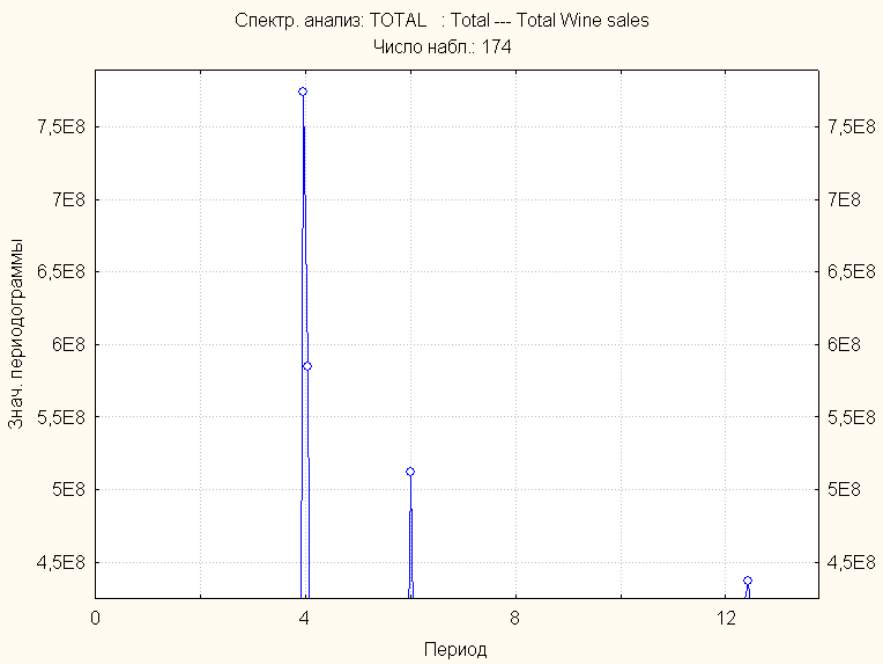


Рисунок 12.47 – Периодограмма для ряда Total в увеличенном виде

Однако, если ряд сгладить с помощью 4 – месячной скользящей средней (устраняющей 4-х месячную сезонность), то периодограмма будет принимать наибольшее значение для 12 месяцев (сделайте это самостоятельно, см. рис.

12.48). Таким образом, мы можем считать, что действительная периодичность ряда Total – 4 месяца, а 12 (точнее 12,1429 – см. рис.12.49) месяцев – это цикличность в данных. Графическое изображение (рис.12.50) в модуле *Census*_1 показывает, что найденная нами закономерность верна.

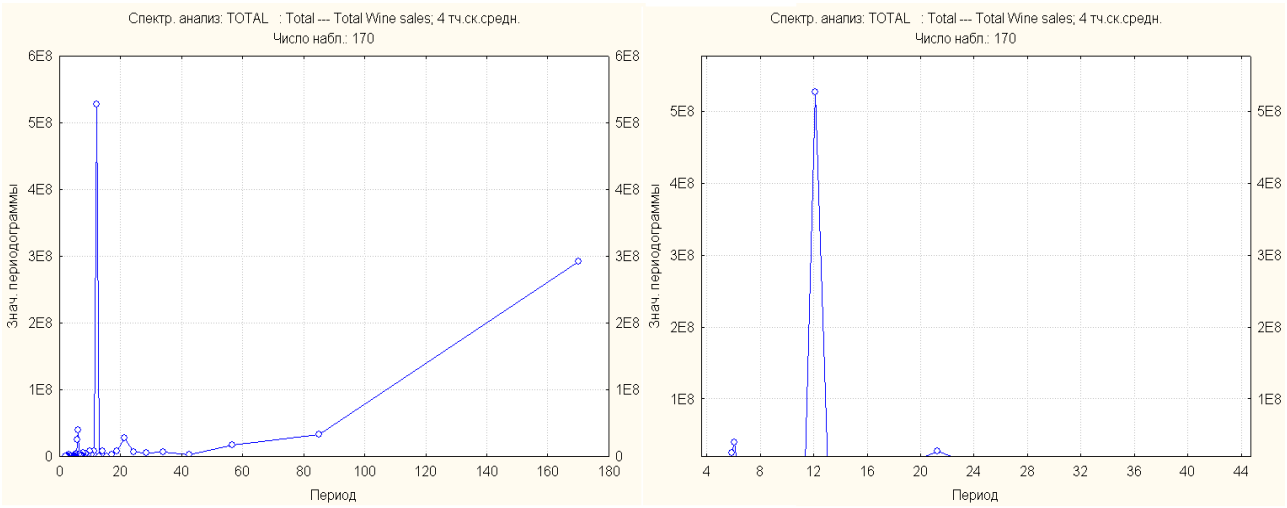


Рисунок 12.48 – Периодограмма для ряда Total после сглаживания 4-х месячной скользящей среднейв стандартном и увеличенном виде

Спектр. анализ: TOTAL : Total --- Total Wine sales; 4 тч.ск (продажи вина.sta)
Число набл.: 170

	Частота	**Период**	Косинус коэфф.	Синус коэфф.	Периодограмма	Плотн.	Хемминг веса
0	0,000000		0,00	0,000	0	143092878	0,035714
1	0,005882	170,0000	-1730,07	663,530	291839059	149385687	0,241071
2	0,011765	85,0000	144,62	-609,810	33386645	89582942	0,446429
3	0,017647	56,6667	-339,08	-303,692	17612089	27124946	0,241071
4	0,023529	42,5000	-111,34	-115,824	2194070	8400437	0,035714
5	0,029412	34,0000	-200,21	-215,060	7338479	6097446	
6	0,035294	28,3333	-265,00	-12,932	5983372	6983619	
7	0,041176	24,2857	139,16	-231,152	6187690	11315059	
8	0,047059	21,2500	-483,81	-294,203	27253474	15865588	
9	0,052941	18,8889	-302,56	9,209	7788347	11053804	
10	0,058824	17,0000	-186,97	56,872	3246424	4594723	
11	0,064706	15,4545	-25,93	19,711	90174	2961710	
12	0,070588	14,1667	-287,54	-84,769	7638446	22517534	
13	0,076471	13,0769	-73,50	30,734	539479	129552488	
14	0,082353	12,1429	-2427,34	560,406	527513808	238021358	
15	0,088235	11,3333	305,17	72,121	8358302	131911145	
16	0,094118	10,6250	150,27	-111,537	2976959	24087572	
17	0,100000	10,0000	182,41	-238,942	7681135	4821575	
18	0,105882	9,4444	74,97	-107,498	1460027	2926989	
19	0,111765	8,9474	-56,45	-69,543	681952	1981997	
20	0,117647	8,5000	217,73	54,028	4277553	2452573	
21	0,123529	8,0952	75,14	-31,010	561710	2665589	
22	0,129412	7,7273	162,56	-191,026	5347892	3206905	
23	0,135294	7,3913	-115,10	99,133	1961427	2622681	
24	0,141176	7,0833	58,68	125,712	1635965	1705976	
25	0,147059	6,8000	97,23	-69,528	1214467	1235766	
26	0,152941	6,5385	-45,25	64,946	532596	2665364	

Рисунок 12.49 – Результаты спектрального анализа после 4-месячного сглаживания

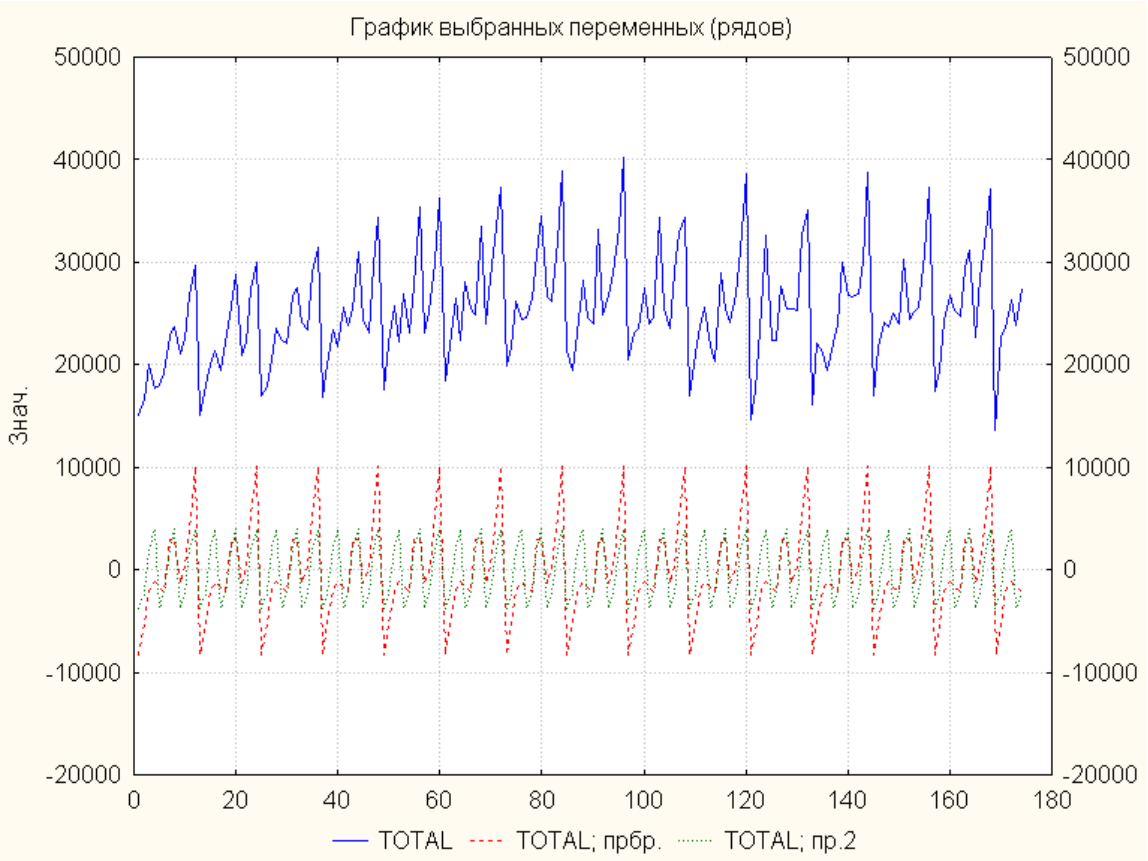


Рисунок 12.50 –Графики переменной Total, 4-х месячной сезонности и 12-и месячной цикличности

Задание

Рассмотрите примеры временных рядов продаж по вариантам – файл временные ряды.xls. Загрузите исходные данные (предварительно сформируйте даты в указанном диапазоне), проанализируйте на наличие трендов, сезонностей. Постройте несколько моделей для прогноза и протестируйте их на предварительно оставленных последних точках.

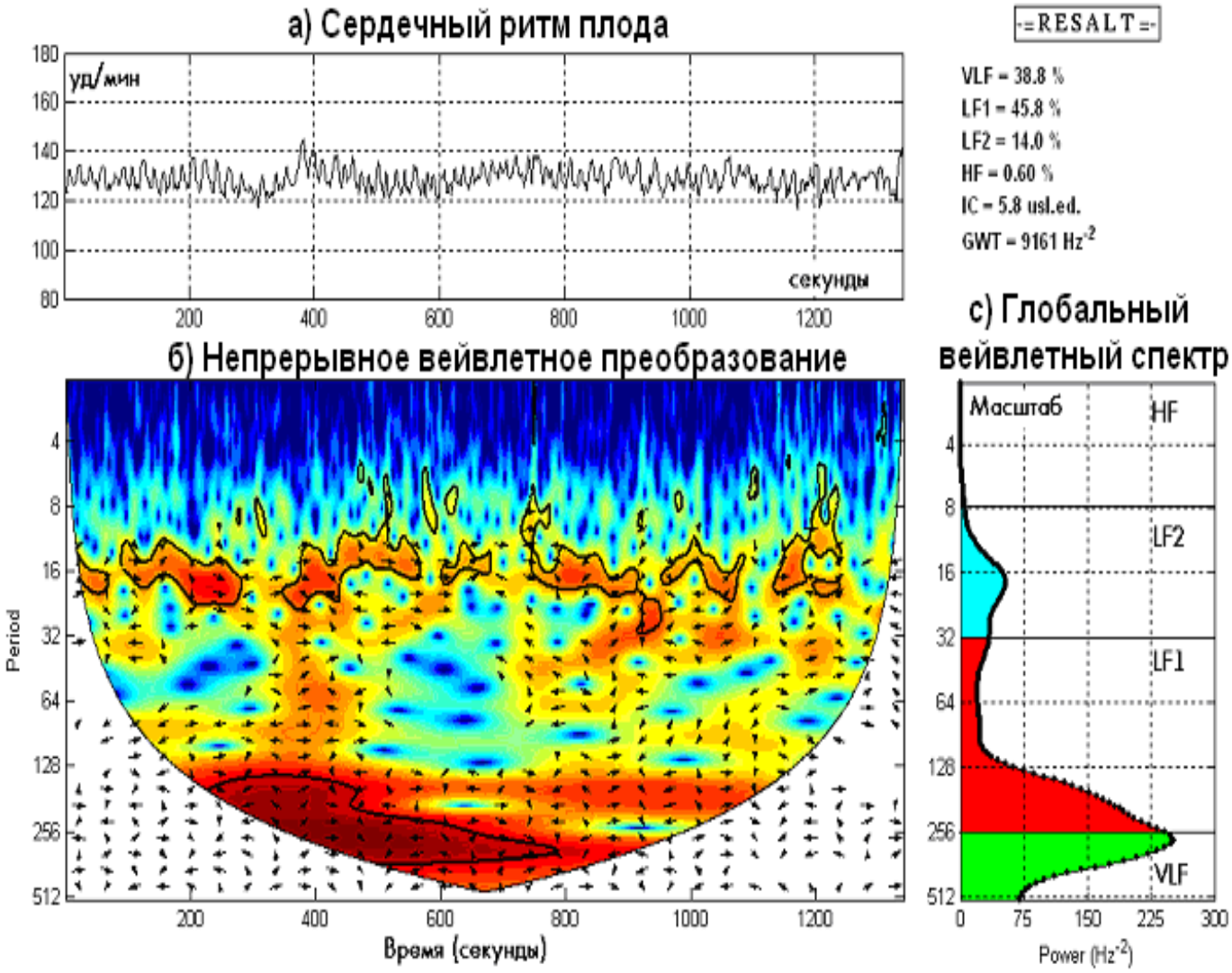
Вариант	Номера рядов		Вариант	Номера рядов		Вариант	Номера рядов	
1	1	3	**11**	17	17	**21**	20	27
2	2	5	**12**	4	18	**22**	24	28
3	6	7	**13**	20	19	**23**	25	29
4	9	8	**14**	24	20	**24**	6	30
5	13	10	**15**	25	21	**25**	9	31
6	17	11	**16**	13	22	**26**	17	32
7	4	12	**17**	2	23	**27**	1	17
8	20	14	**18**	6	24	**28**	13	21
9	24	15	**19**	9	25	**29**	4	15
10	25	16	**20**	1	26	**30**	2	16

Вопросы для самоконтроля

- Какие задачи решаются с использованием анализа временных рядов?
- Опишите основные подходы к построению моделей временных рядов.
- Что такое тренд и сезонность, цикличность?
- Какие способы выделения трендов и сезонностей существуют?
- Что такое автокорреляция?
- Какие временные ряды называют стационарными, а какие нестационарными?
- Какие модели используются при описании стационарных и нестационарных временных рядов?

При всем желании невозможно описать для вас мир таким, каким он вам нравится. Мы способны смотреть на мир лишь через призму сочетания экспериментальных результатов и новых теоретических представлений. Мы убеждены в том, что новая ситуация отражает в какой-то мере ситуацию в деятельности нашего головного мозга.

Илья Пригожин, Изабелла Стенгерс. Порядок из хаоса



Практическое занятие № 13

Оценка сложности временных рядов

Цель работы: ознакомиться с возможностями изучения нелинейных характеристик временных рядов.

Одной из важных величин, относящейся к хаотическим инвариантам, является *корреляционная размерность восстановленного аттрактора* (D_2). На сегодняшний день существует несколько продуктов как коммерческого, так и свободно распространяемого программного обеспечения, предназначенных для вычисления нелинейных характеристик временных рядов. Здесь предложен алгоритм вычисления D_2, для реализации которого применены процедуры, представленные в весьма распространённом проекте TISEAN[9], что делает его доступным широкому кругу пользователей.

Этот пакет представляет собой коллекцию библиотек нелинейных функций, написанных на C и Fortran, разработанных в Институте физики и теоретической химии университета города Франкфурта (авторы – R. Hegger, H. Kantz и T. Schreiber). Набор отдельных "исполняемых" файл-программ работает без графического пользовательского интерфейса (Graphic Users Interface – GUI) и управляется из командной строки после задания параметров вычисления каждой из исполняемых функций. Графическая визуализация содержимого отчетных файлов может быть выполнена в программе MATLAB®Release 7.0 (MathWorks®, США).

Особенность пакета TISEAN состоит в том, что в его состав входят программы для предварительной обработки данных, программы линейного и нелинейного анализа сигналов, а также многое другое. Как уже отмечалось, одним из наиболее распространенных методов исследования хаотических процессов является вычисление корреляционной размерности восстановленного аттрактора (псевдоаттрактора) – D_2. Набор программ, входящих в состав пакета TISEAN позволяет наиболее корректно и с наименьшими затратами времени производить определение данного показателя.

Процедуру вычисления корреляционной размерности D_2 можно разбить на три этапа: 1) определение корреляционного интеграла $C(\varepsilon)$ для различных значений размерности вложения D_{emb}; 2) поиск линейного участка (скейлинга) зависимости $\log C(\varepsilon)$ от $\log \varepsilon$ и определение его наклона для каждого значения D_{emb}; 3) определение оптимального значения D_{emb} и соответствующей ему величины D_2.

Продемонстрируем основные этапы определения корреляционной размерности D_2 хаотической системы на примере дискретного двумерного *отображения Хенона*[10]. Функция, реализующая генератор, имеет следующий вид:

---

[9] Hegger R. et al. Practical Implementation of Nonlinear Time Series Methods. In: The TISEAN package, CHAOS 9, 413, 1999. (http://www.mpipks-dresden.mpg.de/~tisean/TISEAN_2.1/index.html)

[10] Xenon M. A Two – Dimensional Mapping with a Strange Attractor // Commun. Math. Phys., 1976, vol. 50, p. 69 – 77.

$$\begin{cases} X_{t+1} = 1 - aX_t^2 + Y_t \\ Y_{t+1} = bX_t \end{cases}; a = 1{,}4; b = 0{,}3$$

где X_t, Y_t – предшествующие значения хаотической последовательности; X_{t+1}, Y_{t+1} – текущие значение; a и b – параметры отображения.

1. Для моделирования дискретного двумерного отображения Хенона предназначена файл-функция **henon.exe** из пакета TISEAN. Основными переменными, определяющими результат ее вычисления являются: параметры a ('*–A#*') и b ('*–B#*') (по умолчанию равные 1,4 и 0,3, соответственно), число генерируемых отсчетов ('*–l#*'), указатель вывода результата в файл ('*–o#*'). Пример обработки файл-функции **henon.exe** в командной строке (с параметрами – $a = 1{,}4$ и $b = 0{,}3$, число генерируемых отсчетов – 2000) удобно выполнить из интерфейса программы Matlab предварительно указав путь к данной файл-функции:

```
tiseanPath = 'C:\Programme\MATLAB\work\Tisean\';
system([tiseanPath,'henon -B0.3 -A1.4 –l2000 -o']);
x = load('henon.dat');
plot(x(:,1),x(:,2),'.');
```

Результат выводится в выходной файл `henon.dat`, в котором содержится два столбца значений переменных X и Y отображения, соответственно. При начальных условиях: $X_0 = 0$, $Y_0 = 0$ временной ряд переменной Xдля первых 300 шагов показан на рисунке 13.1.

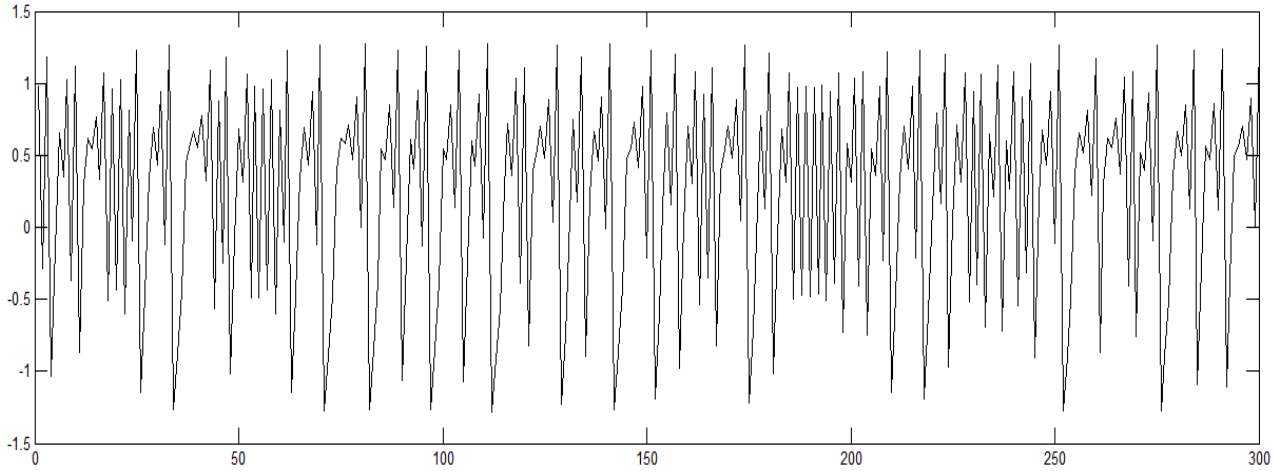


Рисунок 13.1 – Исходный временной ряд – X-координата дискретного отображения Хенона

Временной ряд одной из переменных отображения Хенона демонстрируют явную хаотичность своих колебаний. Фазовый портрет в двумерном пространстве переменных (X,Y) определенно не является случайным – это аттрактор (рис. 13.2). Временные ряды переменных зависят от начальных условий, а фазовый портрет всегда выглядит одинаково. Система притягивается к этой форме, которая и является ее странным аттрактором. Это и есть явление детерминированного хаоса.

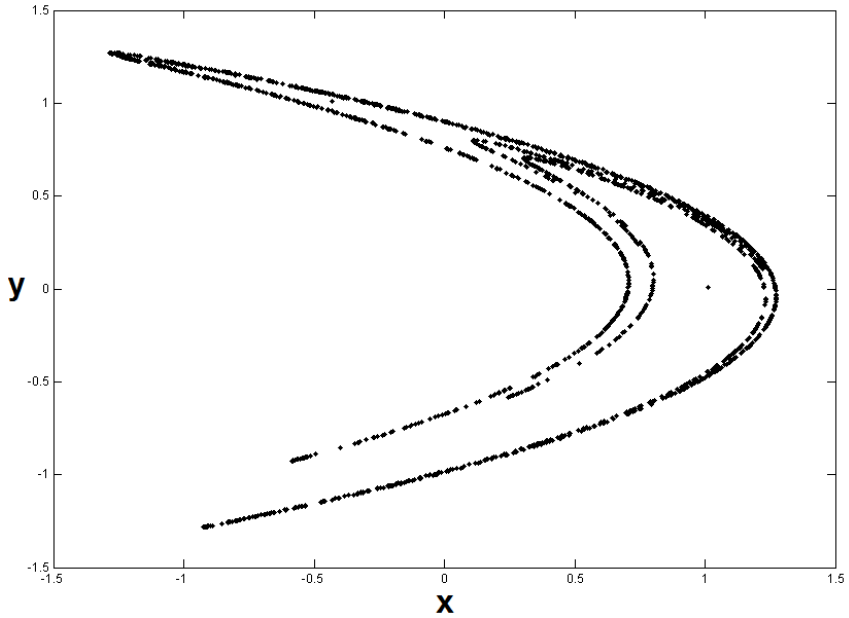


Рисунок 13.2 – Аттрактор отображения Хенона с параметрами a = 1,4 и b = 0,3

Для вычисление корреляционного интеграла $C(\varepsilon)$ прежде всего необходимо задать лаговые и другие переменные: размерность лагового пространства – D_{emb}, время задержки – τ, окно Тейлера (*Theiler window*) – T_w, относительное число ложных ближайших соседей в восстановленном аттракторе (*False Nearest Neighbors*).

2. По первому локальному минимуму функции совместной информации $S(\tau)$ (программа – **mutual.exe**) определяем времени задержки – τ. С помощью файл-функции **mutual.exe** вычисляем функцию совместной информации (или, как ещё иногда её называют, функцию взаимной информации). Эту функцию можно рассматривать как альтернативу автокорреляционной функции $C(\tau)$, но, в отличие от последней здесь учитываются не только линейные, но и нелинейные связи между x_i и $x_{i+\tau}$.

Под совместной информацией подразумевают следующее. Разобьём числовой отрезок, равный размаху амплитуды сигнала, на несколько интервалов и обозначим через p_i вероятность, с которой элемент временного ряда может оказаться в i-м интервале, а через p_j – в j-м. Пусть p_{ij} – совместная вероятность того, что один элемент временного ряда окажется в i-м интервале, а другой, взятый с задержкой τ – в j-м. Тогда функция совместной информации будет выглядеть следующим образом:

$$S(\tau) = -\sum_{i,j} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j}$$

Вариант записи командной строки соответствующей файл-функции из интерфейса программы Matlab выглядит следующим образом:

```
tiseanPath = 'C:\MATLAB\work\myFunction\TISEAN\';
modelPath = 'C:\MATLAB\work\henon.dat';
system([tiseanPath, 'mutual modelPath -D50 -o']);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `modelPath` – путь к месту расположения файла с исходными данными отображения Хенона (`henon.dat`); *'–D#'* – максимальное значение времени задержки (50); *'–o#'* – указатель вывода результатов в файл (по умолчанию – `henon.dat.mut`).

Остальные аргументы данной файл-функции не указаны и соответствуют значениям по умолчанию (см. документацию к программе TISEAN). Вид графика функции совместной информации $C(\tau)$ для рассмотренного примера показан на рисунке 13.3.

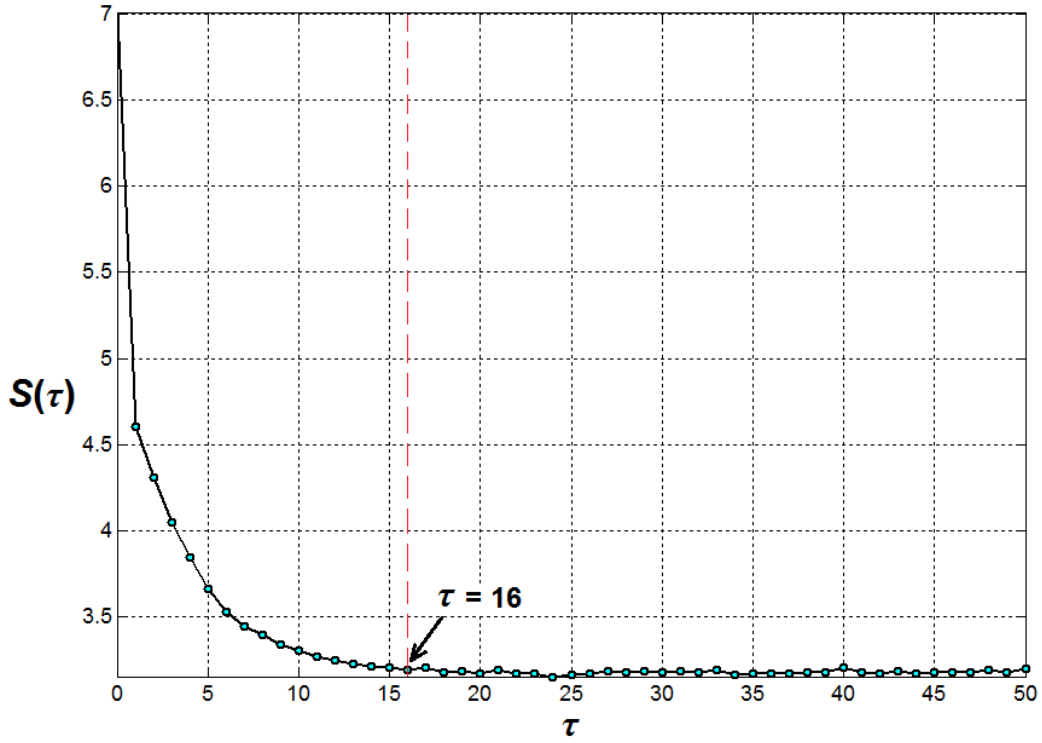


Рисунок 13.3 – График функции совместной информации $S(\tau)$. Стрелкой указан первый локальный минимум ($\tau = 16$)

3. Для вычисления первого локального минимума функции совместной информации предназначена файл-функция **`extrema.exe`**. Вариант запуска данной файл-функции из интерфейса программы Matlab:

```
mutPath = 'C:\Programme\MATLAB\work\henon.dat.mut';
system([tiseanPath, 'extrema mutPath -m2 -w2 -zminima -o']);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `mutPath` – путь к файлу данных взаимной информационной функции (`henon.dat.mut`), *'–m#'* – число столбцов в файле исходных данных (2); *'–w#'* – номер столбца, для которого вычисляются локальные минимумы (2); *'–z#'* – указатель определяющий выбор вычисления локальных минимумов (`minima`) или максимумов (`maxima`); *'–o#'* – указатель вывода результатов в файл (по умолчанию – `henon.dat.mut.ext`). В выходном файле в первых двух столбцах указаны позиции локальных минимумов (третий столбец – расстояние между ними).

В конкретном примере для отображения Хенона первому локальному минимуму функции совместной информации $S(\tau)$ соответствует 16-й отсчет, таким образом, время задержки – τ = 16. Данное значение является величиной лага при вычислении корреляционного интеграла, относительного числа ложных ближайших соседей в восстановленном аттракторе, а также для построения семейства графиков пространственно-временного разделения при определении окна Тейлера.

4. Выбор оптимальной размерности лагового пространства – m. На модельных временных рядах это нетрудно – она известна из справочной информации. При работе с реальными зашумленными временными рядами насыщения зависимости $D_2(m)$ не наступает.

Самое простое, что можно сделать в этом случае – задаться размерностью пространства вложения заведомо большей, чем предполагаемая размерность восстановленного аттрактора.

Опыт исследований показывает, что в зависимости от применяемых методов вычислений и условий эксперимента D_2 принимает значения от 2 до 8. При соответствующем выборе временного сдвига τ и оптимальной размерности пространства вложения m оригинальный и реконструированный аттракторы будут топологически эквивалентны (теорема Такенса о вложении).

5. Конечность длины временного ряда приводит к тому, что при вычислении корреляционного интеграла точки аттрактора, расположенные в непосредственной близости друг от друга, оказываются, как правило, скоррелированными. Это приводит к ошибкам в вычислении корреляционного интеграла $C(\varepsilon)$.

Для того, чтобы избежать этих ошибок, можно в процессе вычислений не принимать в расчет точки, расположенные в исходной последовательности на расстоянии меньше, чем w шагов. Величина T_w называется *окном Тейлера.* Минимальная величина этого окна: $T_w < \tau(2/N)^{2/D_{emb}}$.

Способом определения оптимального размера окна Тейлера является построение семейства *графиков пространственно-временного разделения* (программа – **`stp.exe`**).

Этот график представляет собой кривую равной плотности вероятности того, что две точки временного ряда, находящиеся на расстоянии Δt окажутся в восстановленном аттракторе на расстоянии, не превышающем ε. Таким образом, для того, чтобы построить этот график, необходимо восстанавливать динамику системы в лаговом пространстве.

Следовательно, возникает необходимость задаться лаговыми параметрами – величиной лага τ (τ=16) и размерностью лагового пространства D_{emb} (выбираем фиксированное максимальное значение равное 10). При этом строится семейство кривых, соответствующих разным значениям плотности вероятности.

На рисунке 13.4 показано семейство графиков пространственно-временного разделения для плотностей вероятности, взятых с шагом 0,05. Размер окна Тейлера можно определить как величину δt, соответствующую точке первого локального максимума, общего для всех кривых

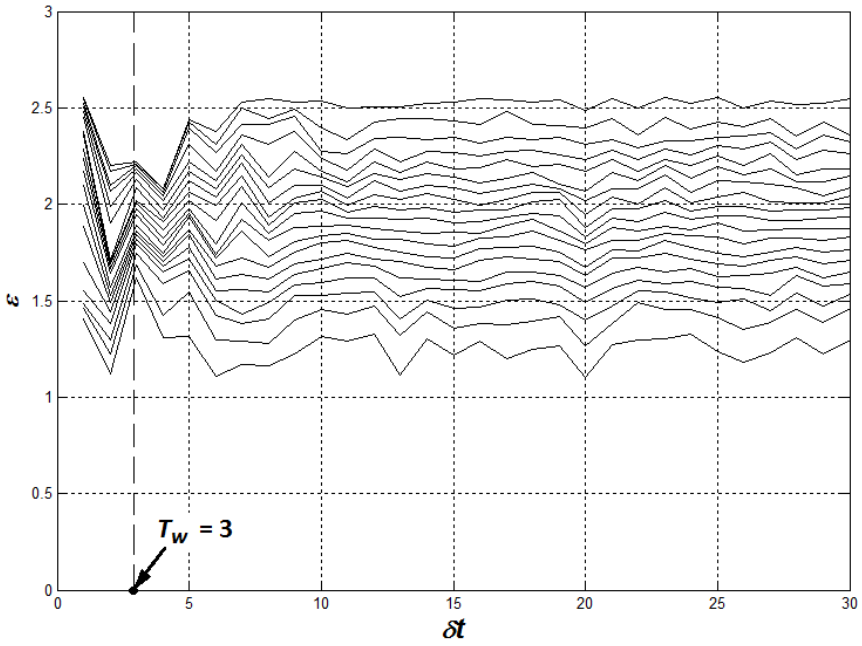


Рисунок 13.4 – Семейство графиков пространственно-временного разделения. Стрелкой указана позиция первого общего локального максимума ($T_w = 3$)

Вариант запуска файл-функции **`stp.exe`** из интерфейса программы Matlab с параметрами: время задержки – $\tau = 16$; максимальная размерность лагового пространства – $m = 10$; '*–o#*' – указатель вывода результатов в файл (`henon.dat_stp`) имеет следующий вид:

```
param_stp = strcat(' –d16', ' -m10, ' -o');
system([tiseanPath, 'stp ', modelPath, param_stp]);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `modelPath` – путь к месту расположения файла с исходными данными отображения Хенона (`henon.dat`); `param_stp`– строка параметров файл-функции. В выходном файле (`henon.dat_stp`) в первом столбце записана величина δt, а во втором – расстояние между точками в аттракторе ε.

6. В процедуре определения корреляционного интеграла для отображения Хенона оптимальное значение окна Тейлера соответствует координате первого локального максимума, общего для всего семейства графиков пространственно-временного разделения. Из рисунка 4.63 видно, что для рассматриваемого процесса этому условию отвечает величина δt, равная трем, т. е. $T_w = 3$.

7. Определение относительного количества ложных ближайших соседей при восстановлении аттрактора в лаговом пространстве выполняется при помощи файл-функции **`false_nearest.exe`**. В основе алгоритма лежит следующий принцип: по исходной временной последовательности ряда восстанавливается аттрактор в лаговом пространстве, после чего для каждой его точки осуществляется поиск её ближайших соседей. После этого аттрактор восстанавливается в лаговом пространстве, размерность которого на единицу больше и снова производится под-

счёт ложных ближайших соседей. Ложными ближайшими соседями называются такие пары точек, расстояние между которыми после приращения размерности лагового пространства значительно увеличилось. Точки аттрактора близкие при одной размерности сильно удаляются друг от друга при ее увеличении. Как и в случае вычисления $C(\varepsilon)$, при расчёте имеется такой параметр, как окно Тейлера.

Вариант запуска программы **`false_nearest.exe`**из интерфейса программы Matlab имеет следующий вид:

```
param_fnn = strcat(' –d16', ' -t3, ' -o');
system([tiseanPath, 'false_nearest ', modelPath, param_stp]);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `modelPath` – путь к месту расположения файла с исходными данными отображения Хенона (`henon.dat`); '*–d*16' – время задержки (лага, $\tau = 16$); '*–t*3' – значение окна Тейлера ($T_w = 3$); '*–o*#' – указатель вывода результатов в файл (`henon.dat.fnn`). В выходном файле результаты расчёта выводятся в виде четырёх столбцов: 1) размерность лагового пространства; 2) относительное количество ложных ближайших соседей; 3) средний размер окрестности; 4) среднее значение квадрата размера окрестности.

График зависимости относительного количества ложных ближайших соседей (*FNN*) от размерности лагового пространства (D_{emb}) показан на рисунке 13.5.

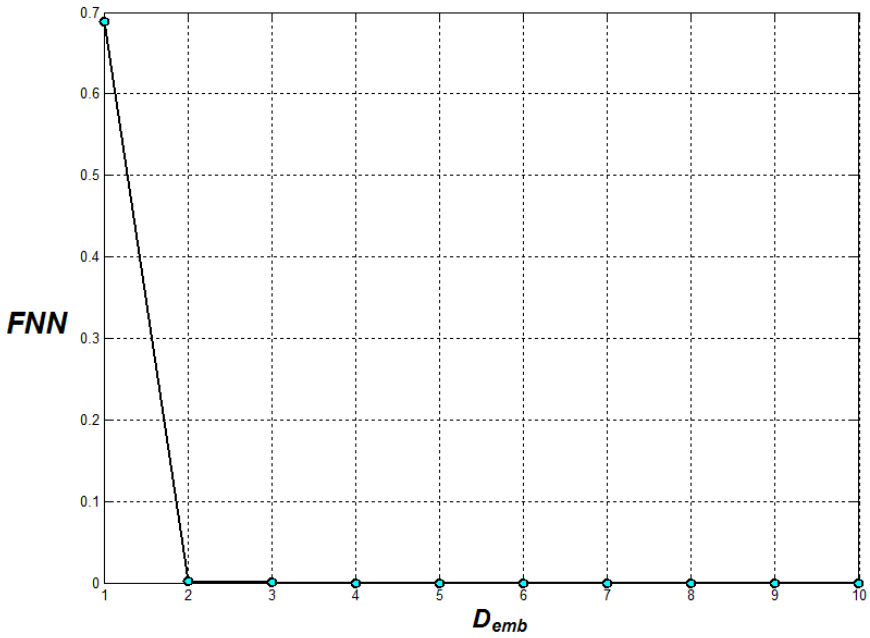


Рисунок 13.5 – Зависимость относительного количества ложных ближайших соседей в аттракторе (*FNN*) от размерности лагового пространства D_{emb}

После определения всех необходимых переменных (время задержки – τ, размерность лагового пространства – m, окно Тейлера – T_w и др.) запускается программа **`d2.exe`**. Эта программа при различных значениях размерности лагового пространства вычисляет *корреляционный интеграл*, локальные наклоны к кривым корреляционных интегралов и функцию корреляционной энтропии. Одна из харак-

терных особенностей этой программы заключается в том, что вычисления можно проводить в ускоренном режиме. В этом режиме при каждом значении ε для расчётов берётся только ограниченное число (как правило – 1000) пар точек, а не все возможные. Если ε настолько мало, что 1000 пар точек не набирается, что вычисления производятся для всех пар.

Вариант запуска программы **`d2.exe`** из интерфейса программы Matlab имеет следующий вид:

```
system([tiseanPath, 'd2 ', modelPath, ' –d16', ' -t3',' -o']);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `modelPath` – путь к месту расположения файла с исходными данными отображения Хенона (`henon.dat`); '*–d16*' – время задержки (лага, $\tau = 16$); '*–t3*' – значение окна Тейлера ($T_w = 3$); '*–o#*' – указатель вывода результатов в файл.

Программа формирует четыре файла с расширениями *.*c2**, *.*d2*, *.*h2* и *.*stat*. В первых трёх файлах содержатся для каждого значения D_{emb} две колонки; в одной из них записаны величины ε, а в другой – значения вычисляемой величины. Файл с расширением *.*c2* содержит величины корреляционных сумм, вычисленные для каждого значения ε – $C(\varepsilon)$; файл с расширением *.*d2* – значения локальных наклонов корреляционных сумм, построенных в двойном логарифмическом масштабе; файл с расширением *.*h2* – величины корреляционной энтропии; файл с расширением *.*stat* – сообщение о текущем состоянии расчётов.

На рисунке 13.6 показаны корреляционный интеграл, вычисленный для размерностей лагового пространства от 1 до 10 (графическое представление данных из файла с расширением *.*c2*).

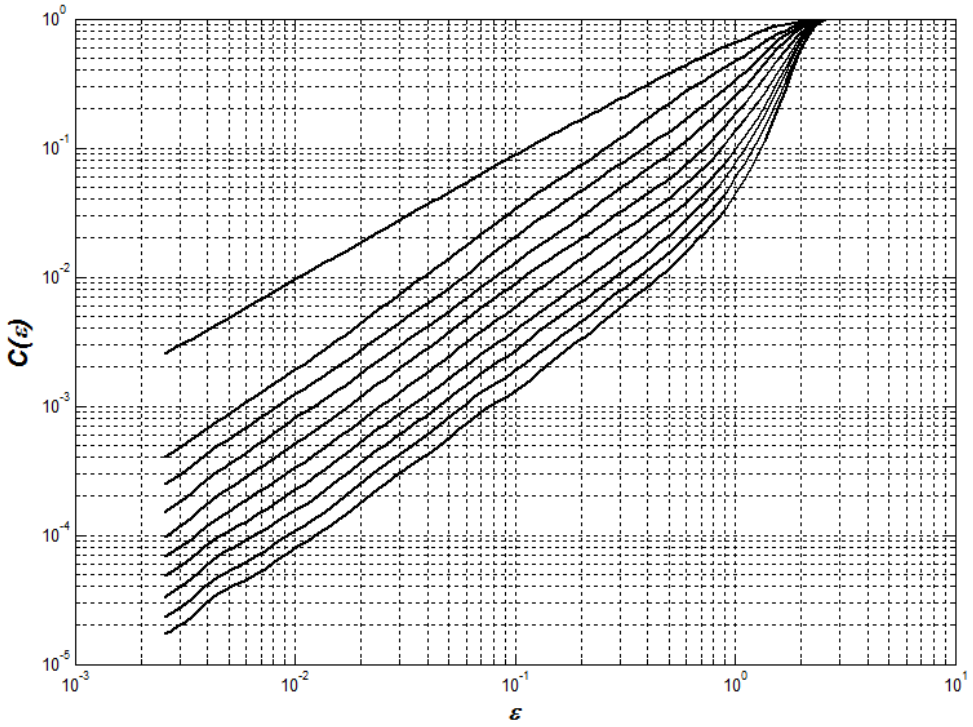


Рисунок 13.6 – Корреляционный интеграл, вычисленный для размерностей лагового пространства от 1 до 10

В принципе, зависимость величины D_2 от размерности вложения m можно получить непосредственно из содержания файла с расширением $*.d2$, где указаны значения локальных наклонов к графикам корреляционного интеграла. Интервал изменения ε, на котором значения вышеупомянутых наклонов формируют плато демонстрируют *графики Рапа* (рис. 13.7).

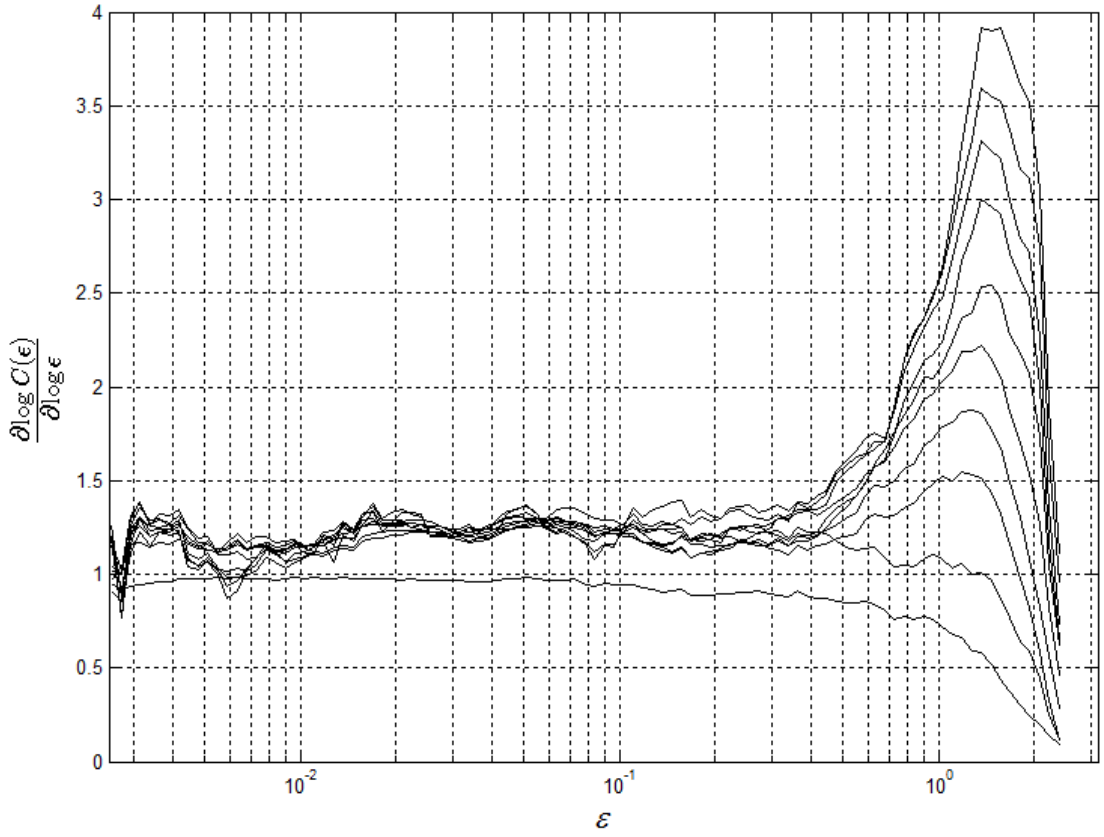


Рисунок 13.7 – Графики Раппа, построенные для аттрактора Хенона в виде локальных наклонов к корреляционным интегралам при различных значениях D_{emb}(от 1 до 10)

Размерность D_2 вычисляется как среднее значение графика Раппа в интервале ε, соответствующем данному плато – в так называемой области измерения (*scalingregion*).

Однако существует ряд дополнительных методов для более эффективной оценки величины корреляционной размерности. В настоящей работе применён *оцениватель Такенса-Тейлера* (*Takens-Theilerestimator*). Соответствующий алгоритм (файл-функция **`c2t.exe`**) считывает из файла, полученного при помощи предыдущей программы (**`d2.exe`**) – зависимость $C(\varepsilon)$ и вычисляет оцениватель Такенса (рис. 13.8). При этом используется формула

$$D_T(\varepsilon) = \frac{C(\varepsilon)}{\int_0^{\varepsilon} dx \frac{C(x)}{x}}.$$

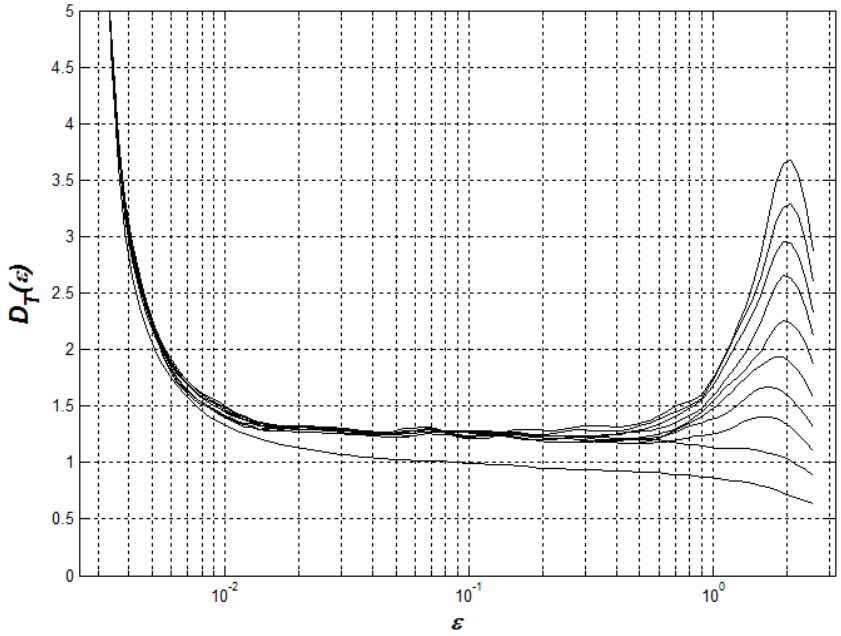


Рисунок 13.8 – Графики Раппа, построенные для аттрактора Хенона при помощи оценивателя Такенса-Тейлера (*Takens-Theilerestimator*) для различных значениях D_{emb}(от 1 до 10)

Вариант запуска программы **`c2t.exe`** из интерфейса программы Matlab имеет следующий вид:

```
c2Path = 'D:\MATLAB\work\myFunction\TISEAN\DBF\henon.dat.c2';
system([tiseanPath, 'c2t ', c2Path, ' -o']);
```

где, `tiseanPath` – путь к каталогу программы TISEAN; `c2Path`– путь к месту расположения файла с расширением **.c2*, полученного после выполнения файл-функции **`d2.exe`**; исходными данными отображения Хенона (`henon.dat`); '*–o#*' – указатель вывода результатов в файл (`henon.dat.c2_t`). Формат выходного файла – такой же, как у выходного файла программы **`d2.exe`**с расширением **.d2* .

Использование оценки Такенса дает на модельном процессе (отображение Хенона) хорошие результаты – вычисленная таким образом корреляционная размерность восстановленного аттрактора Хенона составила 1,256 (принято, что ее численное значение равно 1,26).

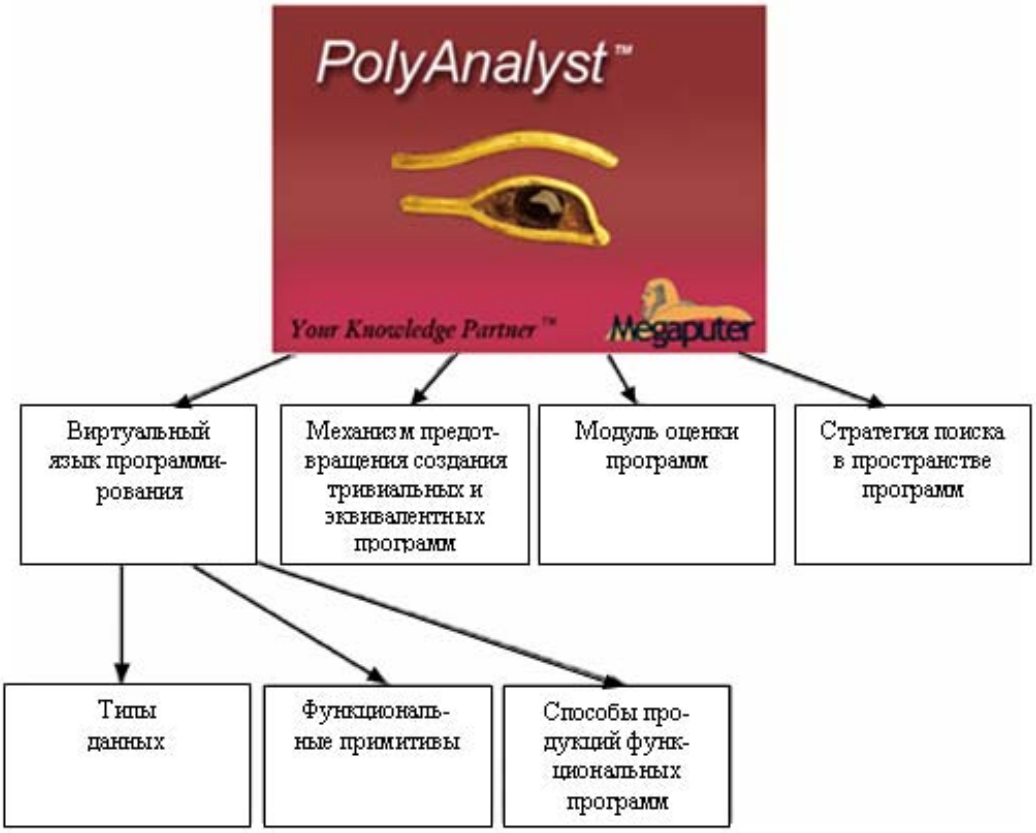
Вопросы для самоконтроля

Дайте пояснение следующим понятиям:

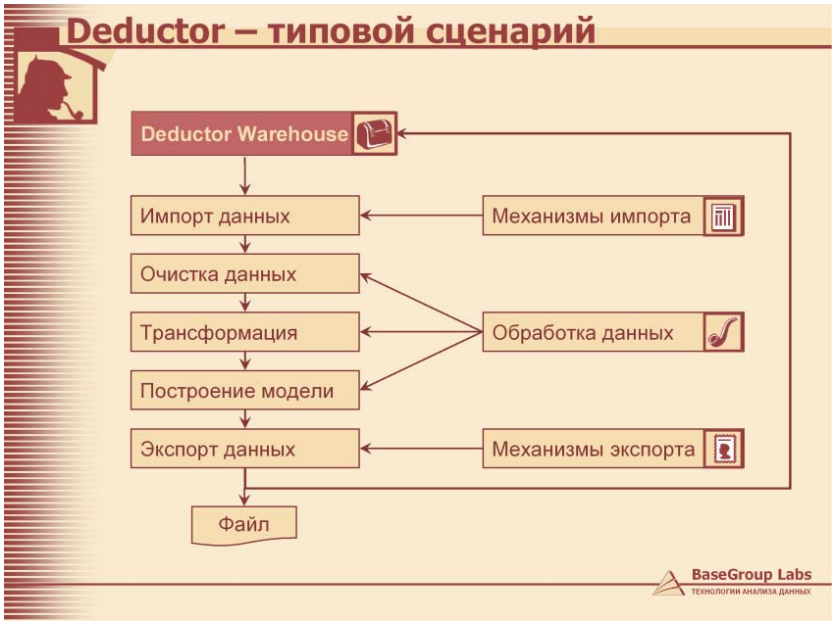
- *корреляционная размерность восстановленного аттрактора*,
- *скелинг,*
- *отображение Хенона*,
- *окно Тейлера,*
- *оцениватель Такенса-Тейлера.*

Ни один статистик не может не признать, что он должен принимать мир таким, каков он есть.

Дж. Тьюки, 1968г.



Data Mining система PolyAnalyst



Аналитическая платформа класса KDD (Knowledge Discovery in Databases) – Deductor

Практическое занятие №14

Знакомство с Data Mining системой PolyAnalyst

Цель работы: Ознакомиться с возможностями DataMining системы PolyAnalyst, основанной на принципах эволюционного программирования. Получить навыки анализа данных с использованием Поиска Зависимостей, Поиска законов, Классификации, Кластеризации, Многопараметрической линейной регрессии.

Теоретические сведения

Создатели первой российской DataMining системы – PolyAnalyst(www.megaputer.ru) изначально отказались от многих классических предположений и результатов математической статистики.

По нашему мнению – основным достоинством системы является отказ от постулирования вида зависимости (обычно линейной). PolyAnalyst формулирует и проверяет гипотезы о виде регрессионной зависимости на внутреннем языке программирования с помощью функциональных примитивов (простейших программ), которые в дальнейшем с помощью трех методов продукций (композиции, итерации/рекурсии и дробно-рациональных выражений) синтезируют многомерные регрессионные модели, что позволяет обнаруживать скрытые зависимости и исследовать данные сложной структуры.

Процесс построения программ строится как эволюция в мире программ (этим подход немного похож на генетические алгоритмы). Результаты представляются в виде понятном пользователю – таблиц, графиков и формул. Следует отметить, что авторы рассматривают методы линейной регрессии и поиска зависимостей в системе PolyAnalyst как дальнейшее развитие методов классического регрессионного анализа.

Для контроля статистической значимости полученных зависимостей применяется рандомизированное тестирование и классические методы.

Окно системы PolyAnalyst, так же как и Statistica, реализовано согласно стандартам программ, работающих в среде Windows. В левой части окна находится иерархическое дерево с папками документов, в которых содержатся исходные данные (World), графики (Graphs), отчеты (Reports) и т.д. (рис.14.1). В этой работе мы рассматриваем русскоязычную версию PolyAnalyst Professional для Windows NT (версия 3.3, рис.14.2), которая по своим возможностям несколько проигрывает 4-5 версиям, но достаточна для понимания идеологии работы системы.

Версия PolyAnalyst 3.3 (единственная русскоязычная и поэтому расматриваемая здесь) включает в себя семь модулей, реализующих различные методы исследования (также называемыми процессами), которые преследуют различные цели и дополняют друг друга. Это[11]:

·CorePolyAnalyst – 'Поиск законов' (FL). (Модуль Core PolyAnalyst ищет скрытые в данных функциональные зависимости и представляет полученные ре-

[11] Справка программы и на сайте www.megaputer.com (ru).

зультаты в символьном виде математических формул, включающих в себя блоки условий. Способность Core PolyAnalyst автоматически строить большое многообразие математических конструкций, которые включают в себя сложные нелинейные алгебраические выражения и функции делает его уникальным инструментом поиска знаний.);

·ARNAVAC – 'Поиск Зависимостей' (FD) (ARNAVAC определяет переменные, которые наиболее сильно влияют на целевую переменную, и оценивает силу обнаруженной зависимости, сокращая, таким образом, пространство поиска для Core PolyAnalyst. Отсеивает бессмысленные и далеко отстоящие значения, оценивает точность, которую может достичь Core PolyAnalyst на исходных данных.

Кроме этого, ARNAVAC сам является ценным инструментом для обнаружения знаний. ARNAVAC представляет полученные им результаты в форме таблиц предсказания. Во многих случаях эти таблицы дают ясную картину отношений между различными компонентами данных. На самом деле, обе формы представления знаний, и символьная, и табличная, дополняют друг друга.

Сложные модели явлений и объектов реального мира обычно требуют различных форм представления знаний для своего описания.);

·Модуль многопараметрической линейной регрессии – 'Линейная Регрессия' (LR). Линейная Регрессия, реализованная в PolyAnalyst основана на очень быстром алгоритме. Как широко распространенный метод статистического исследования, линейная регрессия включена во множество статистических пакетов и электронных таблиц.

Однако, реализация этого метода в PolyAnalyst имеет свои специфические особенности которые делают ее наиболее подходящей для пользователей, не являющимся специалистами в статистике. Эти особенности включают в себя автоматический выбор наиболее значимых независимых переменных и статистически верная оценка значимости полученных результатов.

Нужно заметить, что эта значимость отличается от значимости единичной регрессионной модели, так как в течение одного запуска данного вычислительного процесса может быть проверено большое число регрессионных моделей. Эти две особенности гарантируют, что результаты, полученные даже неопытным пользователем не будут бессмысленны или неправильно поняты.);

·Модули Классификации, и Дискриминации, использующие в своей работе модули CorePolyAnalyst и линейную регрессию (CL и DS) (Метод Классификации, встроенный в PolyAnalyst основывается в своей работе на результаты линейной регрессии, PolyNet Predictor и Core PolyAnalyst. Хотя оба эти метода могут предсказывать только атрибуты числового типа, Классификация строит модели для логических переменных.

Построенное классификационное правило выглядит следующим образом: если значение правила больше, чем порог, то предсказываемая переменная принимает значение истина, иначе – ложь. Правило - это выражение, найденное Поиском Законов или Линейной Регрессией, а порог – число от 0 до 1.

Метод Дискриминации во многом аналогичен методу Классификация. Он применяется для того, чтобы выяснить, чем данные из выбранной таблицы отли-

чаются от остальных данных, включенных в проект, выделить специфические черты, характеризующие некоторое подмножество записей проекта. В основе этого метода также лежат алгоритмы линейной регрессии, PolyNet Predictor либо Core PolyAnalyst.

Отличие от метода Классификация заключается в том, что не нужно выделять целевой параметр – им в этом исследовании является принадлежность или непринадлежность записи к таблице, для которой вызван метод Дискриминация. Поскольку задача этого метода состоит в нахождении отличий выбранной таблицы от всех остальных данных, фактически это исследование проводится на всем множестве данных проекта.);

·Модуль Кластеризации (FC); (Этот метод применяется тогда когда надо выделить в некотором множестве данных компактные типичные подгруппы (кластеры, состоящие из близких по своим характеристикам записей. Причем заранее может быть неизвестно какие переменные нужно использовать для такого разбиения.)

Метод Кластеризация сам находит набор переменных, для которого это разбиение наиболее четко и статистически значимо. Результатом работы этого метода является описание областей (диапазонов значений переменных), характеризующих каждый обнаруженный кластер и разбиение исследуемой таблицы на подмножества, соответствующие кластерам. Если данные являются достаточно однородными по всем своим переменным и не содержат "сгущений" точек в каких-то областях, этот метод не даст результатов.

Надо отметить, что минимальное число обнаруживаемых кластеров равно двум - сгущение точек только в одном месте в нашем методе не рассматривается как кластер. Кроме того, этот метод в большей степени, чем остальные предъявляет требования к наличию достаточного количества записей в исследуемой таблице, а именно, минимальное количество записей в таблице, в которой может быть обнаружено N кластеров, равно (2N-1)4.);

·Модуль PolyNetPredictor (PN). (Этот метод может применяться для решения тех же задач, что и методы Линейная регрессия и Поиск законов. Его существенное отличие заключается в том, что PolyNet Predictor, получая во многих случаях более точные правила, чем два последних метода, не всегда способен представить их в понятном для человека текстовом виде (формулы и т.д.).

Таким образом, он должен применяться тогда, когда главная цель - научиться предсказывать значения какого-либо параметра, а не получить знания о взаимосвязях параметров, составляющих исследуемые данные. По времени работы он занимает промежуточное положение между методами Линейная регрессия и Поиск законов.

Работа метода PolyNet Predictor основана на построении иерархической структуры, подобной нейронной сети. При этом сложность этой сетевой структуры и другие ее параметры подбираются динамически на основе свойств анализируемых данных. Если создаваемая сетевая структура не является слишком сложной, может быть построено эквивалентное ей выражение на языке символических правил системы PolyAnalyst.

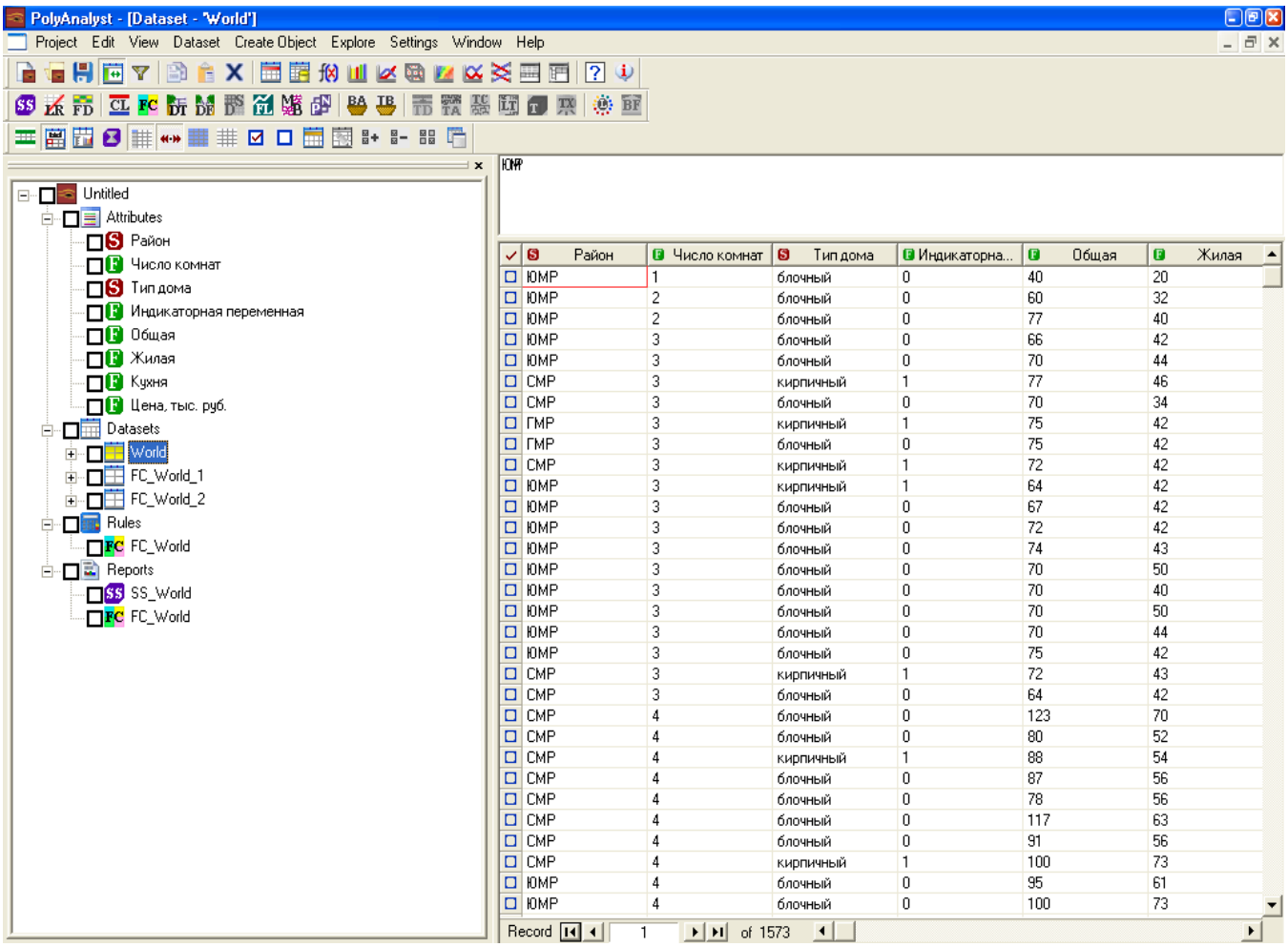


Рисунок 14.1 – PolyAnalyst 5.0

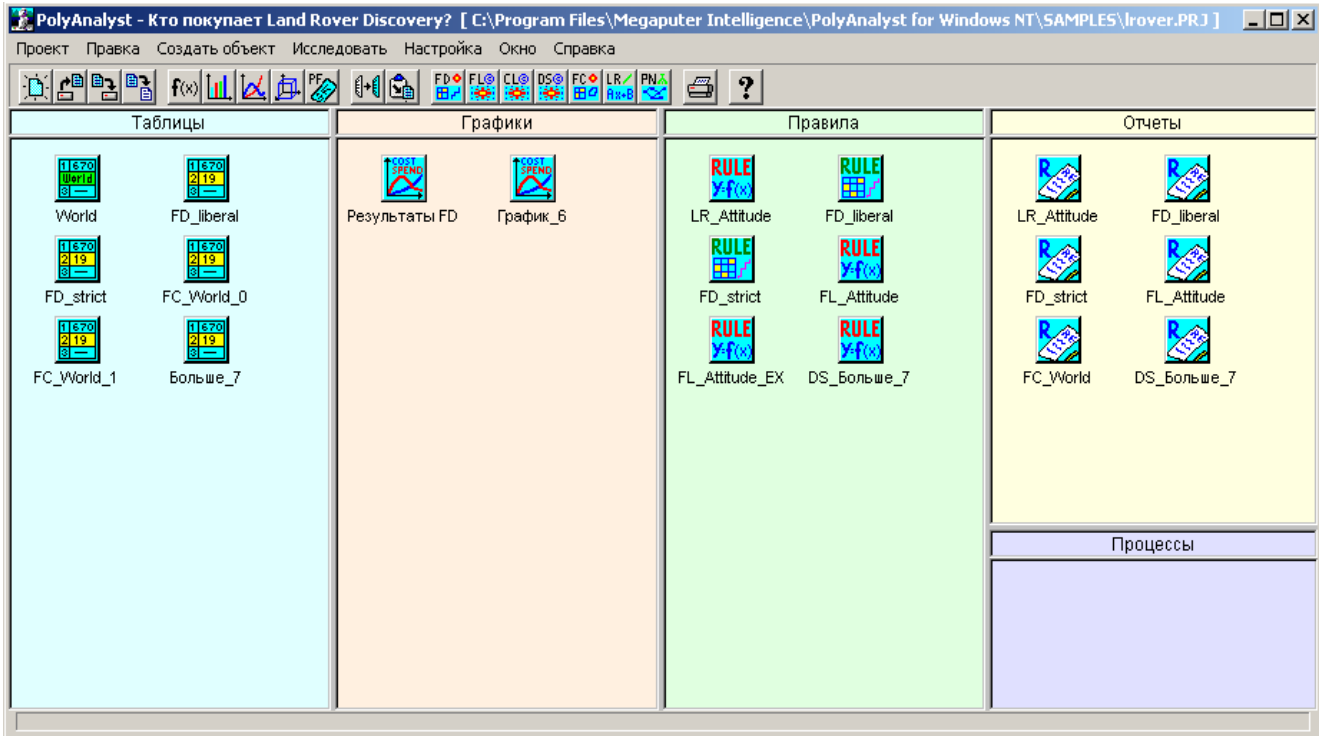


Рисунок 14.2– PolyAnalyst Professional для Windows NT

Следует отметить, что все описанные выше модули (как в принципе и все методы ИАД) используют классические статистические методы – как на этапах поиска и оценки моделей, так и на этапе оценки её адекватности. Например, надёжность полученных результатов основывается на стандартном отклонении (Арсеньев С. Извлечение знаний из медицинских баз данных. –www.megaputer.com (ru)):

$$s_{dev} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-1}}$$

и стандартной ошибке

$$s_{err} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-1)\sigma_y^4}},$$

где y_i – зависимая переменная; $\hat{y}_i$ – соответствующее значение, предсказанное моделью; n – число наблюдений; σ_y^4 – квадрат дисперсии переменной y. Значимость найденной зависимости оценивается с помощью индекса значимости (I_z):

$$I_z = -k \lg(\frac{s_{real}}{s_{rand}}),$$

где s_{real} – стандартное отклонение, полученное на реальных данных, s_{rand} – стандартное отклонение случайных данных, в которых значение результативной переменной случайно перемешано для разных наблюдений, *k=const*. Считается, что результат моделирования значим, если значение индекса значимости больше 2,0.

Стандартный подход к оценке значимости модели – коэффициент детерминации Rsquared (чем ближе он к единице, тем лучше модель).

Визуальный подход к оценке значимости модели заключается в изображении зависимости предсказанных значений ($y_{predicted}$), от реальных (y_{real}) – чем ближе точки лежат к прямой $y_{predicted}=y_{real}$, тем точнее модель описывает данные.

Оценка значимости линейных регрессионных моделей основывается на известной статистике Фишера-Снедекора *F-ratio*:

$$\left[\frac{b_j}{m_{b_j}}\right]^2,$$

где b_j – *j*-ый коэффициент модели, m_{b_j} – стандартное отклонение коэффициента. Этот коэффициент лежит в основе отбора наилучших переменных в уравнение регрессии (обычно переменная включается, если *F-ratio*>2,0).

В качестве примеров 1, 2 рассмотрим стандартные обучающие примеры, поставляемые с системой PolyAnalyst 3.3: Простая и точная зависимость, Прикладное маркетинговое исследование[12].

Анализ данных в системе PolyAnalyst

Пример 1. Создайте проект. Щелкните дважды мышью на пиктограмме PolyAnalyst, чтобы запустить программу. Выберите меню **Проект/Новый**. На экране откроется диалог Выбор источника данных. Выберите файл 1SIMPLE.CSV путем выполнения двойного щелчка на его имени. Появится диалог создания проекта. Назовите проект 1S. В папке таблицы появится папка **World** содержащая исходные данные – искусственно сгенирированную зависимость y от x.

1)Активируйте правой кнопкой мыши таблицу **World**. Выберем в строке меню **Исследовать – Поиск Законов**. В появившемся диалоговом окне дважды щелкните левой кнопкой мыши, чтобы сделать её результативной (зависимой) переменной.

2)Запустите вычислительный процесс. В папке процессы появится пиктограмма **FL-процесса (Найти Закон)**. После получения первого результата откроется окно отчета, в котором отражаются результаты процессы в текущий момент и после окончания поиска появится окончательная формула.

Отчет состоит из 4 частей:

–**текстовый отчет** (содержит наиболее точное правило (формулу),

–**наиболее статистически значимое правило** (формула, таблица),

–**график качества подгонки**,

–**график зависимости предсказанных значений y** (темно синие) **и реальных значений y**(красные) **от номера записи**.

График **Невязок**способствует визуализации распределения ошибок. Он показывает, какой процент предсказанных значений имеет ошибку меньше, чем соответствующее значение на оси X. К примеру, если вы увеличите окно этого графика до максимума, вы увидите, что примерно 27% точек имеют отклонение меньше, чем 0.0001. Чтобы подсчитать это, вы должны вычесть из значения, соответствующему ошибке –0.0001 (36%), значение, соответствующее ошибке 0.0001 (62%).

Пример 2 Для того чтобы лучше понять стиль жизни потенциальных покупателей автомобилей Land Rover менеджер по маркетингу Paul Montopoli организовал специальное социологическое исследование, направленное на изучение их отношения к жизни, интересов и мнений по разным вопросам. Вопросник включал 30 пунктов, включая вопросы по отношению человека к риску, к отечественным и зарубежным продуктам, к привычкам тратить деньги, отношению к своей внешности и семье. Ответы строились по 9-бальной схеме, в которой ответ 1 означает, что человек полностью отвергает данное утверждение, а ответ 9 означает, что он принимает его на 100%. Всего было опрошено 400 человек. Само исследование проводилось независимой компанией с использованием опросных листов в журнале "Car and Driver Business Week". Список вопросов приводится ниже.

- **Q1** У меня очень хорошее физическое состояние;

[12] Неточности, замеченные в примере 2, исправлены

- **Q2** Я всегда одеваюсь по моде, вне зависимости комфортно это или нет;
- **Q3** У меня больше модной одежды, чем у моих друзей;
- **Q4** Я всегда хочу выглядеть несколько отлично от своих друзей;
- **Q5** Жизнь слишком коротка, чтобы не поиграть в нее;
- **Q6** Меня не беспокоит состояние озонного слоя;
- **Q7** Я думаю, что правительство делает слишком много для контроля за загрязнением окружающей среды;
- **Q8** Я считаю, что в основном современное общество устроено правильно;
- **Q9** У меня нет времени на благотворительность;
- **Q10** У нашей семьи сейчас нет слишком больших долгов;
- **Q11** Я всегда предпочитаю платить наличными;
- **Q12** Я слишком много заплатил за «сегодня» так пусть будет, что будет;
- **Q13** Я использую кредитные карты, так как они дают мне отсрочку в платежах;
- **Q14** Я редко использую купоны, когда покупаю;
- **Q15** Процентные ставки сейчас слишком низкие, чтобы я мог покупать все, что мне нравится;
- **Q16** Я всегда более уверен в себе, чем большинство моих друзей;
- **Q17** Я люблю, когда меня считают лидером;
- **Q18** Другие люди всегда обращаются ко мне за помощью;
- **Q19** Дети – самое главное в браке;
- **Q20** Я предпочитаю тихий вечер дома шумным вечеринкам;
- **Q21** Иностранные машины нельзя сравнить со сделанными в Америке;
- **Q22** Правительству следует ограничивать импорт из Японии;
- **Q23** Американцам всегда следует отдавать предпочтение американским товарам;
- **Q24** Я хотел бы совершить кругосветное путешествие;
- **Q25** Я хотел бы порвать с моей текущей жизнью и начать что-нибудь новое;
- **Q26** Обычно я среди первых попробовать новые товары;
- **Q27** Я люблю крепко работать и здорово отдыхать;
- **Q28** Скептические прогнозы обычно неправильны;
- **Q29** Я могу сделать все, что мне захочется;
- **Q30** Пять лет назад мой доход был значительно больше чем сейчас.

В дополнение к этим 30 утверждениям следующее утверждение использовалось как мера отношения человека к машинам Land Rover (по той же 9-бальной шкале): "Я бы, пожалуй, купил модель Discovery, выпущенную Land Rover".
Исходные данные содержаться в текстовом файле lrover.csv. Каждая строка файла включает ответы одного респондента.

1) Выберите пункт меню **Проект/Новый/Из файла...**, выберите файл lrover.csv. В поле **Имя проекта** введите название проекта, например "Кто покупает Land Rover Discovery?",. Оставьте тип данных для всех полей числовым и нажмите **OK**. Сохраните файл проекта. Для этого введите имя файла, например LROVER_R. Теперь новый проект создан, сохранен на диске, и можно приступать к анализу данных.

2) Исследование данных с помощью многопараметрической линейной регрессии.

Во многих случаях бывает полезным начать анализ данных с линейного регрессионного метода. Этот метод в системе **PolyAnalyst** имеет целый ряд особенностей, выгодно отличающих его от линейных регрессионных алгоритмов, применяемых в большинстве статистических пакетов.

Дело в том, что линейная модель, вырабатываемая системой включает только значимые члены, а не все независимые переменные, причем они ранжируются по степени их влияния. Кроме этого линейный вычислительный модуль может работать с категориальным типом данных, что особенно ценно в анализе различного рода социологических данных. Этот метод, также, очень быстр и нагляден.

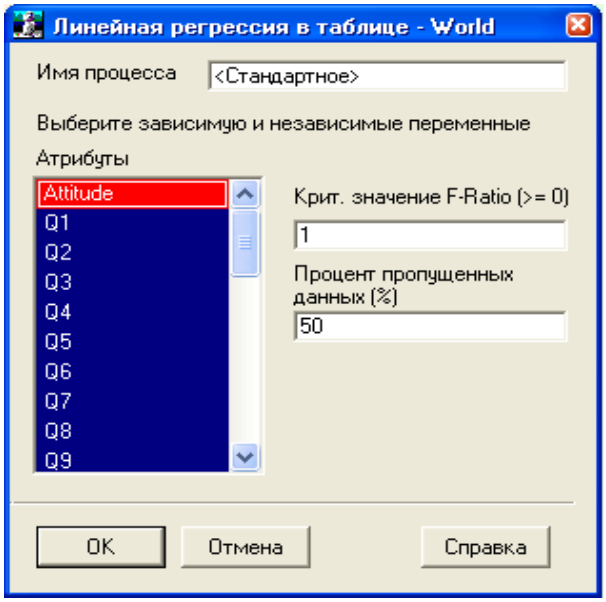


Рисунок 14.3– Диалоговое окно Линейной регрессии

Для запуска исследования нажмем правой кнопкой мышки на таблице **World** и выберем пункт **Исследовать/Линейная Регрессия...**. В диалоге запуска метода дважды щелкнем на поле Attitude, чтобы сделать его целевой переменной (рис.14.3).

Можно не задавать имя отчета, оно будет выбрано автоматически. Вы также можете задать критическое значение F-Ratio для отбраковки незначимых членов и максимальный процент пропущенных значений целевой переменной. По умолчанию эти значения равны 1 и 50% соответственно.

В контейнере активных процессов появилась пиктограмма вычислительного процесса линейной регрессии. Через несколько десятков секунд метод закончит работу и мы увидим отчет, который состоит из текстового окна и нескольких графиков.

Нажав правой кнопкой мышки на пиктограмме отчета, мы можем выбирать компоненты отчета, которые для нас интересны в каждом конкретном случае. Сейчас нас интересует текстовый отчет и график относительного вклада факторов.

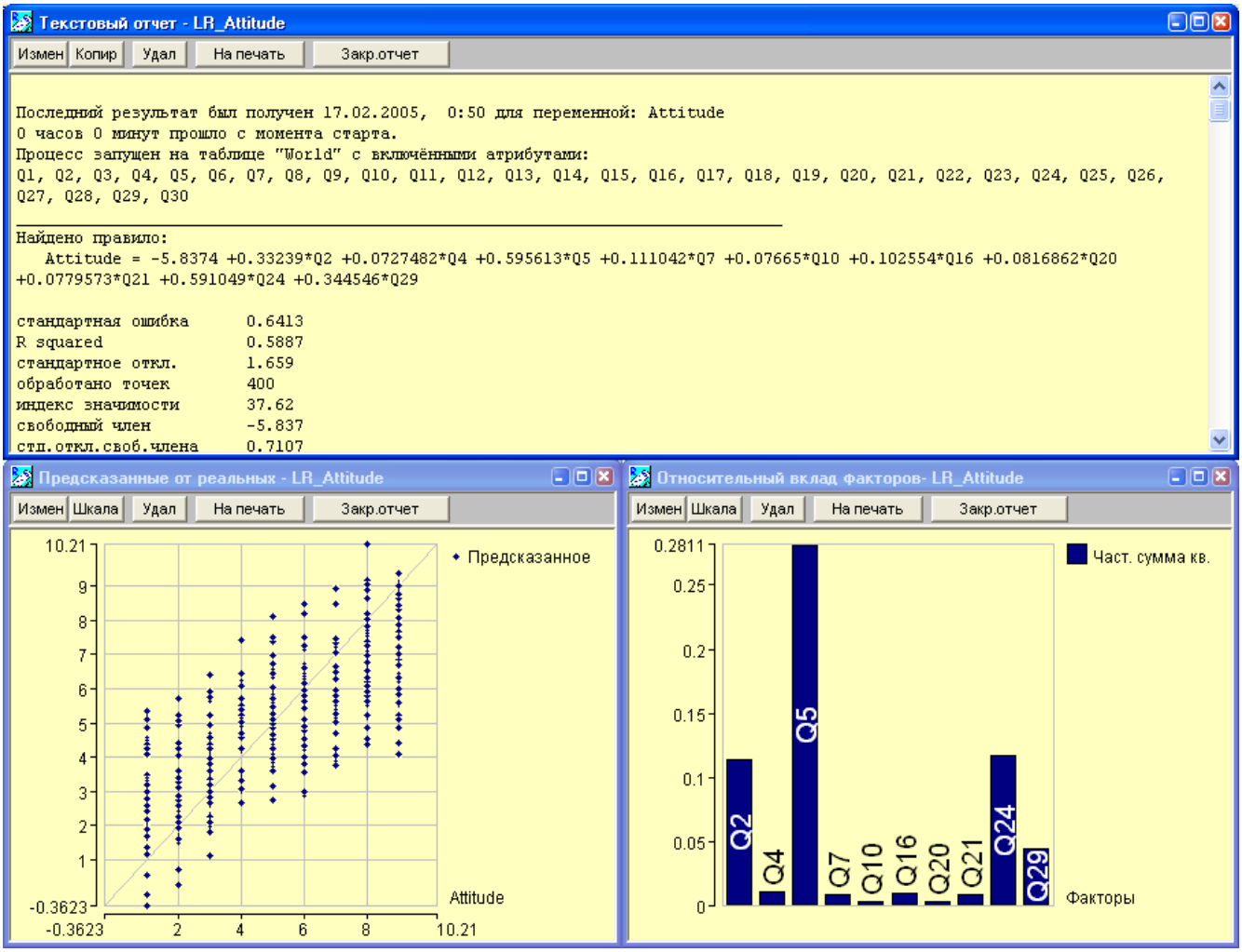


Рисунок 14.4 – Отчёт, полученный после запуска Линейной регрессии

В окне текстового отчета мы видим само найденное линейное выражение для целевой переменной Attitude, общие статистические оценки всей формулы и статистические параметры каждого входящего члена. Мы видим, что, несмотря на довольно большое значение стандартной ошибки линейного правила, 64% (что довольно типично для реальных данных), значимость выражения очень велика (индекс значимости > 37). Это является важнейшим индикатором того, что формула действительно отражает реальные зависимости в данных. Мы также видим, что все члены входят в выражение со знаком плюс, те чем больше значение ответа, тем больше значение результата. Четыре члена дают наибольший вклад так как у них наибольшие значения частичной суммы квадратов. Это: Q2, Q5, Q24 и Q29. Примечательно, что для этих членов и значения F-ratio (показателя значимости) также велики, это говорит о высокой степени независимости этих переменных. Какие же это утверждения?

- **Q2**: Я всегда одеваюсь по моде, вне зависимости комфортно это или нет.
- **Q5**: Жизнь слишком коротка, чтобы не поиграть в нее.
- **Q24**: Я хотел бы совершить кругосветное путешествие.
- **Q29**: Я могу сделать все, что мне захочется.

Таким образом уже линейный вычислительный модуль **PolyAnalyst** позволяет сделать ценные выводы в отношении портрета вероятного покупателя машины Discovery. Этот человек придает большое значение моде, склонен к риску, бесшабашен, готов к приключениям, уверен в себе и возможно богат, так как может позволить себе все, что захочется.

Мы можем оформить эти результаты в виде печатного отчета, выбирая кнопку **На печать**.

Замечание. Возможность печати отсутствует в «пробной» версии.

Для этого нажмем кнопку создания печатных форм. В диалоге создания формы мы введем имя отчета и заполним заголовки и колонтитулы.

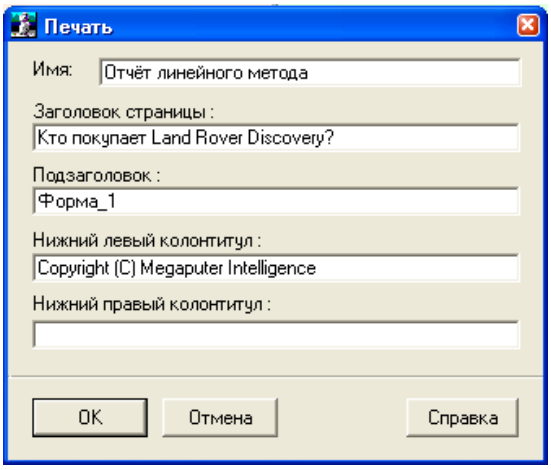


Рисунок 14.5 – Диалоговое окно создания печатной формы

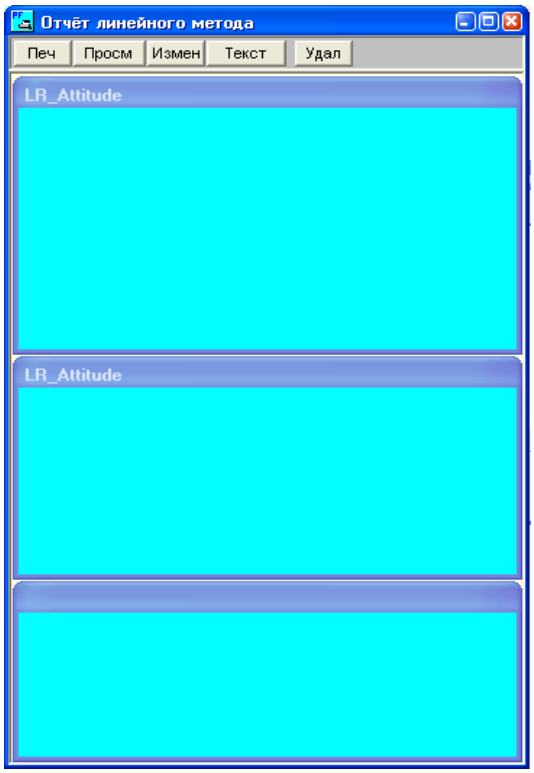


Рисунок 14.6 – Окно формы отчёта

Нажав кнопку **OK**, мы увидим перед собой пустую форму, на которые можно помещать различные объекты, располагая их в удобном для восприятия виде (рис. 14.6).

В последующем эта печатная форма может использоваться как шаблон отчета линейного метода, например, при обработке новых данных. Поместим в форму текстовый отчет, график относительного вклада факторов, а также, используя буфер обмена, фрагменты текста данного урока.

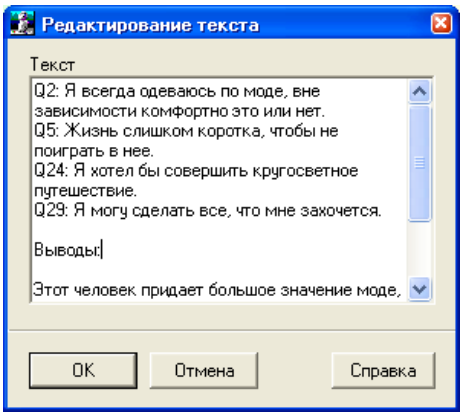


Рисунок 14.7 – Текстовый отчет

Можно контролировать создание печатной формы, нажимая кнопку **Просмотр**.

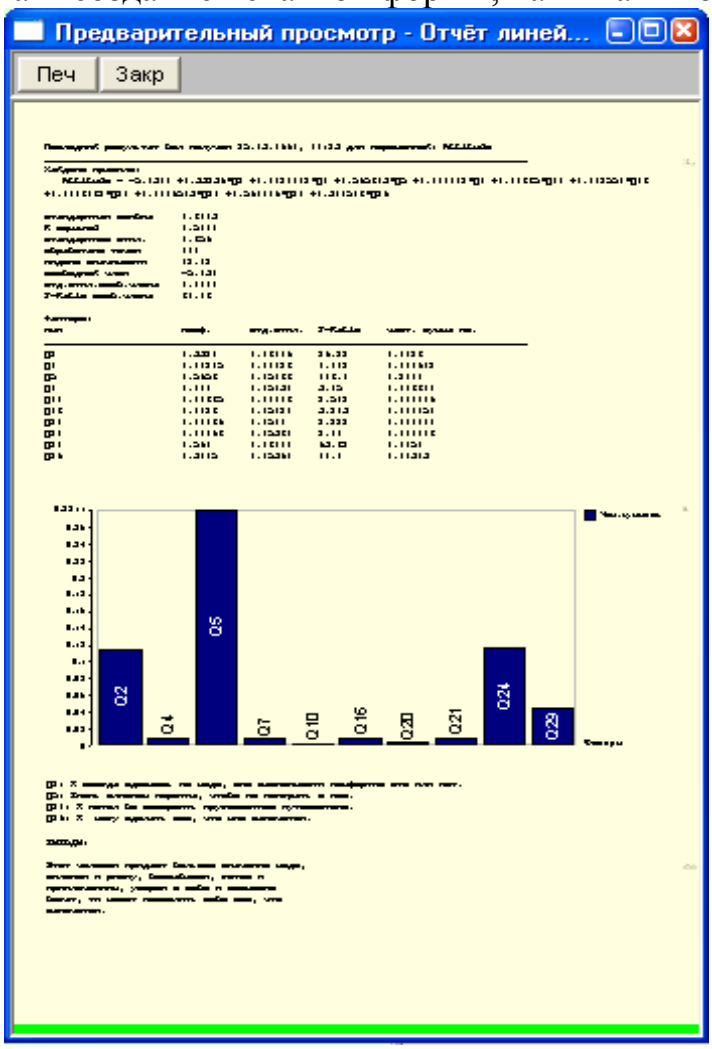


Рисунок 14.8 – Готовый Отчёт Линейного метода

Анализ с помощью **Поиска Зависимостей**. Этот вычислительный модуль предназначен не для нахождения общей формульной зависимости, а для выявления групп записей, имеющих схожие свойства в отношении целевой переменной. Имеются две различные модификации работы метода: жесткий (strict) алгоритм и мягкий (liberal) алгоритм. Жесткий алгоритм находит компактные группы точек, показывающих сильную функциональную зависимость целевой переменной от независимых. Мягкая модификация предназначена для выявления в первую очередь исключений, то есть тех записей, которые сильно отличаются от основной массы. Поиск Зависимостей, также как и Линейная Регрессия, из всего пространства независимых переменных выделяет наиболее влияющие факторы.

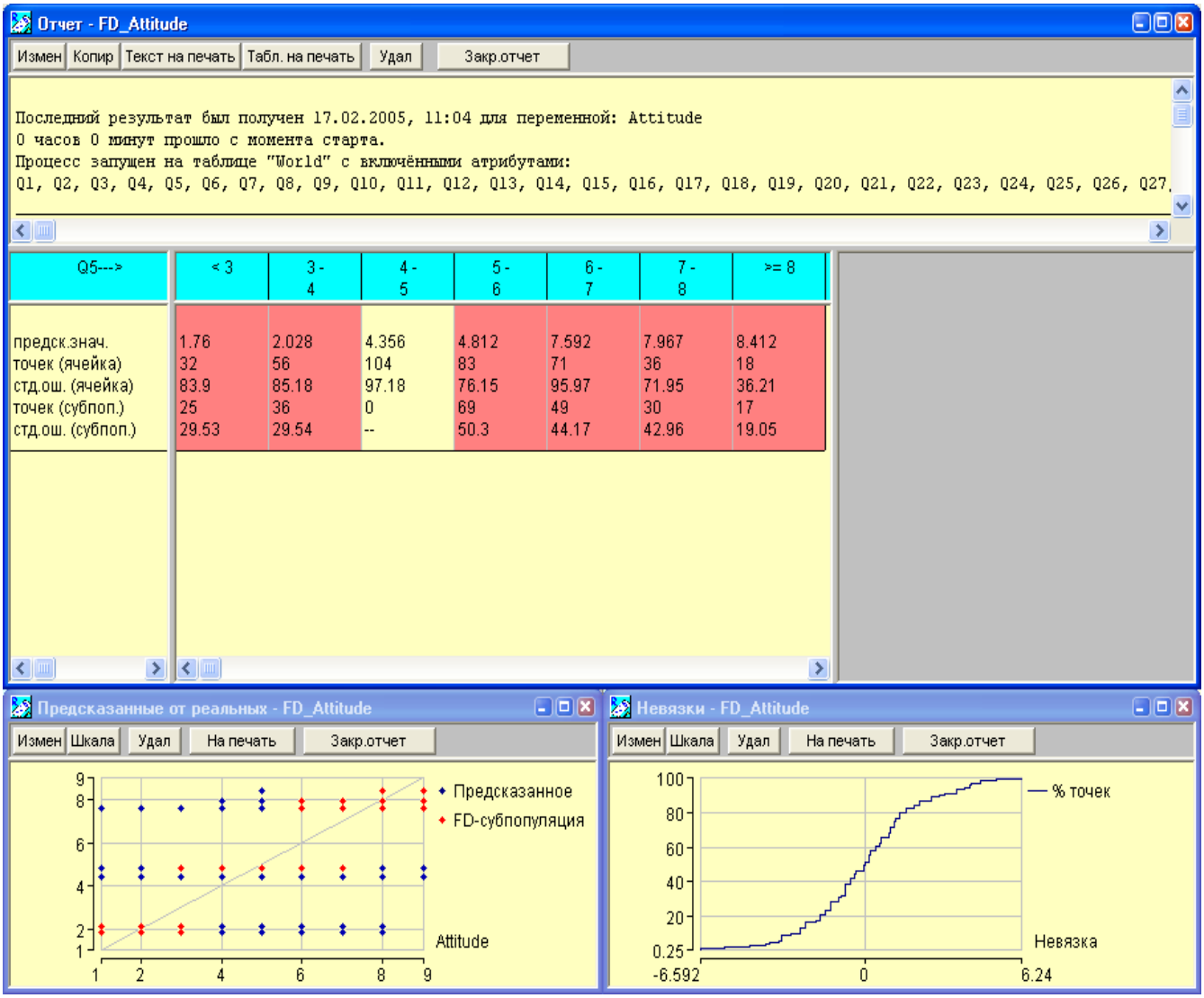


Рисунок 14.9 – Отчёт, полученный после запуска Поиска зависимостей

Сначала проведем исследование, используя жесткий алгоритм. Выберем пункт меню **Исследовать/Поиск зависимостей...**, в диалоге параметров метода введем имя процесса «FD_strict», дважды щелкнем на целевой переменной Attitude, в поле **Лимит времени** введем продолжительность работы метода – 2 минуты (это необязательное ограничение) и нажмем кнопку **OK**. Как обычно, в контейнере

процессов появится пиктограмма активного процесса. И, как только метод закончит свою работу, можно будет заняться анализом его результата, отчета «FD_strict». Отчет состоит из нескольких окон. Верхнее окно содержит текст, в середине экрана располагается так называемая таблица предсказаний, в нижней части помещены графики, помогающие визуально оценить результаты. Вы можете масштабировать эти объекты и располагать их на экране в удобном для вас виде. Каждое окно может быть независимо закрыто, а кнопка **Закрыть отчет** закрывает сразу все окна. Если щелкнуть правой кнопкой мышки на пиктограмме отчета, то можно выбрать для визуализации отдельно любой его элемент.

В данном примере мы имеем дело с дискретными данными, поэтому графики отчета не являются сильно информативными. Их можно закрыть и заняться анализом текстового отчета. Мы видим, что вычислительный модуль выделил в качестве наиболее влияющего фактора переменную Q5, те отношение респондента к жизни. Всего из 400 записей 226 показывают наличие значимой функциональной связи целевой переменной от Q5. Мерой значимости служит логарифм вероятности того, что этот факт является случайным. Вероятность этого здесь исчезающе мала и составляет e^{-371}. Так как это число очень велико по абсолютной величине, то в качестве меры значимости берется логарифм от этого значения. Таким образом Поиск Зависимостей подтвердил факт наибольшего влияния переменной Q5, который был уже обнаружен линейным методом.

В таблице, предсказаний, которая в данном случае является одномерной, мы видим шесть ячеек, закрашенных красным цветом. Эти ячейки содержат параметры областей данных, для которых обнаружены функциональные зависимости. В каждой ячейке имеется 5 чисел. Рассмотрим для примера одну из ячеек. Так, если Q5 больше или равно 8, то всего в нашей исходной таблице имеется 18 записей, у которых Q5>=8, 17 из них выделены в функциональную субпопуляцию. Стандартная ошибка для всей группы составляет 18%, а для субпопуляции 17%. Среднее предсказываемое значение целевой переменной для этой субпопуляции равно почти 8.5 (8.412). Таким образом, мы можем заключить, что 17 респондентов из 400 опрошенных с очень большой вероятностью настроены приобрести машину Discovery.Если условно принять, что человек поставивший бал 9 для целевого утверждения за 100% вероятность того, что он купит Land Rover, то можно составить следующую таблицу вероятностей покупки:

Q5	кол-во человек	вероятность покупки
6-7	49	7.59*100/9=84%
7-8	30	7.97*100/9=88%
8-9	17	8.14*100/9=90%

С другой стороны имеются группа людей, которые относятся к жизни гораздо серьезнее (Q5 меньше 4). Для них предсказываемое значение целевой перемен-

ной также мало (практически меньше 2). Всего таких респондентов 25+36=61. Эти люди, по-видимому, не настроены на покупку Discovery. Поиск Зависимостей автоматически создает новую таблицу (FD_strict), в которую входят только записи, подчиняющиеся найденной зависимости. Обращаясь к базе данных, мы можем идентифицировать этих людей, привлекая дополнительные данные о них и таким образом оптимизировать, например, рекламную кампанию. Вот такие интересные результаты получены методом Поиск Зависимостей.

Теперь попробуем мягкий вариант алгоритма. Выберем корневую таблицу **World** и запустим Поиск Зависимостей в мягком варианте, указав, при этом, имя процесса FD_liberal.

Мы видим, что мягкий алгоритм выделил три наиболее влияющих фактора: Q5, Q28 и Q29 (значимость в мягком алгоритме не вычисляется). В дополнение к факторам, выделенным линейным методом здесь введен новый фактор Q28: "Скептические прогнозы обычно неправильны.", то есть, степень оптимизма человека. Таблица предсказаний теперь является трехмерной и для показа влияния 3-й переменной введен выпадающий список.

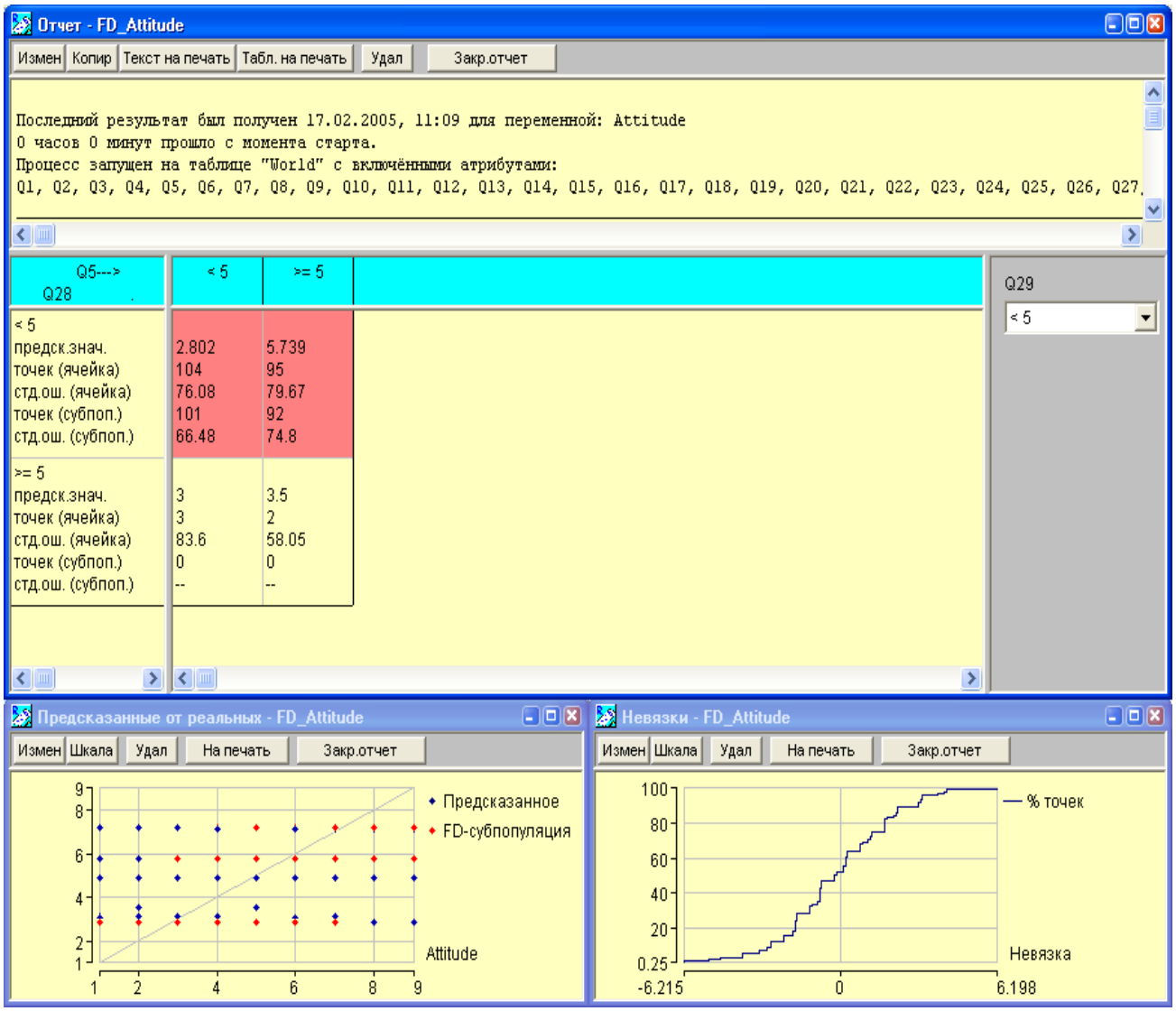


Рисунок 14.10 – Отчёт мягкого алгоритма Поиска зависимостей при Q29>=5

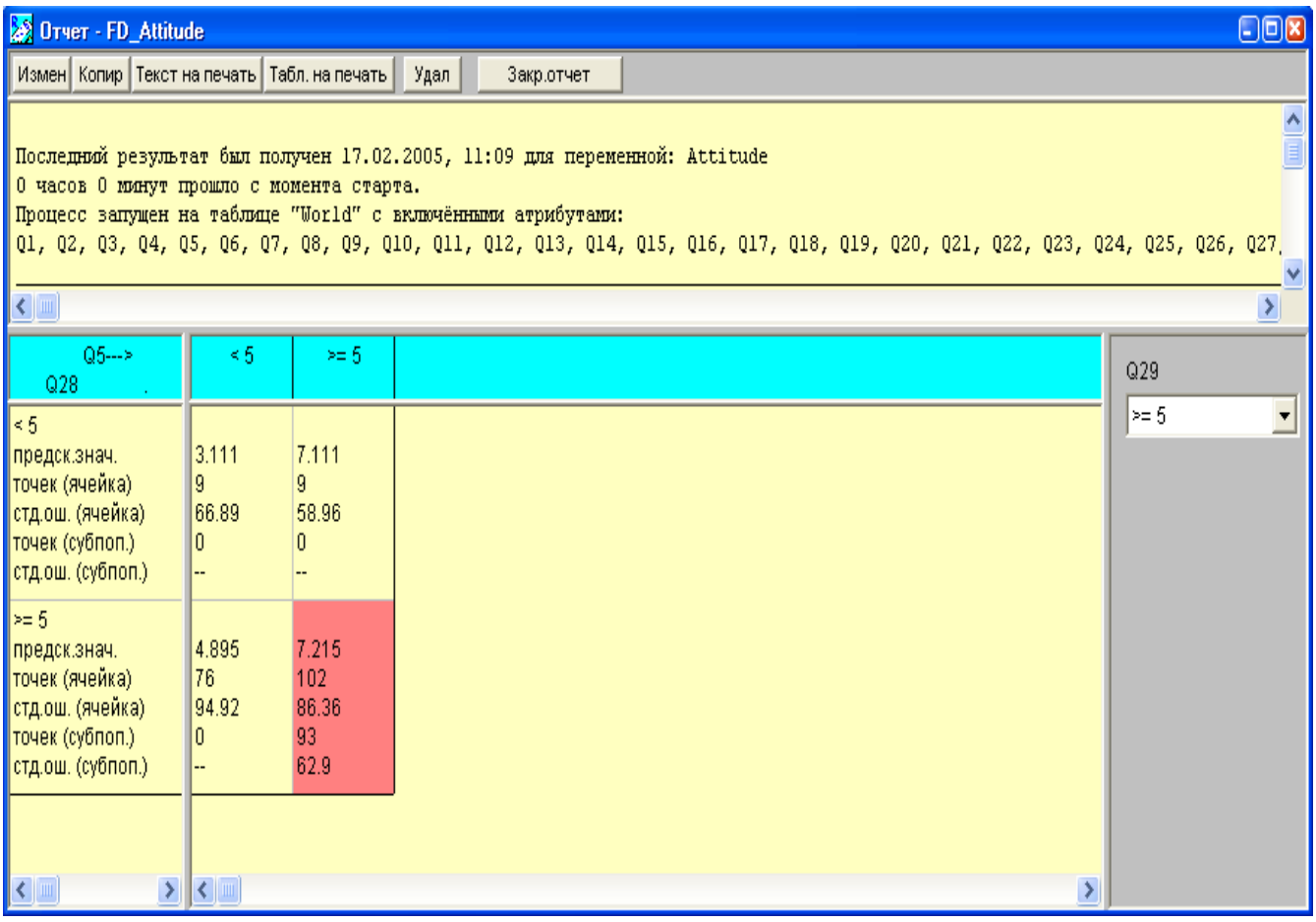


Рисунок 14.11 – Отчёт мягкого алгоритма Поиска зависимостей при Q29<5

Мягкий алгоритм выделил большее количество точек, чем жесткий, 286 против 226. Интересно, что разбиение на группы производится четко по среднему значению 9-бальной шкалы.

Если все три независимые переменные Q5, Q28 и Q29 меньше 5, то имеется группа людей (101 человек), которая достаточно отрицательно относится к идее покупке данного автомобиля. Для них среднее прогнозируемое значение attitude = 2.8. Эти люди склонны к скептическим оценкам происходящего, неуверенные в себе и осторожные в жизни. По-видимому, не стоит тратит рекламные средства на эту группу людей.

Так как она составляет 25% от всего количества респондентов, то это существенная экономия.Прямо противоположную группу составляют люди, поставившие для утверждений Q5, Q28, Q29 оценку больше 5. Эта группа включает 93 человека, что составляет 23% от всех опрошенных, и для нее среднее предсказываемое значение целевой переменной больше 7 (7.215). Эту группу можно считать весьма перспективной при проведении направленной рекламной кампании.

Полученные методом результаты можно наглядно представить в графическом виде. Для этого щелкнем мышкой на кнопке создания 2D-графика  и заполним диалог.

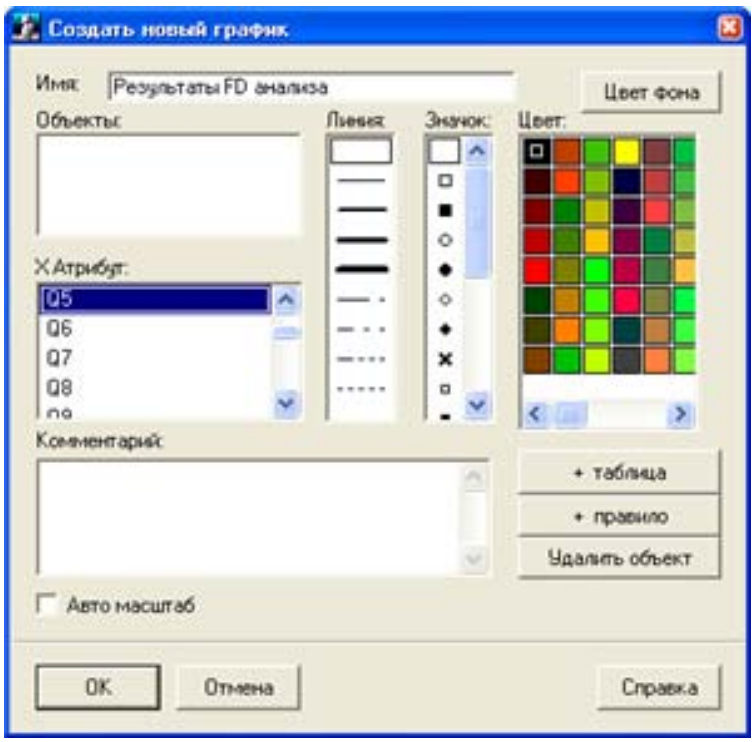


Рисунок 14.12 – Окно создания 2D графиков

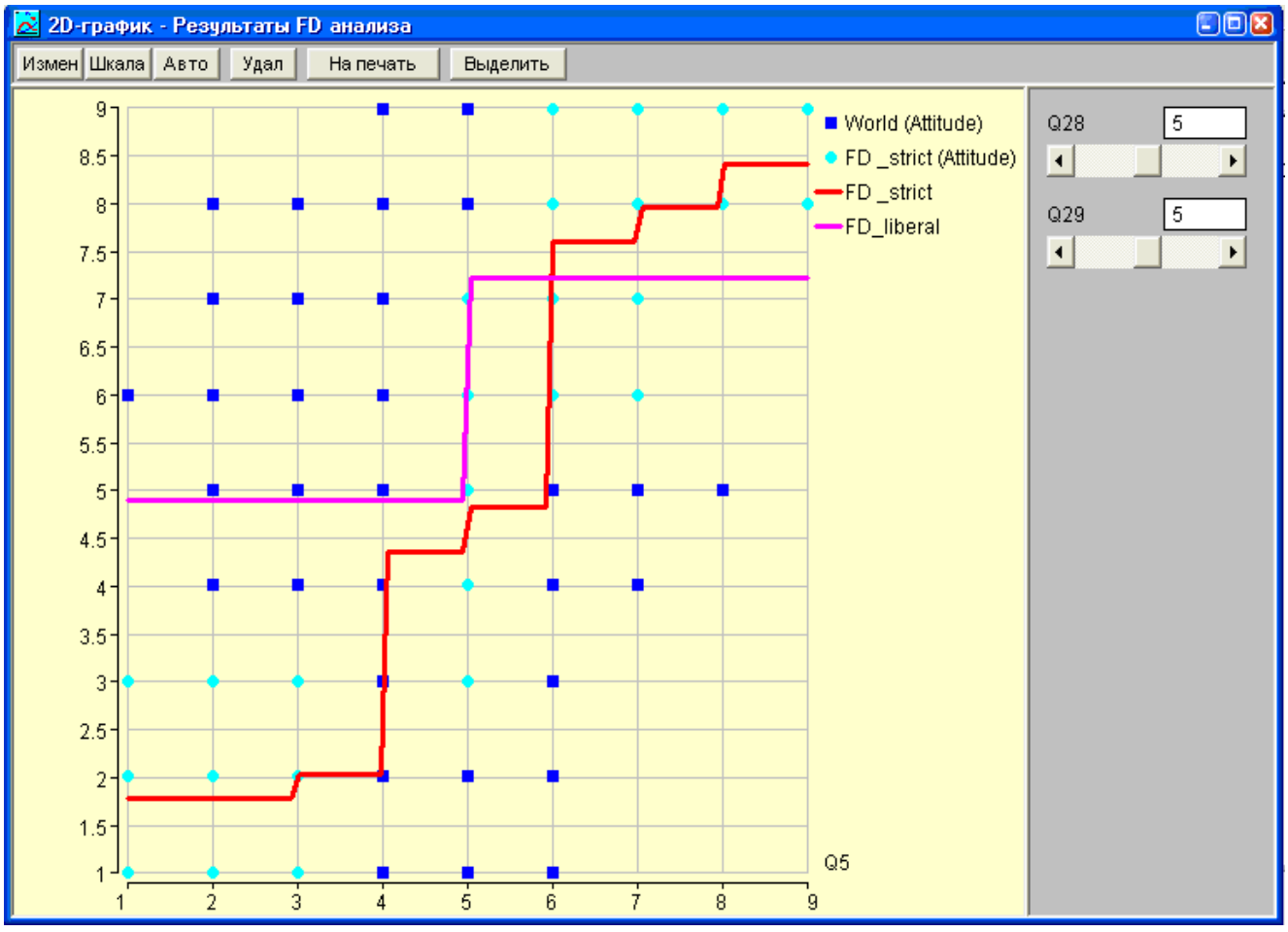


Рисунок 14.13 – 2D-результатов Поиска зависимостей

Введем имя и текстовый комментарий к графику, выберем атрибут по оси X, последовательно щелкая на кнопки "**+таблица**" и "**+правило**" включим таблицы **World** и **FD_strict**, а затем правила **FD_strict** и **FD_liberal**. В качестве атрибута по оси Y выберем целевую переменную attitude. Каждому объекту присвоим свои атрибуты визуализации: цвет и размер маркера тип линии и нажмем **OK**.

Теперь на одном графике мы видим точки данных из двух таблиц и одновременно графики двух табличных правил. Управляя линейками прокрутки, расположенными справа от графика, мы можем изменять значения других переменных, входящих в правило, моделируя графически эффект их влияния.

Исследование с помощью метода Поиск Законов. Это основной модуль системы **PolyAnalyst**, который способен строить и верифицировать на основе данных произвольные нелинейные многопараметрические модели. Удивительным свойством этого метода является то, что вид формул не задается пользователем, а находится автоматически программой в процессе анализа данных. Другой важнейшей чертой метода является корректная оценка значимости моделей, что позволяет решить так называемую проблему подгонки под данные (overfitting).

Для запуска метода как обычно выберем пункт меню **Исследовать/Поиск законов...**, дважды щелкнем на целевой переменной Attitude, введем ограничение времени 2 минуты и нажмем **OK**.

Этот метод работает дольше других и, в зависимости от размера таблицы, содержательности данных и производительности машины, анализ может занимать от нескольких минут до нескольких часов. В процессе работы метода Поиск Законов формирует отчеты каждый раз, когда обнаруживается новое значимое правило. Вы можете наблюдать этот живой процесс и видеть, как постепенно уточняются правила, вырабатываемые системой.

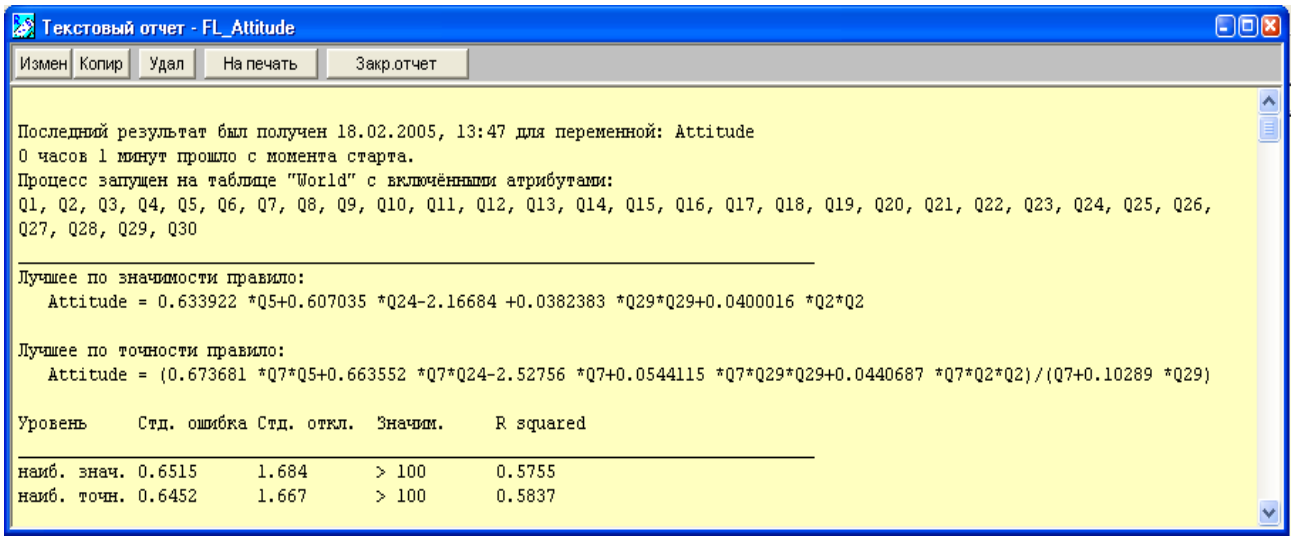


Рисунок 14.14 – Отчёт Поиска законов

Теперь займемся анализом отчета нашего метода. Текстовый отчет содержит два правила. Первое правило имеет гарантируемую значимость как для всего выражения, так и в отношении всех входящих в него членов. Второе правило не обязательно значимо или оно может содержать какие-то незначимые члены. Но точ-

ность второго правила в терминах стандартной ошибки выше. В двух строках ниже располагаются статистические оценки найденных правил. Мы видим, что значимость обоих правил очень высока (>100). В качестве меры значимости в этом методе используется специальный индекс, вычисляемый на основе рандомизированного тестирования. На практике, если значимость больше 3, то этому результату уже можно доверять.

Мы видим, что метод Поиск Законов выделил в качестве наиболее влияющих факторов те же самые переменные Q2, Q5, Q24 и Q29, что и линейный метод. Это является еще одним подтверждением, что эти факторы являются решающими. Примечательно, что переменные Q2 и Q29 входят в формулу во второй степени. Это говорит о том, что высокие значения Q2 и Q29 "усиливают" вклад этих факторов в результат. Построим 2D-график для наглядной визуализации найденных правил. Включим в этот график оба FL-правила, а также линейное правило. Мы видим, что графики всех трех правил весьма близки. Это говорит о том, что для этой задачи линейная модель является вполне адекватной.

Рисунок 14.15 – Графики поиска законов и линейной регрессии

В заключение этапа исследования сформируйте печатный отчет.

Исследование методом кластеризации. Во многих случаях бывает полезно выделять в базе данных компактные подгруппы записей (кластеры), проявляющие похожие свойства. Для этой цели в системе есть специальный вычислительный модуль, производящий поиск многомерных кластеров. В отличие от уже примененных методов этот метод не требует задания целевой переменной.

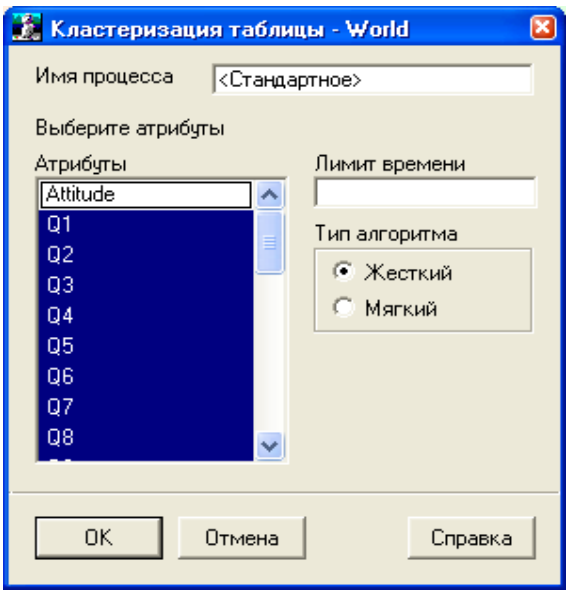


Рисунок 14.16 –Диалоговое окно Кластеризации

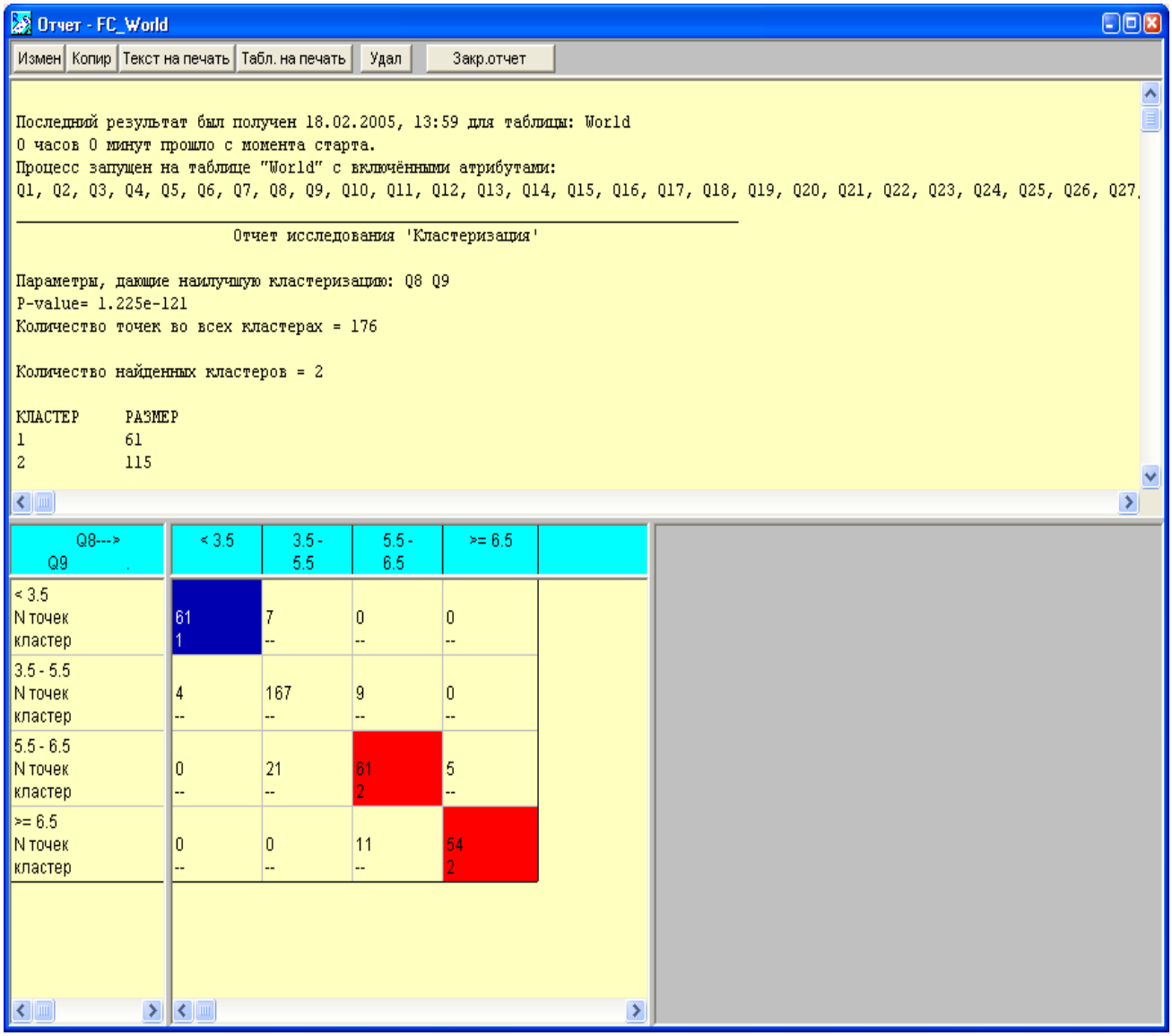


Рисунок 14.17 – Результаты Кластеризации

Для его запуска вызовем пункт меню **Исследовать/Кластеризация...**, в диалоге запуска метода выключим Attitude из исследования, выберем жесткую (strict) модификацию (с мягким алгоритмом вы можете поэкспериментировать самостоятельно) и нажмем **OK**. Этот метод работает быстро и буквально через несколько секунд мы увидим его отчет, который по форме похож на отчет *FD*-метода. Одновременно с отчетом наш метод формирует новые таблицы, в которые входят записи, отнесенные к одному кластеру. Как легко видеть вычислительный модуль обнаружил в данных два кластера, разделяемые по переменным Q8 и Q9.

Посмотрим что же это за утверждения?

- **Q8**: Я считаю, что в основном современное общество устроено правильно.
- **Q9**: У меня нет времени на благотворительность.

В первый кластер вошли люди, считающие, что общество в основном устроено неправильно и поэтому занимающиеся благотворительной деятельностью. Таких записей в нашей базе 61. Ко второму же кластеру отнесены люди, придерживающиеся противоположных взглядов (115 записей).

Интересно узнать отличаются ли эти группы по их отношению к Discovery? Для этого включим в таблицы кластеров целевой атрибут Attitude (кнопка **Измен** в открытой Таблице-FC_World) и вычислим общую статистику для этих таблиц (кнопка **Стат**). В результате получим таблицы, изображённые на рисунке 14.18.

Легко видеть, что у группы людей условно "совестливых" и склонных к благотворительности тяга к покупке Discovery несколько меньше, чем у их оппонентов, 4.6 против 5.4. Наглядно представить этот результат можно с помощью графика гистограмм. Щелкнем на пиктограмму создания гистограммы и заполним диалог.

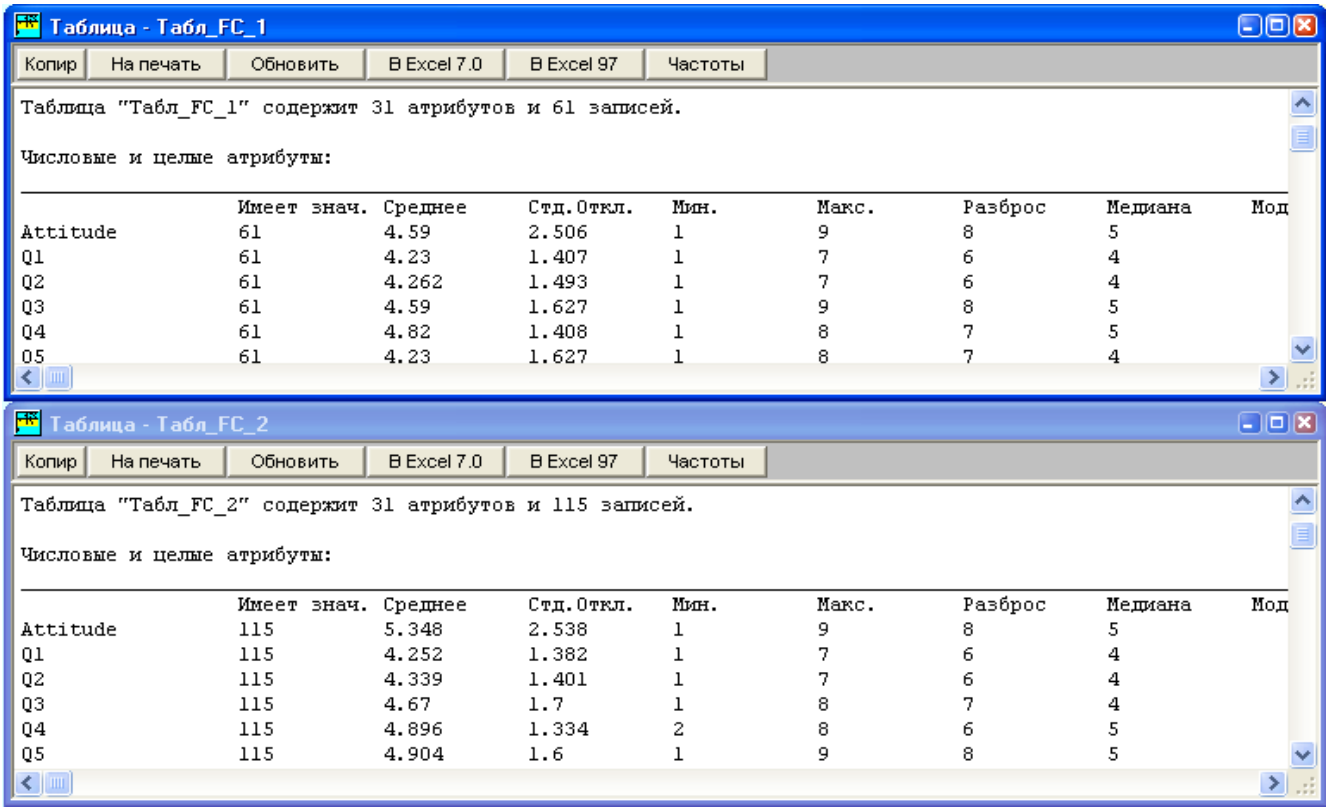


Рисунок 14.18 – Описательные статистики для найденных кластеров

Введем имя графика «Благотворительность», выберем таблицы **FC_world_1** и **FC_world_2**, а в качестве аргумента целевую переменную Attitude и нажмем **OK**. (Предварительно может быть полезным изменить цвета визуализации таблиц, с помощью функции редактирования таблиц.)

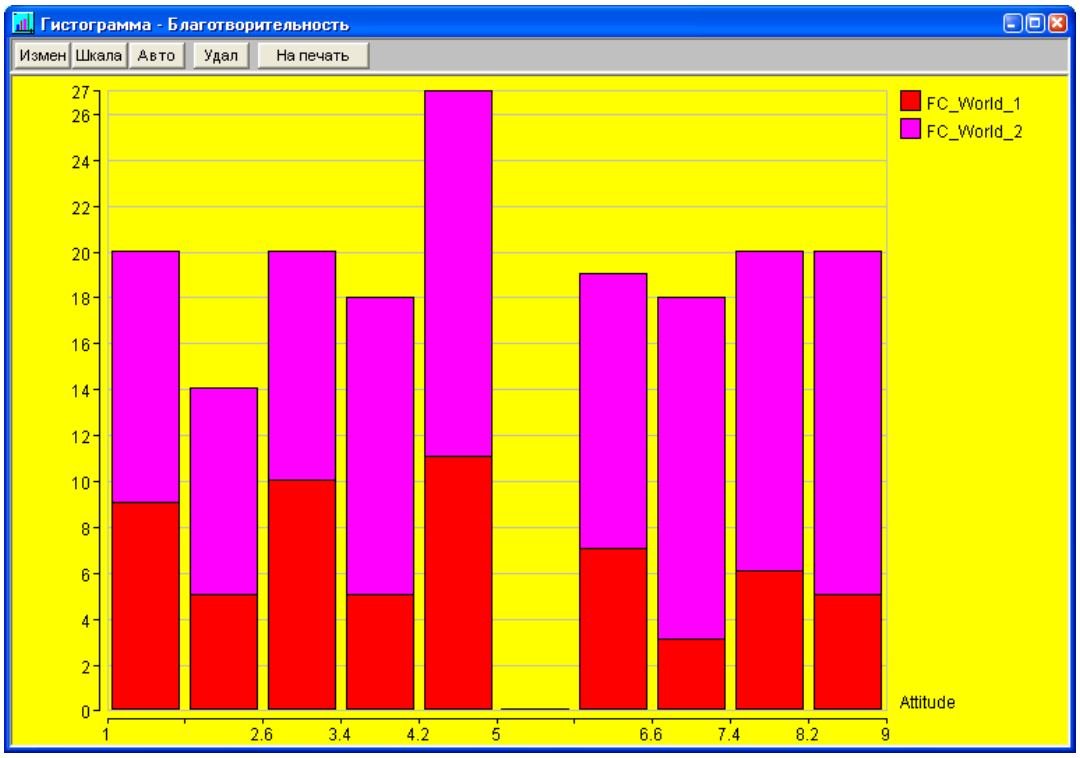


Рисунок 14.19 – Гистограммы кластеров по целевому параметру Attitude

Гистограмма (рис.14.19) наглядно показывает, что людей, поставивших высокий бал для целевой переменной и одновременно не интересующихся благотворительностью явно больше. Таким образом, мы выявили дополнительные интересные черты наших респондентов. Оформите результаты этого этапа в виде печатной формы.

Исследование методом классификации. В качестве последнего этапа исследования проведем анализ данных, используя вычислительный модуль, классифицирующий записи. Этот метод вырабатывает правило, которое позволяет отнести запись таблицы к одному или другому классу. Имеются две модификации этого метода. Собственно классификация, для которой целевая переменная должна иметь логический тип (логический 0 и логическая 1) и дискриминация, позволяющая определить чем и как данная таблица отличается от корневой таблицы **World**. Для дискриминации целевая переменная не нужна.

Сам процесс классификации (дискриминации) производится или Линейной Регрессией, или Поиском Законов.

Поскольку наши данные хорошо подчинятся линейным моделям, мы воспользуемся линейной формой дискриминации. Пусть нас будет интересовать отличия людей, поставивших бал 7 и более для целевого утверждения. Первым делом мы разобьем корневую таблицу в соответствии с этим условием. Щелкнув правой

кнопкой на пиктограмме **World**, вызовем пункт меню **Разбить на равные интервалы.** В поле **От**введем число 7, а в поле **Шаг**– число 10 (рис.14.20).

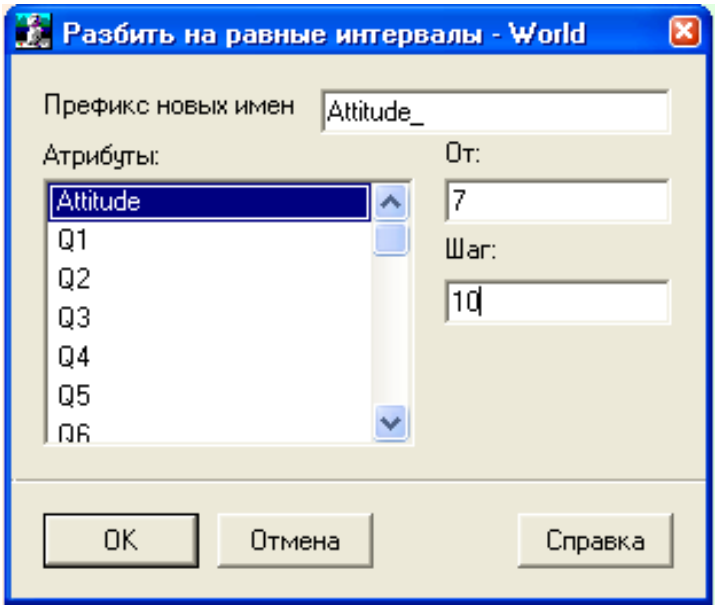


Рисунок 14.20 – Диалоговое окно разбиения на группы

В результате этой манипуляции мы получим две новых таблицы: **Attitude_1** и **Attitude_2**. Первая из них содержит записи, у которых Attitude меньше 7. Эта таблица на для дальнейшей работы не нужна, ее можно просто удалить. Рекомендуется изменить имя второй таблицы, чтобы оно стало более понятным, например "**Больше_7**". Кроме этого надо исключить целевую переменную из таблицы, иначе исследование не будет иметь смысла. После того как это предварительная работа закончена, можно начинать исследование.

Щелкнем правой кнопкой на пиктограмме таблицы **Больше_7** и вызовем пункт меню **Исследовать/Дискриминация**. В диалоге параметров процесса выберем тип процесса –**Линейная Регрессия** и нажмем OK.

Примерно через 1 минуту исследование будет закончено и можно будет проанализировать отчет.

Отчет содержит условие, предсказывающее выражение и количественные характеристики классификации. В нашем случае предсказывающее выражение линейно так как мы воспользовались линейным методом.

Если это предсказывающее выражение больше 0.539, то запись принадлежит таблице **Больше_7**, в противном же случае ее нельзя отнести к этой таблице.

Общая ошибка классификации составляет 16%, иначе говоря, правильная классификация выполняется с вероятностью 84%. Мы также видим, что ошибки классификации для классов 0 и 1 различаются. Более точно выявляются люди, не склонные покупать Discovery (ошибка 8%), те для этого класса людей мы ошибаемся в каждом 12-ом случае. Это важный практический результат, так как он позволяет более чем в 10 раз сократить объем прямой рекламы. Следует отметить высокую достоверность этого результата. Вероятность того, что он случаен составляет исключительно малое число, равное e^{-20}.

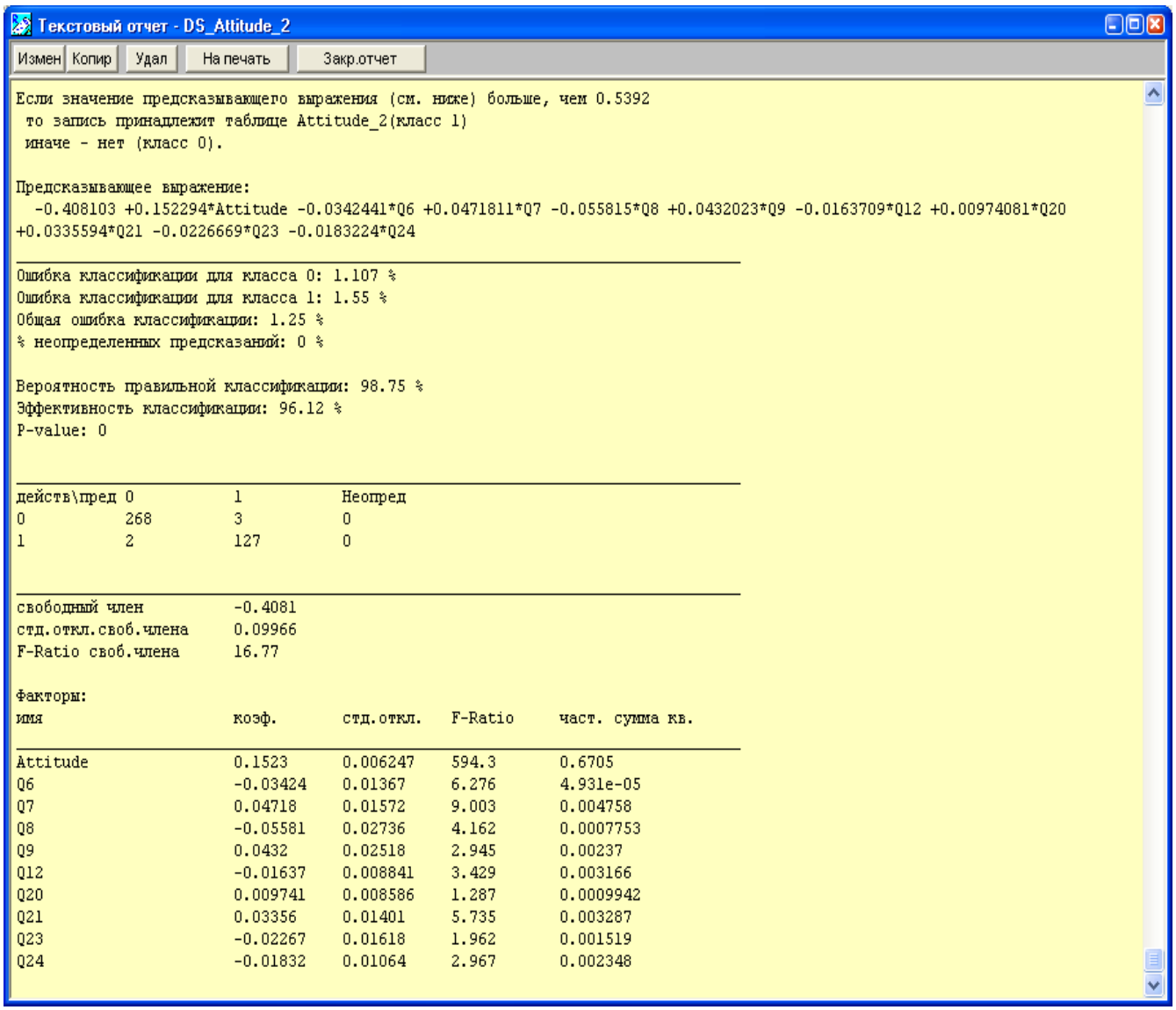


Рисунок 14.21 – Отчёт Классификации

В результате нашего исследования на рабочем месте системы **PolyAnalyst** появилось много новых объектов. Это новые таблицы, графики, правила, отчеты, печатные формы. Весь проект сохраняется в файле и можно продолжить дальнейшую работу в любое время.

Пример 3. Рассмотрим пример анализа стоимости жилья в г.Краснодаре в системе PolyAnalyst. Рассматриваются 1573 квартиры, выставленные на продажу в октябре 2006г. Цель исследования – установить имеющиеся закономерности в ценообразовании на рынке жилья (если таковые есть).

1) Для анализа файла о стоимости жилья в г.Краснодаре, который у нас есть в формате *.xls, необходимо сохранить его в формате *.csv. Для этого в Excel выполните команду **Файл – Сохранить как – Имя файла Nedvig.xls– Тип файла –*. CSV – Сохранить**. В результате сохранится файл Nedvig. csv. (Обратите внимание на то, что текстовые данные не должны содержать пробелов и наименования переменных не разделяться запятыми. Например, значение «40-лет Победы» и имя поля «Цена, тыс. руб.» – неверно. Должно быть: «40-летПобеды» и «Цена тыс. руб.» соответственно. Почему?)

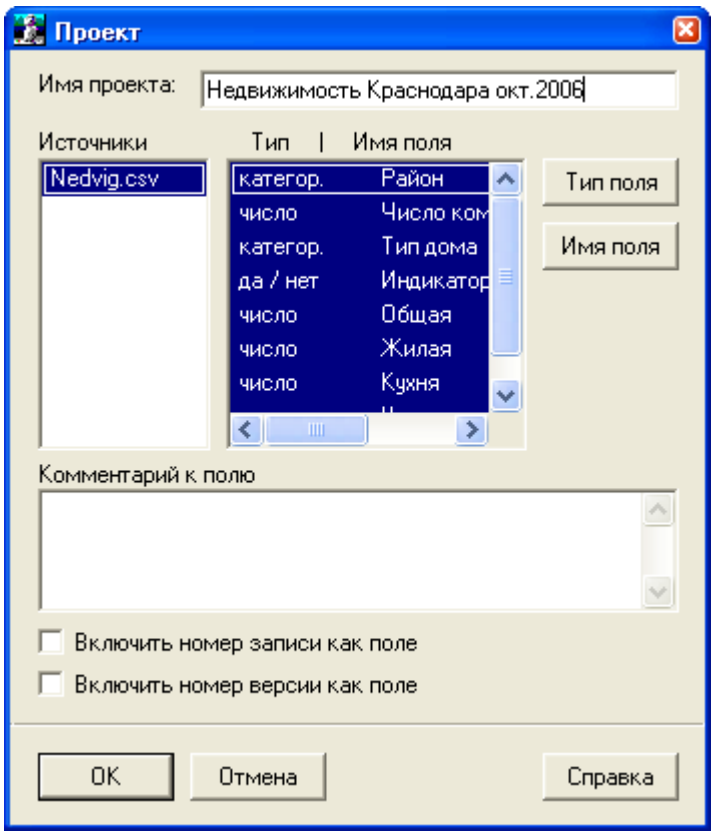


Рисунок 14.22 – Создание проекта недвижимость Краснодара октябрь 2006

Таблица - World

Измен | Нов | Разб | Иссл | Стат | Эксп | Частоты | Удал

No	Район	Число	Тип дома	Индикаторна	Общая	Жилая	Кухня	Цена тыс.
1	ЮМР	1	блочный	0	40	20	10	1596
2	ЮМР	2	блочный	0	60	32	10	1876
3	ЮМР	2	блочный	0	77	40	16	3500
4	ЮМР	3	блочный	0	66	42	9	2075
5	ЮМР	3	блочный	0	70	44	10	2156
6	СМР	3	кирпичный	1	77	46	10	2828
7	СМР	3	блочный	0	70	34	12	2436
8	ГМР	3	кирпичный	1	75	42	15	2489
9	ГМР	3	блочный	0	75	42	15	2489
10	СМР	3	кирпичный	1	72	42	9	2156
11	ЮМР	3	кирпичный	1	64	42	8	1970
12	ЮМР	3	блочный	0	67	42	12	2593
13	ЮМР	3	блочный	0	72	42	9	2178
14	ЮМР	3	блочный	0	74	43	12	2800
15	ЮМР	3	блочный	0	70	50	9	3248
16	ЮМР	3	блочный	0	70	40	8	2576
17	ЮМР	3	блочный	0	70	50	9	3164
18	ЮМР	3	блочный	0	70	44	9	2128
19	ЮМР	3	блочный	0	75	42	15	2541
20	СМР	3	кирпичный	1	72	43	10	2937
21	СМР	3	блочный	0	64	42	9	2904
22	СМР	4	блочный	0	123	70	26	3828
23	СМР	4	блочный	0	80	52	9	2520

Рисунок 14.23 – Таблица исходных данных для анализа –World

С помощью контекстного меню папки World разобьём исходные данные на равные интервалы (рис.14.24).

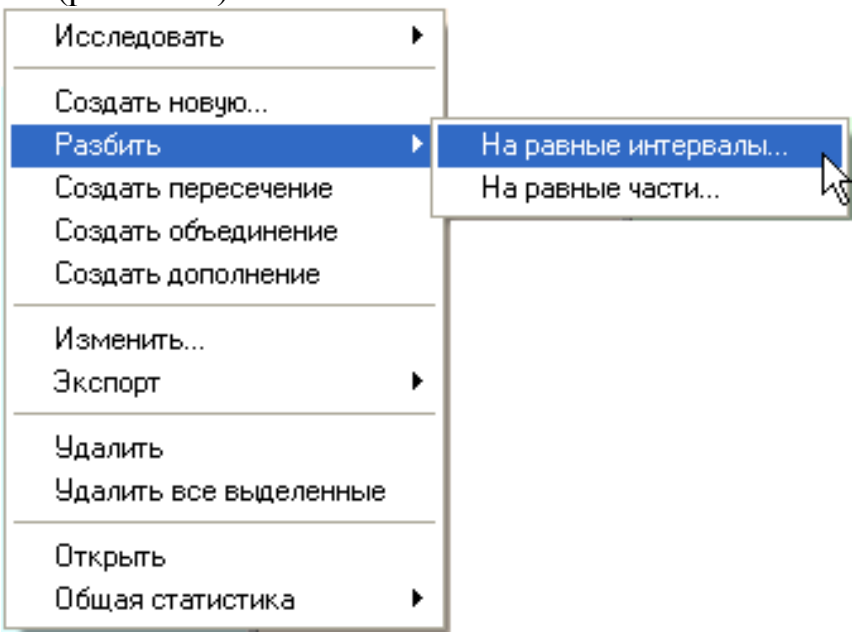


Рисунок 14.24 – Контекстное меню папки World

Признак разбиения – число комнат (рис.14.25).

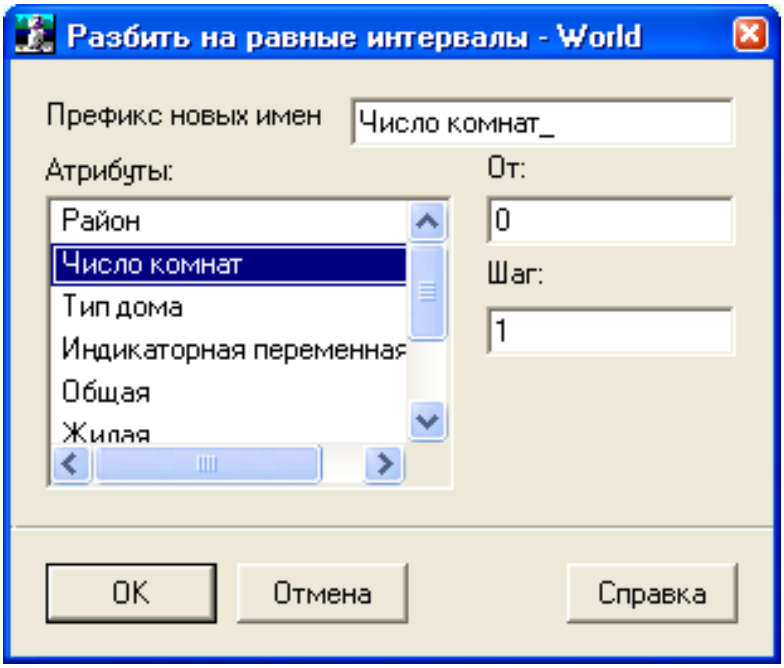


Рисунок 14.25 – Разбиение файла на части по числу комнат

В дальнейшем мы будем работать с таблицей «Число комнат_1», получившейся после разбиения исходной таблицы World.

2) Активировав таблицу «Число комнат_1», запустим процесс многопараметрической линейной регрессии с включёнными атрибутами: Район, Тип дома,

Общая площадь, Жилая площадь, Площадь кухни. В качестве целевого атрибута рассматривалась цена.

Получившееся правило на 70,35% объясняет исходные данные с индексом значимости 87,02 – это означает, что найденное правило можно считать приемлемым. Найденная зависимость показывает, что увеличение площади на 1 квадратный метр ведёт к увеличению стоимости однокомнатной квартиры: Общей – на 10,3399 тыс.руб.; Жилой – на 18,5296 тыс.руб.; Кухни – на 27,3059 тыс.руб. Если район:

– СМР, то это добавляет к стоимости квартиры 104, 849 тыс. руб.;

– ЗИП, то это добавляет к стоимости квартиры 56,114 тыс. руб.;

– КМР, то это добавляет к стоимости квартиры 118,624 тыс.руб.

Наиболее значимым фактором является общая площадь квартиры (рис.14.26). Следует отметить, что указанные районы выделены вследствие того, что исходные данные в основном содержат 1 комнатные квартиры из этих районов.

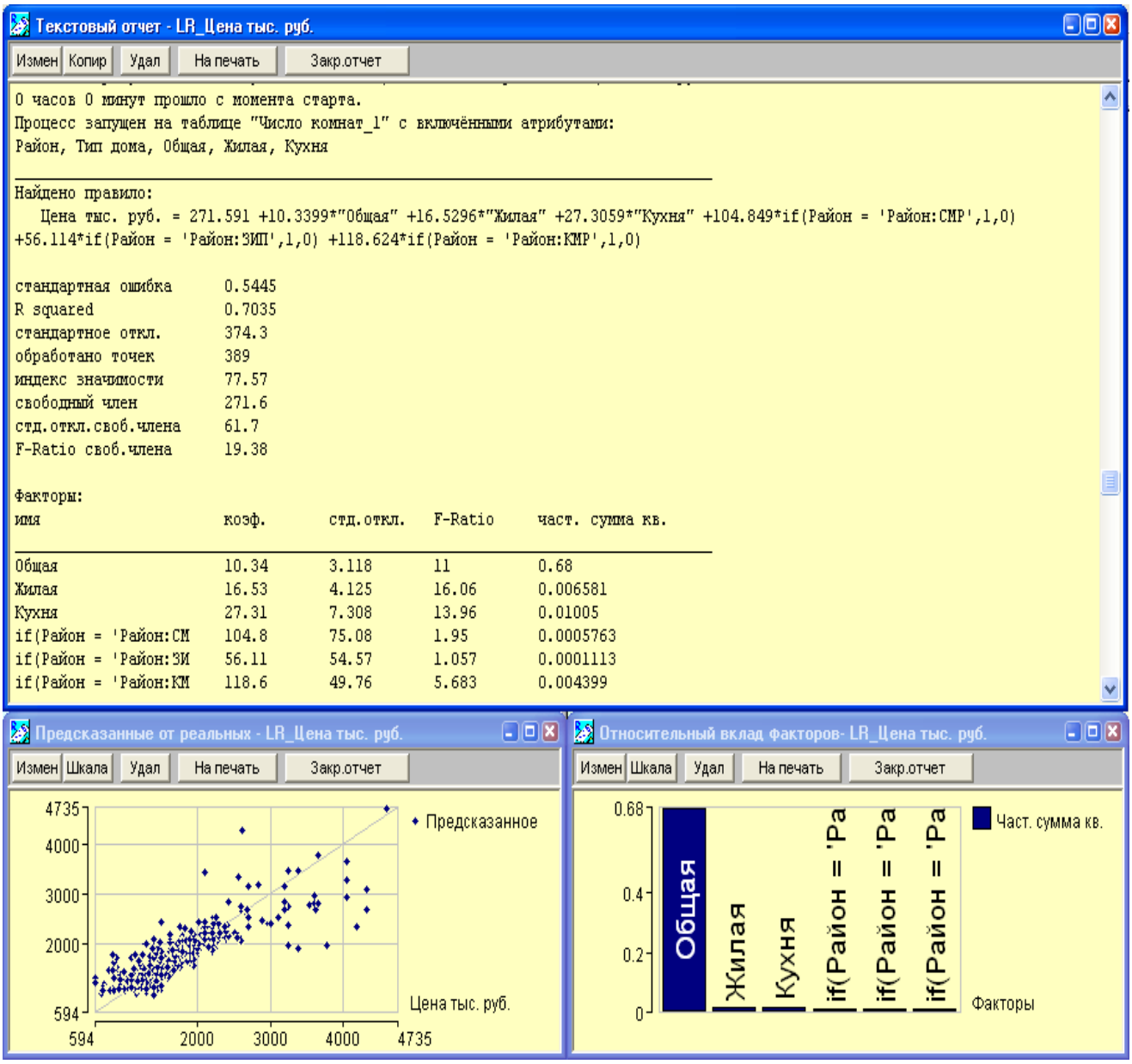


Рисунок 14.26 – Отчёт LR

3) Рассмотрим гистограмму наиболее значимого фактора – общей площади (рис.14.27).

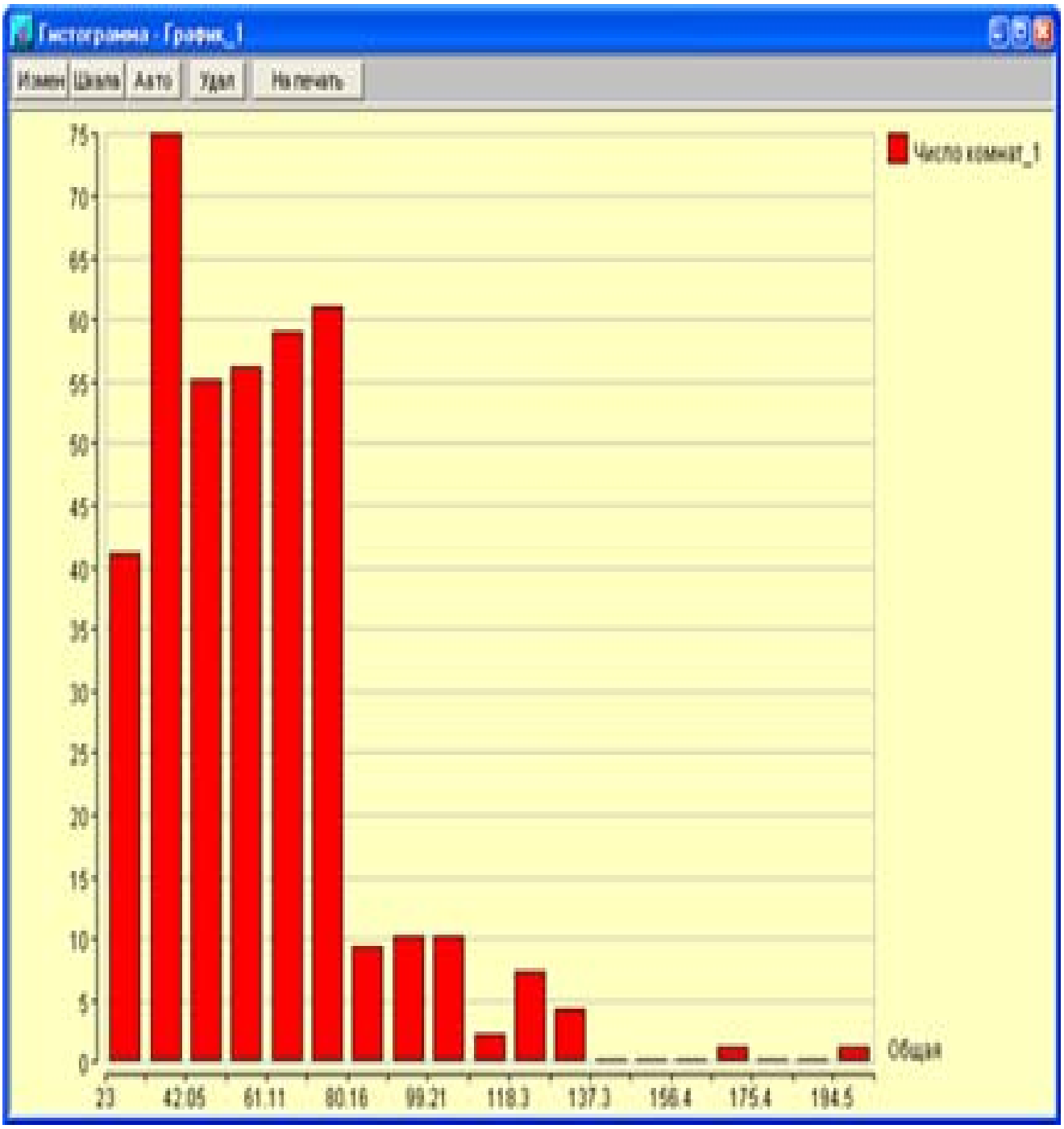


Рисунок 14.27 – Гистограмма Общей площади

Из графика гистограммы, очевидно, что наиболее представительны квартиры с площадью мене 80м^2. С помощью кнопки f(x) создадим два правила: 1) Правило_1 – общая площадь не более 80 м^2 , 2) Правило_2 общая площадь более 80 м^2.

Исследование методом **поиска зависимостей** (рис. 14.28).

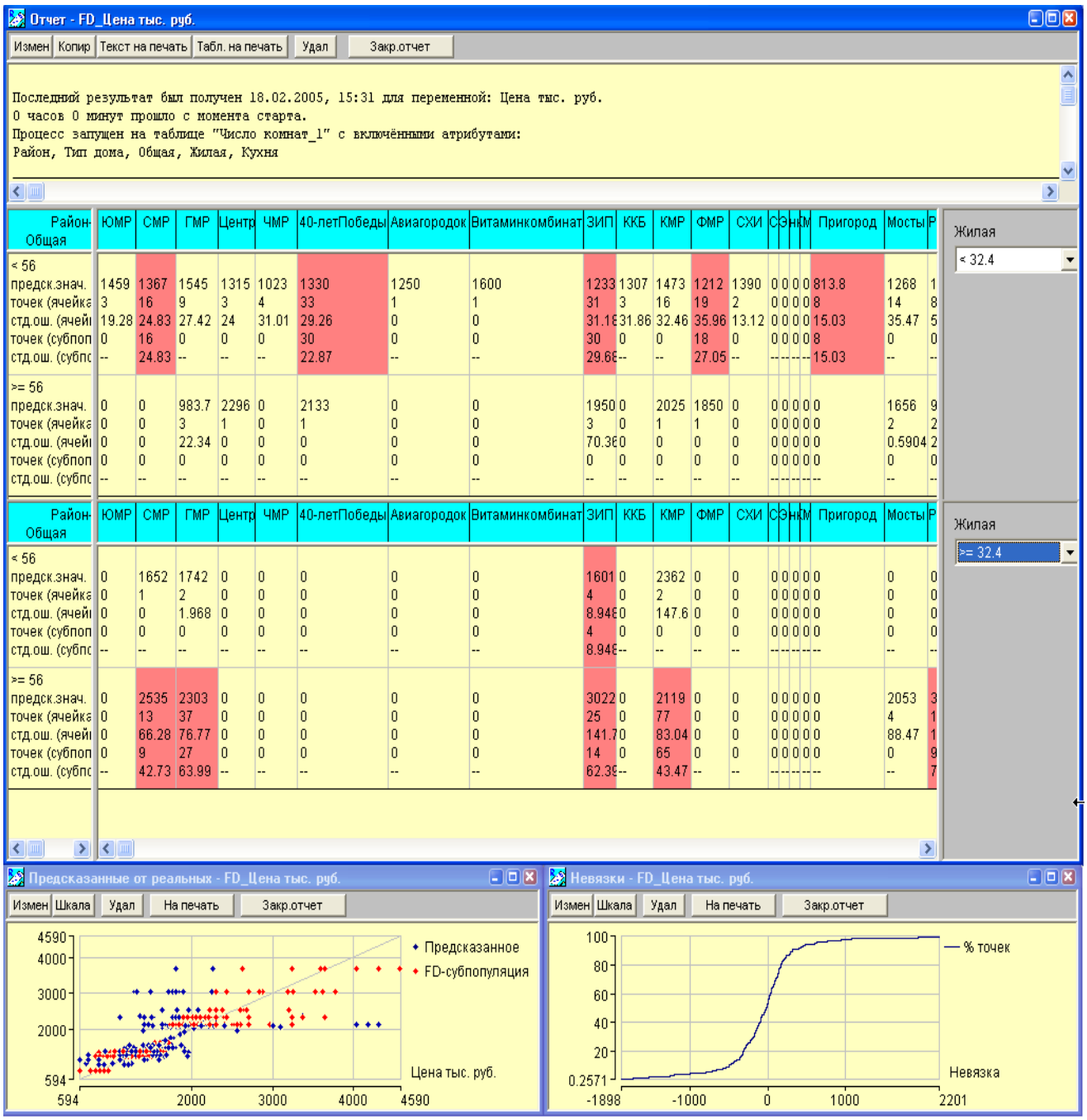


Рисунок 14.28– Отчёт процесса Поиск зависимостей

Наиболее значимыми для оценки стоимости 1 комнатных квартир оказались переменные: Район, Общая площадь (до 56 и больше или равно 56 м^2), Жилая площадь (до 32,4 и больше или равно 32,4 м^2). Этому правилу подчиняются более 59% данных. (Проведите поиск зависимостей с использованием Правила_1 и Правила_2 – с помощью контекстного меню Правил примените их к выделенным таблицам и сделайте выводы!).

Исследование с помощью **кластеризации**. После запуска процесса кластеризации (с включёнными атрибутами: Район, Тип дома, Общая площадь, Жилая площадь, Площадь кухни, Цена – мы получим, что наши данные разбиты на три кластера, характеризующихся различными соотношениями Жилой площади и площади Кухни – рисунок 14.29.

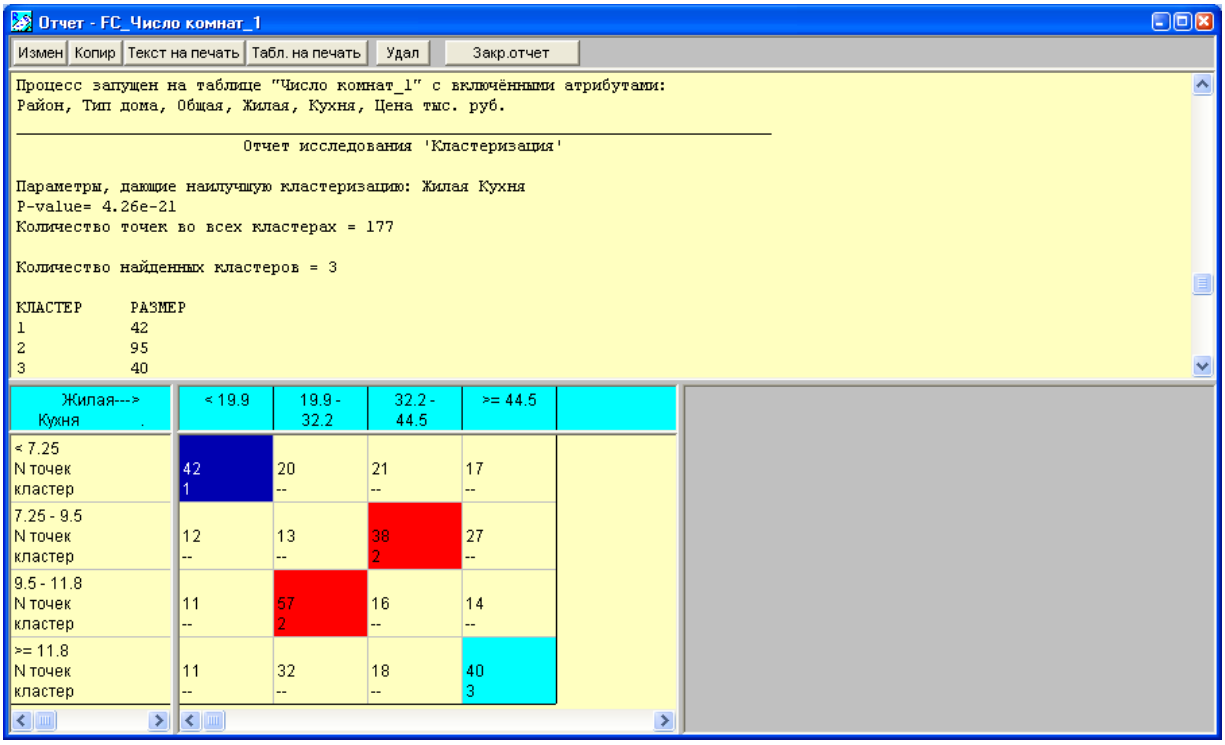


Рисунок 14.29 –Кластеризация данных о стоимости 1 комнатных квартир (жёсткий алгоритм)

Найденную кластеризацию вряд ли можно считать успешной, так как ей подчиняются всего 177 наблюдений из 389. В найденных кластерах естественно попытаться найти зависимость. Так как наиболее значимыми при кластеризации оказались Жилая площадь и площадь Кухни, то запустите процессы линейной регрессии для каждого из кластеров с этими атрибутами и сделайте выводы. Запустим процесс кластеризации с мягким алгоритмом, получим рисунок 14.30.

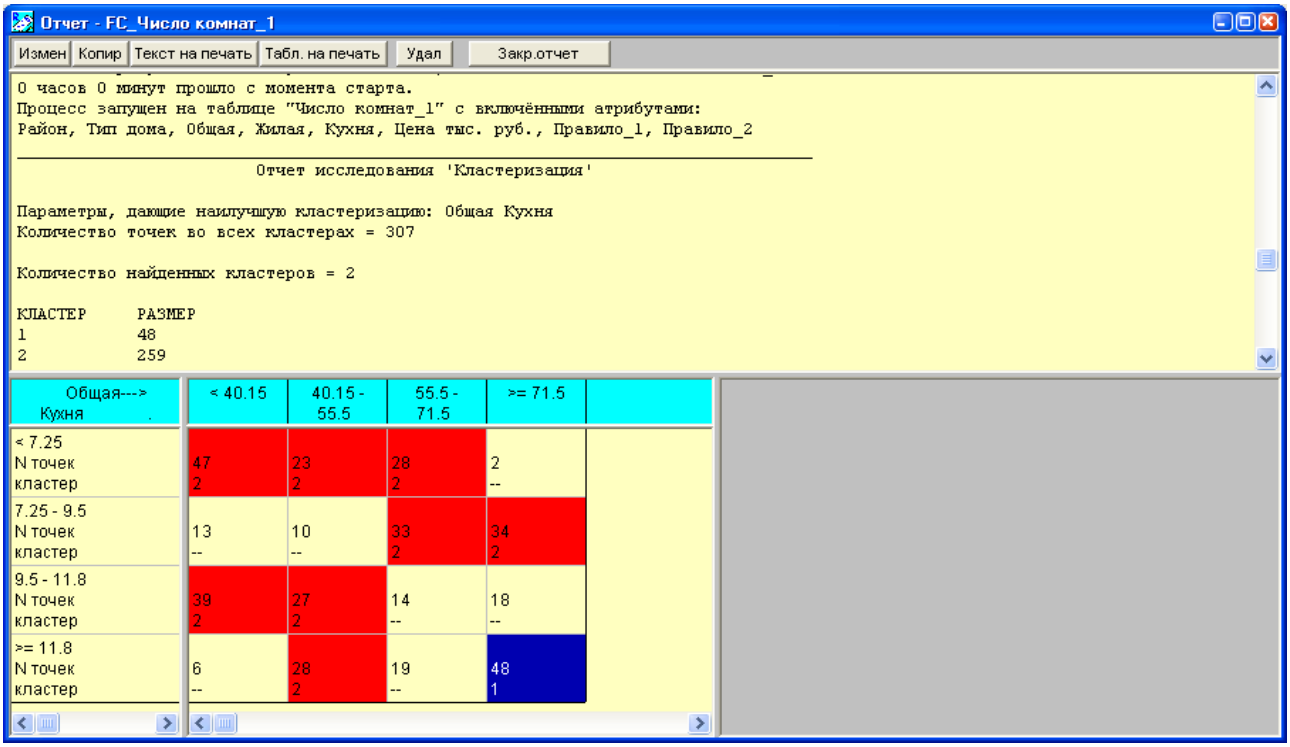


Рисунок 14.30 – Мягкий алгоритм кластеризации

Мягкий алгоритм кластеризации, запущенный практически со всеми переменными и правилами, заданными нами выше, позволил получить разбиение данных на 2 класса. Причём значимыми переменными в этот раз являются Общая площадь и Площадь кухни. Количество точек в кластерах 307 из 389 – это хороший результат. С помощью контекстных меню FC-таблиц рассмотрим описательные статистики – рисунок 14.31.

Таблица "FC_Число комнат_1_1" содержит 8 атрибутов и 48 записей.

Числовые и целые атрибуты:

	Имеет знач.	Среднее	Стд.Откл.	Мин.	Макс.	Разброс	Медиана	Мода
Общая	48	100.9	27.17	72	204	132	91	
Жилая	48	57.83	17.91	26	111	85	54	
Кухня	48	16.82	4.438	12	30	18	16	
Цена тыс. руб.	48	2836	750.6	1500	4590	3090	2619	

Логические (да/нет) атрибуты:

	Имеет знач.	Число 1	Число 0
Правило_1	48	10	38
Правило_2	48	33	15

Категориальные атрибуты:

	Имеет знач.	Мода	Разл.знач.
Район	48	Рос-кая	9
Тип дома	48	каркасный	3

Таблица "FC_Число комнат_1_2" содержит 8 атрибутов и 259 записей.

Числовые и целые атрибуты:

	Имеет знач.	Среднее	Стд.Откл.	Мин.	Макс.	Разброс	Медиана	Мода
Общая	259	51.24	15.51	23	100	77	49	
Жилая	259	29.88	12.63	8	70	62	23	
Кухня	259	8.783	2.613	5	19	14	9	
Цена тыс. руб.	259	1556	448.9	594	3375	2781	1498	

Логические (да/нет) атрибуты:

	Имеет знач.	Число 1	Число 0
Правило_1	259	253	6
Правило_2	259	6	253

Категориальные атрибуты:

	Имеет знач.	Мода	Разл.знач.
Район	259	КМР	16
Тип дома	238	кирпичный	4

Рисунок 14.31 – Описательные статистики кластеров, полученных с помощью мягкого алгоритма

Первый кластер содержит 48 значений квартир в каркасных домах со средней площадью 100,9 ($м^2$) в основном, удовлетворяющих Правилу_2 (общая площадь более 80) с достаточно большим разбросом значений. Второй Кластер имеет 259 значений со средней площадью 51,24 ($м^2$) из которых 253 удовлетворяют Правилу_1 и большинство находятся в кирпичных домах (наибольшее значение моды).

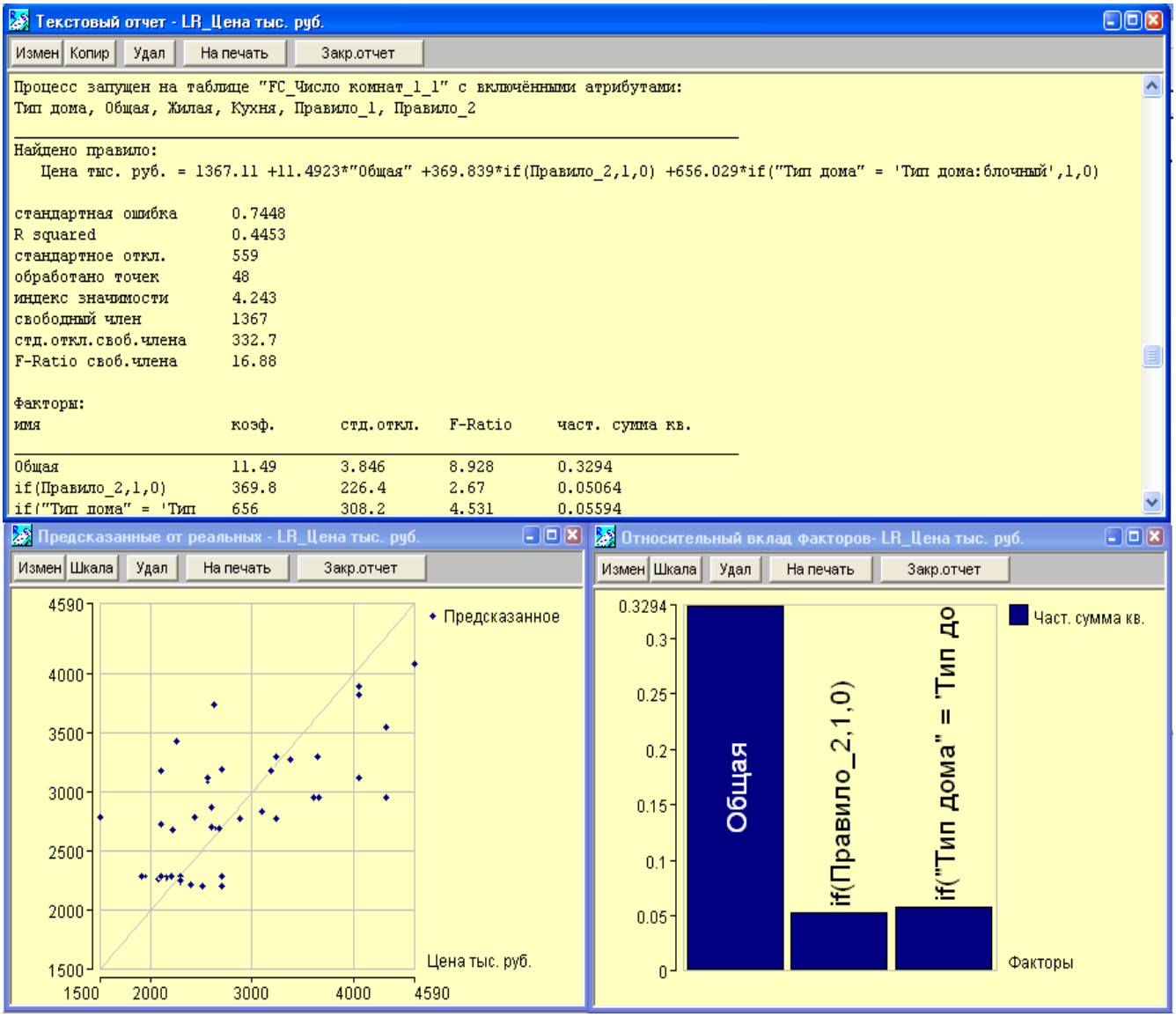


Рисунок 14.32 – Результат применения LR для 1 кластера

Для первого кластера многопараметрическая линейная регрессия (рис. 14.32) показывает, что каждый метр общей площади увеличивает цену на 11,4923 тыс. руб. Если площадь более 80 $м^2$ (Правило_2), то цена возрастает на 369,839 тыс.руб. Если тип дома блочный , то это добавляет к цене 656,029. Малые значения Rsugared (0,4453) и индекса значимости 4,243 говорят о малой значимости полученного результата.

Найденное правило отображает зависимость в имеющихся данных и в силу того, что используемая выборка мала – не может отражать истинное положение на рынке жилья.

Для второго кластера (рис. 14.33) увеличение общей площади на 1 прибавляет к стоимости 21,7411 тыс.руб, уменьшение площади кухни на 1м отнимает от стоимости 11,0166 тыс. руб. Если площадь менее 80 м^2 (Правило_1), то цена падает на 227,9 тыс.руб. Значения Rsugared (0,6277) и индекса значимости 77,28 говорят о значимости полученного результата. Построим так же для найденных кластеров модели поиска законов (лимит времени 1 минута). Найденные правила изобразим при помощи 2D графика – рисунок 14.34.

Замечание.

1. Следует отметить, что найденные правила могут быть улучшены при увеличении файла исходных данных и отбрасывании 1 комнатных квартир с большой площадью.
2. Численные значения некоторых результатов могут отличаться в зависимости от параметров используемого компьютера.
3. Идеология эволюционных алгоритмов моделирования, используемая в системе PolyAnalist, оперирующая с популяциями частных моделей исследуемой системы восходит к 60-м годам XX века – Дж. Фогелю (эволюционное моделирование) и академику А.Г. Ивахненко (Метод группового учёта аргументов – МГУА). В настоящее время она продолжает развиваться и использоваться не только при анализе табличных данных, но и, например, в теории идентификации сложных нелинейных систем большой размерности [Пащенко Ф.Ф. Введение в состоятельные методы моделирования систем. Ч.2. Идентификация нелинейных систем. – М.: Финансы и статистика, 2007.]

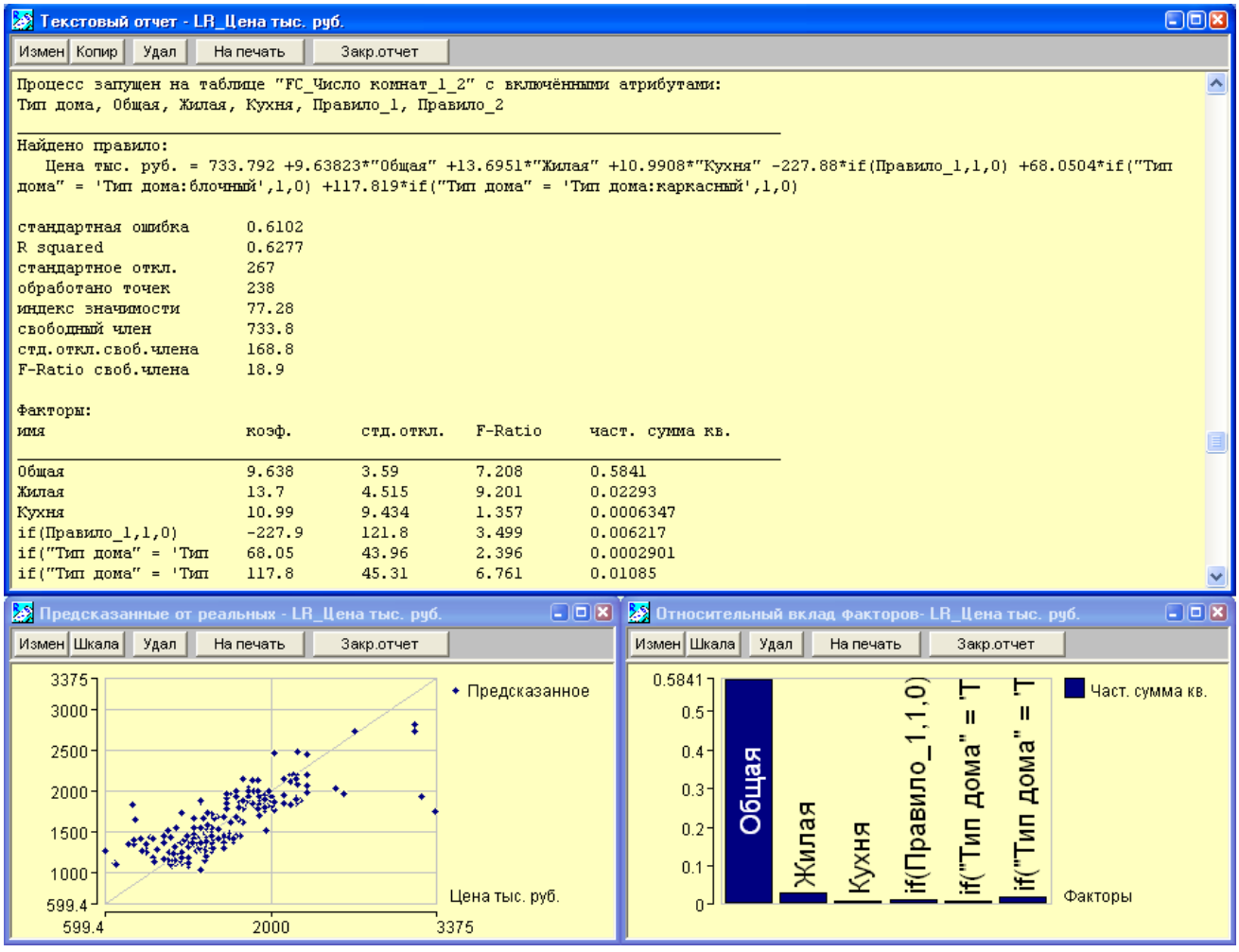


Рисунок 14.33 – Результат применения LR для 2 кластера

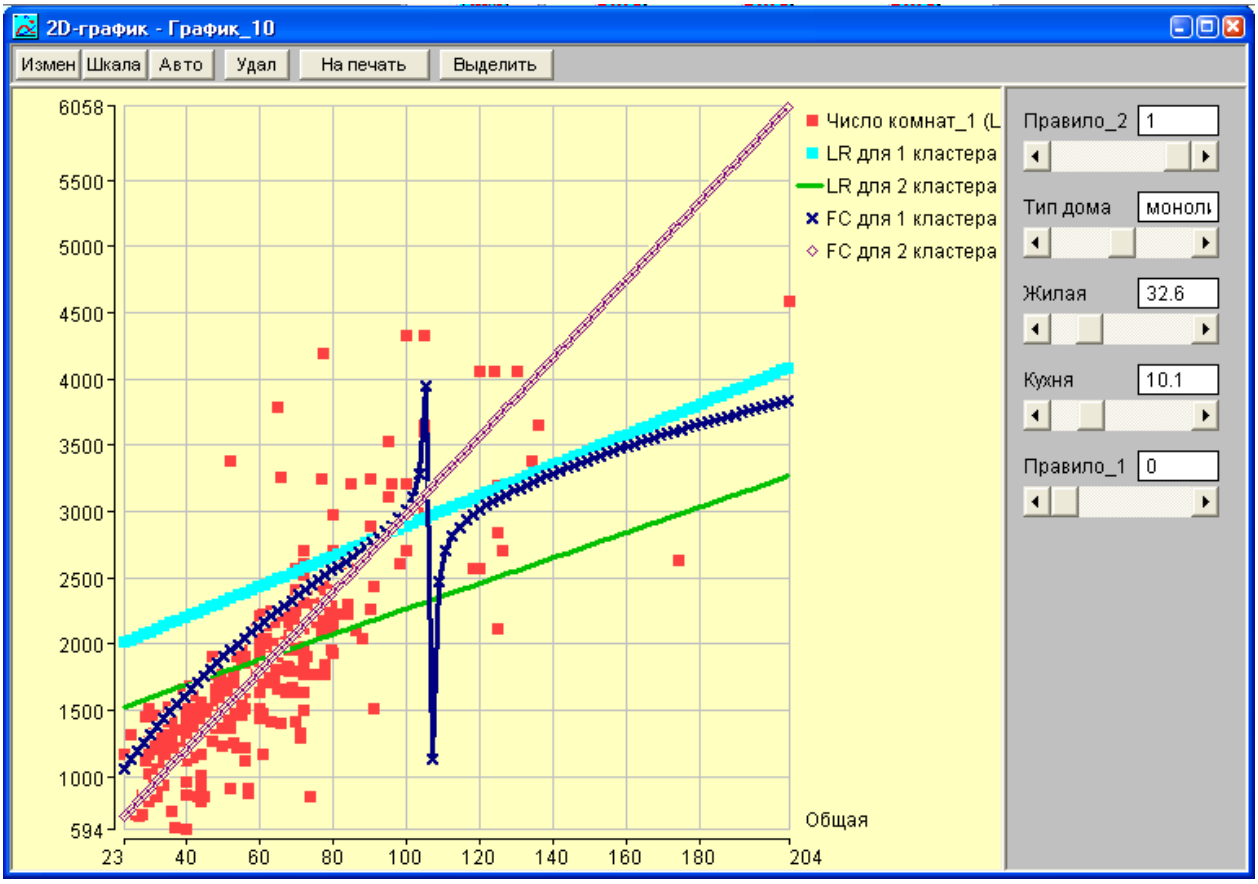


Рисунок 14.34 – 2D– график, найденных правил и исходных данных

Задание. В соответствии со своим вариантом без учёта района (см. работу №8), проанализируйте данные о стоимости жилья.

Вопросы для самоконтроля

- Опишите основные принципы идеологии эволюционного программирования.
- Какие инструменты анализа (процессы) используются в третьей версии PolyAnalyst? Поясните особенности их работы и задачи, которые они позволяют решать.
- Что общее и в чём отличие методов, реализованных в инструментах анализа системы PolyAnalyst и соответствующих классических методов.

Практическое занятие №15

Знакомство с аналитической платформой Deductor. Хранилища данных

Из-за огромного количества информации очень малая часть ее будет когда-либо увидена человеческим глазом. Наша единственная надежда понять и найти что-то полезное в этом океане информации – широкое применение методов Data Mining.
Г. Пятецкий-Шапиро

Цель работы – ознакомиться с архитектурой, основными частями и пользовательским интерфейсом Deductor, получить навыки создания сценариев обработки и визуализации данных, создания и наполнения хранилища данных.

Теоретические сведения

Deductor – это аналитическая платформа класса KDD (Knowledge Discovery in Databases) и Data Mining, основа для создания законченных прикладных решений в области анализа данных.

Реализованные в Deductor технологии позволяют на базе единой архитектуры пройти все этапы построения аналитической системы от создания хранилища данных до автоматического подбора моделей и визуализации полученных результатов, в частности, в виде OLAP кубов, таблиц, диаграмм, гистограмм, карт, графов и т.д.

Аналитическая платформа Deductor состоит из трех компонентов:

1. Многомерного хранилища данных Deductor Warehouse;
2. Аналитического приложения Deductor Studio;
2. Средства тиражирования знаний Deductor Viewer.

Deductor Warehouse – многомерное хранилище данных, аккумулирующее из разных источников всю необходимую для анализа предметной области информацию. Использование единого хранилища позволяет обеспечить непротиворечивость данных, их централизованное хранение и автоматически обеспечивает всю необходимую поддержку процесса анализа данных. Deductor Warehouse оптимизирован для решения именно аналитических задач, что положительно сказывается на скорости доступа к данным.

Deductor Studio – это программа, предназначенная для анализа информации из различных источников данных. Она реализует функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать и без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является совместное использование с Deductor Warehouse.

В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обра-

ботки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону.

Deductor Viewer – это облегченная версия Deductor Studio, предназначенная для отображения построенных в Deductor Studio отчетов. Она не включает в себя механизмов создания сценариев, но обладает полноценными возможностями по их выполнению и визуализации результатов.

Deductor Viewer является средством тиражирования знаний для конечных пользователей, которым не требуется знать механику получения результатов или изменять способы их получения.

Взаимодействие компонентов аналитической платформы Deductor проиллюстрировано на рисунке 9.1.

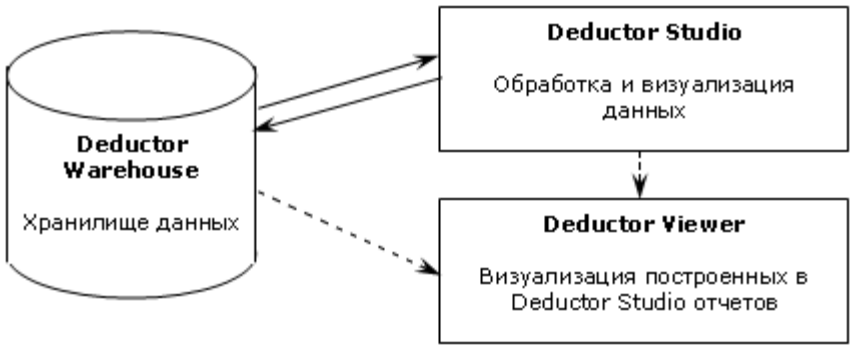


Рисунок 15.1 – Компоненты платформы Deductor

Вся работа в Deductor сводится к использованию 6 Мастеров:

- Мастер подключений;
- Мастер импорта;
- Мастер обработки;
- Мастер визуализации;
- Мастер экспорта.

Deductor не имеет собственных средств для ввода данных. Предполагается, что данные в табличном виде находятся в каком-то источнике. Мастер импорта и экспорта обеспечивают взаимодействие со всеми возможными источниками и приемниками данных, для которых существуют стандартные механизмы доступа (ODBC, ADO и т.п.).

Обработка и *визуализация* – еще две атомарные операции с данными в Deductor. Под обработкой понимаются любые манипуляции над набором данных: от самых простых (например, сортировка) до сложных (построение модели нейронной сети). Обработчик можно представить в виде «черного ящика», на вход которого подается набор данных, а на выходе формируется преобразованный набор данных (рис. 15.2).

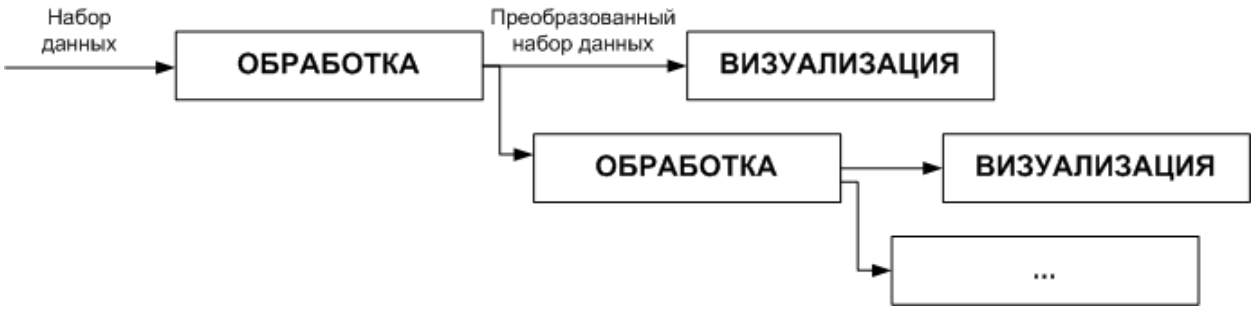


Рисунок 15.2 – Обработка и визуализация

Реализованные в Deductor обработчики покрывают основную потребность в анализе данных и произвольном манипулировании данными (очистка, слияние, объединение, фильтрация).

Любой набор данных можно визуализировать каким-либо доступным способом или несколькими способами, поскольку визуализация помогает интерпретировать построенные модели.

Сценарий представляет собой иерархическую последовательность обработки и визуализации наборов данных (дерево). Сценарий реализует встроенный язык визуального моделирования и состоит из узлов. Сценарий всегда начинается с импорта набора данных из произвольного источника. После импорта может следовать произвольное число обработчиков любой степени глубины и вложенности. \

Каждой операции обработки соответствует отдельный узел дерева, или объект сценария. Набор данных служит механизмом, соединяющим все объекты сценария. Можно сказать, что сценарий – наиболее естественный с точки зрения аналитика способ представления этапов построения модели. Это позволяет быстро создавать модели, обладающие большой гибкостью и расширяемостью, сравнивать несколько моделей.

На рисунке 15.3 изображен пример сценария.

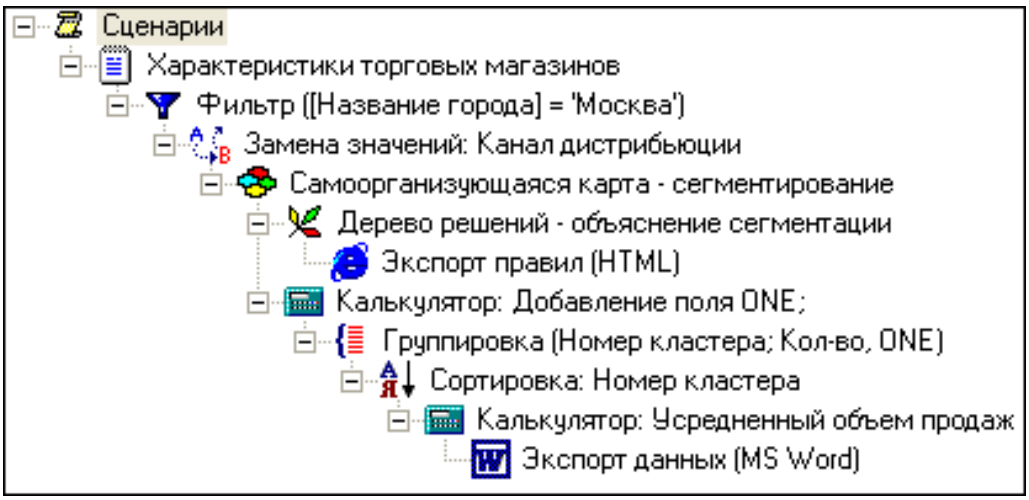


Рисунок 15.3 – Пример сценария в Deductor

Непосредственно для работы со сценарием используются три мастера. Рассмотрим их подробнее.

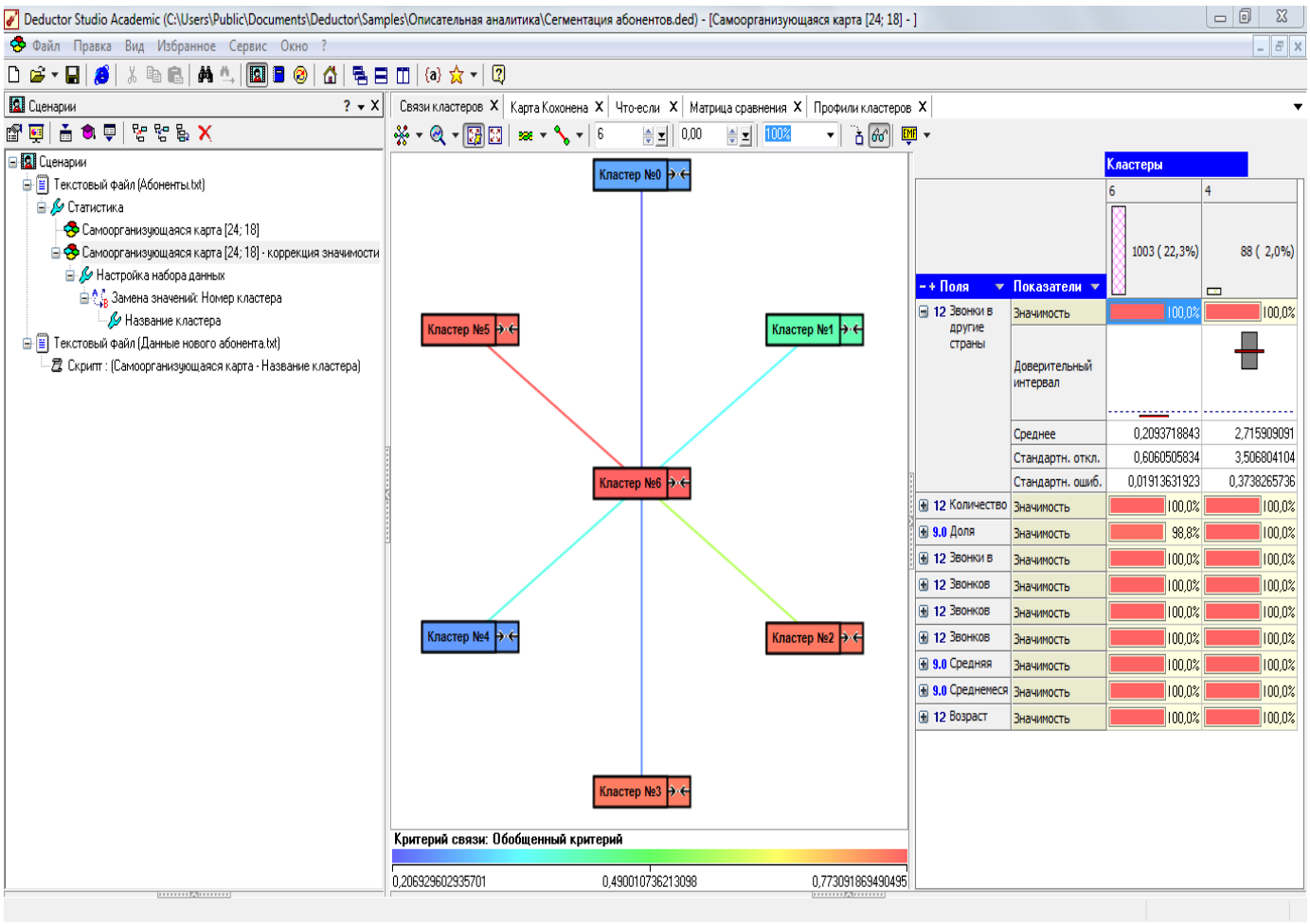


Рисунок 15.4 – Главное окно Deductor Studio 5.3

Мастер импорта предназначен для автоматизации получения данных из любого источника, предусмотренного в системе. На первом шаге мастера импорта открывается список всех предусмотренных в системе типов источников данных. Число шагов мастера импорта, а также набор настраиваемых параметров отличается в зависимости для разных типов источников.

Мастер обработки настраивает параметры выбранного узла-обработчика.

Мастер визуализации позволяет в пошаговом режиме выбрать и настроить наиболее удобный способ представления данных. В зависимости от узла, в результате которого была получена ветвь сценария, список доступных для него видов отображений будет различным. Например, после построения деревьев решений их можно отобразить с помощью визуализаторов «Деревья решений» и «Правила». Эти способы отображения не доступны для других обработчиков.

Мастер экспорта позволяет в пошаговом режиме выполнить экспорт данных в файлы и базы данных наиболее распространенных форматов, в том числе и в Deductor Warehouse.

Для настройки подключений к любым внешним источникам и приемникам данных используется *Мастер подключений*.

Интерфейс главного окна Deductor Studio 5.3 приведен на рисунке 15.4. Слева расположено окно с вкладками «Сценарии» (для визуального моделирования

потоков данных и узлов), «Отчеты» и «Подключения» (список доступных подключений). Справа отображаются визуализаторы.

Веб-сервисы

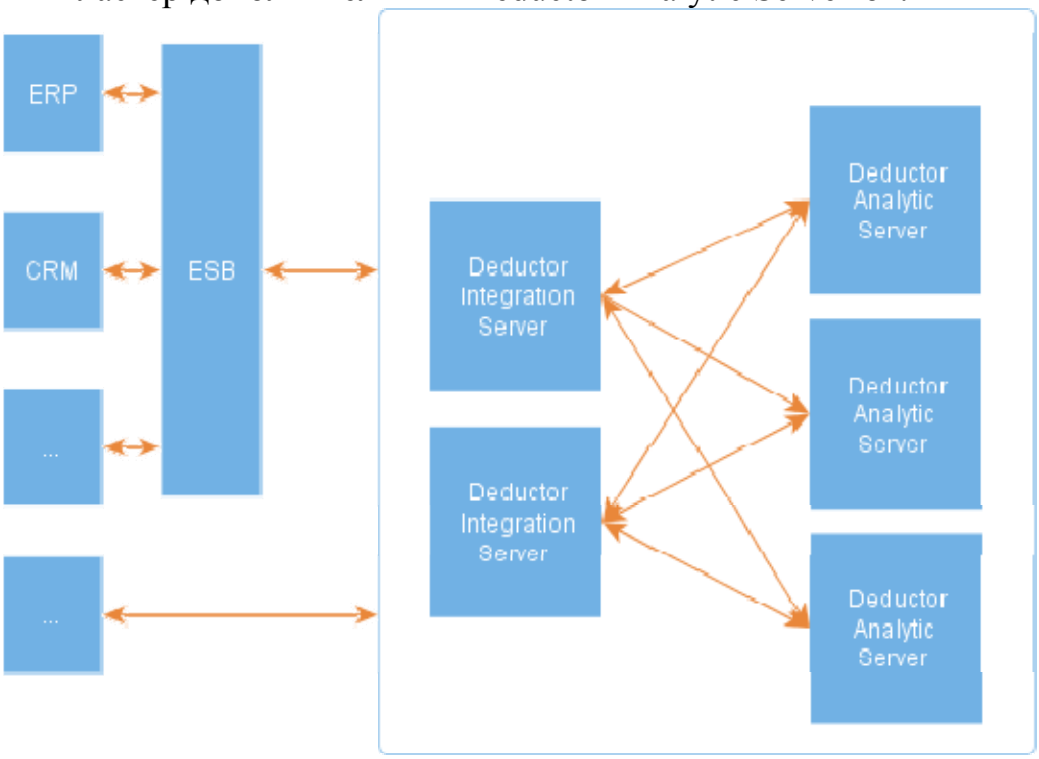
Deductor 5.3 позволяет полноценно интегрироваться с веб-сервисами.

Платформа позволяет работать в режиме клиента для любого веб-сервиса, имеющего WSDL-описание. Обращения к внешнему сервису могут производиться на любом этапе обработки. Связывание и настройка XML-запросов производится без программирования, при помощи мастеров.

В состав Deductor включен новый серверный компонент – Deductor Integration Server, который является веб-сервисом. Таким образом, результат любой аналитической обработки становится доступным для всех продуктов, взаимодействующих при помощи обмена XML-запросами. WSDL-описание формируется автоматически, без программирования и использования дополнительных инструментов.

Масштабируемая архитектура

Использование Deductor Integration Server позволяет строить отказоустойчивые системы, поддерживающие автоматическую балансировку нагрузки, горячую замену аналитических серверов и повышение производительности обработки за счет включения в кластер дополнительных Deductor Analytic Server-ов.



Обработчики

По сравнению с предидущей версией прогрпммы, серьезно переработан блок очистки данных. Вместо одного обработчика "Парциальная обработка" появилось несколько модулей:

- Оценка качества данных
- Заполнение пропусков
- Редактирование выбросов

- Спектральная обработка

Обработчик "Оценка качества данных" предназначен для проведения профайлинга и аудита данных с целью определения степени их пригодности для решения задач анализа по объективным критериям. Выполнив единственную операцию, пользователь может сразу увидеть "масштаб бедствия" и наметить способы улучшения качества данных.

Добавлены новые обработчики:

- Сэмплинг. Построение репрезентативной выборки. Варианты сэмплинга: случайный, равномерный, стратифицированный, пользовательский, отбор со смещением.
- Разбиение данных на обучающее и тестовое множество. Обеспечивает возможность строить Data Mining модели на идентичных выборках.
- Конечные классы. Расчет оптимальных способов квантования, с удобной визуализацией, расчетом показателей качества разбиения, возможностью ручной правки конечных классов.
- Масштабируемые алгоритмы кластеризации: CLOPE, EM.
- Декомпозиция временных рядов. Выделение тренда, сезонной составляющей и остатка, с возможностью удобной ручной правки полученных коэффициентов.
- Нечеткая фильтрация данных и Изменение переменных

Доработаны и значительно улучшены имеющиеся обработчики:

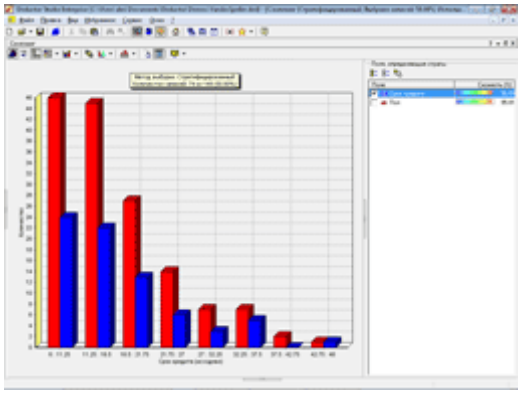
- Факторный анализ: методы вращения варимакс и квартимакс;
- Логистическая регрессия: пошаговые методы отбора, внесение поправок на априорные вероятности, взвешенная регрессия, расчет баллов скоринговых карт, взаимодействия второго уровня на основе кросс-переменных.
- Линейная регрессия: пошаговые методы отбора переменных.
- Калькулятор: повторное использование полей, обращение по абсолютным адресам, новые функции.
- Групповая обработка: упрощение процесса построения сценариев

Визуализаторы

Многочисленные улучшения и новые возможности в диаграмме, OLAP-кубе и кросс-диаграмме.

Новые визуализаторы:

Оценка качества данных



Сэмплинг

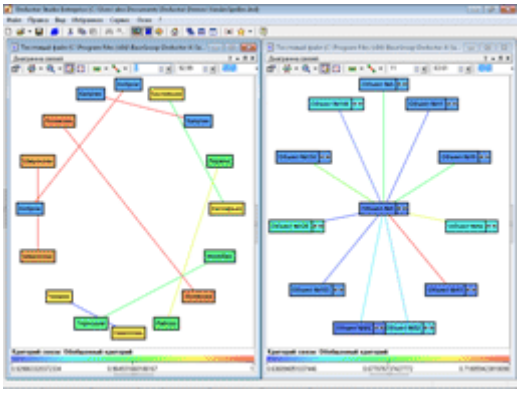
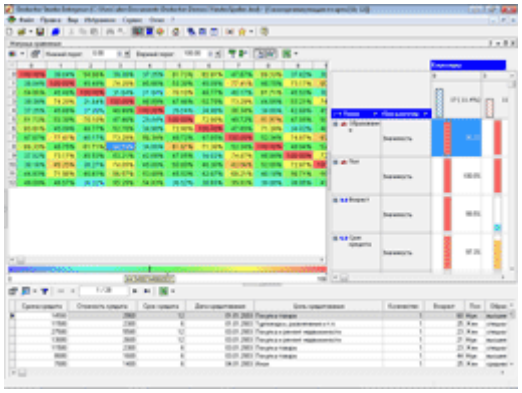
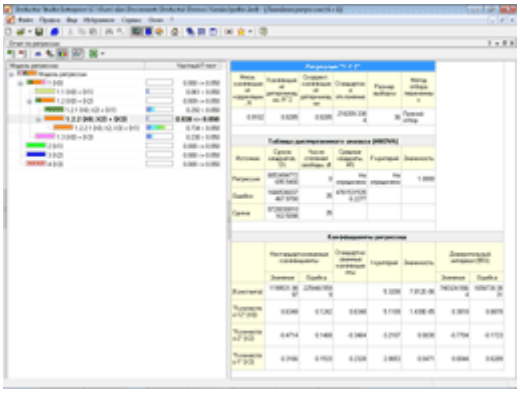


Диаграмма связей



Матрица сравнения



Отчет по регрессии

Настройка тренда и сезонных индексов

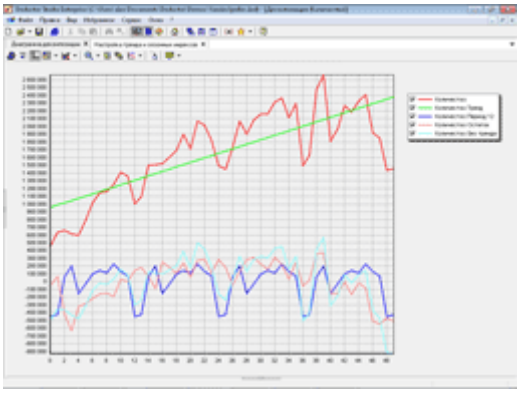
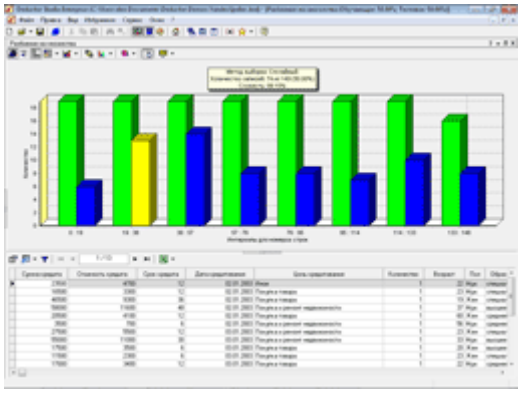
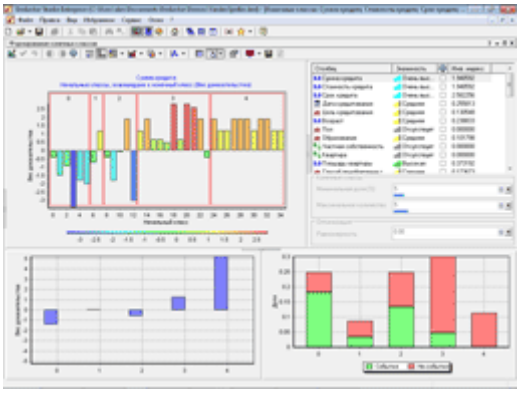


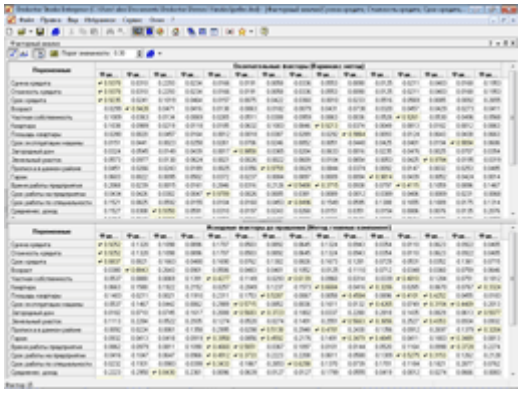
Диаграмма декомпозиции



Разбиение на множества



Конечные классы



Факторный анализ

Существенная переработка визуализатора ROC-кривая, который переименован в Качество классификации с включением диаграмм Lift-кривых и расчетом индекса Gini.

Импорт и экспорт данных

Deductor Academic отныне, помимо csv-файлов и хранилища данных на базе Firebird, поддерживает быстрый "родной" формат Deductor Data File.
Добавлена поддержка импорта/экспорта данных из/в файлы MS Excel 2007, 2010.
Улучшена работа с платформой 1С: поддержка импорта 1С:Предприятия 8.2, а также построение запросов к 1С:Предприятие 8.0, 8.1, 8.2.
Добавлена возможность импорта данных из CRM-систем:

- Terrasoft CRM
- Microsoft Dynamics CRM

Переработан импорт и экспорт из XML-документов на основе хранилища XSD-схем.

Новый уровень аналитики

Deductor 5.3 поднимает возможности аналитической обработки на новый уровень. Поддержка веб-сервисов и новых источников данных позволяет проще интегрировать систему в разнородное программное окружение. Теперь аналитика не ограничивается только внутрикорпоративными данными, любой внешний веб-сервис может быть встроен в конвейер принятия решений. Сам Deductor тоже может стать источником данных для других систем.
Новые обработчики и визуализаторы значительно упрощают процесс анализа: автоматический перебор вариантов обработки, выбор и предложение оптимальных способов очистки, удобная визуализация результатов анализа.
Включение многих новых обработчиков и изменение существующих, в значительной степени направлены на повышение уровня автоматизации работы аналитика. Они позволяют строить гибкие, универсальные, но при этом простые для понимания и поддержки сценарии обработки.
В новой версии значительное внимание уделено повышению скорости обработки больших объемов данных: добавлены новые масштабируемые Data Mining алгоритмы, оптимизирована работа существующих обработчиков. Применение Deductor Integration Server позволяет производить аналитические расчеты на кластере серверов, что значительно снижает время отклика и повышает отказоустойчивость комплекса.

Хранилище данных

Хранилище данных Deductor Warehouse – это специально организованная база данных, ориентированная на решение задач анализа данных и поддержки принятия решений, обеспечивающая максимально быстрый и удобный доступ к информации.

На практике в компаниях часто бывает так, что информация вроде бы где-то есть, и ее даже много, но она не структурирована, не согласована, разрознена, не всегда достоверна, ее практически невозможно найти и получить в едином формате. Для устранения этого противоречия (когда при физическом наличии данных и

даже их избытке фактически информация для анализа отсутствует) создается хранилище данных. Это позволяет превратить все многообразие накопленных в организации данных в ценную для бизнеса информацию. Таким образом, хранилище представляет собой специальную базу данных, в которую по определенному регламенту (например, 1 раз в сутки) выгружаются данные из одной или сразу нескольких учетных систем (1С-бухгалтерия, собственные источники и т.п.).

Назначение хранилища данных – своевременно обеспечить аналитика всей информацией, необходимой для проведения анализа, построения моделей и принятия решений. Цель хранилища данных – не анализ данных, а подготовка данных для анализа и их *консолидация*.

Хранилище данных Deductor Warehouse (рис. 15.5) включает в себя потоки данных, поступающих из различных источников, и специальный семантический слой, содержащий так называемые метаданные (данные о данных).

Семантический слой позволяет преобразовать реляционное (табличное) представление данных к многомерному.

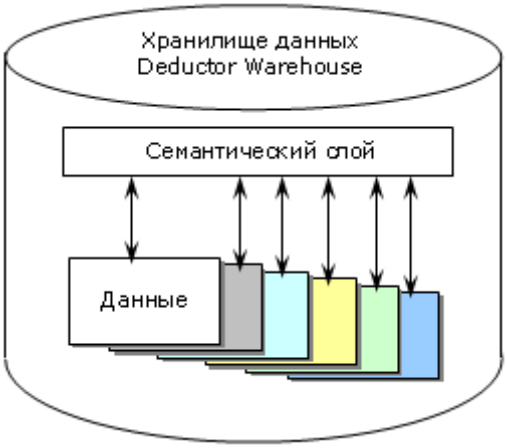


Рисунок 15.5 – Хранилище данных

Запрос к хранилищу данных осуществляется непосредственно через семантический слой, который через внутреннюю систему команд (скрытую от пользователя) подбирает запрашиваемую информацию из многообразия хранимых данных. Работу семантического можно сравнить с работой библиотекаря, который по просьбе читателя достает с разрозненных полок нужные книги, раскрывая их на нужных страницах.

Благодаря семантическому слою и многомерному представлению данных работа с данными из хранилища Deductor Warehouse осуществляется в терминах предметной области (в бизнес-терминах), что является очень удобным для пользователя. От пользователя не требуется знания структуры хранения данных и языка запросов. Он работает с привычными ему терминами бизнес-среды – *отгрузка*, *товар*, *клиент*.

Хранилище Deductor Warehouse включает в себя определенным образом связанные между собой данные (таблицы из разных источников), и семантический слой, где хранятся данные о данных. Все данные в хранилище Deductor Warehouse хранятся в структурах типа «снежинка», где в центре расположены таблицы фактов, а «лучами» являются измерения, причем каждое измерение может ссылаться

на другое измерение. Именно эта схема чаще всего встречается в хранилищах данных (рис. 15.6).

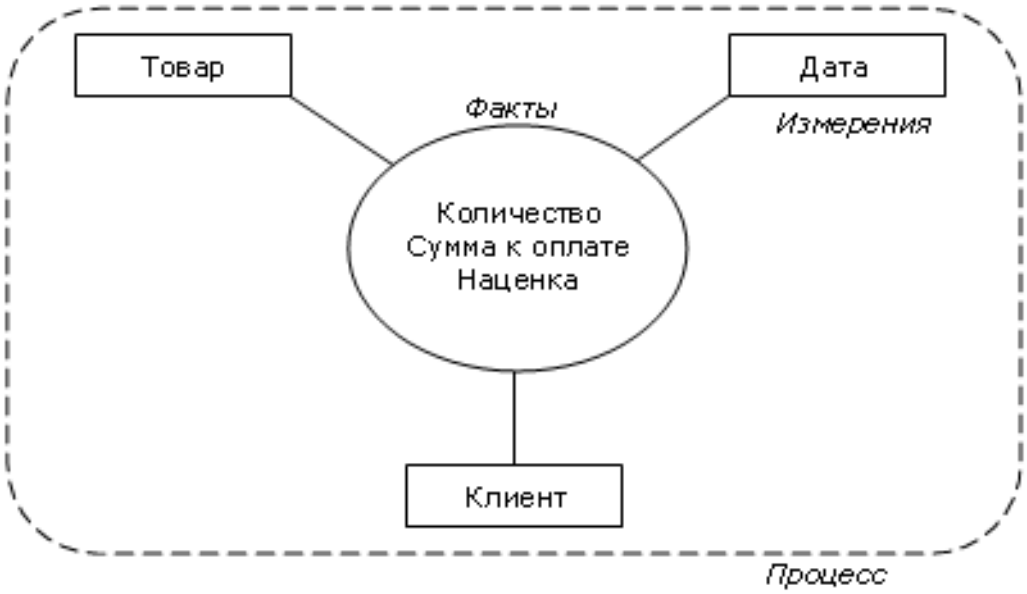


Рисунок 15.6 – Структура ХД

Сведения о том, какие данные являются фактами, а какие измерениями задаются пользователем на этапе проектирования структуры ХД и хранятся в семантическом слое. Такая архитектура хранилища наиболее адекватна задачам анализа данных. Каждая «снежинка» называется процессом и описывает определенное действие, например, продажи товара, отгрузки, поступления денежных средств и прочее.

В Deductor Warehouse может одновременно храниться множество процессов, имеющих общие измерения, например, измерение «Товар», фигурирующее в процессах «Поступления» и «Отгрузка».

В упрощенном варианте все данные в процессе обязательно должны быть определены как измерение, атрибут либо факт (рис. 15.7).

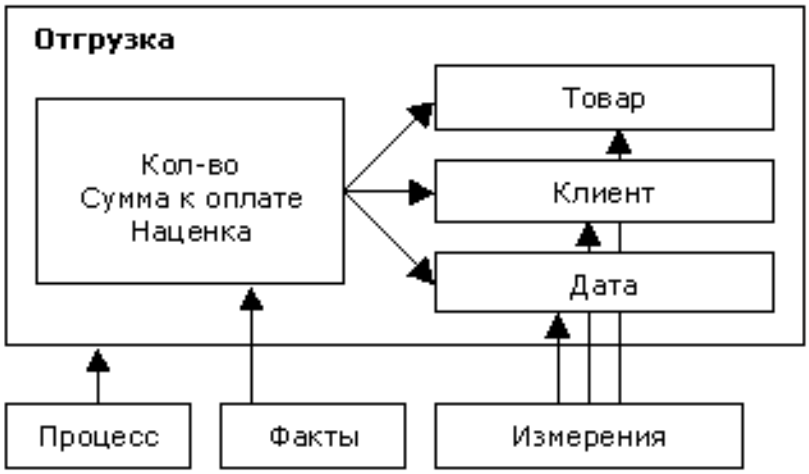


Рисунок 15.7 – Проектирование структуры ХД

Измерение – это последовательность значений одного из анализируемых параметров. Например, для параметра «время» это последовательность календарных дней, для параметра «регион» – список городов. Каждое значение измерения может быть представлено координатой в многомерном пространстве процесса. Например, *товар*, *клиент*, *дата*.

Атрибут является свойством измерения (точки в пространстве). Атрибут как бы скрыт внутри другого измерения и помогает пользователю полнее описать исследуемое измерение. Например, для измерения товар атрибутами могут выступать цвет, вес, габариты товара.

Факт – значение, соответствующее измерению. Факты – это данные, отражающие сущность события. Как правило, фактами являются численные значения, например, сумма и количество отгруженного товара, скидка.

Некоторые бизнес-понятия (соответствующие измерениям в хранилище данных) могут образовывать иерархии, например, *Товар* может включать *Продукты питания* и *Лекарственные препараты*, которые, в свою очередь, подразделяются на группы продуктов и лекарств и т. д. В этом случае первое измерение содержит ссылку на второе, второе – на третье и т.д. Иногда для повышения скорости доступа к данным отказываются от иерархии измерений. И схема «снежинка» превращается в более простую схему под названием «звезда».

Принадлежность данных к типу (измерение, ссылка на измерение, атрибут или факт) содержится в семантическом слое хранилища. Обратим внимание на то, что:

- таблицы *измерений* содержат только справочную информацию (коды, наименования и т.п.) и ссылки на другие измерения при необходимости;
- таблица *процесса* содержит только факты и коды измерений (без их атрибутов).

Проиллюстрируем это на следующем примере – рассмотрим данные по истории продаж различных товаров (на рисунке 15.8 представлены фрагменты таблиц).

Здесь в таблице процесса хранится информация о значениях измерений (как правило, это код измерения) и значениях фактов. На рисунке 15.8 в таблице процесса в первой строке содержится информация, что 05.06.2006 г. клиент №3 приобрел товар №386 в количестве 100 шт. на сумму 25 500, при этом наценка составила 3 825. Кто такой клиент №3 и что за товар №386 он приобрел, в таблице процесса не указано.

Информация с описанием (атрибутами) клиентов и товаров находится в таблицах измерений, которые можно сравнить со словарями, хранящими справочную информацию по измерениям. «Дата» является измерением без атрибутов, и поэтому она присутствует только в таблице процесса.

Перед тем, как загружать таблицу процесса, необходимо загрузить *все* измерения.

Рассмотрим на примере взаимоотношение процесса, измерений, атрибутов и фактов.

Для анализа работы сети аптек данные поставляются в 4-х таблицах: *Товары*, *Группы*, *Отделы* и *Продажи* (табл. 15.1-15.4).

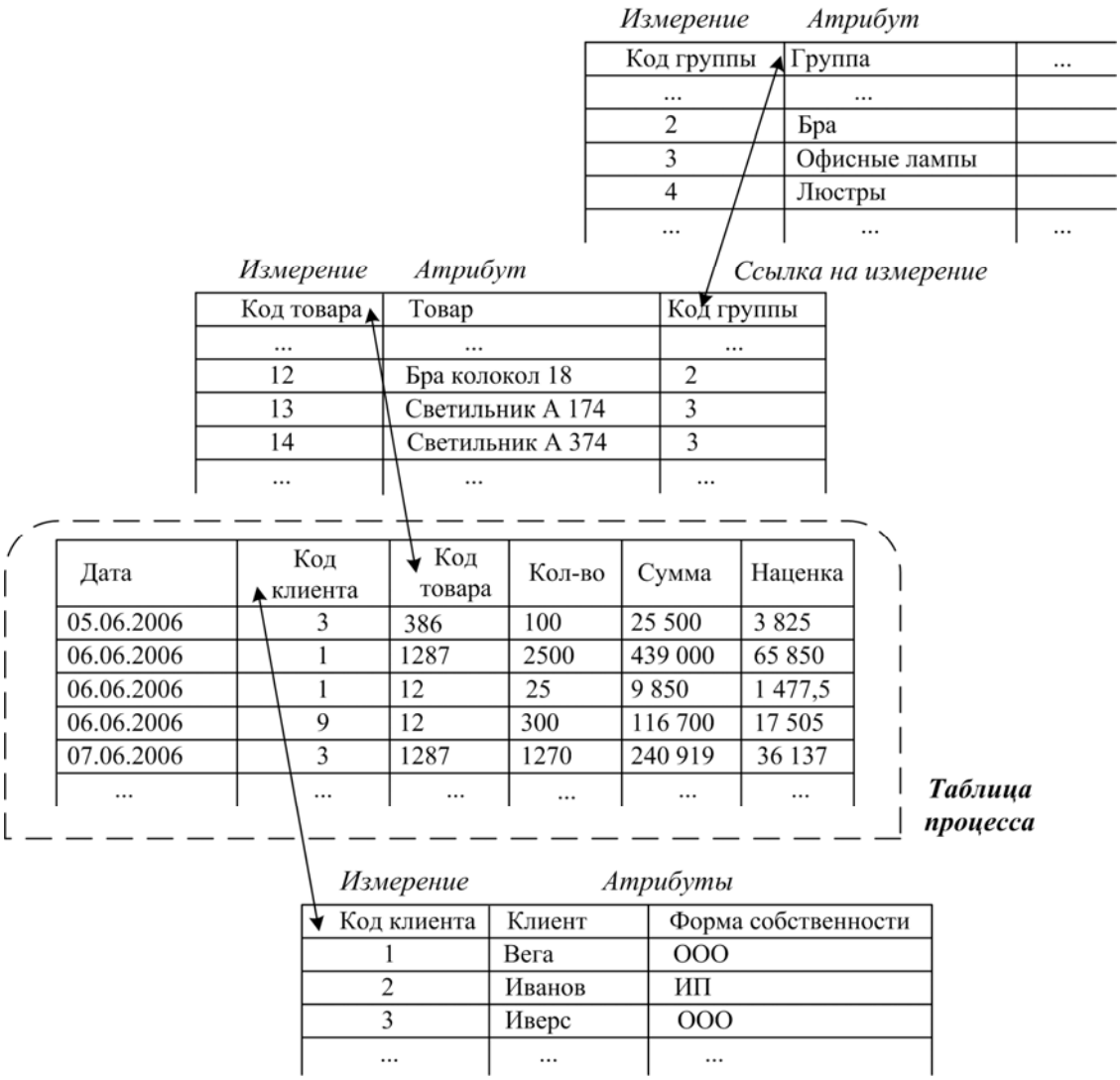


Рисунок 15.8 – Пример схемы «снежинка»

Таблица 15.1 – Товарные группы

Код группы	**Наименование группы**
33	Иммуномодуляторы
48	Общетонизирующие средства и адаптогены
50	Местные анестетики
108	Микро- и макроэлементы
198	Витамины и витаминоподобные средства
223	Желчегонные средства и препараты желчи
247	Антисептики и дезинфицирующие средства
320	Биологически активные пищевые добавки

Таблица 15.2 – Товары (фрагмент)

Код товара	**Наименование товара**	**Код группы**
774	Альмагель	1
810	Иммунорм	33
824	Ревит	198
898	Настойка пустырника	48

Таблица 15.3 – Отделы

Код отдела	**Наименование отдела**
1	Аптека 1
2	Аптека 2
3	Аптека 3

Таблица 15.4 – Продажи (фрагмент)

Дата	**Код отдела**	**Код товара**	**Час покупки**	**Количество**	**Сумма**
01.01.2006	1	31052	13	1	56.5
01.01.2006	1	36259	16	1	72.48
01.01.2006	1	40315	15	1	15.84
01.01.2006	1	40315	15	3	47.52
01.01.2006	**3**	**810**	14	**1**	**163.50**

Покажем, какие данные являются измерениями, какие атрибутами, а какие фактами, и что представляет собой процесс?

В табл. 15.1 код группы является измерением, а наименование группы его атрибутом.

В табл. 15.2 код товара является измерением, наименование товара его атрибутом, а код группы – ссылкой на одноименное измерение.

В табл. 15.3 код отдела является измерением, а наименование отдела его атрибутом.

В табл. 15.4 *Дата* является измерением, *Отдел*, *Код товара* и *Код группы* как было сказано выше – измерения, *Час покупки* – измерение, *Количество* и *Сумма* – факты. То есть таблица 15.4 является описанием процесса продаж в трех аптеках.

Взаимоотношение измерений, атрибутов и фактов внутри процесса продаж в трех аптеках (см. последнюю строку таблице 15.4, выделенную жирным шрифтом) проиллюстрировано на рис. 15.9.

В силу того, что визуально можно представить только трехмерное пространство, на рисунке показано взаимодействие трех измерений (*Дата*, *Отдел* и *Код товара*). В рассмотренном примере измерений гораздо больше. Каждое новое может быть представлено новой осью.

Обратим внимание на то, что *Час покупки* не может быть фактом, т.к. комбинация оставшихся 3-х измерений (Дата, Код отдела, Код товара) уникально не определяет точку в многомерном пространстве: в один и тот же день может быть продано несколько одинаковых товаров в одном и том же отделе.

Создание нового хранилища данных

Deductor позволяет создавать хранилища данных на трех СУБД: InterBase/FireBird, Microsoft SQL Server, Oracle начиная с версии 9. Выбор той или иной СУБД часто зависит от многих критериев: стоимость, производительность, сложность администрирования и др. Интерфейс проектирования структуры хранилища данных не зависит от выбранной СУБД, поэтому в данном пособии хранилище

данных будет создаваться на FireBird, с которым, что важно, можно работать и локально (при помощи библиотеки **fbclient.dll**).

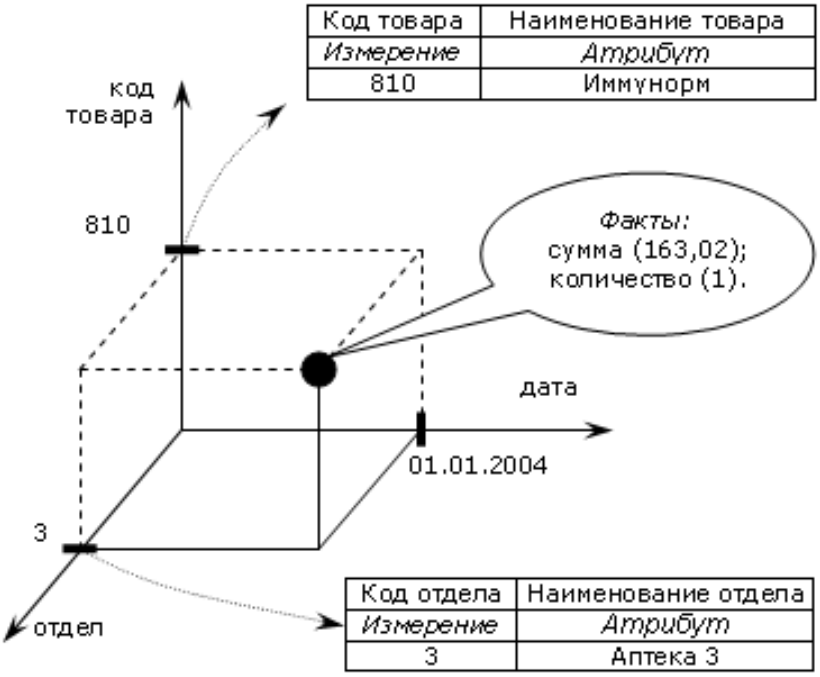


Рисунок 15.9 – Измерения, атрибуты и факты внутри процесса продаж

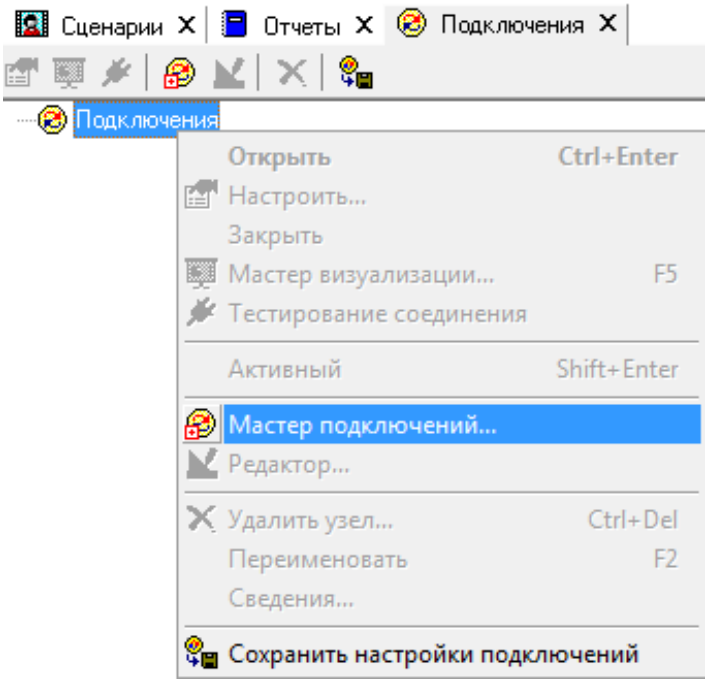


Рисунок 15.10 – Создание (подключение) хранилища данных

Замечание. Возможность работы с хранилищами данных на СУБД MS SQL Server и Oracle доступна только в Deductor Enterprise.

Для создания нового хранилища данных или подключения к существующему в Deductor Studio необходимо перейти на закладку **Подключения** и запустить **Мастер подключений** (рис. 15.11).

На экране появится первый шаг Мастера (рис. 15.11), в котором нужно выбрать тип источника (приемника), к которому нужно подключиться. Нас интересует Deductor Warehouse.

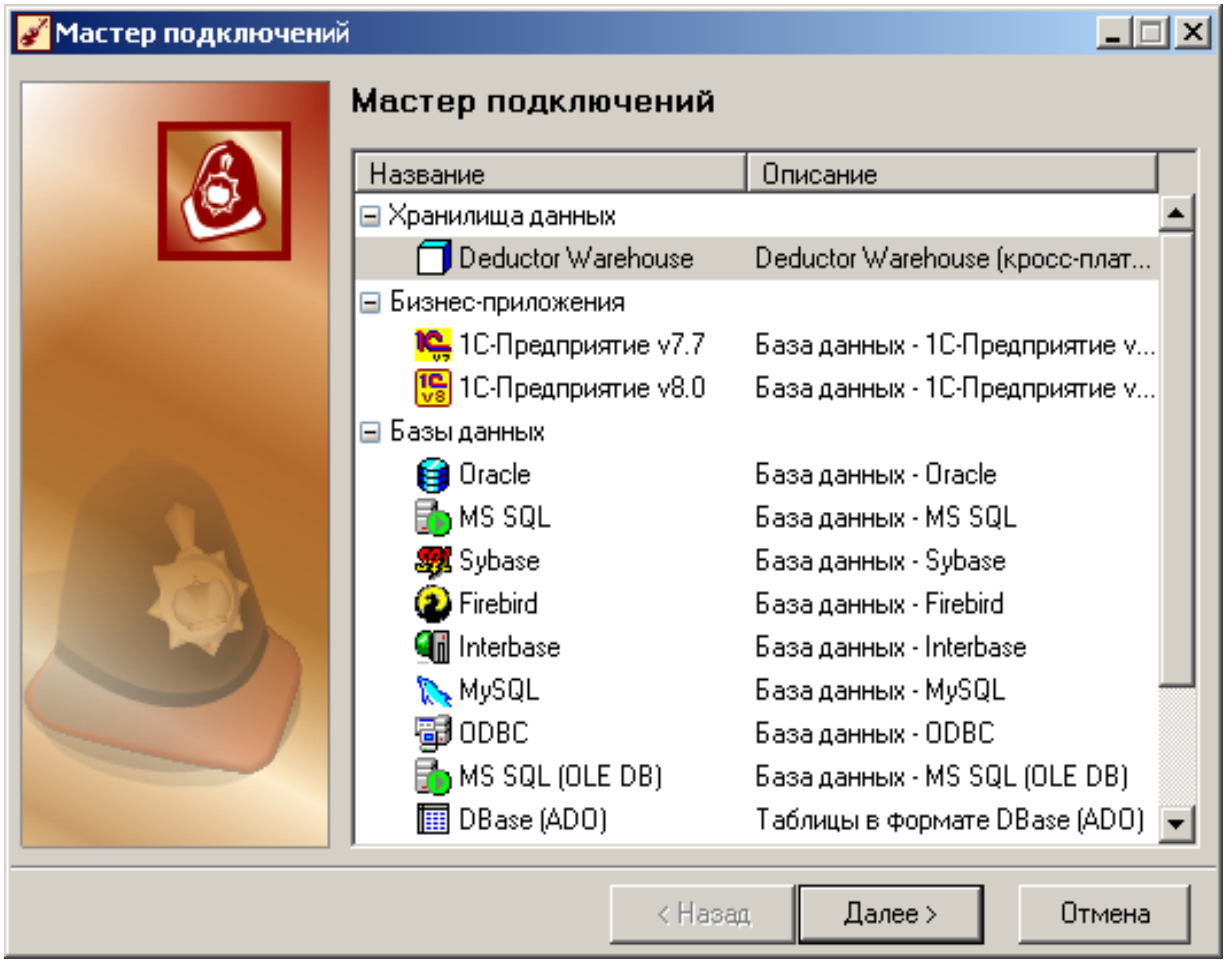


Рисунок 15.11 – Окно выбора типа подключения

На следующем шаге из единственно доступного в списке типа базы данных выберем Firebird и перейдем на третий шаг Мастера. В нем зададим параметры базы данных, в которой будет создана физическая и логическая структура хранилища данных (рис. 15.12):

- База данных – C:\Users\...\Часть2\15\farrma.gdb (или любой другой путь на диске);
- Логин – sysdba, пароль – masterkey;
- Установить флажок **Сохранять пароль**.

На следующей вкладке (рис.15.13) выберем последнюю версию для работы с ХД Deductor Warehouse 6.

На следующем шаге при нажатии на кнопку  **Создать файл базы данных с необходимой структурой метаданных** по указанному ранее пути будет создан файл farma.gdb (появится сообщение об успешном создании). Это и есть пустое хранилище данных, готовое к работе.

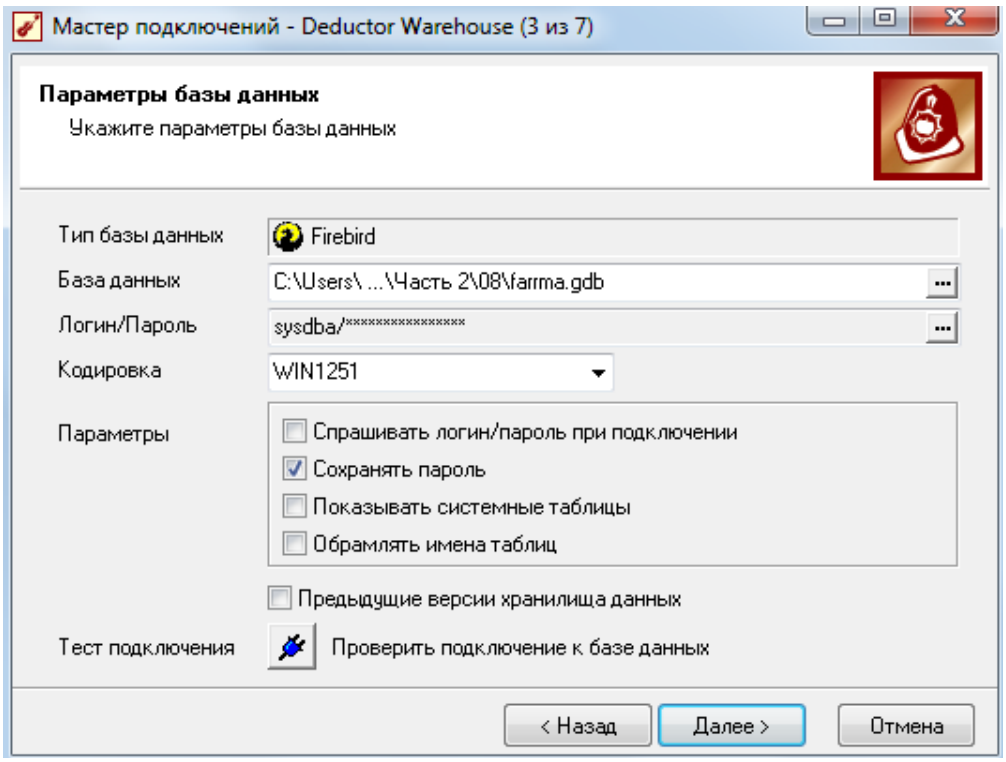


Рисунок 15.12 – Установка параметров базы данных

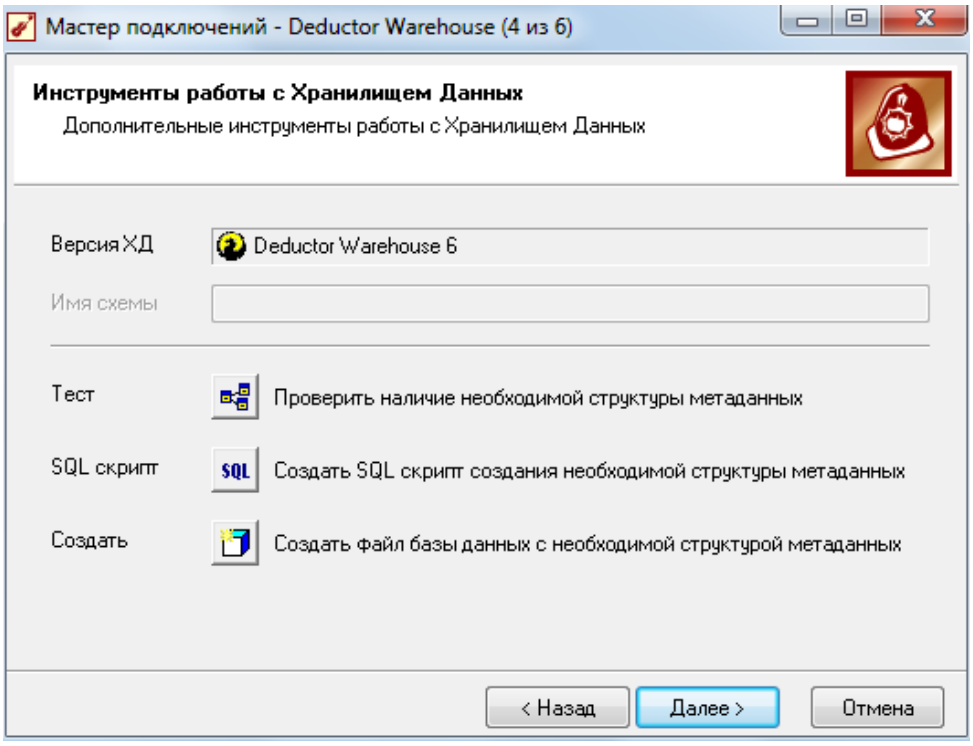


Рисунок 15.13 – Выбор версии хранилища данных

На последних двух шагах осталось выбрать визуализатор для подключения (здесь это **Сведения** и **Метаданные**) и задать имя, метку и описание для нового хранилища (рис. 15.14).

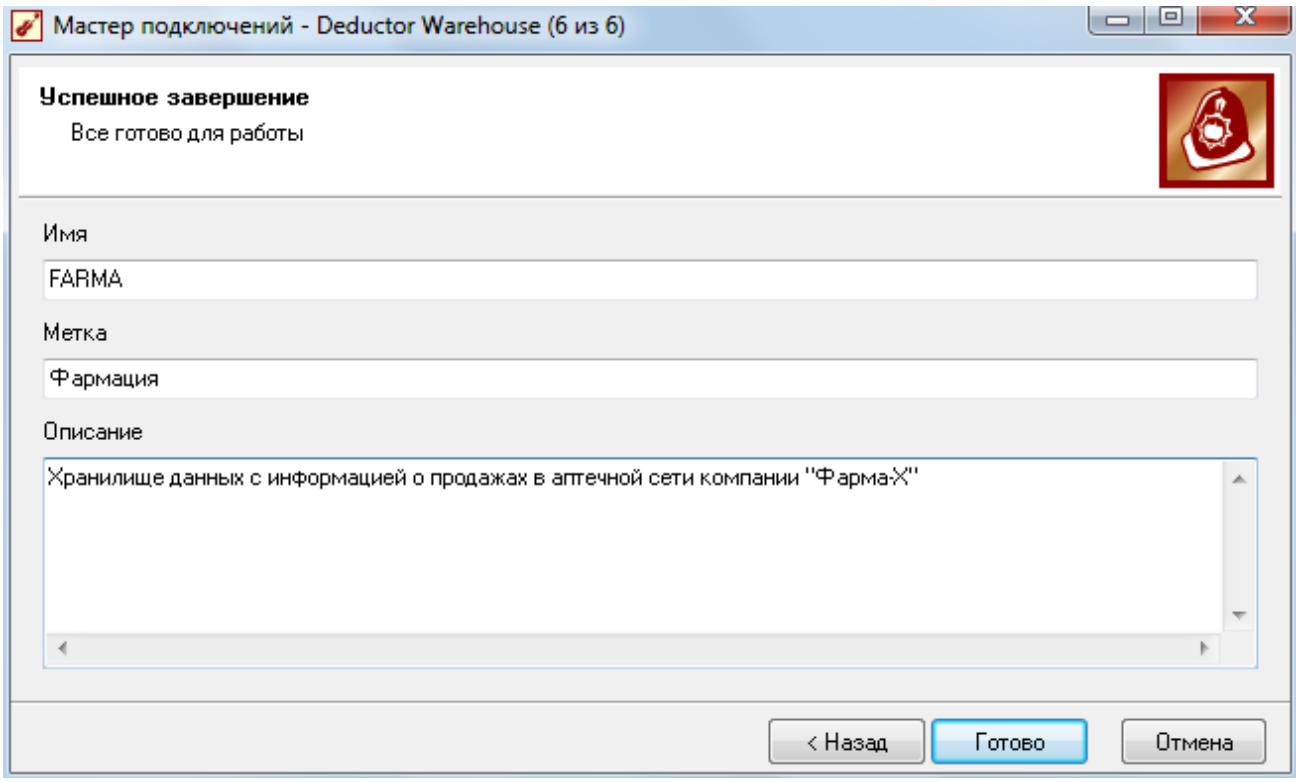


Рисунок 15.14 – Настройка семантики узла подключения

Имя хранилища может быть введено только латинскими буквами.

После нажатия на кнопку **Готово** на дереве узлов подключений появится метка хранилища (рис. 15.15).

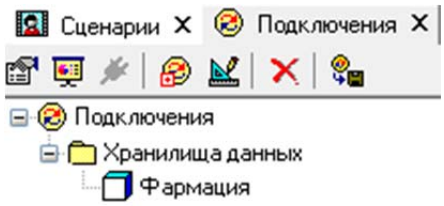


Рисунок 15.15 – Хранилище данных «Фармация»

Для проверки доступа к новому хранилищу данных воспользуйтесь кнопкой . Если спустя некоторое время появится сообщение «Тестирование соединения прошло успешно», то хранилище готово к работе. Сохраните настройки подключений, выбрав в контекстном меню одноименный пункт.

Если соединение по какой-либо причине установить не удалось, то будет выдано сообщение о соответствующей ошибке. В этом случае нужно проверить параметры подключения хранилища данных и при необходимости внести в них изменения (используйте для этого действие **Настроить подключение**).

Проектирование структуры хранилища данных

После создания хранилища необходимо спроектировать его структуру, т.к. в пустом хранилище нет ни одного объекта (процессов, измерений, фактов). Для это-

го предназначен Редактор метаданных, который вызывается кнопкой на вкладке **Подключения**.

В открывшемся окне Конструктора, выберем кнопку - разрешить редактировать, встав на узле **Измерения**, при помощью кнопки **Добавить** добавим в метаданные первое измерение *Код группы* со следующими параметрами:

- Идентификатор – GR_ID;
- Имя – Группа.Код;
- Тип данных – целый.

Имя – это семантическое название объекта хранилища данных, которое будет отображаться пользователю, работающему с ХД.

Проделаем аналогичные действия для создания всех остальных измерений, взяв параметры из табл. 15.5.

Таблица 15.5 – Параметры измерений

Измерение	**Идентификатор**	**Имя**	**Тип данных**
Код группы	GR_ID	Группа.Код	целый
Код товара	TV_ID	Товар.Код	целый
Код отдела	PART_ID	Отдел.Код	целый
Дата	S_DATE	Дата	дата/время
Час покупки	S_HOUR	Час	целый

В результате структура метаданных нашего хранилища будет содержать 5 измерений (рис. 15.16).

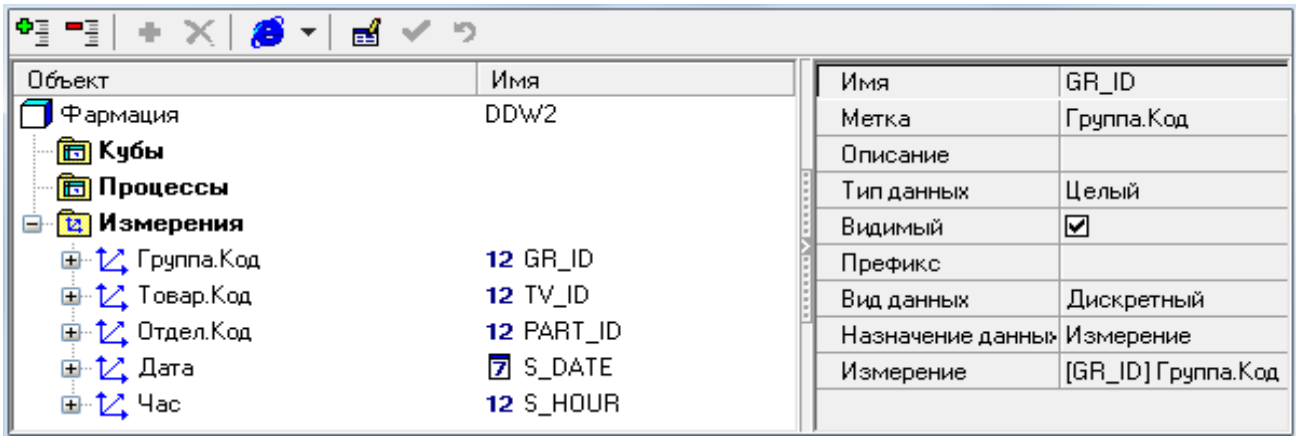


Рисунок 15.16 – Структура метаданных хранилища

К каждому измерению, кроме *Дата* и *Час*, теперь добавим по текстовому атрибуту – это будут *Группа.Наименование*, *Товар.Наименование*, *Отдел.Наименование* соответственно.

Каждое измерение может ссылаться на другое измерение, реализуя тем самым иерархию измерений.

В нашем случае измерение «Товар.Код» ссылается на «Группу.Код». Эту ссылку и установим путем простого добавления (ссылка на измерение отображается иконкой). Результат работы иллюстрирует рисунок 15.17.

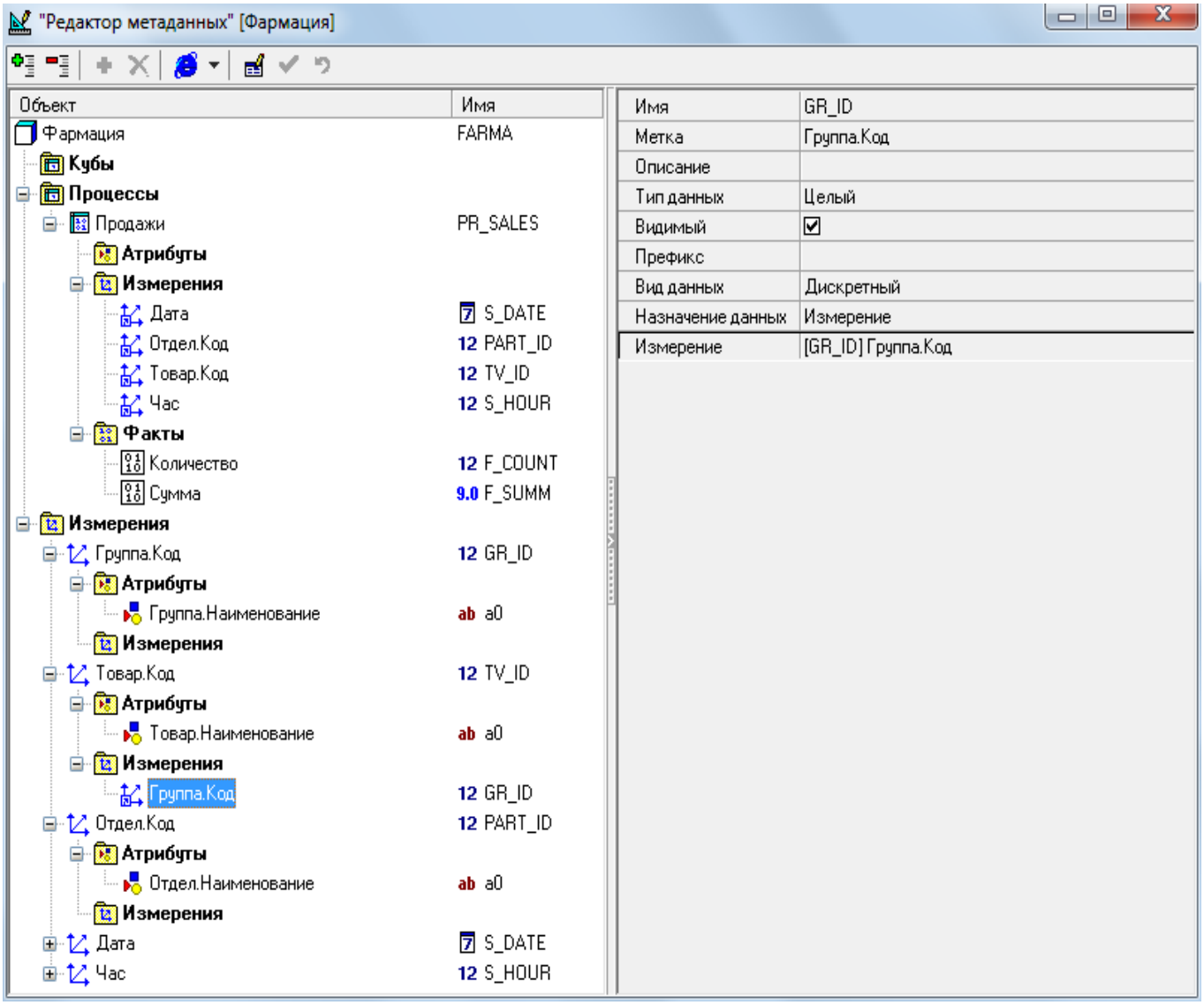


Рисунок 15.17 – Проектирование структуры ХД

После того, как все измерения созданы, приступают к формированию процесса. Назовем его «Продажи» (PR_SALES) и добавим в него ссылки на 4 существующих измерения: *Дата*, *Отдел.Код*, *Товар.Код*, *Час (*кнопка ✚). Кроме них в нашем процессе присутствуют два факта: *Количество* и *Сумма*, причем первый – целочисленный, второй – вещественный (Рисунок 15.18).

На этом проектирование структуры и метаданных ХД закончено, выберем кнопку ✔ - принять изменения и закроем окно Редактора

Загрузка информации в хранилище

При загрузке данных в хранилище сначала загружаются измерения со своими атрибутами, и только после этого загружаются данные в процесс.

В нашем распоряжении имеются 4 текстовых файла:

- groups.txt – товарные группы;
- produces – товары;
- stores.txt – отделы;
- sales.txt – продажи товаров по дням.

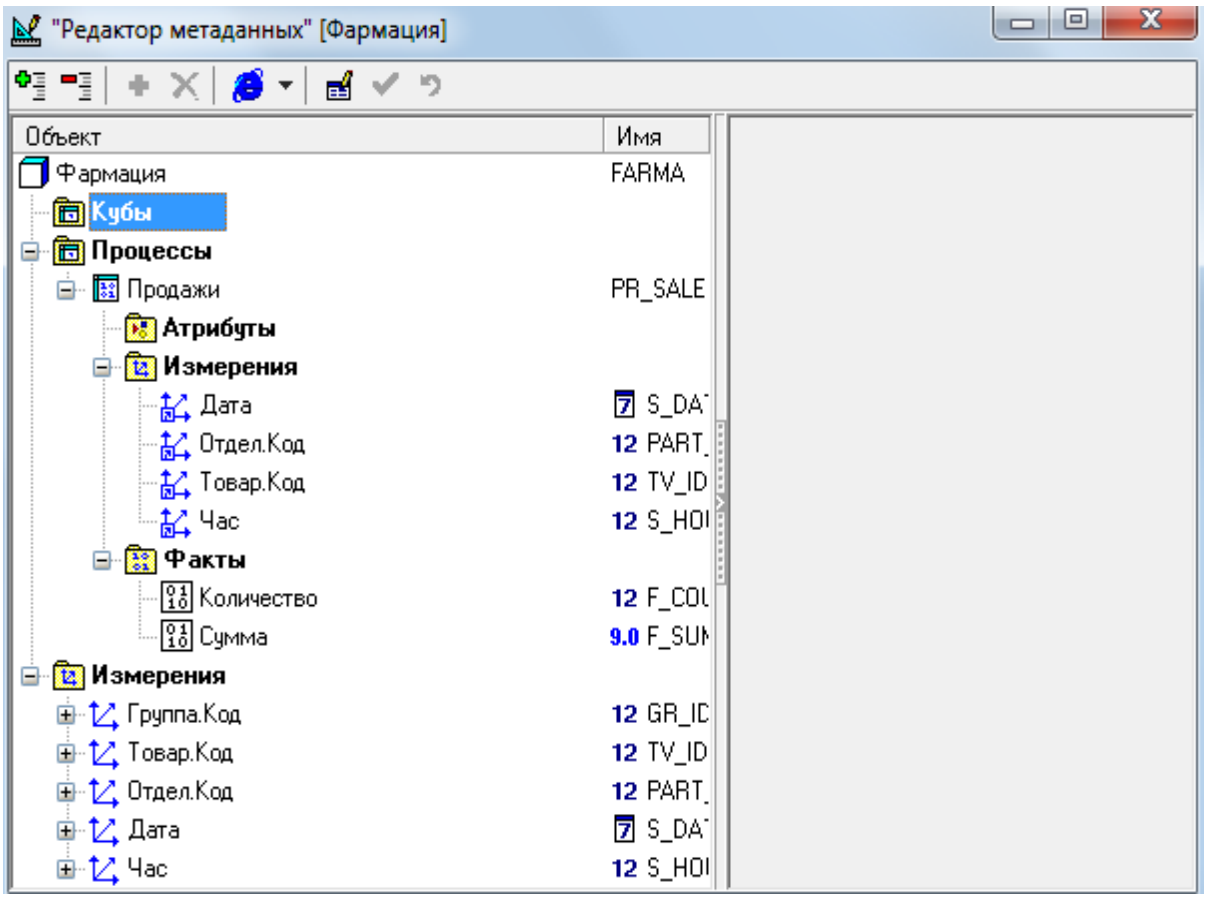


Рисунок 15.18 – Создание метаданных процесса

Последовательно импортируем все 4 файла в Deductor. Покажем шаги по импорту только для первого файла, поскольку последовательность действий для остальных файлов одинаковая. Для этого в Deductor Studio начнем работу на вкладке **Сценарии**. Запустим **Мастер импорта** и выберем источник – Text (текстовый файл с разделителями), как это показано на рисунке 15.19.

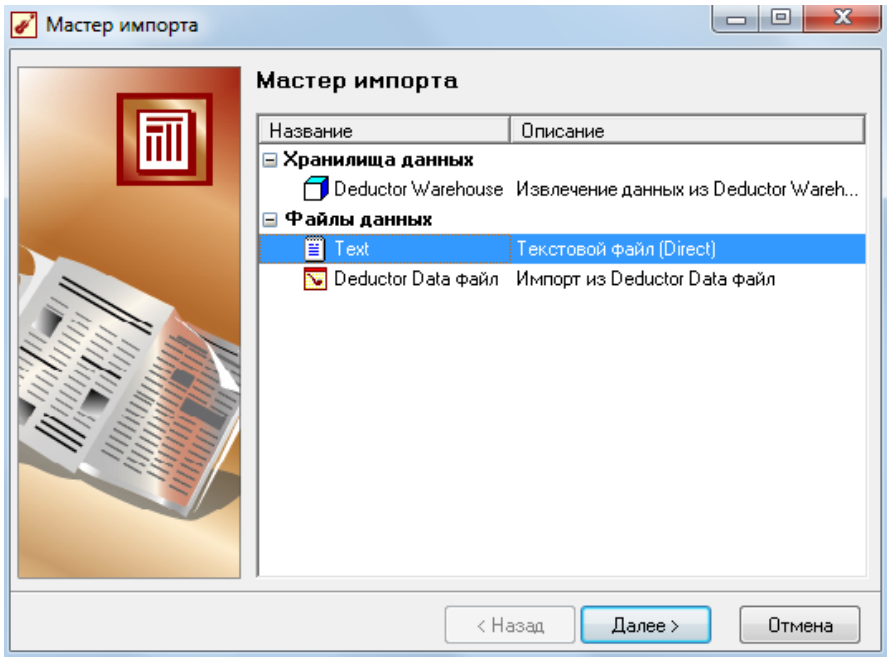


Рисунок 15.19 – Первый шаг Мастера импорта

На следующем шаге в строке укажем имя файла для импорта – groups.txt, причем лучше использовать относительный путь (рис. 9.20). Это означает, что файл должен находиться в той же папке, что и файл со сценарием Deductor.

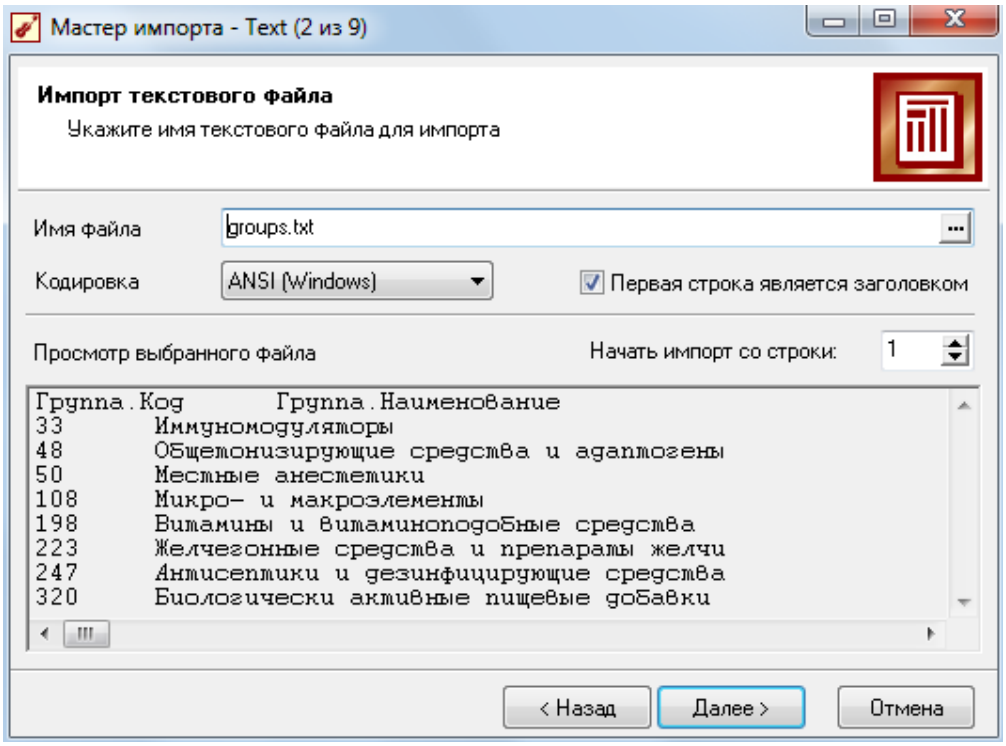


Рисунок 15.20 – Выбор текстового файла для импорта

На третьей вкладке расположены параметры импорта, специфичные для текстовых файлов. Оставим все установки по умолчанию (рис. 15.21).

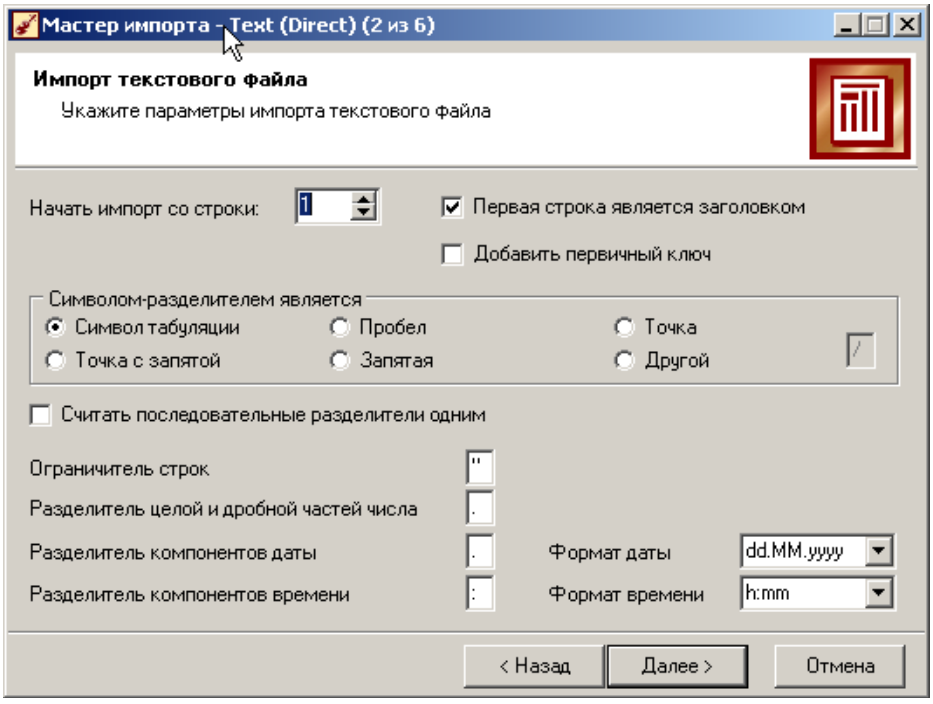


Рисунок 15.21 – Параметры импорта

Замечание. Для корректного импорта текстовых файлов в качестве разделителя целой и дробной частей числа и в качестве разделителя компонентов даты установим знак « . » – точка.

Все остальные настройки шагов Мастера можно оставить по умолчанию, нажимая кнопки «Далее», но на шаге 6 следует провести настройку измерений в соответствии с таблицей 15.5.

Проделав это же самое для оставшихся 3-х файлов, получим сценарий, состоящий из 4-х веток. По умолчанию будет предложен визуализатор Таблица, который отобразится справа (рис. 15.22).

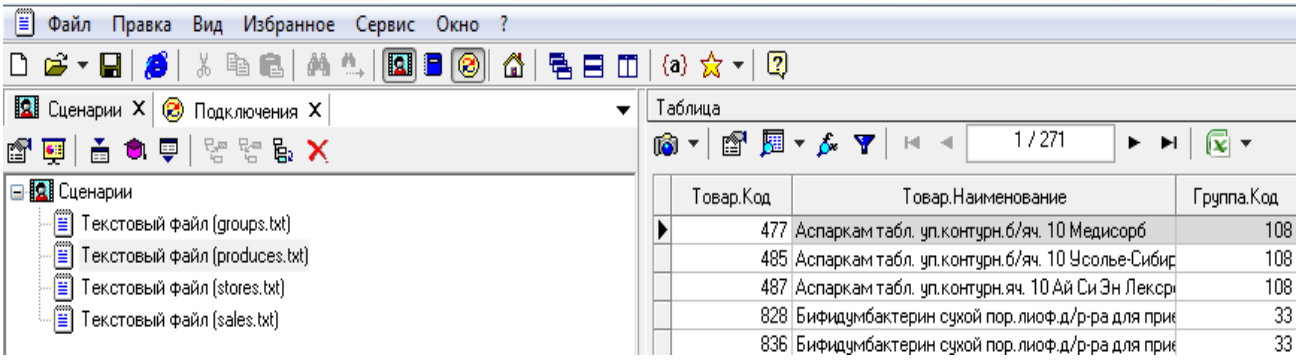


Рисунок 15.22 – Сценарий в Deductor

Теперь, после импорта, можно приступать к загрузке данных в ХД. Первыми следуют таблицы измерений, и только в конце – таблица процесса sales.txt. Менять порядок веток сценария можно при помощи кнопок CTRL+↑ и CTRL+↓.

Покажем последовательность загрузки данных в измерение снова на примере первого измерения *Группа.Код*. Для этого, встав на первом узле, вызовем **Мастер экспорта**. Из списка типа приемников выберем Deductor Warehouse (рис.15.23).

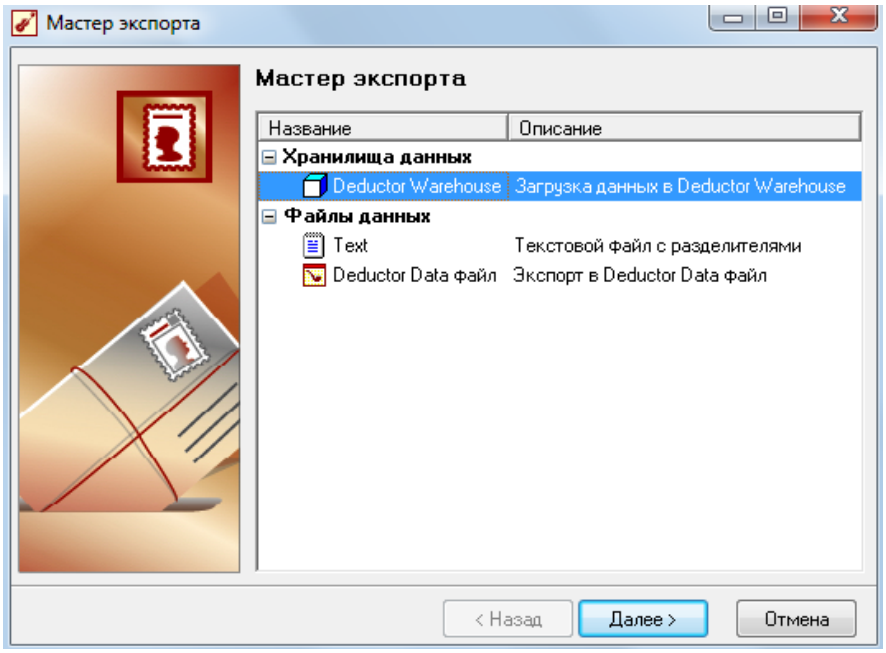


Рисунок 15.23 – Экспорт в хранилище данных

На следующей вкладке из списка доступных хранилищ укажем нужное нам ХД под названием «Фармация». Далее требуется указать, в какое именно измерение будет загружаться информация. Это *Группа. Код* (рис.15.24).

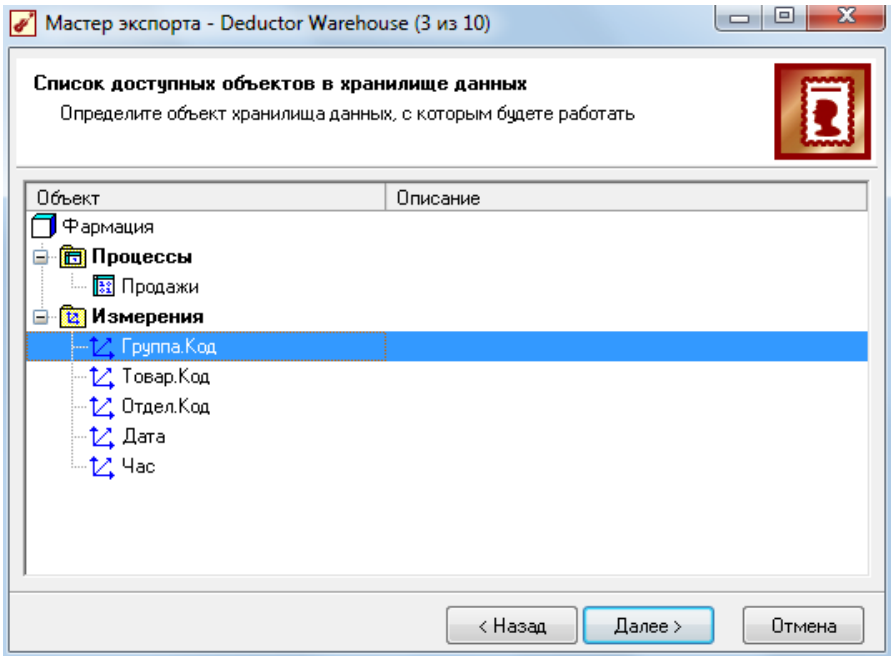


Рисунок 15.24 – Выбор объекта для экспорта

Последнее, что осталось, это установить соответствие элементов объекта в хранилище данных с полями входного источника данных (т.е. таблицы groups.txt). В случае если имена полей и метки в семантическом слое хранилища данных совпадают (для этого при загрузке рекомендовалось настроить поля в соответствии с таблицей 15.5), делать ничего не нужно (рисунок 15.25).

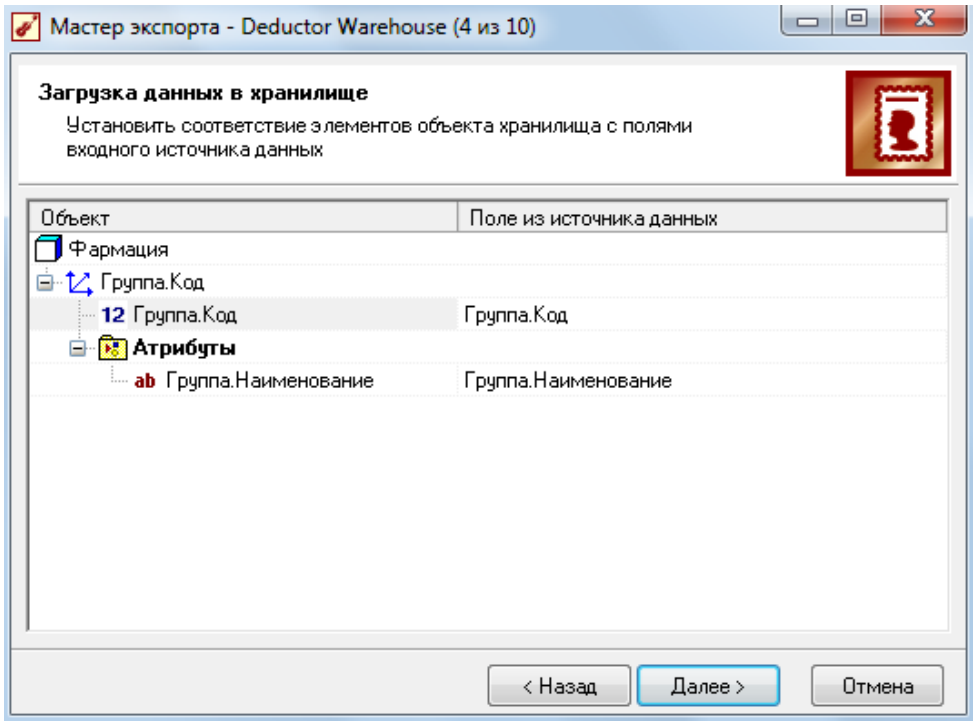


Рисунок 15.25 – Настройка соответствий полей

Нажатие кнопки «Пуск» на следующем шаге загрузит в измерение данные. При этом старые данные, если они были, будут обновлены новыми.

Проделав аналогичные действия еще для двух измерений – *Отдел.Код*, *Товар.Код*, мы получим следующий сценарий (рис.15.26).

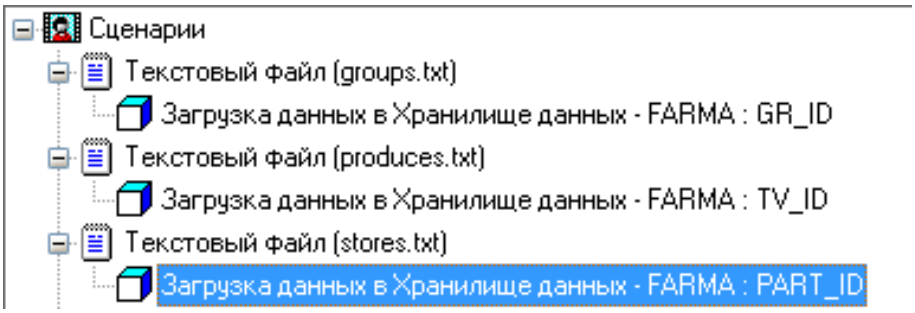


Рисунок 15.26 – Фрагмент сценария загрузки данных в ХД

Загрузка измерений на этом не заканчивается – еще остались два измерения (без атрибутов): *Дата* и *Час*. Но они содержатся в таблице процесса. Для их загрузки применим обработчик **Группировка**. Он позволяет объединить записи, которые содержат одинаковые значения. В нашем случае в полях *Дата* и *Час* таблицы процесса содержатся много одинаковых значений, т.к. это продажи. Сгруппируем сначала по полю *Дата* и загрузим в хранилище (в обработчике **Группировка** выставим единственное измерение – *Дата*, все остальные поля – неиспользуемые), после чего выполним то же самое для поля *Час* (рис. 15.27).

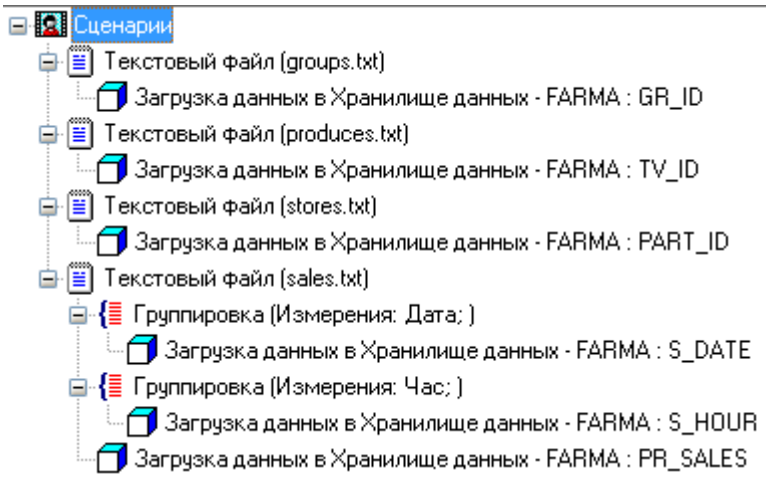


Рисунок 15.27 – Окончательный сценарий загрузки

Теперь, когда все измерения загружены (т.е. определены все координаты в многомерном пространстве), и можно загружать данные в процесс *Продажи*. В отличие от загрузки измерений в Мастере экспорта появляются два специфических шага. На одном из них нужно указать измерения, по которым необходимо удалять данные из хранилища (рис. 15.28).

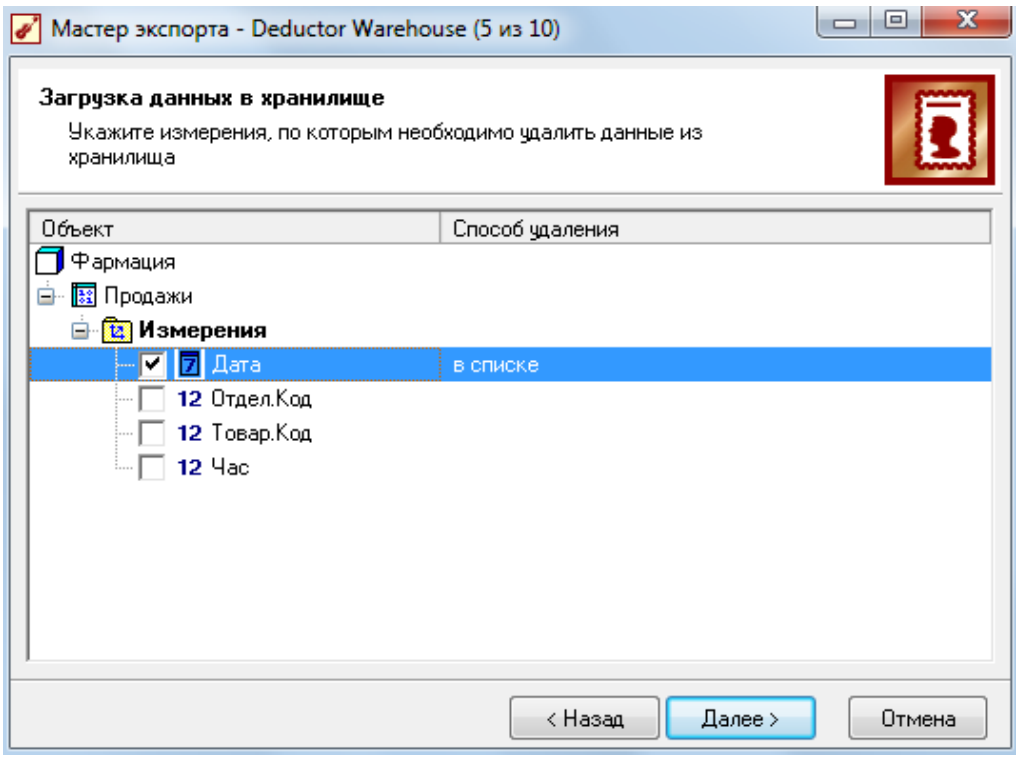


Рисунок 15.28 – Параметры для контроля непротиворечивости информации в ХД

Это требуется для контроля непротиворечивости информации: мы указываем выполняемое действие в ситуации, когда в хранилище загружается информация, которая совпадает по значениям из нескольких измерений. Вариантов может быть два: удалить «старые» данные и загрузить новые, либо запретить удаление и оставить то, что уже было загружено ранее. В нашем случае, когда установлено измерение *Дата* на удаление, при повторной загрузке в процесс *Продажи*, из него будут удалены и загружены заново данные на те даты, которые совпадают в источнике и в хранилище. Например, если в хранилище есть данные о том, что на 01.03.2004 данному клиенту продано данного товара в количестве 1000, а теперь загружается количество 1200, то реально будет храниться именно 1200. Правила, в каких случаях удалять, в каких оставлять информацию, диктуются бизнес-процессами деятельности компаний.

На последней странице настроек лучше оставить настройки по умолчанию (рис.15.29).

Сохраните файл сценария под именем **load.ded** в той же папке, где находятся текстовые файлы таблиц.

В результате всех вышеописанных действий будет:

- создано и наполнено данными хранилище данных;
- создан сценарий загрузки информации из источников в ХД.

Обратим внимание на то, что сценарий загрузки *не привязан* непосредственно к данным, он привязан к их структуре, т.е. в нем смоделирована последовательность действий, которые нужно сделать для загрузки информации в ХД: имена файлов-источников, соответствие полей и т.д. Один раз созданный сценарий впоследствии используется повторно для пополнения хранилища данных. Для этого

только нужно выгрузить новую информацию о продажах и измерениях в текстовые файлы. Как правило, эти процедуры проводятся по регламенту в нерабочее время (например, ночью) с использованием пакетного режима. Настройка пакетного режима является прерогативой системного администратора.

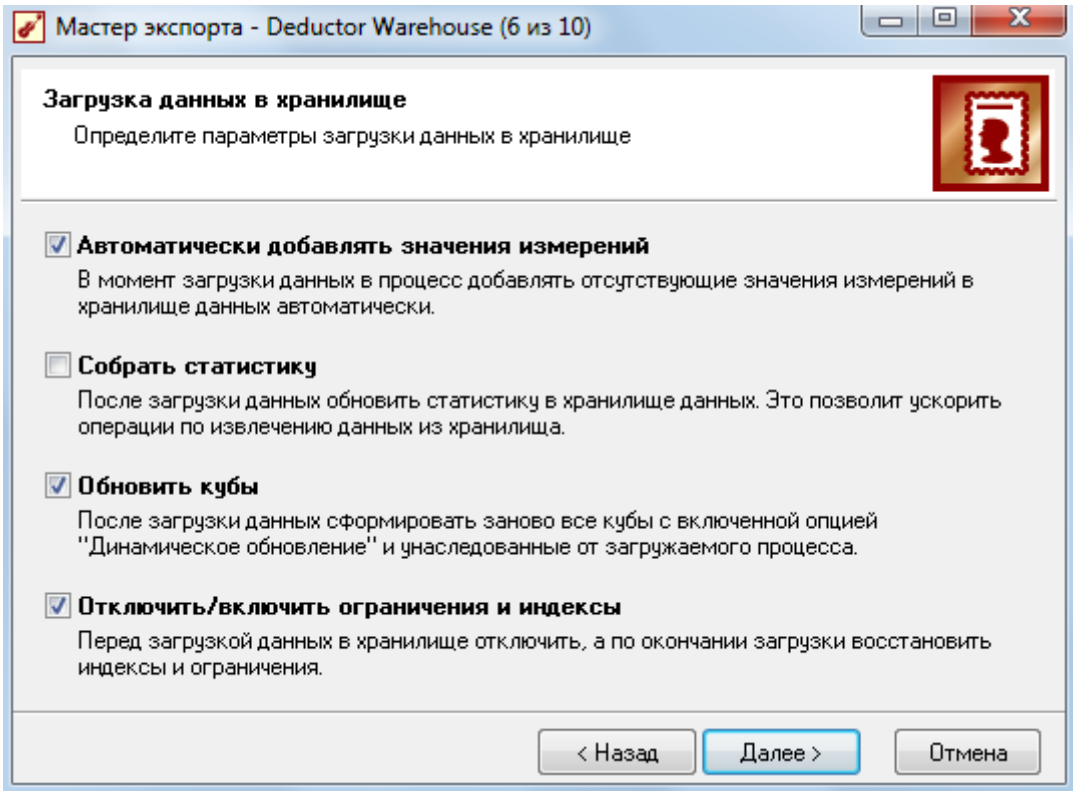


Рисунок 15.29 – Вспомогательные параметры загрузки в процесс

Извлечение информации из хранилища

Процесс получения данных из хранилища осуществляется через **Мастер импорта**. Осуществим импорт данных из процесса *Продажи* за последние 3 месяца. Для этого необходимо выполнить следующие действия.

1. С помощью Мастера импорта выберите тип источника данных – Deductor Warehouse, на следующем шаге – ХД *Фармация*, а затем – процесс *Продажи*.
2. Определите, какие измерения и атрибуты из выбранного на предыдущем шаге процесса должны быть импортированы (рис.15.30). Заметим, что внутри измерения *Товар.Код* появилась возможность доступа к измерению *Группа.Код*.
3. Определите импортируемые факты и виды их агрегаций (рис.15.31). В большинстве случаев требуется агрегация в виде суммы.

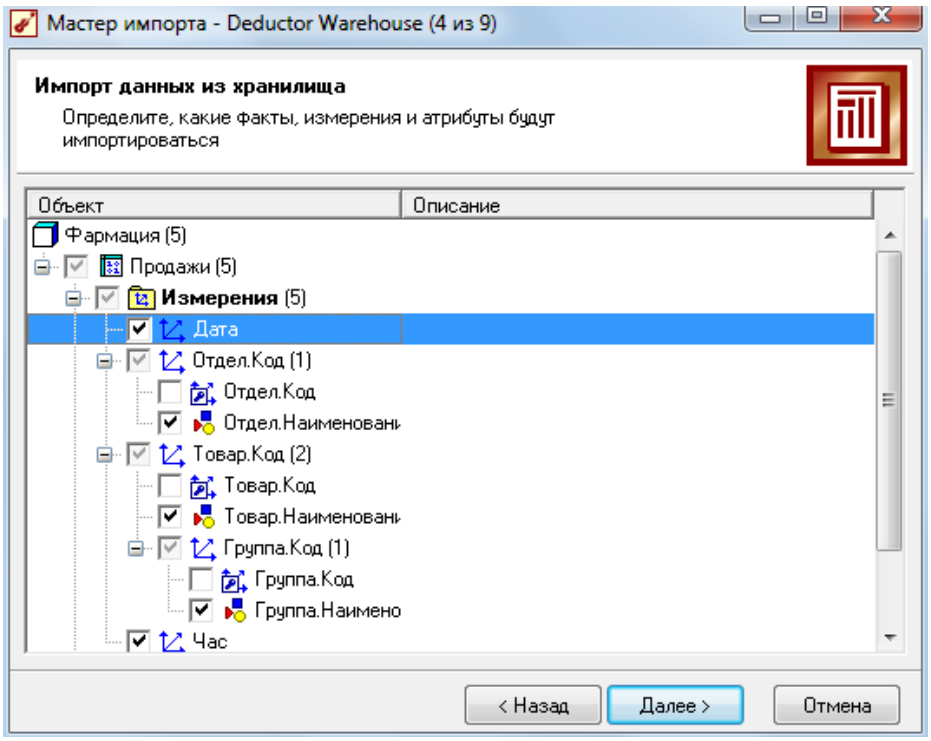


Рисунок 15.30 – Выбор импортируемых измерений и атрибутов

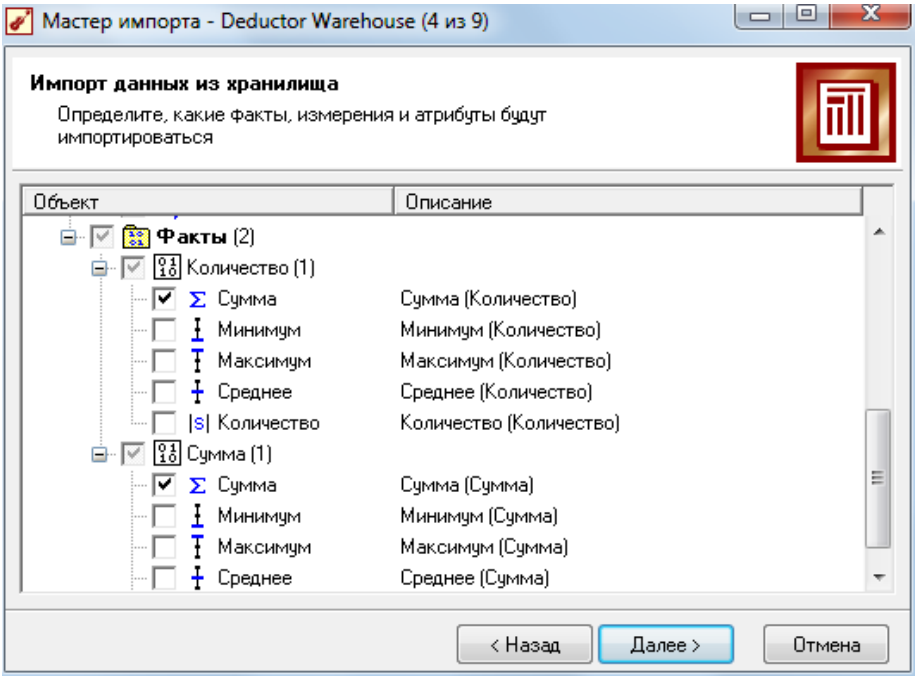


Рисунок 15.31 – Выбор импортируемых фактов

4. Определите срезы для выбранных измерений. Это целесообразно делать при большом количестве значений измерения, так как позволяет загрузить с сервера, на котором расположено ХД, только интересующие пользователя значения измерения и тем самым сэкономить время загрузки данных. На рисунке 15.32 приведен пример среза «Все продажи за последние 3 месяца от имеющихся данных».

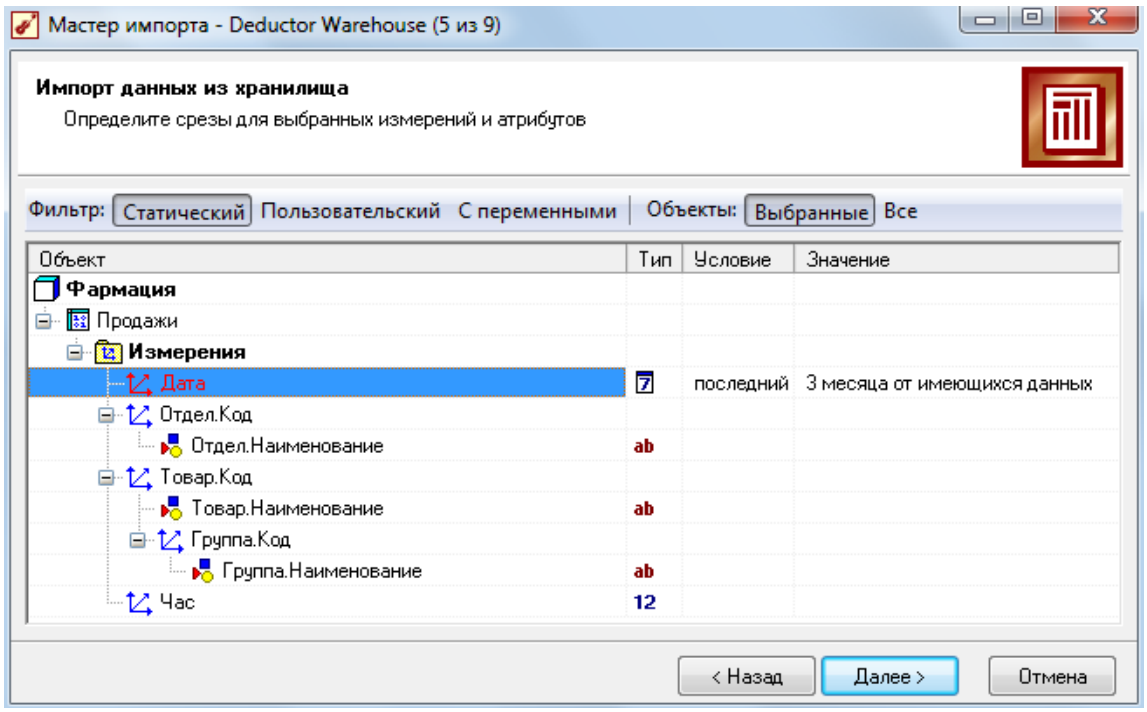


Рисунок 15.32 – Выбор среза из ХД

Для результирующего набора данных определите способ его отображения.

Задание

1. Повторите все действия, описанные выше: создайте пустое хранилище данных *Фармация*, спроектируйте структуру ХД и загрузите в него информацию из следующих текстовых файлов: **groups.txt**, **produces.txt**, **stores.txt**, **sales.txt**. Результатом работы должен стать сценарий загрузки **load.ded**.
2. Убедитесь, что в хранилище загружена вся информация о продажах.
3. Импортируйте информацию о продажах из ХД, включая атрибуты товара. Установите следующие срезы:
   - «кроме последнего периода 1 месяц от имеющихся данных»;
   - срез только по одной любой товарной группе.

Вопросы для самоконтроля

- Для чего предназначено хранилище данных?
- Из каких частей состоит хранилище данных?
- Для чего нужен семантический слой в хранилище данных?
- Что представляет собой структура «снежинка»?
- Назовите примеры объектов хранилища данных.

Практическое занятие №16

Многомерные отчеты и OLAP

OLAP – технология обработки информации, включающая составление и динамическую публикацию отчетов и документов.
Wikipedia

Цель работы – освоить и закрепить навыки создания хранилища данных и извлечения из него информации, построения многомерных отчетов и их анализа.

Теоретические сведения

Многомерный анализ данных

Механизм **OLAP** является на сегодня одним из популярных методов анализа данных. OLAP (англ.: OnLine Analytical Processing) – технология оперативной аналитической обработки данных, обеспечивающая возможность многомерного анализа данных.

Основное назначение OLAP – поддержка аналитической деятельности, произвольных (не регламентированных) запросов лиц, принимающих решения. На основе OLAP строятся системы поддержки принятия решений и системы подготовки отчетов.

OLAP-анализ может быть применен для построения отчетности, а также для первичной проверки возникающих гипотез.

В процессе принятия решений аналитик генерирует некоторые гипотезы. Для превращения этих гипотез в законченные решения они должны быть проверены. Проверка гипотез осуществляется на основании информации об анализируемой предметной области. Как правило, наиболее удобным способом представления такой информации для человека является зависимость между некоторыми параметрами. Например, зависимость объемов продаж от региона, времени, категории товара и т.п. Другим примером может служить зависимость количества выздоравливающих пациентов от применяемых средств лечения, возраста и т.п.

В процессе анализа данных, поиска решений часто возникает необходимость в построении зависимостей между различными параметрами. Кроме того, число таких параметров может варьироваться в широких пределах. Традиционные средства анализа, оперирующие данными, которые представлены в виде таблиц реляционной базы данных, не могут в полной мере удовлетворить таким требованиям. Чаще всего данные по различным параметрам анализируемого процесса хранятся в разрозненных таблицах и нужно затратить немало времени, чтобы свести их в единое целое. При этом увидеть зависимость между параметрами зачастую очень сложно.

В OLAP-системах разрозненная информация представляется в виде *многомерного куба*, которым можно легко манипулировать, извлекая срезами нужную информацию (рис. 16.1).

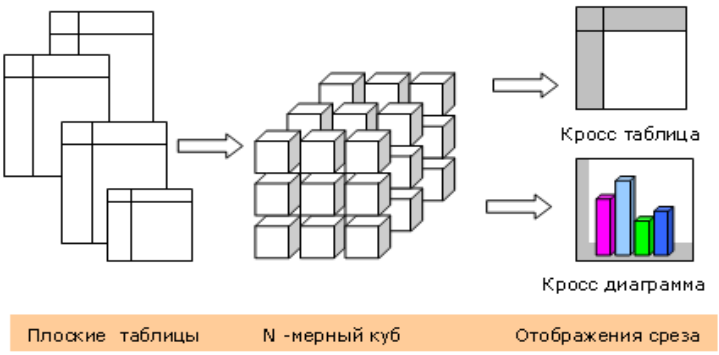


Рисунок 16.1– Технология OLAP

Проиллюстрируем идею OLAP-куба на простом примере.

Пусть руководителя интересуют объемы продаж за некоторый период, к примеру, за последние два месяца. Компания продает лекарственные средства и имеет 3 аптеки.

Первые два простейших вопроса, на которые нам сразу же хотелось бы иметь ответы, – это объемы продаж по товарным группам и объемы продаж товаров по каждой аптеке за каждый месяц.

Очевидно, что «ответ» на каждый из этих вопросов будет оформлен в виде двумерной таблицы. В первом случае строками и столбцами этой таблицы соответственно будут названия товарных групп, месяцы и суммы, а во втором - названия аптек и суммы.

Однако анализировать информацию в таком виде неудобно. Возникает потребность «соединить» данные нескольких таблиц. В итоге в таком отчете будет фигурировать три равноправных аналитических измерения (аптека, товарная группа и месяц), и вместо двумерных таблиц появляется трехмерная модель представления данных, так называемый куб (рис. 16.2).

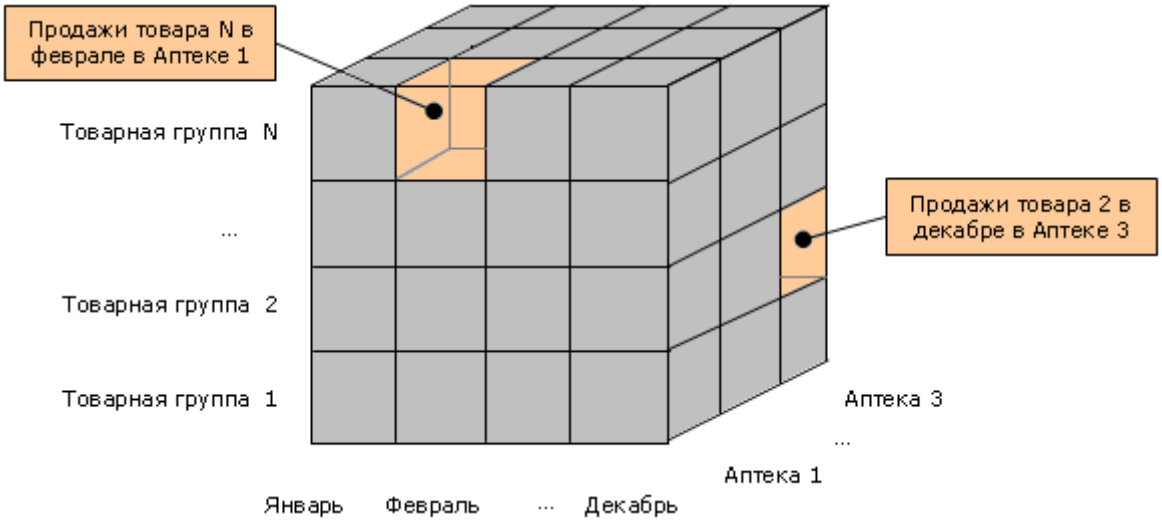


Рисунок 16.2 – Трёхмерная модель представления данных (куб)

Такая модель представления данных позволяет получать нужную для аналитика информацию, производя соответствующие сечения (срезы) OLAP-куба. Такие срезы исходного куба представляются на экране в виде кросс-таблицы и кросс-диаграммы.

Кросс-таблица отличается от обычной (плоской) таблицы наличием уровней вложенности, например, разбиение строк (столбцов) на подстроки (подстолбцы).

Рассмотрим срез куба из хранилища *Фармация* (см. предыдущую лабораторную работу) по продажам за последние два месяца во всех трех аптеках по всем товарным группам. Получим следующий многомерный отчет (рис. 16.3).

Нет необходимости пытаться произвести геометрическую интерпретацию OLAP-куба с размерностью более 3-х измерений. Тем более что речь идет не о реальном, а об информационном пространстве, а само понятие «многомерный куб» есть не что иное, как служебный термин, используемый для описания метода. В принципе, используемое число измерений может быть любым. Каждое новое измерение (товар, час покупки) будет представлено новой осью.

Следует отметить, что задача расчета и визуализации куба с большим числом измерений, во-первых, является трудоемкой с точки зрения ее выполнения на компьютере, и, во-вторых, ее осмысление и интерпретация результатов аналитиком может быть затруднена. Поэтому, с методической точки зрения, сложные задачи, требующие анализа данных большой размерности, следует по возможности сводить к нескольким более простым. Как правило, человек не способен одновременно анализировать больше 5-7 измерений.

			Отдел.Наименование			
Дата (Год)	**Дата (Месяц)**	**Группа.Наименование**	Аптека 1	Аптека 2	Аптека 3	**Итого:**
2004	11 Ноябрь	Антисептики и дезинфицирующие средства	10 734.91	5 327.69	11 431.80	27 494.40
		Биологически активные пищевые добавки	300.75	33.59	854.08	1 188.42
		Витамины и витаминоподобные средства	8 354.38	5 956.64	14 079.23	28 390.25
		Желчегонные средства и препараты желчи	76.47	42.61	299.46	418.54
		Иммуномодуляторы	11 322.86	4 574.72	17 272.79	33 170.37
		Местные анестетики	888.74	118.99	979.22	1 986.95
		Микро- и макроэлементы	564.34		511.68	1 076.02
		Общетонизирующие средства и адаптогены	1 171.41	362.52	971.02	2 504.95
		Итого:	33 413.86	16 416.76	46 399.28	96 229.90
	12 Декабрь	Антисептики и дезинфицирующие средства	13 102.61	7 421.62	12 358.47	32 882.70
		Биологически активные пищевые добавки			33.25	33.25
		Витамины и витаминоподобные средства	7 893.09	6 812.91	11 763.64	26 469.64
		Желчегонные средства и препараты желчи	185.64	57.57	107.94	351.15
		Иммуномодуляторы	9 911.07	5 971.85	14 197.24	30 080.16
		Местные анестетики	447.52	411.66	870.08	1 729.26
		Микро- и макроэлементы	623.73		1 004.41	1 628.14
		Общетонизирующие средства и адаптогены	432.83	690.13	2 786.23	3 909.19
		Итого:	32 596.49	21 365.74	43 121.26	97 083.49
	Итого:		66 010.35	37 782.50	89 520.54	193 313.39
Итого:			66 010.35	37 782.50	89 520.54	193 313.39

Рисунок 16.3 – Пример OLAP-отчета

OLAP в Deductor

В Deductor OLAP – это визуализатор «Куб» с двумя типами отчетов внутри: кросс-таблица и кросс-диаграмма. Кросс-диаграмма представляет собой диаграмму заданного типа, построенную на основе кросс-таблицы. Основное отличие кросс-диаграммы от обычной диаграммы в том, что она однозначно соответствует теку-

щему состоянию кросс-таблицы и при любых ее изменениях изменяется соответственно.

Рассмотрим порядок настройки OLAP-отчета на примере ХД *Фармация* по работе сети аптек. Поставим задачу построить многомерный отчет, отражающий динамику сумм продаж по месяцам года в разрезе групп и аптек. В нашем распоряжении имеется только измерение *Дата*, а для построения отчета требуется измерение *Месяц*. Месяц года из даты можно получить, применив к узлу импорта из хранилища обработчик **Дата и время**. В параметрах обработчика зададим для поля *Дата* тип разбиения **Год** и **Месяц**, как это показано на рисунке 16.4.

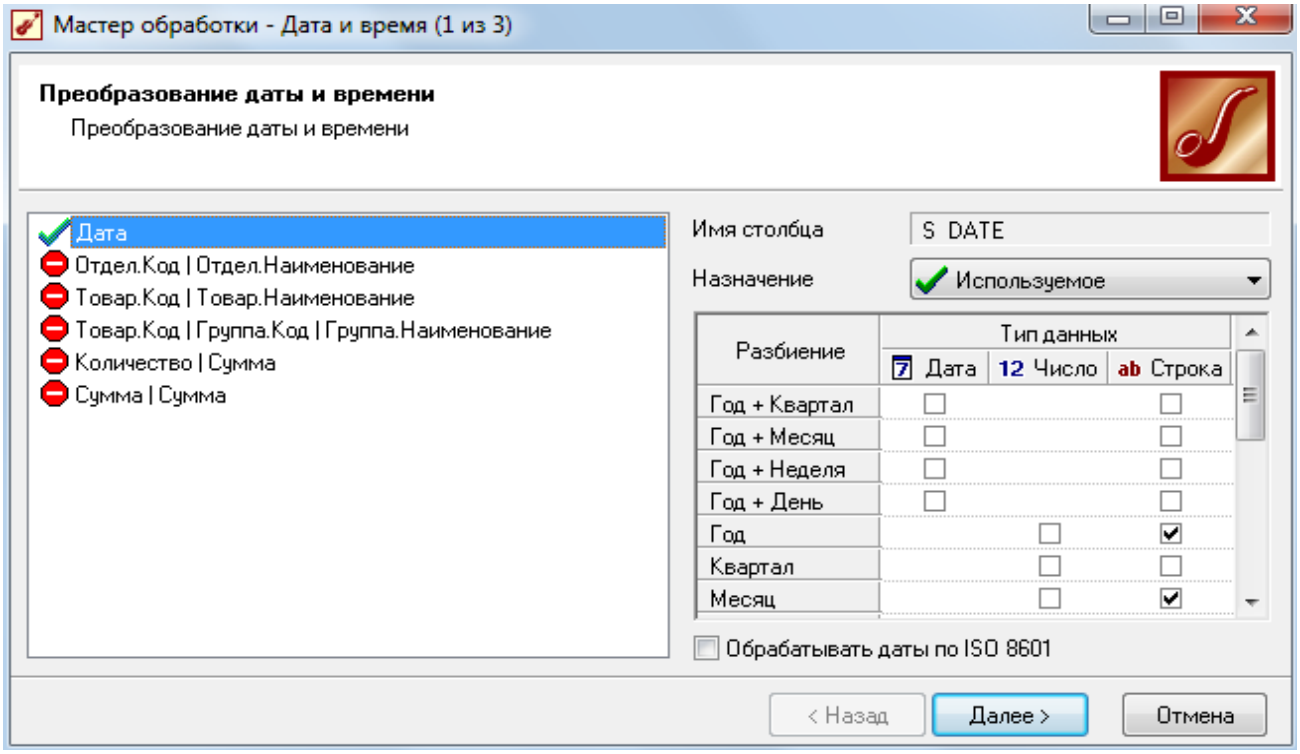


Рисунок 16.4 – Обработчик «Дата и время»

В результате работы данного обработчика в выходном наборе будет создано 2 новых столбца с метками *Дата (Год)* и *Дата (Месяц)*, а сценарий будет состоять из двух узлов (рис.16.5).

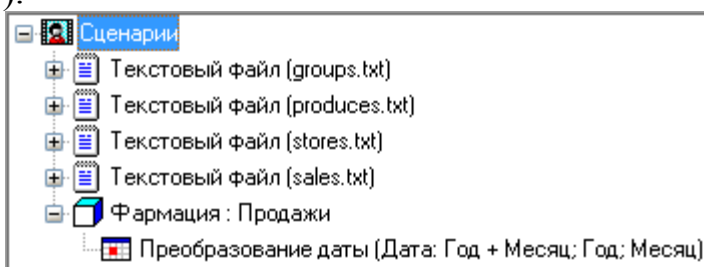


Рисунок 16.5 – Фрагмент сценария

Дальше активируем **Мастер визуализации** (нажать кнопку на панели инструментов, или пункт **Мастер визуализации** во всплывающем меню), после чего выбрать способ отображения данных в виде куба (рис. 16.6).

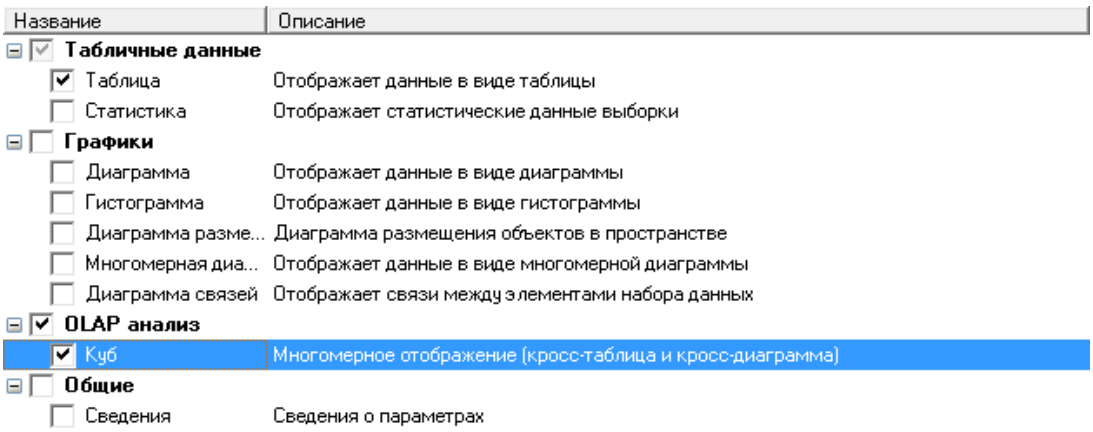


Рисунок 16.6 – Выбор способа отображения данных в виде куба

2. Произвести настройку назначений полей куба, то есть указать измерения и факты (рис. 16.7).

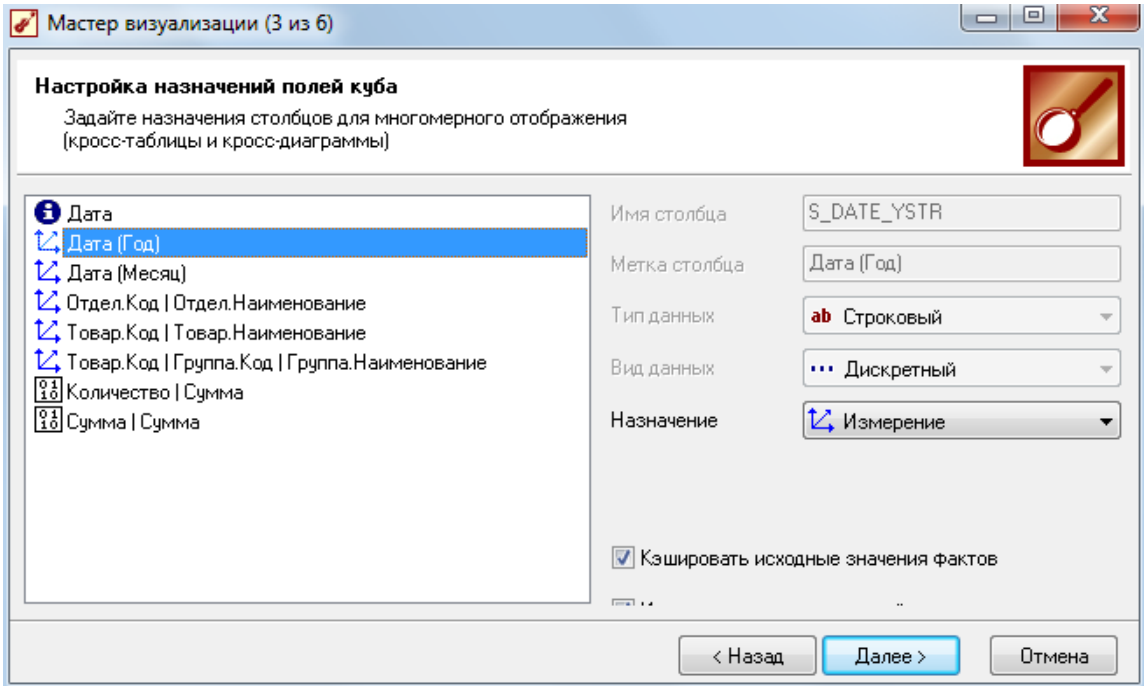


Рисунок 16.7– Настройка назначений полей куба

В данном случае измерения – это *Дата (Месяц)*, *Дата (Год)*, *Отдел.Наименование, Товар.Наименование* и *Группа.Наименование*, а факты – *Количество* и *Сумма* проданных товаров (с агрегацией «Сумма»). Информационное поле – *Дата* – не будет отображаться при построении кросс таблицы и кросс-диаграммы, но будет доступно в детализации.

На следующем шаге нужно задать размещение измерений по строкам/столбцам (рис.16.8).

На последнем шаге определяем, какие факты отображать в кросс-таблице на пересечении измерений (рис.16.9).

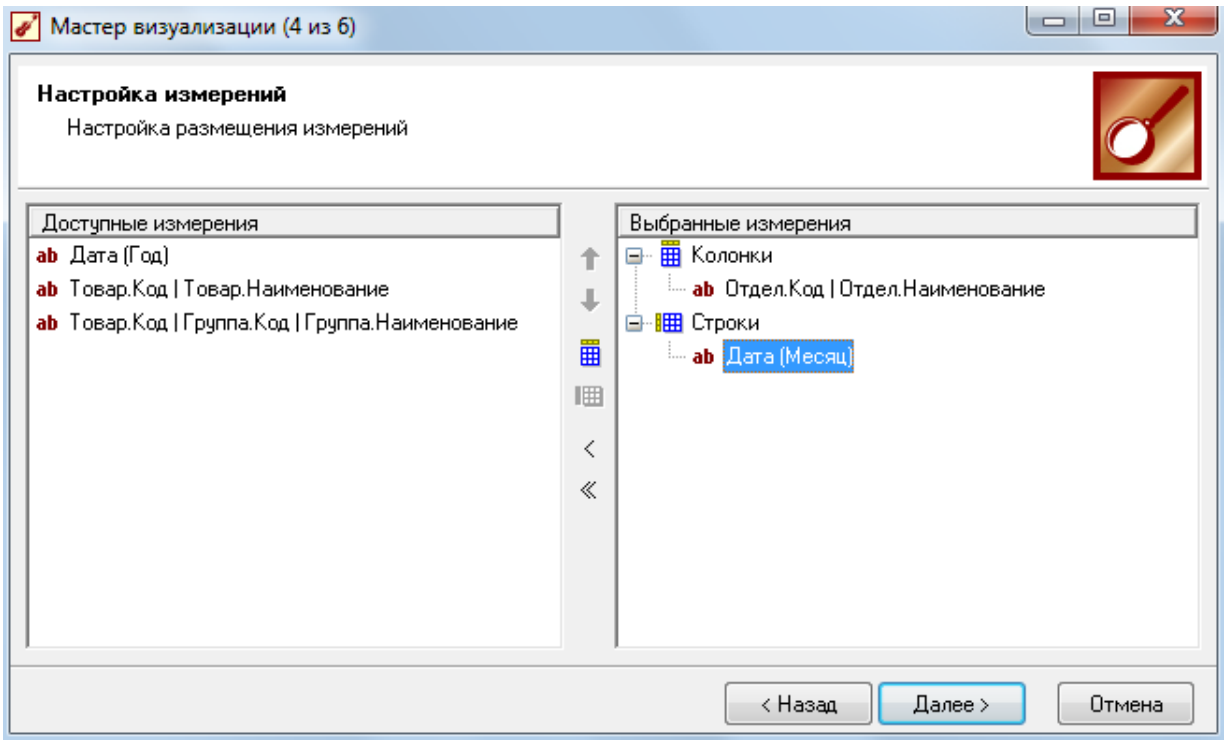


Рисунок 16.8 – Настройка размещений полей куба

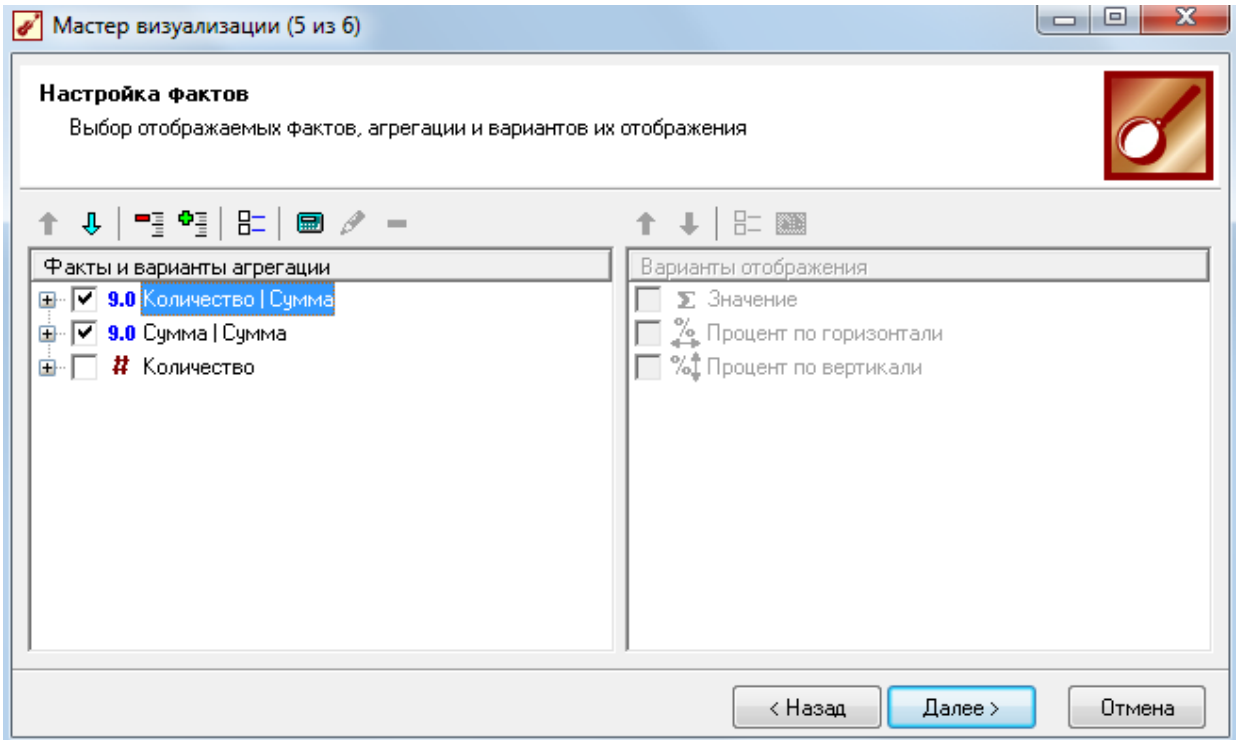


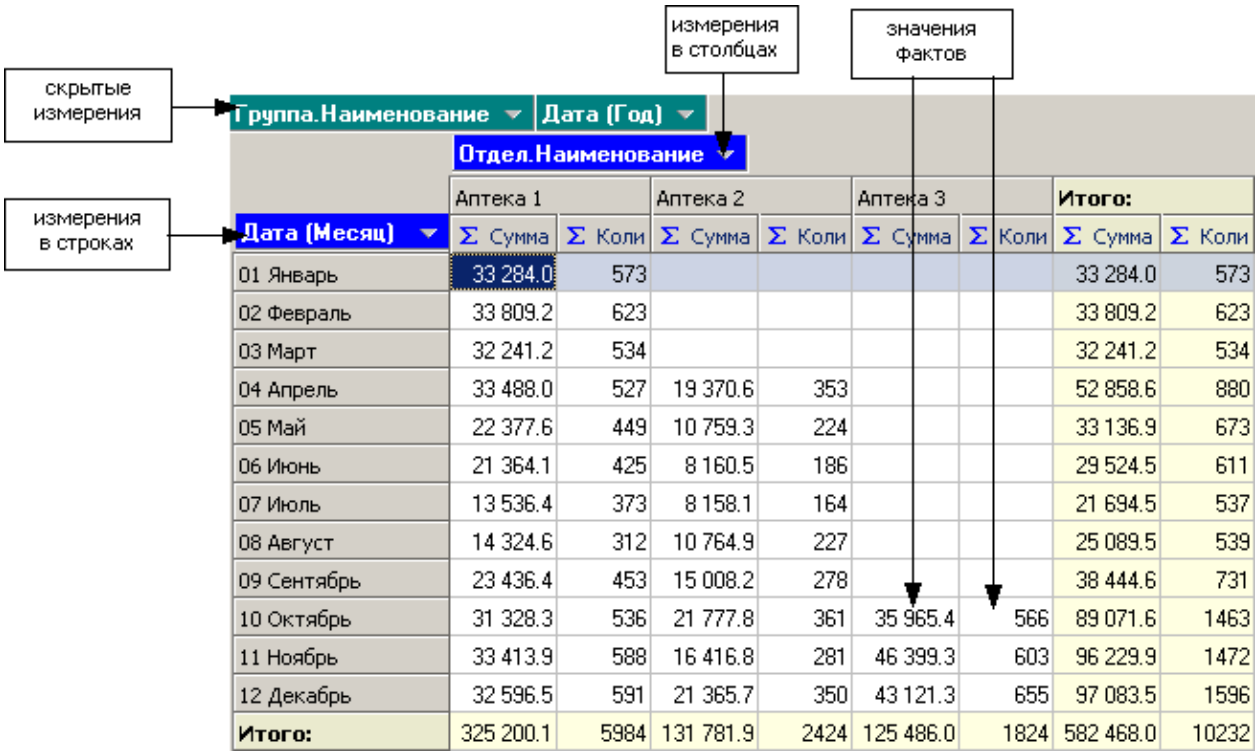
Рисунок 16.9 – Настройка отображения фактов

Для отображения фактов предусмотрено пять способов объединения (агрегирования) фактов в кросс таблице:

- Сумма – вычисляется сумма объединяемых фактов;
- Минимум – среди всех объединяемых фактов в таблице отображается только минимальный;

- Максимум - среди всех объединяемых фактов в таблице отображается только максимальный;
- Среднее – вычисляется среднее значение объединяемых фактов;
- Количество – в кросс таблице будет отображаться количество объединенных фактов.

В результате для нашего примера получим следующую кросс таблицу (Рисунок 16.10).



Группа.Наименование / Дата (Год)								
Отдел.Наименование	Аптека 1		Аптека 2		Аптека 3		Итого:	
Дата (Месяц)	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли
01 Январь	33 284.0	573					33 284.0	573
02 Февраль	33 809.2	623					33 809.2	623
03 Март	32 241.2	534					32 241.2	534
04 Апрель	33 488.0	527	19 370.6	353			52 858.6	880
05 Май	22 377.6	449	10 759.3	224			33 136.9	673
06 Июнь	21 364.1	425	8 160.5	186			29 524.5	611
07 Июль	13 536.4	373	8 158.1	164			21 694.5	537
08 Август	14 324.6	312	10 764.9	227			25 089.5	539
09 Сентябрь	23 436.4	453	15 008.2	278			38 444.6	731
10 Октябрь	31 328.3	536	21 777.8	361	35 965.4	566	89 071.6	1463
11 Ноябрь	33 413.9	588	16 416.8	281	46 399.3	603	96 229.9	1472
12 Декабрь	32 596.5	591	21 365.7	350	43 121.3	655	97 083.5	1596
Итого:	325 200.1	5984	131 781.9	2424	125 486.0	1824	582 468.0	10232

Рисунок 16.10 – Результат OLAP-анализа в виде кросс-таблицы

Измерения в кросс-таблице изображаются специальными полями. Поля синего цвета показывают измерения, участвующие в построении кросс таблицы. Поля темно-зеленого цвета указывают на скрытые измерения, не участвующие в построении кросс-таблицы.

Фильтрация данных в кросс таблице может производиться двумя способами:

- по значениям фактов;
- по значениям измерений, путем непосредственного выбора значений из списка, или отбора их по условию. Фильтрация выполняется отдельно по каждому измерению.

Вернемся к нашему примеру, рассмотренном при построении кросс-таблицы для анализа работы сети аптек. Пусть имеются данные по продажам в сети аптек за один год (рис.16.11).

В данной кросс таблице представлены измерения:

- *Отдел.Наименование* – Аптека 1, Аптека 2 и Аптека 2;
- *Дата (Месяц)* – месяцы работы отделов (01 Январь, 02 Февраль и т.д.);
- *Группа.Наименование* – названия групп лекарственных препаратов, присутствующих в продаже (Антисептики, Витамины и т.п.).

и факты – сумма и количество проданных медикаментов.

При этом *Дата (Месяц)* и *Отдел.Наименование* являются рабочими измерениями, а *Группа.Наименование* – скрытым измерением.

Для фильтрации данных в кросс таблице необходимо во всплывающем меню или на панели инструментов нажать на кнопку  , после чего будет открыто окно селектора (рис. 16.12).

Группа.Наименование

	Отдел.Наименование							
	Аптека 1		Аптека 2		Аптека 3		Итого:	
Дата (Месяц)	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли	Σ Сумма	Σ Коли
01 Январь	33 284.0	573					33 284.0	573
02 Февраль	33 809.2	623					33 809.2	623
03 Март	32 241.2	534					32 241.2	534
04 Апрель	33 488.0	527	19 370.6	353			52 858.6	880
05 Май	22 377.6	449	10 759.3	224			33 136.9	673
06 Июнь	21 364.1	425	8 160.5	186			29 524.5	611
07 Июль	13 536.4	373	8 158.1	164			21 694.5	537
08 Август	14 324.6	312	10 764.9	227			25 089.5	539
09 Сентябрь	23 436.4	453	15 008.2	278			38 444.6	731
10 Октябрь	31 328.3	536	21 777.8	361	35 965.4	566	89 071.6	1463
11 Ноябрь	33 413.9	588	16 416.8	281	46 399.3	603	96 229.9	1472
12 Декабрь	32 596.5	591	21 365.7	350	43 121.3	655	97 083.5	1596
Итого:	325 200.1	5984	131 781.9	2424	125 486.0	1824	582 468.0	10232

Рисунок 16.11 – Кросс-таблица

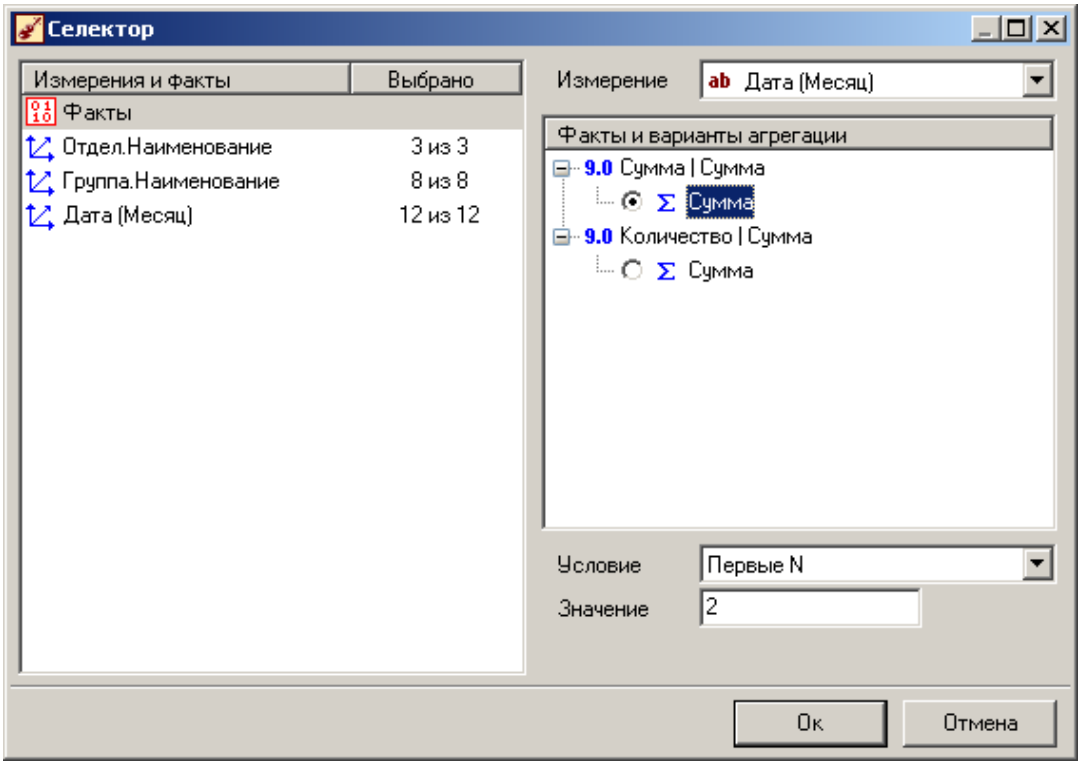


Рисунок 16.12 – Окно селектора

Это пример окна для фильтрации данных **по значениям фактов**. Слева отображаются все измерения кросс таблицы и поле «Факты», означающее фильтрацию по фактам. Справа находятся элементы:

- Измерение. Фильтрация подразумевает, что в таблице останется лишь часть значений некоторого измерения. Это поле как раз и задает измерение, значения которого будут отфильтрованы;
- Факт. В кросс-таблице может содержаться один и более фактов. Фильтрация будет происходить по значениям выбранного здесь факта.
- Агрегация – можно выбрать функцию агрегации, в соответствии с которой следует выполнить отбор записей. В результате будут выбраны только те записи, агрегированные значения которых удовлетворяют выбранному условию;
- Условие – условие отбора записей по значениям выбранного факта.

Поле «условие» может принимать различные значения, перечислим некоторые из них:

- Первые N. Значения измерения сортируются в порядке убывания факта и выбираются первые N значений измерений. Таким образом, можно, например, находить лидеров продаж – первые 10 наиболее продаваемых товаров, или первые 5 наиболее удачных дней;
- Последние N. Значения измерения сортируются в порядке убывания факта и выбираются последние N значений измерений. Например, 10 наименее популярных товаров;
- Доля от общего. Значения измерения сортируются в порядке убывания факта. В этой последовательности выбирается столько первых значений измерения, сколько в сумме дадут заданную долю от общей суммы. Например, можно отобрать клиентов, приносящих 80% прибыли, или товары, дающие 50 % объема продаж;
- Диапазон. Результатом отбора будут записи, для которых значение соответствующего факта лежит в заданном диапазоне;
- Больше. Будут отобраны записи, значение соответствующего факта, для которых будет больше указанного значения;
- Меньше. Будут отобраны записи, значение соответствующего факта, для которых будет меньше указанного значения.

Отдел.Наименование	Дата (Месяц)	
Группа.Наименование	Σ Сумма	Σ Коли
Антисептики и дезинфицирующие средства	171 903.8	4186
Биологически активные пищевые добавки	3 740.9	24
Витамины и витаминоподобные средства	200 440.5	2290
Желчегонные средства и препараты желчи	3 454.7	134
Иммуномодуляторы	160 210.4	2024
Местные анестетики	14 353.0	621
Микро- и макроэлементы	10 614.9	573
Общетонизирующие средства и адаптогены	17 749.6	380
Итого:	582 468.0	10232

Рисунок 16.13 – Кросс-таблица перед фильтрацией данных

Рассмотрим пример. Пусть нам нужно определить товарные группы, приносящие 80% выручки. Исходная кросс таблица содержит 8 товарных групп (рис. 16.13).

Применим к ней селектор (рис.16.14).

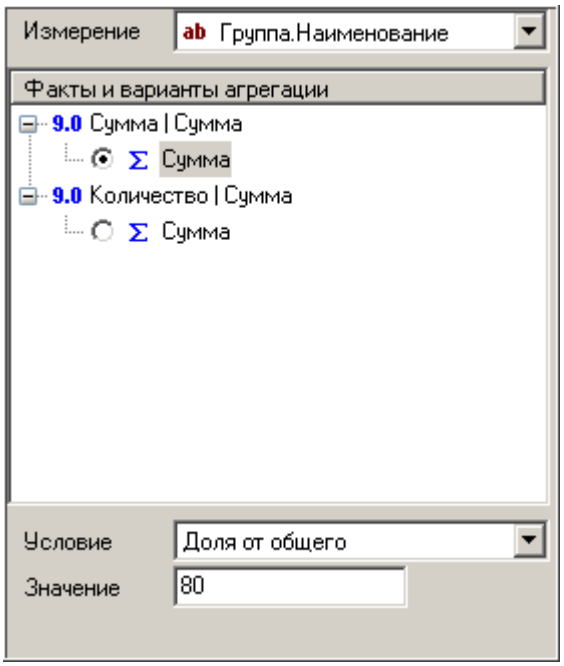


Рисунок 16.14 – Селектор

В результате получим товарные группы, приносящие основной доход. Такую выборку можно получить по любому факту. В данном примере – это сумма. Если отфильтровать по количеству, то получим товары, пользующиеся наибольшим спросом.

Разработка системы аналитической отчетности

В процессе работы специалист-аналитик выполняет множество операций над анализируемыми данными. Результаты его работы могут быть интересны широкому кругу лиц – так называемым конечным пользователям, которым не обязательно вникать в последовательность действий аналитика, знать особенности математического аппарата и методов, применяемых при анализе данных. Для представления результатов анализа для конечных пользователей может быть использована аналитическая отчетность.

Аналитическая отчетность (отчеты) – это одно из средств визуализации и консолидации результатов анализа данных для конечного пользователя. Аналитическая отчетность обеспечивает быстрый доступ к результатам анализа, не требуя от пользователя навыков анализа данных и работы в пакете Deductor. При работе с отчетами пользователь не видит сценарий анализа данных, ему доступны только конечные результаты (выдержки) из работы аналитика.

Для создания аналитической отчетности необходимо в меню **Вид** выбрать пункт **Отчеты**, или нажать соответствующую кнопку на панели инструментов. В результате в рабочей части экрана появится панель **Отчеты**.

Отчеты строятся в виде древовидного иерархического списка, каждым узлом которого является отдельный отчет или папка, содержащая несколько отчетов. Каждый узел дерева отчетности связан со своим узлом в дереве сценария. Для каждого отчета настраивается свой способ отображения (таблица, гистограмма, кросс таблица, кросс диаграмма и т.п.). Это удобно, так как несколько отчетов могут быть связаны с одним узлом дерева сценария.

Для создания нового отчета необходимо нажать на кнопку на панели инструментов или выбрать соответствующую команду **Добавить узел** из всплывающего меню. В результате откроется окно **Выбор узла** в котором следует выделить узел дерева сценария, где содержится нужная выборка данных и щелкнуть по кнопке **Выбрать**.

Для создания новой папки для хранения отчетов необходимо нажать на кнопку на панели инструментов или выбрать соответствующую команду **Добавить папку** из всплывающего меню. Чтобы поместить отчет в папку, нужно перед созданием отчета (команда **Добавить узел**) выделить эту папку.

В результате использования перечисленных операций получим дерево отчетов. Пример изображен на рисунке 16.15.

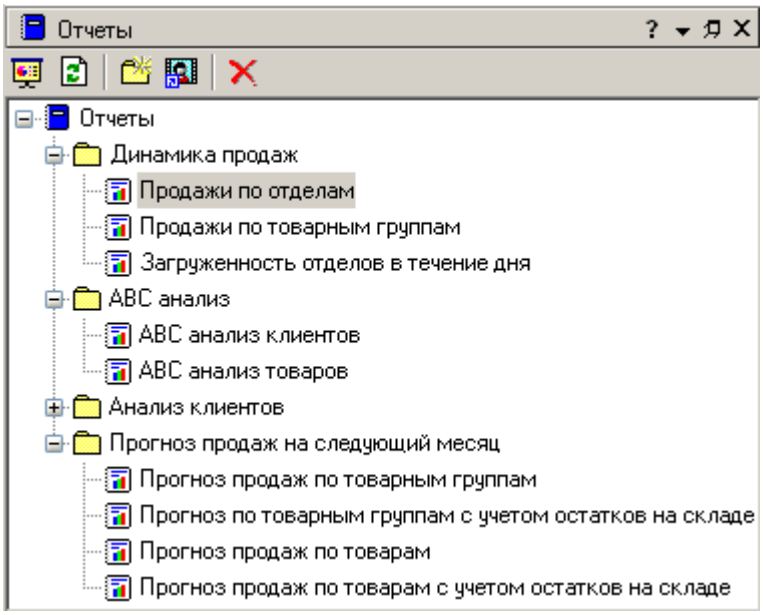


Рисунок 16.15 – Панель отчетов

Задание

Разработайте систему аналитической отчетности на основе созданного в предыдущей лабораторной работе ХД *Фармация*. Для этого напишите в Deductor Studio сценарий обработки данных и назовите его **olap.ded**. Выберите любые 5-7 отчетов из списка, приведенного ниже. Кроме кросс-таблицы, самостоятельно изучите кросс-диаграмму.

1. Построить кросс-таблицу и кросс-диаграмму по трем измерениям (отдел, месяц года, товарная группа), в ячейках которого отображается сумма и объем (количество проданных единиц продукции) продаж за все периоды, имеющиеся в ХД.

Какая торговая точка приносит наибольшую сумму продаж? Какая товарная группа имеет максимальную сумму продаж? Постройте кросс-диаграмму сумм продаж: общие продажи, продажи по торговым точкам, продажи по товарным группам.

2. То же, что в п.1, но за последние три месяца от имеющихся данных.

3. То же, что в п.1, но за последние три недели от имеющихся данных.

4. Найти сумму максимальной и средней стоимости покупки за последний месяц от имеющихся данных.

5. Сформировать многомерный отчет и график загруженности торговых точек по времени суток и торговым точкам. На какие часы приходятся пики продаж?

6. То же, что в п. 5, но за три месяца от имеющихся данных.

7. Сформировать многомерный отчет и график загруженности торговых точек по дням недели.

8. То же, что в п. 7, но за последний месяц от имеющихся данных.

9. Сформировать многомерный отчет и график загруженности торговых точек по дням месяца. Постройте линию тренда.

10. То же, что в п. 9, но за последние три месяца от имеющихся данных.

11. 20 самых продаваемых товаров.

12. То же, что в п. 11, но за последние три недели от имеющихся данных.

13. 10 самых продаваемых товаров по воскресеньям.

14. 5 самых популярных товаров в каждой товарной группе.

15. То же, что и п. 14, но за последнюю неделю.

16. Товары, дающие 50% объема продаж.

17. То же, что и п. 16, но за последние 3 месяца от имеющихся данных.

18. То же, что и п. 16, но за последнюю неделю.

19. 10 самых продаваемых товаров с 18 до 21 часа.

20. 10 товаров, пользующихся наименьшим спросом осенью.

21. Товары, дающие 50 % объема продаж в летние месяцы.

Вопросы для самоконтроля

- Дайте определение термину «OLAP-анализ»?
- Укажите назначение и цель OLAP-анализа.
- Что представляет собой OLAP-куб?
- Какие операции с кубом можно осуществлять?

Практическое занятие №17

Искусственные нейронные сети. Многослойный персептрон

Не имеет значения, похожи ли на самом деле в работе нейронные сети на мозг. Значение имеет лишь то, что у данных теоретических моделей можно математически обосновать наличие способностей к переработке информации.
К. Мид

Цель работы – освоить принципы работы с искусственными нейронными сетями в Deductor на примере аппроксимации нелинейной многомерной функции.

Теоретические сведения

Искусственные нейронные сети

Появление первых концепций искусственных нейронных сетей (ИНС), или просто нейронных сетей (НС), восходит к 1940-м годам, когда исследователи МакКаллок и Питс делали опыты по моделированию работы биологических нейронов. Несмотря на медицинскую основу этих исследований (изучалось строение мозга), в последствии оказалось, что принципы работы нейронов лягут в основу новых методов решения технических проблем, выходящих далеко за рамки науки о жизни.

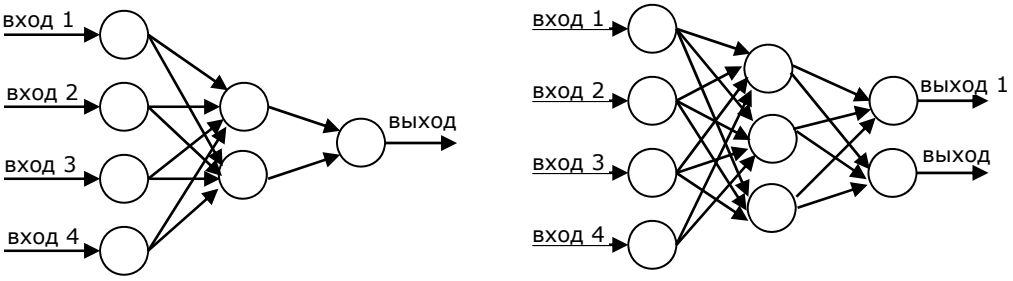
В течение 1960-х и 1970-х гг. развитие вычислительных возможностей ЭВМ позволило ученым реализовать первые прототипы моделей нейронов МакКаллока-Питса. В 1982 Хопфилд предложил метод подстройки связей между нейронами, основанный на обратном распространении ошибки.

К концу 80-х гг. теория искусственных нейронных сетей сформировалась окончательно, а стремительное развитие и популяризация персональных компьютеров дала возможность обрабатывать большие массивы данных посредством нейронных сетей за приемлемое время. Сегодня НС широко применяются в качестве универсального средства моделирования сложных систем и процессов, а нейросетевые алгоритмы встраиваются в коммерческие программные продукты. В числе популярных бизнес-задач, которые успешно решают НС, следующие: распознавание речи и текста, обнаружение мошенничеств с кредитными картами, и, конечно же, кредитный скоринг.

Нейронные сети в Data Mining используются для решения задач классификации и регрессии. Для решения этих задач получили распространение нейронные сети прямого распространения.

Что представляет собой нейронная сеть прямого распространения (ее еще называют *многослойным персептроном*)? Нейронная сеть более сложна в понимании, чем, к примеру, линейная регрессия. Рисунок 17.1 иллюстрирует примеры сетей. НС состоит из совокупности узлов (нейронов), соединен-

ных между собой связями. Существует три типа узлов: входной, скрытый и выходной. Соединяющая их связь имеет вес (числовой параметр). Направление соединяющих линий соответствуют направлению прохождения сигнала. Каждый узел является своеобразным обрабатывающим модулем. Входные узлы формируют первый слой сети. В большинстве нейронных сетей каждому входному узлу соответствует один входной атрибут (возраст, пол, доход и т.д.). Перед обработкой исходное значение входного признака должно быть отмасштабировано (часто в диапазон от -1 до 1).



а *б*

Рисунок 17.1– Примеры нейронных сетей

Скрытые узлы – это узлы, расположенные в промежуточных слоях. Скрытый узел получает входной сигнал от узлов предыдущего слоя. В этом узле объединяются все сигналы, с ними производятся некоторые вычисления, и результат обработки снова подается на вход узлов следующего слоя.

Выходные узлы соответствуют зависимым, или предсказываемым переменным. Нейронная сеть может иметь несколько выходных узлов, как это показано на Рисунок 9,б. Однако почти всегда сеть с несколькими выходами можно представить набором сетей с одним выходным узлом. Выходом нейронной сети является вещественное число в диапазоне от 0 до 1.

В режиме прогнозирования нейронная сеть работает довольно просто: поступающие сигналы подаются на входы и «прогоняются» через сеть, в результате чего на выходе генерируется рассчитанное значение. Оно подвергается операции денормализации в исходное значение (для непрерывных атрибутов) или в исходное состояние (для дискретных атрибутов).

Как видно на рисунке 17.1 структура нейронных сетей может быть различной. На Рисуноке 17.1,а представлена НС с одним скрытым слоем и одним выходом, на рисунке 17.1,б – с двумя выходами. В скрытом слое содержится 2 и 3 нейрона соответственно. Каждый нейрон скрытого слоя соединен со всеми нейронами предыдущего слоя. Наличие скрытого слоя крайне важно, поскольку это позволяет моделировать нелинейные зависимости между входами и выходами сети.

Существуют нейронные сети, в которых сигнал проходит не только в прямом направлении, но и в обратном, в структуре их связей присутствуют замкнутые циклы, однако, такие сети реже применяются на практике. Искусственные нейронные сети прямого распространения, в которых присутствует хотя бы один скрытый слой, еще называют многослойным персептроном.

После того как сформирована архитектура нейронной сети (задано число слоев и нейронов в каждом слое), запускается процесс обучения сети. Он заключается в нахождении оптимальных значений весовых коэффициентов. Это вычислительный процесс, который может длиться продолжительное время и работает так. Вначале веса инициализируются, как вариант, случайными числами. Дальше на каждой итерации все примеры из обучающей выборки «прогоняются» через нейронную сеть с текущими весовыми коэффициентами и рассчитываются выходы сети. После этого вычисляется ошибка. На основе информации об ошибке по специальному правилу корректируются веса нейронной сети. Конкретное правило зависит от выбранного алгоритма обучения.

Остановимся немного на понятии активационной функции. Каждый нейрон в сети представляет собой элементарный блок обработки. К нейрону поступают входные сигналы, и генерируется выходной сигнал. В блоке происходит суммирование сигналов, определенные вычисления над ними и преобразование в выходное возбуждение (активация). Процесс очень похож на работу биологического нейрона.

Рисунок 17.2 демонстрирует структуру нейрона. Он содержит две функции: суммирование входов и расчет выхода. Операция суммирования объединяет все входные сигналы в один. Самый распространенный способ это сделать – использовать взвешенную сумму (линейную комбинацию входных сигналов и их весов). Выходное значение рассчитывается через выражение для активационной функции.

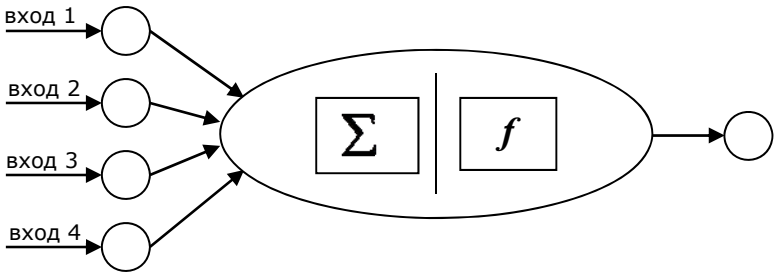


Рисунок 17.2 – Нейрон как элементарный процессор

Если провести аналогию с функционированием биологического нейрона, то обнаружится, что его выходное возбуждение подчиняется следующему поведению: небольшие изменения во входном сигнале иногда вызывают значительные изменения в выходном сигнале, и наоборот. Это свойство позволяет нейронной сети моделировать нелинейные зависимости. Несколько математических функций удовлетворяют такому поведению. Наиболее известные – сигмоида и гипертангенс. Их аналитические формулы следующие:

Sigmoid: $f = 1/(1+e^{\alpha})$

Tanh: $f = (e^{\alpha} - e^{-\alpha})/(e^{\alpha} + e^{-\alpha})$,

где a – параметр крутизны функции, f – выходное значение.

На рисунке 11.3 изображены графики активационных функций. Выходное значение сигмоидной функции изменяется в диапазоне от 0 до 1, а гипертангенс – от -1 до 1.

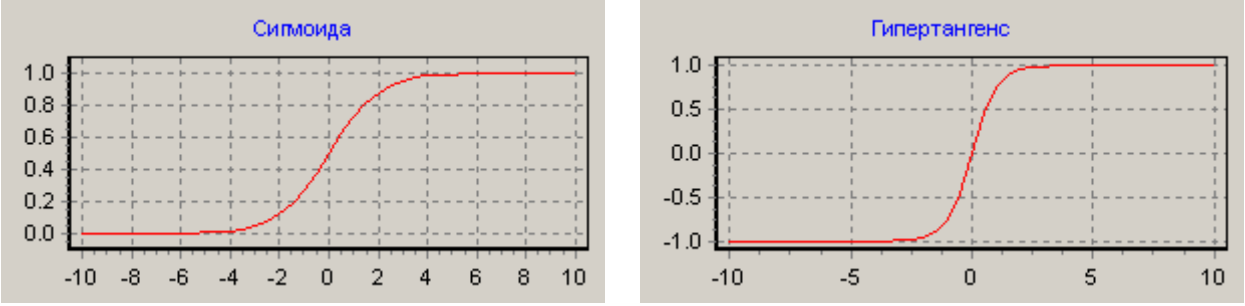


Рисунок 17.3 – Функции активации

В нейросетевой модели перед аналитиком встает вопрос определения архитектуры нейронной сети, в частности, количество скрытых слоев и нейронов в них.

При выборе архитектуры рекомендуется руководствоваться следующими правилами.

1. Количество скрытых слоев больше 2-х на практике используется редко.

2. Если две обученные нейросети имеют одинаковый порядок ошибок обучения и обобщения, то предпочтение следует отдать той нейросети, которая проще (т.е. содержит меньше скрытых слоев и нейронов).

3. Количество примеров обучающей выборки должно быть в 1,5-2 раза больше числа связей (весов). В противном случае количество подбираемых параметров будет равно либо меньше числа прецедентов, что статистически незначимо, и нейронная сеть просто «запомнит» все примеры.

Количество нейронов в скрытых слоях (при ограничениях в п.3) можно приблизительно рассчитать по следующей формуле: $c \cdot \sqrt{n \cdot m}$, где n – число входных нейронов, m – число выходов сети, c – константа (по умолчанию c=4).

Другим важным вопросом является определение момента окончания обучения. Дело в том, что слишком долгое обучение может привести к «переобучению сети», которое выражается в детальной адаптации параметров нейронной сети (весов) к любым нерегулярностям в обучающих данных. Этот эффект часто наблюдается при использовании сети с чрезмерным количеством весов. Для предупреждения переобучения в обучающем множестве выделяется область контрольных данных (тестовое множество), которые в процессе обучения применяются для оперативной проверки фактически набранного уровня обобщения.

Погрешностью обучения называется ошибка (как правило, среднеквадратическая) на обучающем множестве, погрешностью обобщения – ошибка на тестовом наборе. Истинная цель обучения состоит в таком подборе архитектуры и параметров сети, которые обеспечат минимальную погрешность

распознавания тестового множества данных, не участвовавшего в обучении. Эксперименты показали, что погрешность обучения при увеличении количества итераций монотонно уменьшается, тогда как погрешность обобщения снижается только до определенного момента, после чего начинает расти.

Алгоритм обратного распространения

Рассмотрим более детально один из самых известных классических алгоритмов обучения нейронной сети – алгоритм обратного распространения ошибки (back propagation). Он применяется для обучения многослойного персептрона – многослойной искусственной нейронной сети прямого распространения. Как уже говорилось выше, такая сеть состоит из входного и выходного слоев, а также из нескольких внутренних (скрытых) слоев.

Входной слой имеет размерность входного вектора $\mathbf{x} = [x_1,...,x_n]$. Обычно размерность вектора **x** увеличивают еще на единицу, добавляя $x_0 = 1$. Это делается для включения величины смещения функции активации в множество весовых коэффициентов. Каждый нейрон первого скрытого слоя (k=1) осуществляет суммирование входящих сигналов:

$$u_i^1 = \sum_{j=0}^{n} w_{ij}^1 x_j, \quad i = \overline{1, N_1}.$$

Выходной сигнал нейрона преобразуется с помощью функции активации

$$z_i^k = G(u_i^k),\ i = \overline{1, N_k};\ k = \overline{1, K_c},$$

где N_k – число нейронов в k-м слое; K_c – число слоев.

В качестве функции активации используется сигмоида

$$G(s) = \frac{1}{1 + \exp(-\beta \cdot s)}.$$

Производная от этой функции выражается через значения самой функции:

$$\frac{dG}{ds} = \beta \cdot G(s)\big(1 - G(s)\big).$$

Выходные преобразованные сигналы суммируются на последующем слое и, так далее, до последнего выходного слоя:

$$u_i^k = \sum_{j=0}^{N_{k-1}} w_{ij}^k z_j^{k-1}, \quad z_i^k = G(u_i^k),\ i = \overline{1, N_k},\ k = \overline{1, K_c}, \tag{17.1}$$

так, что

$$\mathbf{z}^0 = \mathbf{x}, \quad \mathbf{y} = \mathbf{z}^{K_c}. \tag{17.2}$$

Построенная таким образом нейронная сеть содержит весовые коэффициенты $w_{ij}^k,\ i = \overline{1, N_k},\ j = \overline{0, N_{k-1}},\ k = \overline{1, K_c}$, требующие определения в процессе обучения.

Для обучения используется система данных, представляющая собой набор наблюдаемых точек $(\mathbf{x}^j, \mathbf{f}^j),\ j = \overline{1, p}$, где **x**, **f** – входной вектор и вектор функции соответственно. Система данных из p точек делится на две выборки: обучающую $(\mathbf{x}^j, \mathbf{f}^j),\ j = \overline{1, h}$ и проверочную $(\mathbf{x}^j, \mathbf{f}^j),\ j = \overline{h+1, p}$. Весовые коэффициенты нужно подобрать таким образом, чтобы они обеспечили мини-

мальное отклонение рассчитываемых в сети значений **y** от имеющихся **f**, т.е. давали бы минимум целевой функции

$$F(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{m}(y_i - f_i^q)^2 \Rightarrow \min. \qquad (17.3)$$

Здесь **W** – вектор коэффициентов $w_{ij}^k, i = \overline{1, N_k}, j = \overline{0, N_{k-1}}, k = \overline{1, K_c}$, q – номер предъявляемой для обучения пары из выборки $(\mathbf{x}^q, \mathbf{f}^q), q = \overline{1, h}$.

Для решения задачи (17.3) применим метод наискорейшего спуска.

1. Нулевое приближение $\mathbf{W}^0$ задается случайным образом на промежутке (0,1).
2. В точке $\mathbf{W}^0$ вычисляется градиент функции $\mathbf{g}^0 = \nabla F(\mathbf{W}^0)$.
3. На *t*-ом шаге с помощью одномерного поиска в направлении $\mathbf{g}^t$ находим минимум $F(\mathbf{W})$, определяемый величиной λ_*, $\lambda_* = \arg\min\left(F(\mathbf{W}^t - \lambda \mathbf{g}^t)\right)$. В результате находится точка $\mathbf{W}^{t+1} = \mathbf{W}^t - \lambda_* \mathbf{g}^t$. В упрощенном варианте величина λ_* задается пользователем из диапазона (0,1).
4. Алгоритм заканчивается, когда величина модуля $\mathbf{g}^t$ не станет меньше заданного малого числа.

Для выполнения алгоритма наискорейшего спуска на каждой итерации необходимо вычислять составляющие градиента функции $F(\mathbf{W})$, величины $\frac{\partial F}{\partial w_{ij}^k}$, $i = \overline{1, N_k}, j = \overline{0, N_{k-1}}, k = \overline{1, K_c}$. Они определяются по алгоритму **обратного распространения ошибки**.

Запишем $\frac{\partial F}{\partial w_{ij}^k}$ в следующем виде:

$$\frac{\partial F}{\partial w_{ij}^k} = \frac{\partial F}{\partial z_i^k}\frac{\partial z_i^k}{\partial u_i^k}\frac{\partial u_i^k}{\partial w_{ij}^k}. \qquad (17.4)$$

По правилу дифференцирования сложной функции распишем три сомножителя в правой части выражения (11.4):

$\frac{\partial u_i^k}{\partial w_{ij}^k} = z_j^{k-1}$, согласно (11.1),

$$\frac{\partial z_i^k}{\partial u_i^k} = \frac{dG(u_i^k)}{du_i^k} = \beta G(u_i^k)\left(1 - G(u_i^k)\right) = \beta \cdot z_j^k (1 - z_i^k),$$

$$\frac{\partial F}{\partial z_i^k} = \sum_{j=1}^{N_{k+1}} \frac{\partial F}{\partial z_j^{k+1}}\frac{\partial z_j^{k+1}}{\partial z_i^k} = \sum_{j=1}^{N_{k+1}} \frac{\partial F}{\partial z_j^{k+1}}\frac{\partial z_i^{k+1}}{\partial u_j^{k+1}}\frac{\partial u_j^{k+1}}{\partial z_i^k} = \sum_{j=1}^{N_{k+1}} \frac{\partial F}{\partial z_j^{k+1}}\frac{dG(u_j^{k+1})}{du_j^{k+1}} w_{ij}^{k+1}. \qquad (5)$$

Обозначим $\delta_i^k = \frac{\partial F}{\partial z_i^k}\frac{dG(u_i^k)}{du_i^k}$. Тогда из (17.5) получим:

$$\delta_i^k = \left[\sum_{j=1}^{N_{k+1}} \delta_j^{k+1} w_{ij}^{k+1}\right]\frac{dG(u_i^k)}{du_i^k}, \qquad (17.6)$$

и выражение для составляющих градиента целевой функции примет вид:

$$\frac{\partial F}{\partial w_{ij}^k} = \delta_i^k z_j^{k-1}. \qquad (17.7)$$

Целевая функция вычисляется при этом по выражению

$$F(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{m}\left[G\left(\sum_{j=0}^{N_{Kc-1}} w_{ij}^{Kc} z_j^{Kc-1} \right) - f_i^q \right]^2. \qquad (17.8)$$

Алгоритм обучения многослойного персептрона

Из обучающей выборки системы данных $(\mathbf{x}^j, \mathbf{f}^j), j = \overline{1,h}$ случайным образом выбирается пара $(\mathbf{x}^q, \mathbf{f}^q)$. При значениях весовых коэффициентов $\mathbf{W}^0$ проводится расчет выходных сигналов сети по формулам (17.1), (17.2). Из соотношения (17.8) вычисляем:

$$\frac{\partial F}{\partial w_{ij}^{Kc}} = (y_i - f_i^q) z_j^{Kc-1} \frac{dG(u_i^{Kc})}{du_i^{Kc}} \quad \text{и} \quad \delta_i^{Kc} = (y_i - f_i^q)\beta y_i (1 - y_i).$$

Далее по формулам (11.6), (11.7), для $k = K_c$-1 до $k = 1$, вычисляются все составляющие градиента целевой функции $\frac{\partial F}{\partial w_{ij}^k}$, $i = \overline{1, N_k}, j = \overline{0, N_{k-1}}, k = \overline{1, Kc}$. Затем по алгоритму наискорейшего спуска находится следующее приближение $\mathbf{W}^t$, и итерации повторяются до приемлемого значения ошибки δ_i^{Kc}. Таким образом, данный алгоритм функционирует до тех пор, пока ошибка не станет меньше заданной, т. е. $\delta_i^{Kc} \le \delta$.

Суммарная ошибка сети оценивается по выражению (17.8).

В процессе обучения нейронная сеть проверяется на тестовой выборке и приобретает возможность прогнозирования и обобщения.

Пример работы многослойного персептрона

Рассмотрим решение задачи регрессии с помощью многослойного персептрона на примере прогнозирования результата умножения двух чисел. Для этого потребуется файл **multi.txt**, который содержится в директории демопримера Deductor. В файле содержится таблица со следующими полями: **Аргумент1**, **Аргумент2** – множители, **Произведение** – их произведение.

Импортировав данные из файла, можно посмотреть результат умножения, используя визуализатор **Таблица** (рис. 17.4).

	Аргумент1	Аргумент2	Произведение
▶	1	0	0
	0	1	0
	3	0	0
	0	3	0
	5	0	0
	0	5	0
	7	0	0
	0	7	0

Рисунок 17.4 – Фрагмент набора данных

Пусть необходимо построить модель прогноза умножения, подавая на вход которой два множителя получать на выходе их произведение. Для этого

нужно, находясь на узле импорта, открыть **Мастер обработки**. В нем выбрать обработчик **Нейросеть** и перейти к следующему шагу мастера. На втором шаге мастера необходимо установить назначение полей **Аргумент1** и **Аргумент2** как входные, а поле **Произведение** – как выходное (рис.17.5).

На следующем шаге предлагается настроить разбиение исходного множества данных на обучающее и тестовое. Оставим все опции по умолчанию (рис.17.6).

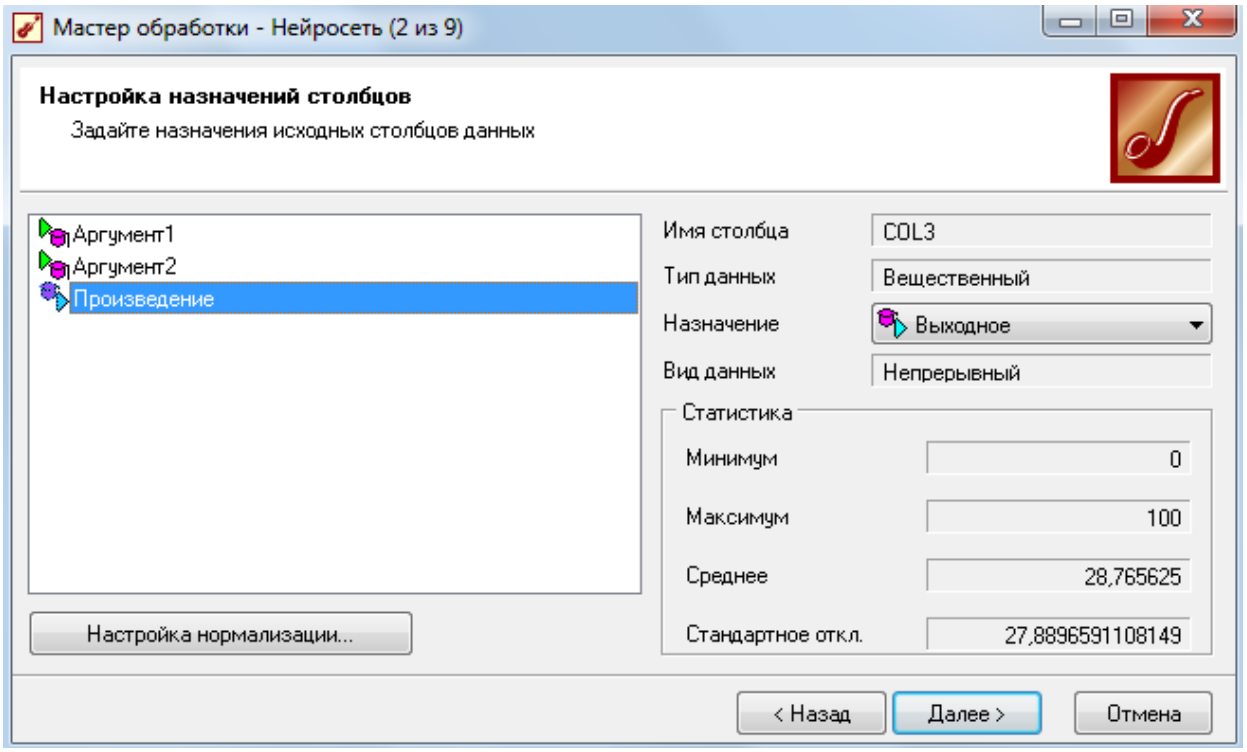


Рисунок 17.5 – Задание входов и выходов

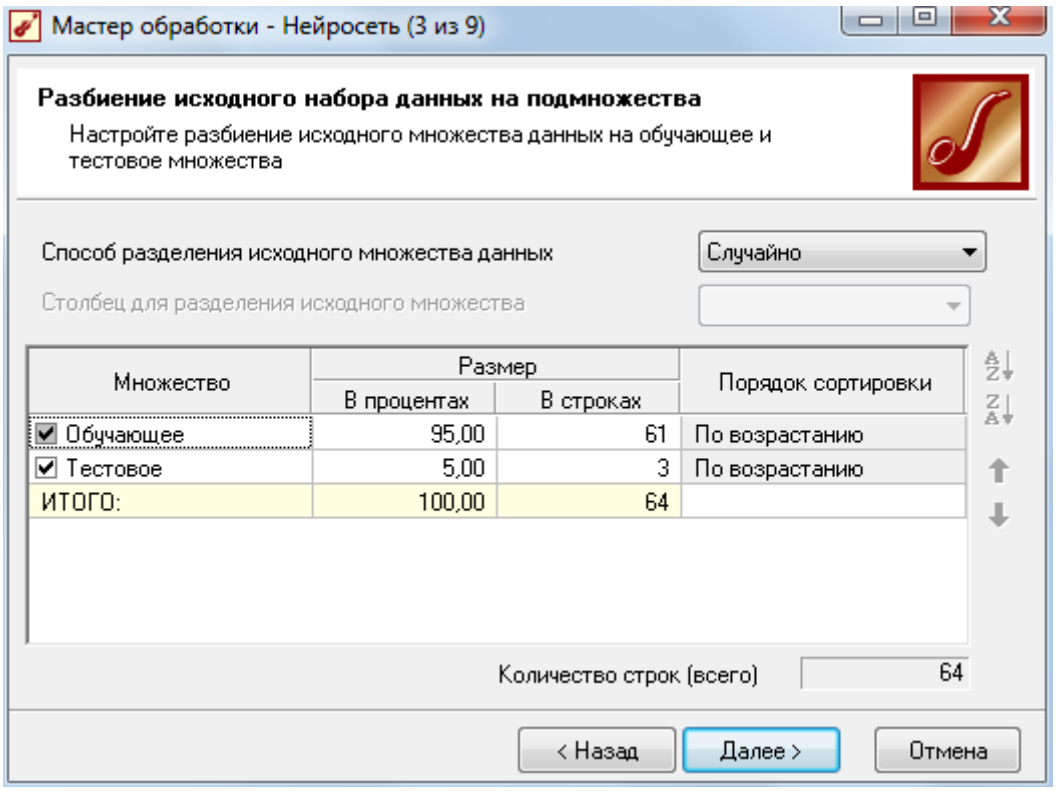


Рисунок 17.6 – Разбиение выборки на обучающую и тестовую

Третий шаг Мастера отвечает за архитектуру многослойного персептрона и параметры активационной функции. Для нашей задачи вполне достаточно одного скрытого слоя с двумя нейронами (рис.17.7).

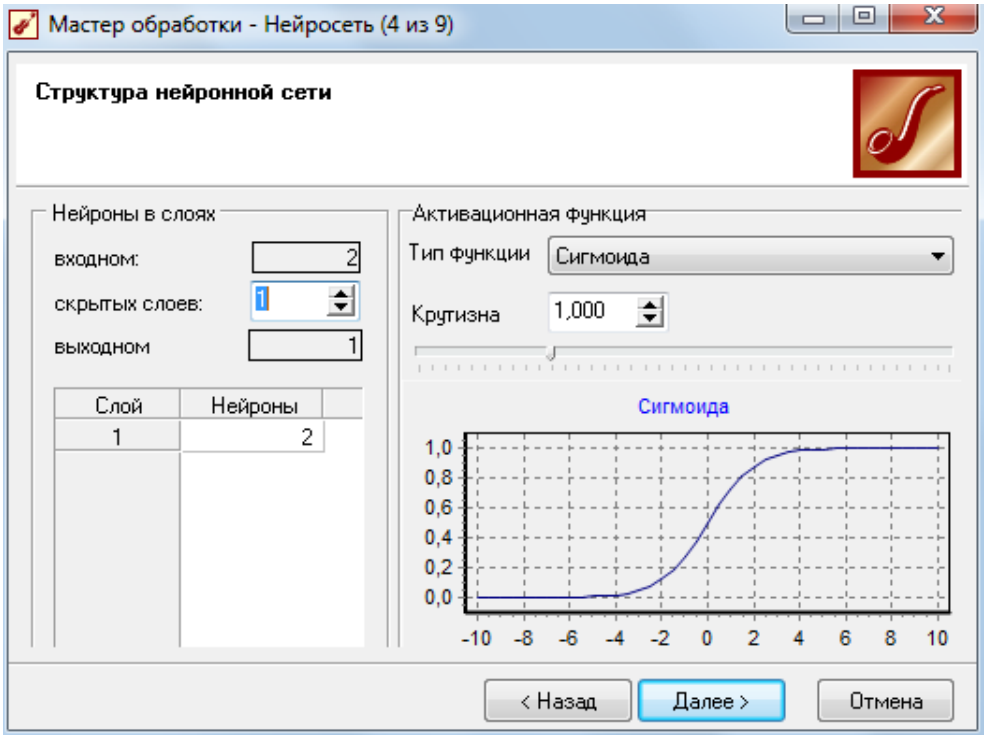


Рисунок 17.7 – Структура нейронной сети

На четвертом шаге нужно выбрать алгоритм обучения многослойного персептрона и его параметры. Выберем рассмотренный выше «Back Propagation». Коэффициенты, отвечающие за скорость и момент обучения, оставим без изменений (рис.17.8).

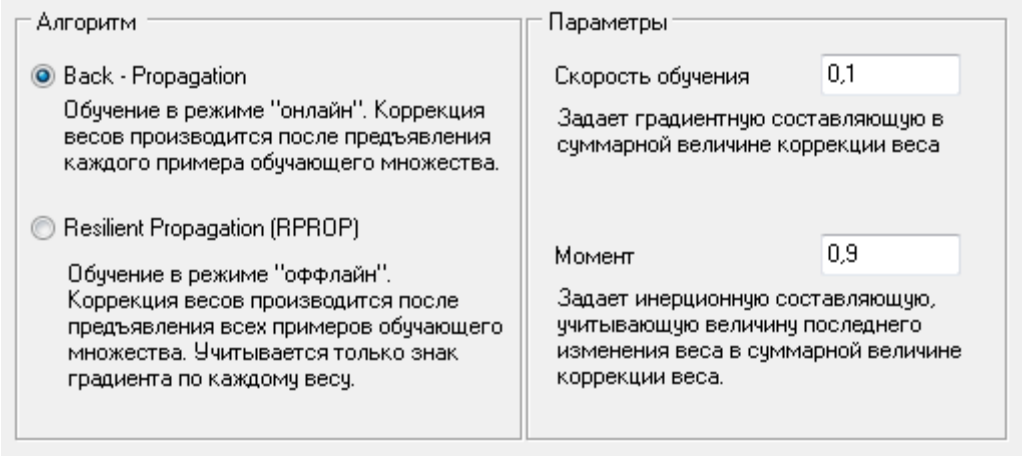


Рисунок 17.8 – Параметры алгоритмов обучения

Следующий шаг предлагает настроить условия остановки обучения. Укажем, что следует считать пример распознанным, если ошибка меньше 0.005, и также укажем условие остановки обучения при достижении эпохи 10000.

Следующий шаг мастера предлагает запустить процесс обучения и наблюдать в процессе обучения величину ошибки, а также процент распознанных примеров (рис.17.9). Параметр **Темп обновления** отвечает за то, через какое количество эпох обучения выводится данная информация.

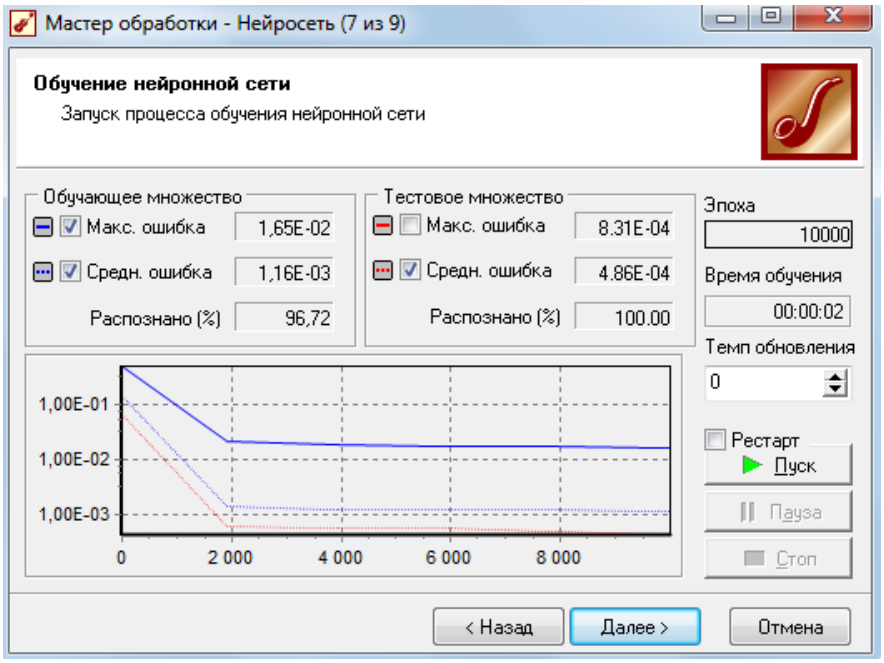


Рисунок 17.9 – Обучение нейронной сети

После обучения сети, в качестве визуализаторов выберем: **Граф нейросети**, **Диаграмма рассеяния**, **Что-если** (рис.17.10).

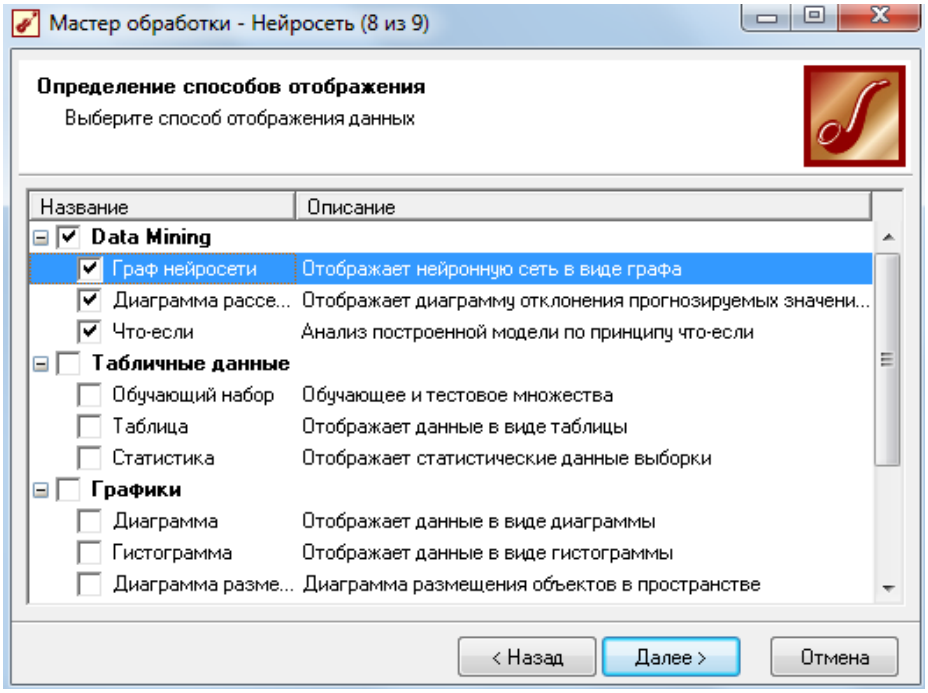


Рисунок 17.10 – Список доступных способов визуализации

Визуализатор **Граф нейросети** позволяет графически представить нейронную сеть со всеми ее нейронами и синаптическими связями (рис. 17.11). При этом можно увидеть не только структуру нейронной сети, но и значения весов, которые принимают те или иные нейроны. В зависимости от веса нейрона он отображается определенным цветом, а соответствующее значение можно определить по цветовой шкале, расположенной внизу окна.

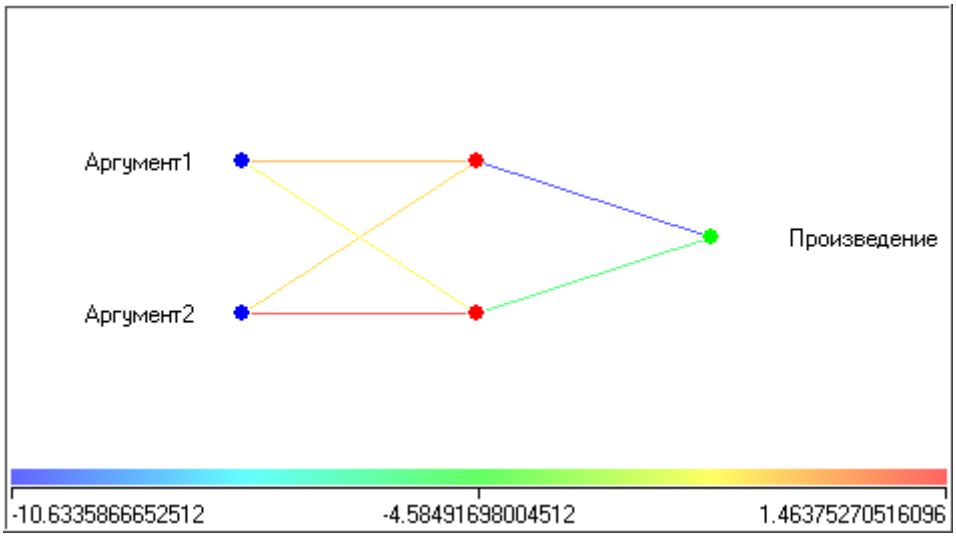


Рисунок 17.11 –Граф нейросети

Оценить качество модели позволяет диаграмма рассеяния, которая показывает рассеяние прогнозируемых данных относительно эталонных (рис.17.12).

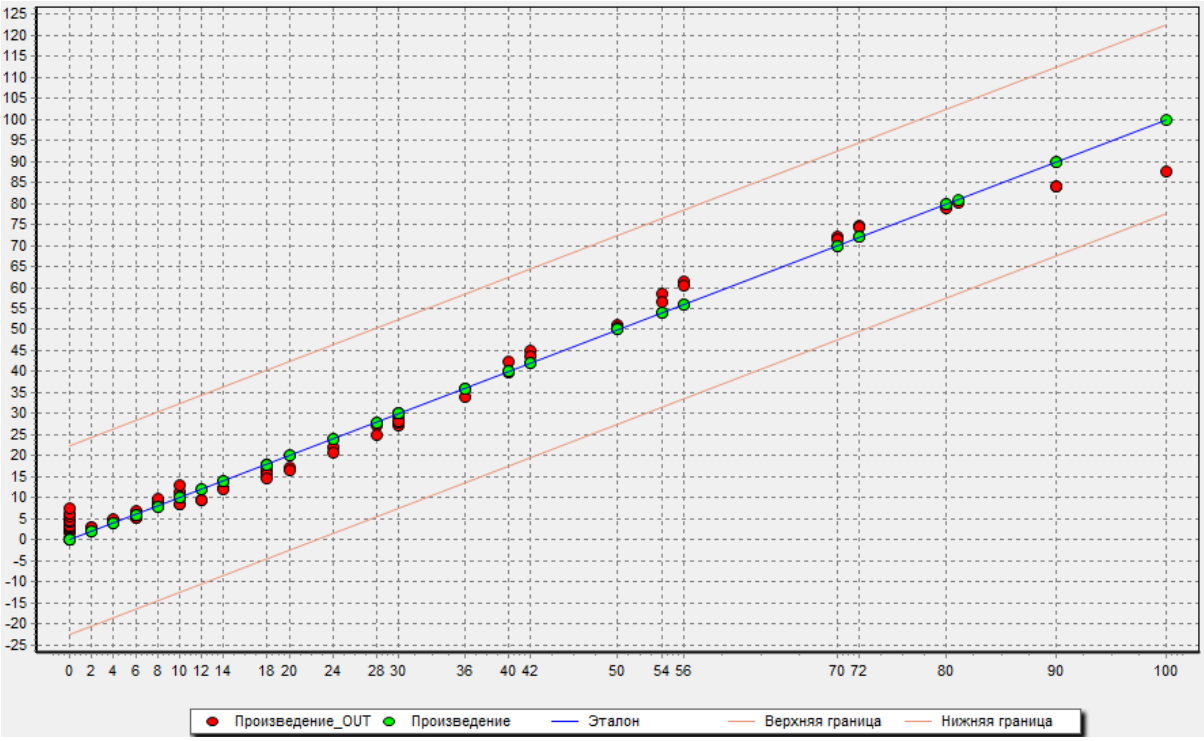


Рисунок 17.12 – Диаграмма рассеяния

На диаграмме отображаются выходные значения для каждого из примеров обучающей выборки, координаты которых по оси X – это значение выхода на обучающей выборке (эталон), а по оси Y – значение выхода, рассчитанное обученной моделью на том же примере. Прямая диагональная линия представляет собой ориентир (линию идеальных значений). Чем ближе точка к этой линии, тем меньше ошибка модели.

Визуализатор «Что-если» позволит провести эксперимент, введя любые значения множителей **Аргумент1** и **Аргумент2** и рассчитав результат их произведения. Так, в обучающей выборке не было примера, когда **Аргумент1=5** и **Аргумент2**=7. Рассчитаем этот результат многослойным персептроном в **Что-если** (рис. 17.13). Получим число, равное 36.06, что очень близко к истине.

Поле	Значение
Входные	
9.0 Аргумент1	7
9.0 Аргумент2	5
Выходные	
9.0 Произведение	36,0575935058045

Рисунок 17.13 –Что-если

Задание

Решите задачу регрессии нейронной сетью на примере аппроксимации многомерной нелинейной функции согласно варианту.

1. $f = \frac{x_1 + x_2}{x_3} + x_4 x_5$.
2. $f = x_1 - 20\sin(x_2) + 5x_3 + \frac{x_4}{e^{x_5}}$.
3. $f = \frac{x_1 x_2^2}{\sqrt{x_3}} + \sin(x_4 x_5)$.
4. $f = x_1 - x_2 - x_3 + x_4 x_5^2$.
5. $f = 0.5\cos(x_1 + x_2)^2 + \frac{1}{(x_3 + x_4)^2} + x_5$
6. $f = 5x_1 + \cos(x_2 + \sqrt{x_3}) + \sin(x_4 + \frac{x_5}{2})$.
7. $f = 3\cos(x_1 x_2) + 2\sin x_3 + \ln x_4 + 10x_5^2$.
8. $f = \sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2}$
9. $f = x_1 + 2\cos(x_2) + x_3^2 + \sqrt{x_4} + \sin x_5$.

Систему данных объемом *N*=200 получите с помощью равномерного случайного распределения в областях определения переменных $x_i \in [1;3], i = 1,..,5$. Это можно сделать, например, в табличном процессоре MS Excel. Полученную последовательность векторов разделите на два множество: обучающее и тестовое. Подберите архитектуру нейросети и обучите ее. Оцените качество нейросетевой модели.

Вопросы для самоконтроля

- Какие задачи можно решать при помощи многослойного персептрона?
- Как формируется обучающая выборка для решения задачи аппроксимации функции?
- Как нормируются обучающие данные?
- Какие имеются эмпирические правила подбора количества скрытых слоев, нейронов, объема обучающей выборки и коэффициента обучения?
- В чем недостатки алгоритма обратного распространения ошибки?

Практическое занятие №18

Логистическая регрессия и деревья решений в задаче кредитного скоринга

Среди преимуществ скоринга – снижение потерь по займам, усиление лояльности, экономия времени на сбор просроченной задолженности и культивирование навыков открытого количественного анализа в помощь руководителям.
Марк Шнайдер, CGAP

Цель работы – освоить принципы работы с искусственными нейронными сетями в Deductor на примере аппроксимации нелинейной многомерной функции.

Теоретические сведения

Кредитный скоринг

На протяжении последних нескольких лет российский рынок розничного кредитования населения переживает стадию стремительного развития. Потенциал этого рынка оценивается экспертами в несколько миллиардов долларов в год. Это стимулирует все новые и новые банки выходить на сегмент розничного кредитования. Банками запускаются программы потребительского, ипотечного, автокредитования, кредитных карт. Появились коммерческие банки, специализирующиеся исключительно на кредитовании населения: Русский стандарт, Росбанк и другие.

Несмотря на это эксперты признают, что сегодня методики оценки заемщика не поспевают за ростом рынка потребительского кредитования. И этому имеется несколько причин.

Во-первых, в РФ отсутствует единая база о предоставленных займах, в которой скапливается вся информация о добросовестных и недобросовестных заемщиках. Процесс формирования кредитных бюро в России только начался, и закончится через несколько лет. И на этом пути присутствуют трудности: слабое развитие банковского сектора, нежелание банков разглашать информацию. Во-вторых, во многих банках не накоплены реальные кредитные истории в области потребительских займов. Наконец, большинство западных скоринговых методик не подходят по причине социально-экономической направленности портретов заемщиков в России. Последнее означает, что в отсутствие единой базы кредитных историй первостепенными факторами при принятии решения о выдаче кредита становятся социально-экономические: образование, возраст, должность, уровень доходов и т.п.

В таких условиях комплексная оценка кредитоспособности заемщика в условиях ограниченности временных ресурсов становится трудновыполнимой задачей. Как следствие, банком назначается единая процентная ставка, в которую закладываются риски невозврата, из-за чего добросовестный заемщик несет повышенные издержки. Эти издержки весьма велики – годовые эффективные процентные ставки

по экспресс-кредитам сегодня нередко достигают отметки в 40-60%. А в условиях конкурентной борьбы на рынке кредитов выиграет тот, кто сумеет адекватно оценить риск выдачи займа за минимальное время и предложить низкую процентную ставку. При решении этой задачи не обойтись без скоринга – методики автоматической оценки кредитоспособности заемщика по набору его характеристик. Иными словами – скоринг позволяет *управлять* рисками в розничном кредитовании.

Кредитным скорингом, или просто *скорингом*, называется быстрая, точная и устойчивая процедура оценки кредитного риска, основанная на математической модели. Эта модель соотносит уровень кредитного риска с параметрами, характеризующими заемщика. Как бы ни была сложна модель, ее работа всегда к разделению потенциальных заемщиков на два класса – тех, кому кредит выдать можно, и тех, кому он «противопоказан».

История скоринга связана с именем Дюрана – американского финансиста, который впервые разработал балльную модель для оценки заемщика по совокупности его имущественных и социальных параметров (возраст, пол, профессия и т.д.). Преодолев границу некоторого порога, заемщик считался кредитоспособным. Поэтому под скорингом традиционно понимается балльная, или рейтинговая методика оценки заемщика. Статистическим алгоритмом автоматического расчета баллов скоринговой карты сегодня является *логистическая регрессия*. Кроме нее, сегодня в скоринге популярен еще один метод машинного обучения – *деревья решений*. Рассмотрим их более подробно.

Логистическая регрессия

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной. Бинарная логистическая регрессия, как следует из названия, применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения). Иными словами, с помощью логистической регрессии можно оценивать вероятность того, что событие наступит для конкретного испытуемого (больной/здоровый, возврат кредита/дефолт и т.д.).

Как известно, все регрессионные модели могут быть записаны в виде формулы:

$$y = F(x_1, x_2, ..., x_n).$$

Например, во множественной линейной регрессии предполагается, что зависимая переменная является линейной функцией независимых переменных, т.е.:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n.$$

Можно ли ее использовать для задачи оценки вероятности исхода события? Да, можно, вычислив стандартные коэффициенты регрессии. Например, если рассматривается исход по займу, задается переменная y со значениями 1 и 0, где 1 означает, что соответствующий заемщик расплатился по кредиту, а 0, что имел место дефолт. Однако здесь возникает проблема: множественная регрессия не «знает», что переменная отклика бинарна по своей природе. Это неизбежно приведет к модели с предсказываемыми значениями большими 1 и меньшими 0. Но такие зна-

чения вообще не допустимы для первоначальной задачи. Таким образом, множественная регрессия просто игнорирует ограничения на диапазон значений для *y*.

Для решения проблемы задача регрессии может быть сформулирована иначе: вместо предсказания бинарной переменной, мы предсказываем непрерывную переменную со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением следующего регрессионного уравнения (логит-преобразование):

$$P = \frac{1}{1+e^{-y}},$$

где P – вероятность того, что произойдет интересующее событие; e – основание натуральных алгоритмов 2,71…; y – стандартное уравнение регрессии.

Зависимость, связывающая вероятность события и величину y, показана на следующем графике (рис.18.1):

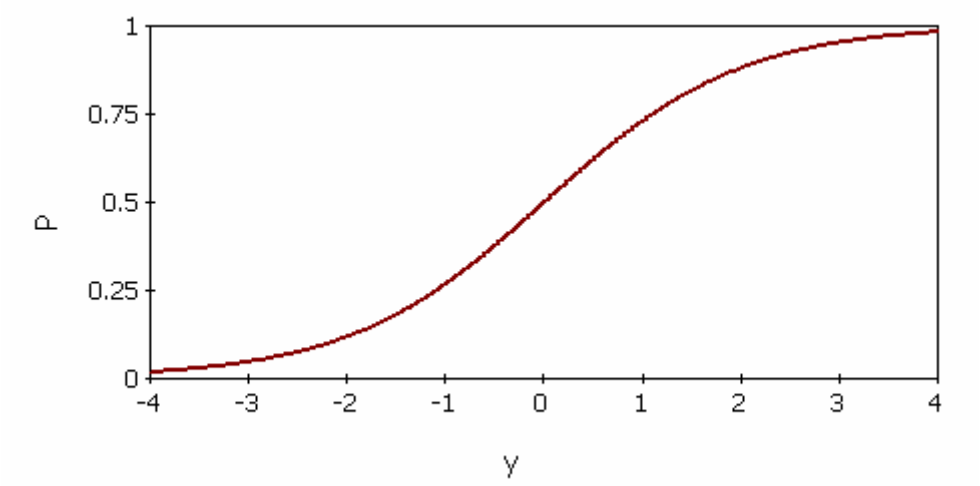


Рисунок 18.1 – Логистическая кривая

Поясним необходимость преобразования. Предположим, что мы рассуждаем о нашей зависимой переменной в терминах основной вероятности P, лежащей между 0 и 1. Тогда преобразуем эту вероятность P:

$$P' = \log_e\left(P/(1-P)\right).$$

Это преобразование обычно называют логистическим или логит-преобразованием. Теоретически P’ может принимать любое значение. Поскольку логистическое преобразование решает проблему об ограничении на 0-1 границы для первоначальной зависимой переменной (вероятности), то эти преобразованные значения можно использовать в обычном линейном регрессионном уравнении. А именно, если произвести логистическое преобразование обеих частей описанного выше уравнения, мы получим стандартную модель линейной регрессии.

Существует несколько способов нахождения коэффициентов логистической регрессии. На практике часто используют *метод максимального правдоподобия*. Он применяется в статистике для получения оценок параметров генеральной совокупности по данным выборки. Основу метода составляет *функция правдоподобия* (likehood function), выражающая плотность вероятности (вероятность) совместного появления результатов выборки Y_1, Y_2,…,Y_k:

$$L(Y_1, Y_2, ...Y_k; \theta) = p(Y_1; \theta) \cdot ... \cdot p(Y_k; \theta).$$

Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра θ принимается такое значение $\Theta = \Theta(Y_1,...,Y_k)$, которое максимизирует функцию L.

Нахождение оценки $\Theta = \Theta(Y_1,...,Y_k)$ упрощается, если максимизировать не саму функцию L, а натуральный логарифм $\ln(L)$, поскольку максимум обеих функций достигается при одном и том же значении θ:

$$L^*(\mathbf{Y},\theta) = \ln(L(\mathbf{Y},\theta)) \to \max .$$

В случае бинарной независимой переменной, которую мы имеем в логистической регрессии, выкладки можно продолжить следующим образом. Обозначим через P_i вероятность появления единицы: $P_i = \Pr ob(Y_i = 1)$. Эта вероятность будет зависеть от $\mathbf{X}_i\mathbf{W}$, где $\mathbf{X}_\text{i}$ – строка матрицы регрессоров, $\mathbf{W}$ – вектор коэффициентов регрессии:

$$P_i = F(\mathbf{X}_i\mathbf{W}),\ F(z) = \frac{1}{1+e^{-z}}.$$

Логарифмическая функция правдоподобия равна:

$$L^* = \sum_{i\in I1} \ln P_i(\mathbf{W}) + \sum_{i\in I0} \ln(1-P_i(\mathbf{W})) = \sum_{i=1}^{k} \left[Y_i \ln P_i(\mathbf{W}) + (1-Y_i)\ln(1-P_i(\mathbf{W}))\right],$$

где $I0$, $I1$ – множества наблюдений, для которых Y_i=0 и Y_i=1 соответственно.

Можно показать, что градиент $\mathbf{g}$ и гессиан функции H правдоподобия равны:

$$\mathbf{g} = \sum_i (Y_i - P_i)\mathbf{X}_i ,$$

$$H = -\sum_i P_i(1-P_i)\mathbf{X}_i^T\mathbf{X}_i \le 0 .$$

Гессиан всюду отрицательно определенный, поэтому логарифмическая функция правдоподобия всюду вогнута. Для поиска максимума можно использовать метод Ньютона, который здесь будет всегда сходиться:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - (H(\mathbf{W}_t))^{-1}\mathbf{g}_t(\mathbf{W}_t) = \mathbf{W}_t - \Delta\mathbf{W}_t .$$

На самом деле, логистическую регрессию можно представить в виде однослойной нейронной сети (рис.18.2).

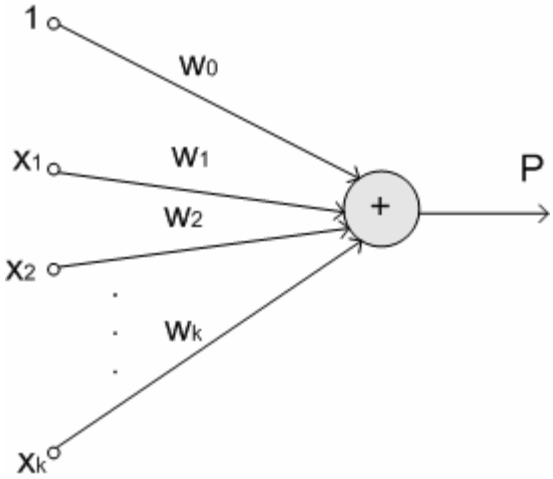


Рисунок 18.2 – Представление логистической регрессии в виде нейронной сети

Как известно, однослойная нейронная сеть может успешно решить лишь задачу линейной сепарации. Поэтому возможности по моделированию нелинейных зависимостей у логистической регрессии отсутствуют. Однако, как показывается далее, для оценки качества классификации логистической регрессии существует эффективный инструмент ROC-анализа, что является несомненным преимуществом.

***ROC*-анализ**

ROC-кривая (*Receiver Operator Characteristic*) – кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении. Название пришло из систем обработки сигналов. Поскольку классов два, один из них называется классом с положительными исходами, второй – с отрицательными исходами. ROC-кривая показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. В терминологии ROC-анализа первые называются истинно положительным, вторые – ложно отрицательным множеством. При этом предполагается, что у классификатора имеется некоторый параметр, варьируя который, мы будем получать то или иное разбиение на два класса. Этот параметр часто называют порогом, или точкой отсечения (cut-off value). В логистической регрессии порог отсечения изменяется от 0 до 1. В зависимости от него будут получаться различные величины *ошибок I и II рода*.

Табл. 18.1. Таблица классификации

	Фактически	
Модель	положительно	отрицательно
положительно	*TP*	*FP*
отрицательно	*FN*	*TN*

Для понимания сути ошибок I и II рода рассмотрим четырехпольную таблицу классификации (табл. 18.1), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам.

TP (True Positives) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);

TN (True Negatives) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);

FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи. Например, если мы прогнозируем вероятность того, что заем-

щик добросовестно будет погашать долг, то положительным событием будет класс «Хороший заемщик», а отрицательным – «Плохой заемщик».

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах:

Доля истинно положительных примеров (*True Positives Rate*):

$$TPR = \frac{TP}{TP + FN} \cdot 100\% .$$

Доля ложно положительных примеров (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \cdot 100\% .$$

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (*Sensitivity*) – это и есть доля истинно положительных случаев:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\% .$$

Специфичность (*Specificity*) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$Sp = \frac{TN}{TN + FP} \cdot 100 .$$

Заметим, что FPR=100–Sp.

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

ROC-кривая строится следующим образом:

1. Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом dx (например, 0.01) рассчитываются значения чувствительности Se и специфичности Sp. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.

2. Строится график зависимости: по оси Y откладывается чувствительность Se, по оси X – 100%–Sp (сто процентов минус специфичность), или, что то же самое, FPR – доля ложно положительных случаев.

В результате вырисовывается некоторая кривая (рис.18.3).

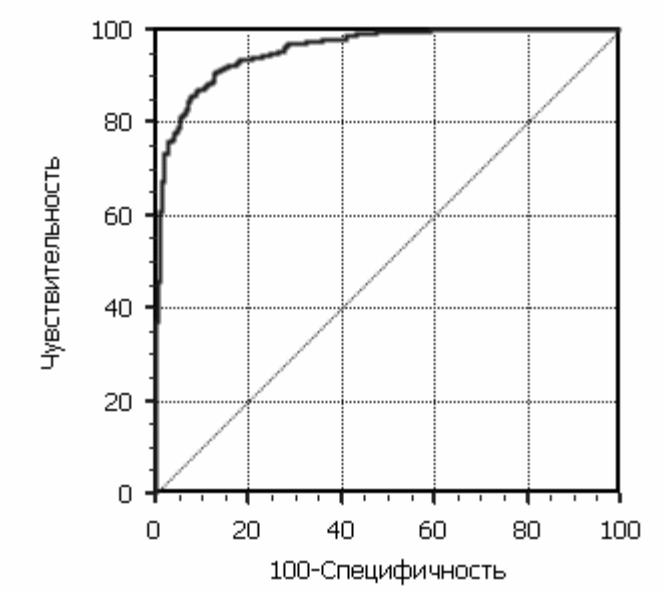


Рисунок 18.3 – Кривая *ROC*

График часто дополняют прямой *y=x*.

Для идеального классификатора график *ROC*-кривой проходит через верхний левый угол, где доля истинно положительных случаев составляет 100% или 1.0 (идеальная чувствительность), а доля ложно положительных примеров равна нулю. Поэтому чем ближе кривая к верхнему левому углу, тем выше предсказательная способность модели. Наоборот, чем меньше изгиб кривой и чем ближе она расположена к диагональной прямой, тем менее эффективна модель. Диагональная линия соответствует «бесполезному» классификатору, т.е. полной неразличимости двух классов.

При визуальной оценке *ROC*-кривых расположение их относительно друг друга указывает на их сравнительную эффективность. Кривая, расположенная выше и левее, свидетельствует о большей предсказательной способности модели. Так, на рисунке 12.4 две *ROC*-кривые совмещены на одном графике. Видно, что модель «А» лучше.

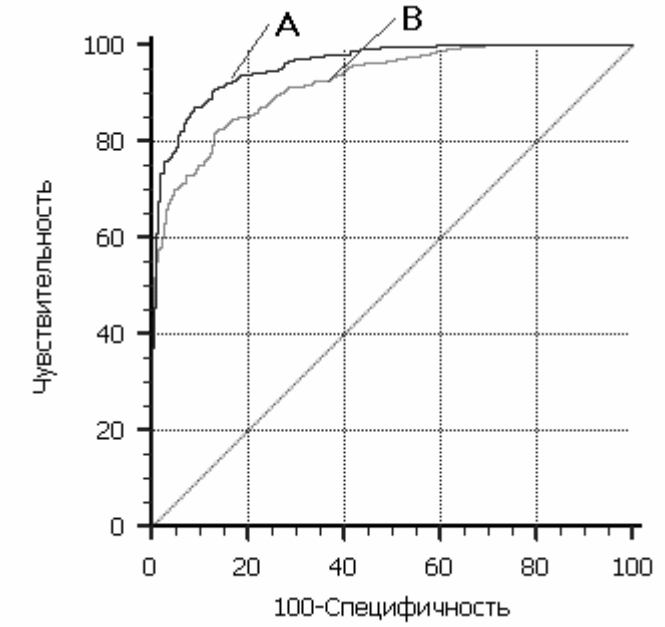


Рисунок 18.4 – Сравнение ROC-кривых

Визуальное сравнение кривых *ROC* не всегда позволяет выявить наиболее эффективную модель. Своеобразным методом сравнения *ROC*-кривых является оценка площади под кривыми. Теоретически она изменяется от 0 до 1.0, но, поскольку модель всегда характеризуются кривой, расположенной выше положительной диагонали, то обычно говорят об изменениях от 0.5 («бесполезный» классификатор) до 1.0 («идеальная» модель). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху – экспериментально полученными точками (рис. 18.5). Численный показатель площади под кривой называется *AUC* (Area Under Curve). Вычислить его можно, например, с помощью численного метода трапеций.

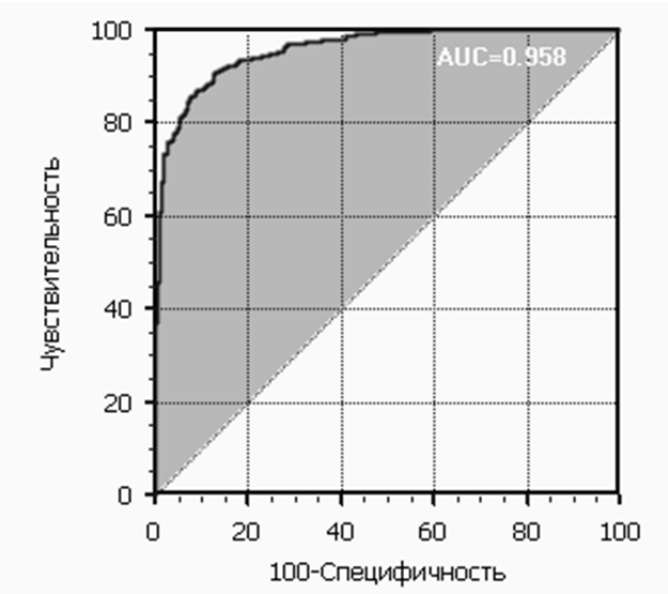


Рисунок 18.5 – Площадь под ROC-кривой

С большими допущениями можно считать, что чем больше показатель *AUC*, тем лучшей прогностической силой обладает модель. Однако следует знать, что:

- показатель *AUC* предназначен скорее для сравнительного анализа нескольких моделей;
- *AUC* не содержит никакой информации о чувствительности и специфичности модели.

Идеальная модель обладает 100% чувствительностью и специфичностью. Однако на практике добиться этого невозможно, более того, невозможно одновременно повысить и чувствительность, и специфичность модели. Компромисс находится с помощью порога отсечения, т.к. пороговое значение влияет на соотношение *Se* и *Sp*. Можно говорить о задаче нахождения *оптимального порога отсечения* (*optimal cut-off value*).

Порог отсечения нужен для того, чтобы применять модель на практике: относить новые примеры к одному из двух классов. Для определения оптимального порога нужно задать критерий его определения, т.к. в разных задачах присутствует своя оптимальная стратегия. Критериями выбора порога отсечения могут выступать:

- Требование минимальной величины чувствительности (специфичности) модели. Например, нужно обеспечить чувствительность теста не менее 80%. В этом случае оптимальным порогом будет максимальная специфичность (чувствительность), которая достигается при 80% (или значение, близкое к нему «справа» из-за дискретности ряда) чувствительности (специфичности).
- Требование максимальной суммарной чувствительности и специфичности модели, т.е. $C = \max_k (Se_k + Sp_k)$.
- Требование баланса между чувствительностью и специфичностью, т.е. когда $\mathrm{Se} \approx \mathrm{Sp}$: $C = \min_k |Se_k - Sp_k|$.

В последнем случае порог есть точка пересечения двух кривых, когда по оси X откладывается порог отсечения, а по оси Y – чувствительность или специфичность модели (рис. 18.6).

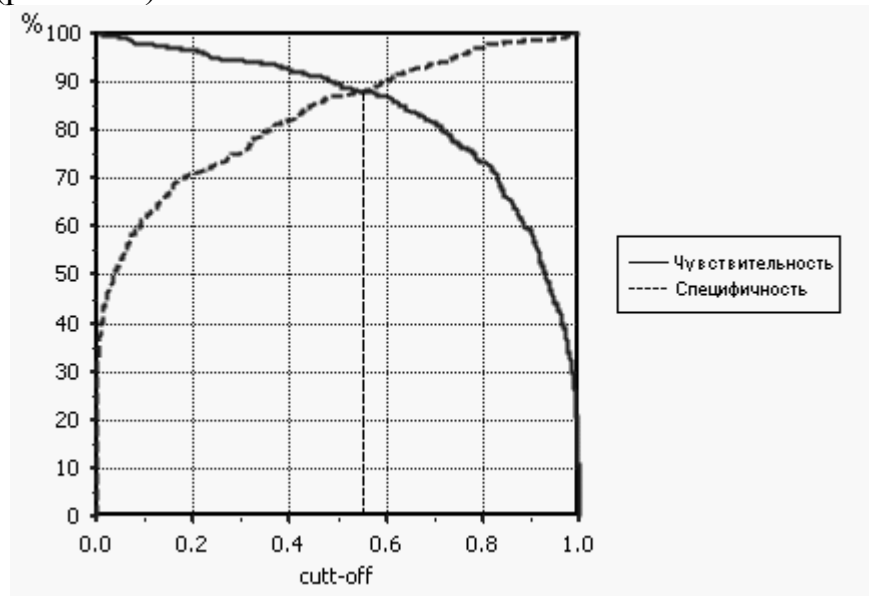


Рисунок 18.6 – «Точка баланса» между чувствительностью и специфичностью

Существуют и другие подходы, когда ошибкам I и II рода назначается вес, который интерпретируется как цена ошибок. Но здесь встает проблема определения этих весов, что само по себе является сложной, а часто не разрешимой задачей.

Деревья решений

Деревья решений (decision trees), наверное, можно назвать самой популярной технологией data mining. Мировой опыт показывает, что деревья решений демонстрируют отличные результаты при решении задачи оценки кредитоспособности кандидата на получение кредита.

В основе деревьев решений лежит идея рекурсивного разбиения множества объектов на подмножества таким образом, чтобы значения зависимой переменной в каждом подмножестве были как можно более однородными. Для этого каждый раз разбиение осуществляется по какой-либо одной независимой переменной, которая делает его наилучшим. По окончании такого рекурсивного процесса получается дерево решений – набор правил в иерархической структуре.

Есть несколько преимуществ использования деревьев решений: быстрота построения и легкость интерпретации, и некоторые другие. Каждый путь от вершины до листа дерева (конечного узла) образует правило. В режиме предсказания новый объект «прогоняется» сквозь дерево правил и «оседает» в каком-либо листе; это образно можно сравнить с падением шарика в пинболе.

Существуют несколько методов построения деревьев решений. Например, можно использовать различные формулы для определения варианта разбиения. Форма дерева может быть бинарной (каждый узел дерева имеет двух потомков), или небинарной. Что касается глубины дерева, здесь также имеются различные способы ее контроля: можно построить дерево с полной глубиной, а затем отсечь часть узлов, либо априори ограничить максимальную глубину дерева.

Наиболее популярным алгоритмом построения дерева решений считается ID3 и его усовершенствованный алгоритм C4.5. C4.5 может работать с числовыми, категориальными, пропущенными и зашумленными данными. Рисунок 18.7 демонстрирует пример фрагмента дерева решений для классификации заемщиков.

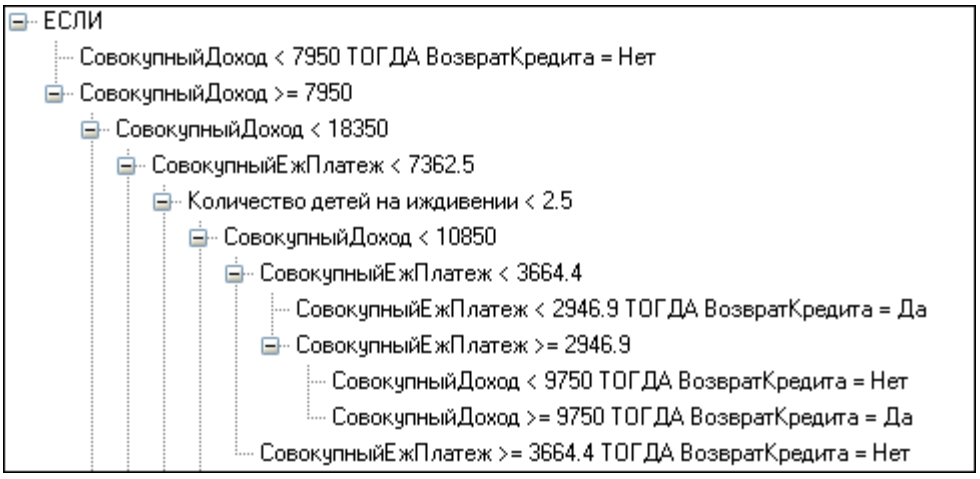


Рисунок 18.7 – Пример фрагмента дерева решений

Раскроем основные концепции деревьев решений на конкретном примере скоринга с использованием понятия *энтропии*. Пусть имеется 3000 заемщиков с известными кредитными исходами. Для простоты изложения ограничимся тремя переменными из табл. 18.2: пол (мужской, женский), состоит в браке (да, нет), количество лет проживания в регионе (до 1 года, 1-3 года, свыше 3 лет). Зависимой переменной является «Возврат кредита»: да – в случае успеха и нет – в случае дефолта.

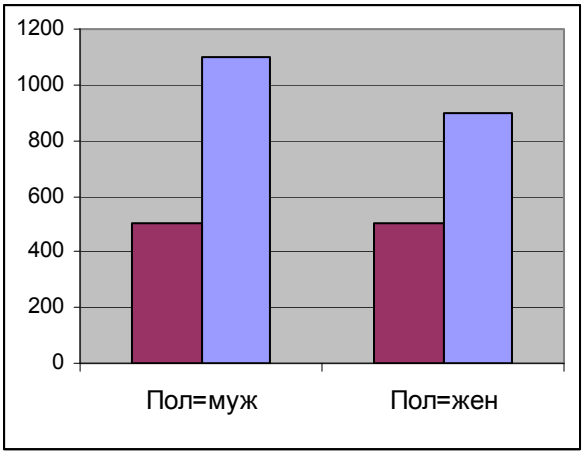
Первым шагом в построении дерева решений является формирование таблицы частот, как это показано на рисунке 18.8. Каждая колонка в таблице есть пара «атрибут-значение» входных атрибутов (переменных, факторов кредитоспособности), каждая строка – возможное состояние зависимой переменной. В ячейках таблицы находится количество комбинаций из входной и зависимой переменной. Например, из таблицы следует, что из 2000 человек, вернувших ссуду, 1700 состояли в браке, 300 – не состояли в нем, и так далее. После таблицы для наглядности построены графики на основе данных из таблицы частот.

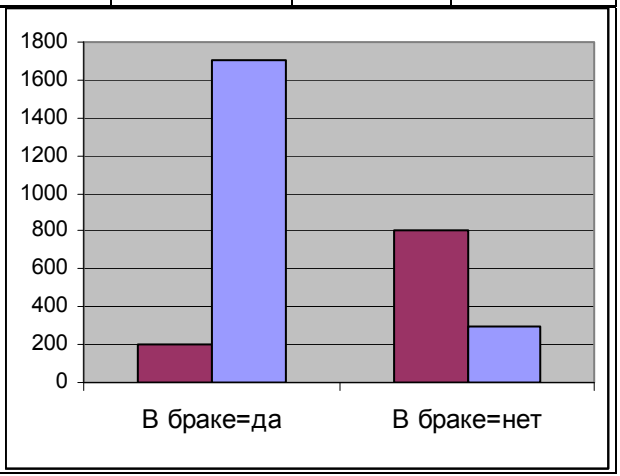
Алгоритм построения дерева решений начнет свою работу с самого верха. Необходимо выбрать такой атрибут, чтобы после разбиения полученные подмножества состояли из заемщиков, принадлежащих к одному классу, или были максимально приближены к этому. Из гистограмм на рисунке 12.8 интуитивно понятно, что наилучшим атрибутом является атрибут «Состоит в браке». Когда «Состоит в браке» равно «да», то синий столбик имеет значительную высоту по сравнению с красным, наоборот, когда «Состоит в браке=нет» – выше красный столбик.

Естественно, что алгоритм дерева решений не способен напрямую анализировать графики частот атрибутов; необходим более формализованный критерий. Таким критерием, в частности, является энтропия, или количество информации.

Примечание. Понятие энтропии широко используется в теории информации. Энтропия представляет собой некоторую меру «неопределенности», связанную с появлением некоторого события. Чем выше энтропия, тем больше неопределенность появления данного события. Чем больше число различных состояний атрибута и чем меньше отличаются друг от друга их вероятности, тем больше энтропия.

		Пол		Состоит в браке		Количество лет проживания		
		муж	жен	да	нет	до 1 года	1-3 года	свыше 3 лет
Возврат кредита	Да	1100	900	1700	300	400	700	900
	Нет	500	500	200	800	600	200	200





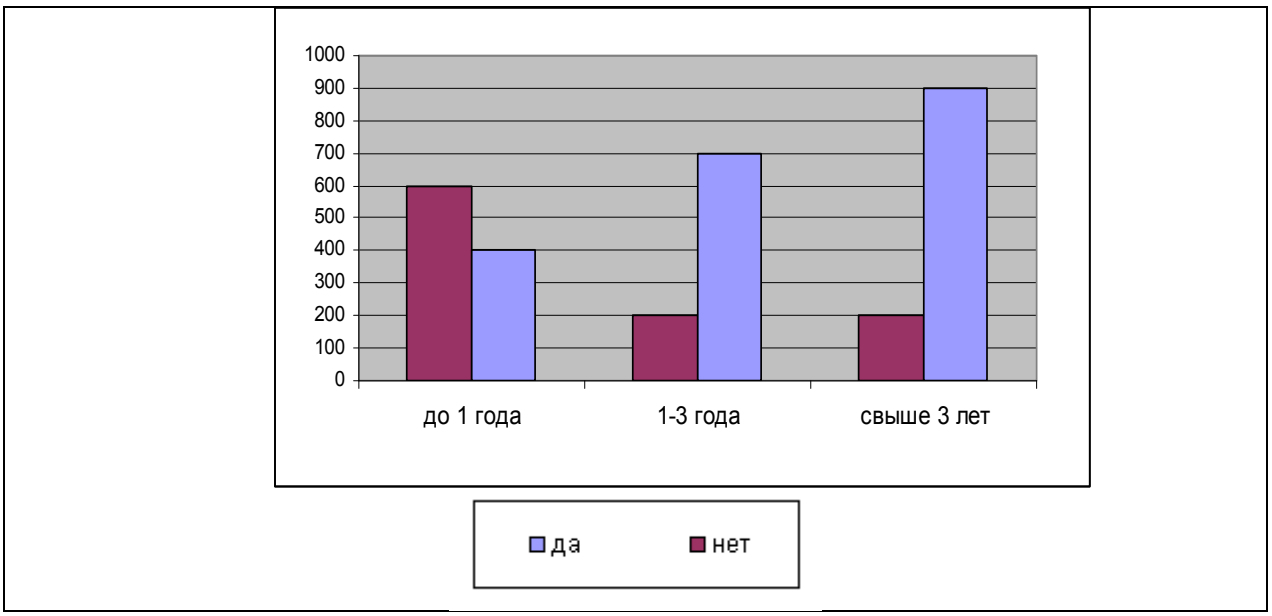


Рисунок 18.8 – Выбор наилучшего атрибута для разбиения

Итак, нам нужно найти математическую формулу для оценки «чистоты» набора данных после применения разбиения по какому-либо атрибуту. Эта формула должна удовлетворять следующим условиям:

- если в кредитной истории все случаи возврата кредита принадлежат к одному классу (т.е. имеют одинаковый исход, либо «да, либо «нет»), то энтропия равна нулю;
- если в кредитной истории присутствует одинаковое количество случаев для каждого класса, то энтропия имеет максимальное значение.

Эта формула имеет следующий вид (обозначим энтропию через *H*):

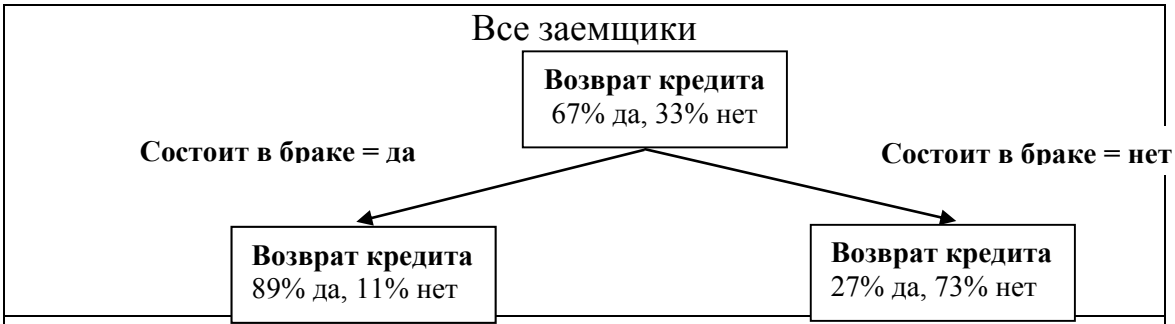
$$H(p_1, p_2, ..., p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 ... - p_n \log_2 p_n,$$

где $p_1, p_2, ..., p_n$ – вероятности появления каждого состояния в зависимой переменной, причем $p_1 + p_2 + ... + p_n = 1$.

В нашем примере только два состояния зависимой переменной, да или нет, *n*=2. Используя приведенную выше формулу, рассчитаем энтропии для разбиений по каждому из трех атрибутов:

- Разбиение по атрибуту «Пол»:
  H(1100,500) + *H*(500,900) = 0.896 + 0.940 = 1.836.
- Разбиение по атрибуту «Состоит в браке»:
  H(1700,200) + *H*(300,800) = 0.485 + 0.845 = 1.330.
- Разбиение по атрибуту «Количество лет проживания»:
  H(400,600) + *H* (700,200) + *H* (900,200) =0.97 + 0.764 + 0.684 = 2.418.

Проведя анализ этих вычислений, мы придем к выводу, что разбиение по атрибуту «Состоит в браке» обладает наименьшей энтропией, т.е. ведет к наибольшему снятию неопределенности, следовательно, является наилучшим в данном случае. Мы получим атрибут «Состоит в браке» в корне дерева. Далее алгоритм продолжит процесс разбиения для каждого узла и будет это делать до тех пор, пока не выполнится условие останова. Рисунок 18.9 демонстрирует новую таблицу частот для множества «Состоит в браке={«Да»}».



		Пол		Состоит в браке		Количество лет проживания		
		муж	жен	да	нет	до 1 года	1-3 года	свыше 3 лет
Возврат кредита	да	1000	700	1700	0	250	650	800
	нет	150	50	200	0	100	50	50

Рисунок 18.9 – Продолжение процесса построения дерева решений

Под каждым узлом на рисунке 18.9 находятся процентные значения – частоты появления зависимой переменной в каждом подмножестве, полученным соответствующим разбиением. Если уже после первого шага прекратить строить дерево, то конечный узел превратится в лист. В качестве следствия правила будет выбрано значение зависимой переменной с максимальной частотой, если значения частот одинаковые – то любая переменная.

Эта максимальная частота есть так называемая достоверность каждого узла. *Достоверность* – это количество правильно классифицированных данным узлом примеров. В нашем примере она равна 89% для узла «Состоит в браке=да» и 73% для узла «Состоит в браке=нет». Их можно интерпретировать как вероятность возврата, или как рейтинговый балл заемщика.

Если остановить построение дерева на рисунке 18.9, то получим простое дерево решений, содержащее 2 правила (табл. 18.3).

Таблица 18.3 – Список правил дерева решений

N	Условие	Решение (Возврат кредита)	Поддержка		Достоверность	
				Кол-во	%	Кол-во
1	Состоит в браке = да	Да	63	1900	89	1700
2	Состоит в браке = нет	Нет	37	1100	73	800

В таблице появился еще один параметр – поддержка. *Поддержка* – это общее количество примеров, классифицированных данным узлом дерева.

В простейшем случае разбиение дерева решений можно вести до тех пор, пока в узле имеется более одного примера, однако, такое делать нецелесообразно. Прямой связи между числом узлов, ветвистостью дерева и качеством предсказания нет. Построенное «до конца» дерево решений может «запомнить» все примеры из

обучающей выборки, и очевидна бессмысленность такой скоринговой модели: ее предсказательная способность на новых данных будет низкой. Это явление получило название *переобучения* дерева.

Для предотвращения переобучения существуют различные приемы. Рассмотрим два из них. В первом способе рекурсивный процесс разбиения очередного узла продолжается до тех пор, пока в нем содержится не менее $k>1$ примеров, либо не меньше заданной априори величины поддержки (что, по сути, одно и то же). Конкретные величины k или поддержки зависят от специфики задачи и объема данных, и даже для задачи скоринга не существует какой-то одной рекомендуемой величины. При втором способе применяется отсечение веток дерева, что может регулироваться специальным параметром, от которого будет зависеть глубина финального дерева. Заметим, что первый и второй способы могут применяться одновременно.

При построении дерева решений имеется еще одна проблема – количество возможных значений входных атрибутов может быть большим. Например, почтовый индекс заемщика может принимать тысячи значений. Для дерева решений это не будет проблемой: оно обработает и проигнорирует этот атрибут. Но иногда в этом поле может содержаться полезная информация, например, аппликанты, проживающие в одинаковом районе, могут иметь более высокий кредитный риск по сравнению с другими. Для решения этой задачи используют прием группировки, в результате которой уменьшается количество значений атрибута. Хорошо себя зарекомендовала следующая группировка: на основе статистики по набору данных отобрать m<<n (m значительно меньше n) наиболее часто встречающихся состояний атрибутов, а все остальные объединить в один. Естественно, это базируется на гипотезе о том, что часто встречаемые состояния атрибутов вносят б*о*льший вклад в формирование зависимой переменной.

Замечание

Для дискретного поля при построении дерева решений рекомендуется использовать не более 10 значений. Если их больше, то лучше объединить часть значений в одно, либо совсем исключите это поле из участия в работе алгоритма.

В качестве входных полей дерева решений используются как дискретные, так и непрерывные поля, такие как возраст, доход, ежемесячный платеж по ссуде и так далее. Разбиение по непрерывному полю, как и по дискретному, осуществляется с применением принципа энтропии, но имеет свою специфику, т.к. вводятся операции сравнения (больше, меньше). Проиллюстрируем работу алгоритма с непрерывным полем на примере.

В табл. 18.4 находятся данные пары значений «Доход заемщика–Возврат кредита». Доход заемщика в данном случае – непрерывное поле. По нему и будем проводить разбиение.

Таблица 18.4 – Пример с непрерывным атрибутом

		Доход, USD				
Возврат кредита	да	500	750	600	200	200
	нет	120	50	80	40	50

Первое, что сделаем – упорядочим прецеденты по возрастанию дохода, убрав повторяющиеся (табл. 18.5).

Таблица 18.5 – Упорядоченные значения доходов заемщиков

Доход, USD							
40	50	80	120	200	500	600	750

Разбиение проводим по правилу половинного деления, кандидатами на разбиение будут два соседних значения атрибута. Рассчитав энтропию для каждого такого варианта по аналогии с дискретными атрибутами, выберем оптимальное разбиение с минимальной энтропией (см. табл. 18.6).

Таблица 18.6 – Расчет энтропии для разбиений

			Возврат кредита		
			да	нет	Энтропия
Разбиение (Доход)	45	<45	0	1	0.99
		>=45	5	4	
	65	<65	0	3	0.86
		>=65	5	2	
	100	<100	0	4	0.65
		>=100	5	1	
	160	<160	5	0	0
		>=160	0	5	
	350	<350	2	5	0.86
		>=350	3	0	
	550	<550	3	5	0.95
		>=550	2	0	
	675	<675	4	5	0.99
		>=675	1	0	

Первое разбиение приведет нас к двум правилам (табл. 18.7).

Таблица 18.7 – Список правил дерева решений

N	Условие	Решение (Возврат кредита)	Поддержка		Достоверность	
			%	Кол-во	%	Кол-во
	Доход < 160	Нет	50	5	100	5
	Доход >= 160	Да	50	5	100	5

Иногда атрибут можно подать на вход алгоритма дерева решений как в дискретном, так и в непрерывном виде. Например, атрибут «Количество иждивенцев»={0,1,2,3}. При использовании алгоритма C4.5 лучше, если это возможно, сделать выбор в пользу непрерывного поля.

Конкретные реализации алгоритмов деревьев решений более сложны и содержат массу тонкостей, вместо энтропии может быть взят другой критерий, однако, основная идея о рекурсивном разбиении остается неизменной.

Пример построения скоринговых моделей в *Deductor*

Построим модели логистической регрессии и дерева решений в *Deductor*. Для этого воспользуемся файлом **loans_demo.txt**, который идет в комплекте к практикуму. Этот файл содержит так называемую кредитную историю, т.е. информацию о заемщиках и о качестве обслуживания ими долга.

Вообще говоря, информация о заемщиках-физлицах и кредитных договорах хранится в АБС – автоматизированной банковской системе. Там же хранятся графики погашений, даты погашений, просрочки, суммы просрочек, проценты и так далее. Получить для скоринга таблицу, содержащую параметры заемщиков и информацию о характере погашений и наличии просрочек представляет собой отдельную задачу, и будем считать, что она уже выполнена и результат представлен текстовым файлом.

Следующая задача, которую нужно решить – это выработать правила, по которым мы будем относить заемщика к одному из двух классов («хороший» или «плохой»), используя информацию о просрочках. Просрочка измеряется, как правило, в днях. В Российской Федерации кредитные организации при определении категории заемщика руководствуются Положением N 254-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам, по ссудной и приравненной к ней задолженности». В частности, в нем указано, что для физических лиц «…обслуживание долга признается плохим, если.. имеются просроченные платежи по основному долгу и (или) по процентам в течение последних 180 календарных дней… свыше 60 календарных дней». Однако кредитной организации никто не мешает выработать свои собственные правила для классификации заемщиком с учетом кредитной политики банка и других факторов. Например, в экспресс-кредитовании на малые суммы просрочки до 5 дней могут не учитываться, вместо 60 дней может быть взято 90 и так далее.

В нашем файле последний столбец, характеризующий качество обслуживания долга заемщика, представлен полем «Число просрочек свыше 60 дн.». Остальные поля, кроме информационного Код, представляют социально-экономические характеристики заемщиков: возраст, пол, доход и т.п. Создадим новый проект в Deductor и импортируем в него этот файл (рис. 18.10).

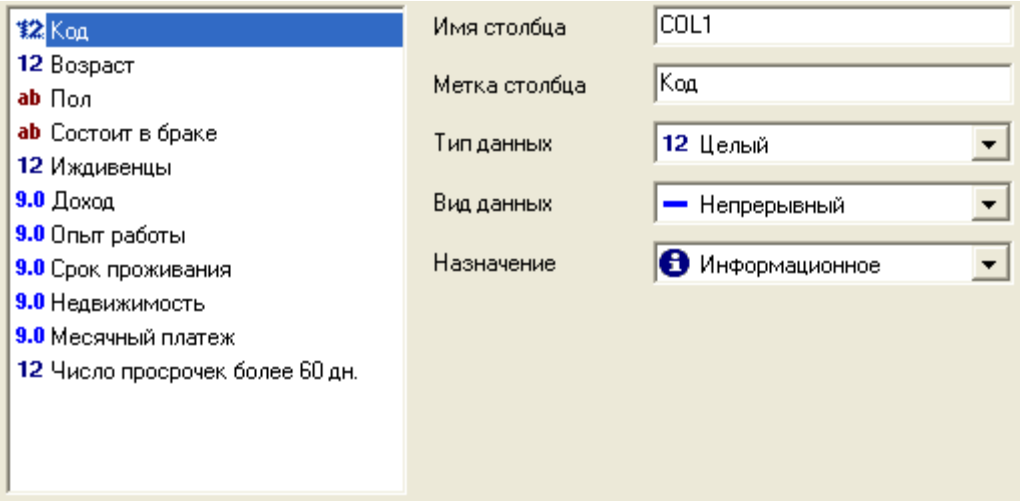


Рисунок 18.10 – Импорт файла

Получим из поля **Число просрочек более 60 дн.** новое поле **Класс заемщика**. Для этого с помощью обработчика Калькулятор создадим строковое поле и в строке функции напишем (рис. 18.11):

IF(COL11>0;"Плохой";"Хороший")

Теперь все готово для построения скоринговой модели. Добавим в ветку сценария обработчик **Логистическая регрессия**. На первом шаге зададим входные и выходные значения столбцов, как это показано на рисунке 18.12. Поле **Код** будет информационным, **Число просрочек более 60 дн.** – неиспользуемым, **Класс заемщика** – выходным. Остальные поля будут входными.

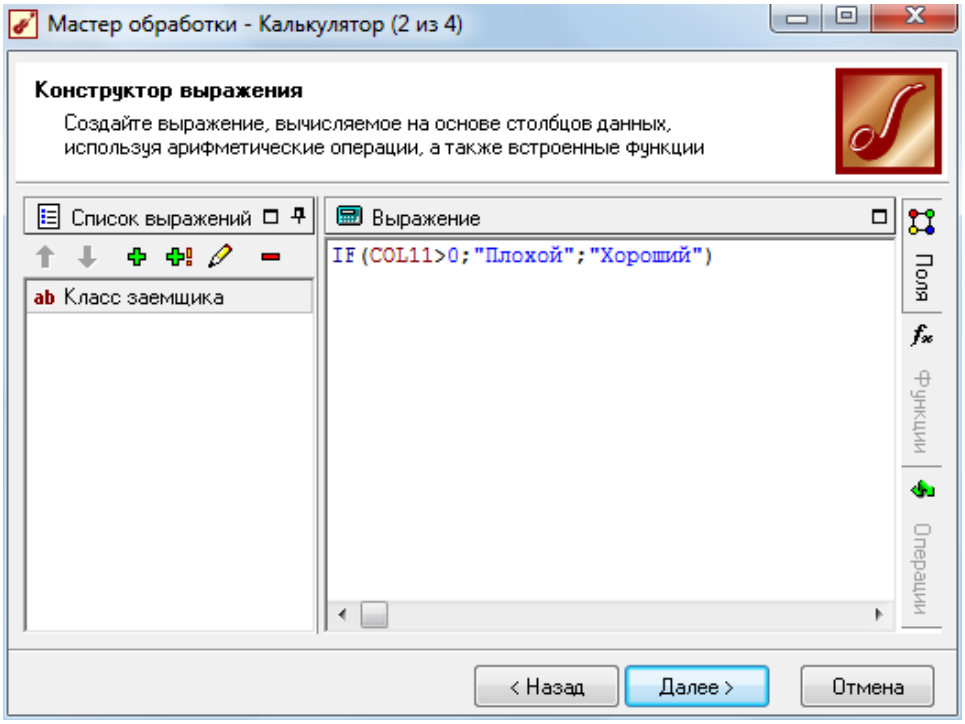


Рисунок 18.11 – Обработчик Калькулятор

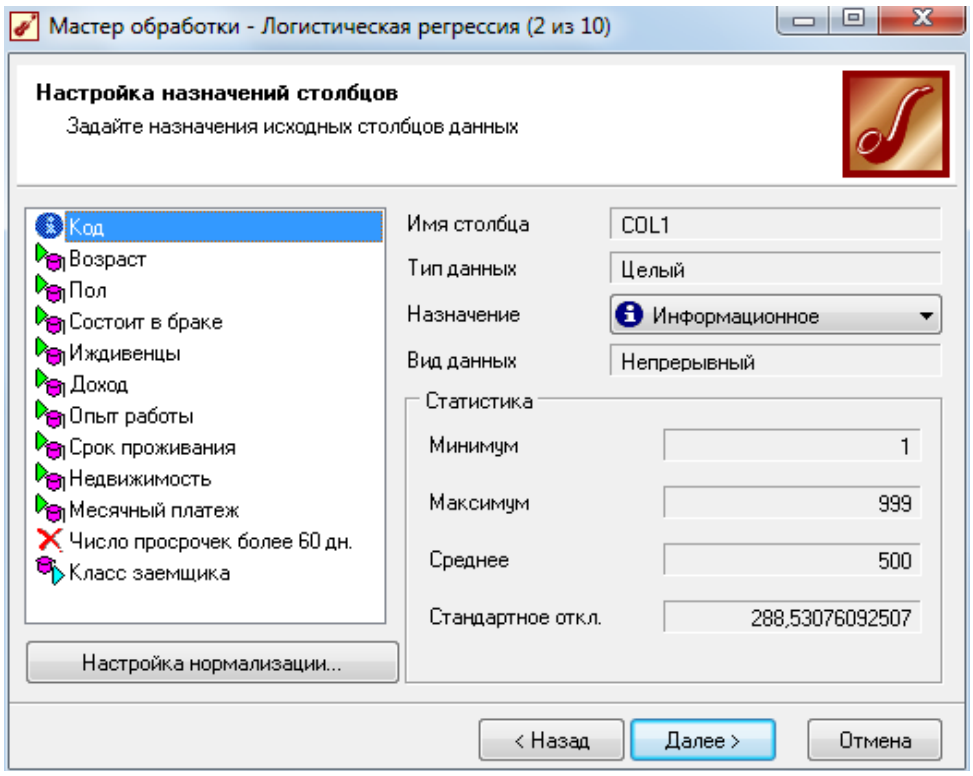


Рисунок 18.12 – Настройка параметров столбцов

Нажав на кнопку **Настройка нормализации**, появится следующее диалоговое окно (рис.18.13). Для логистической регрессии в нем настраиваются:

- Способы кодирования дискретных входных полей (битовая маска или уникальные значения);
- Значения положительного и отрицательного события для выходного поля.

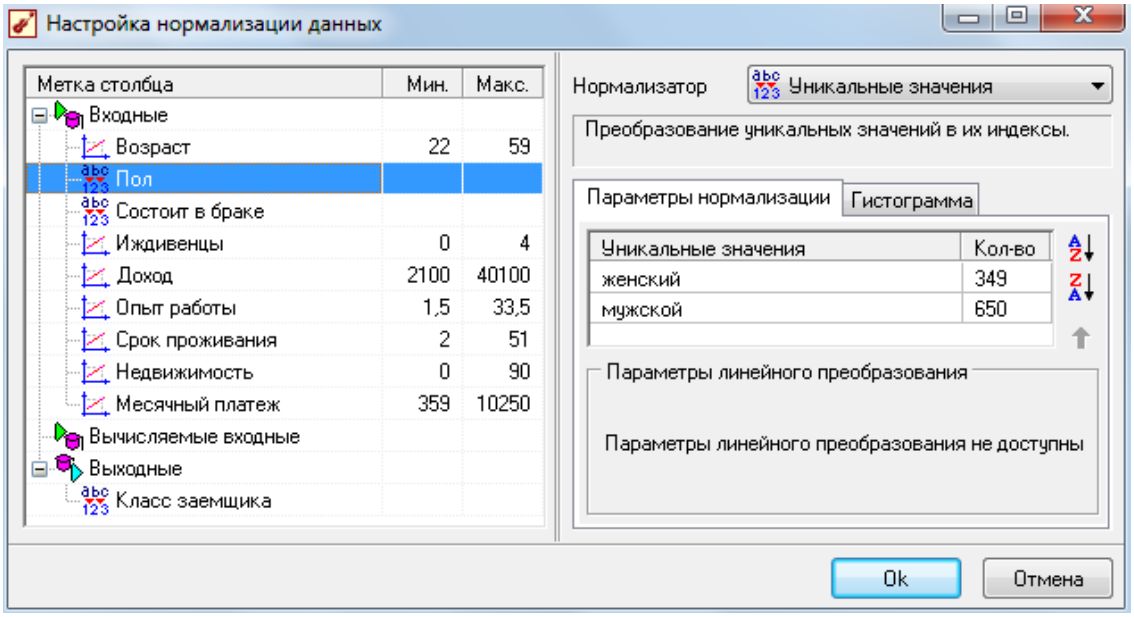


Рисунок 18.13 – Настройка нормализации

В нашем случае имеется два входных дискретных поля – **Пол** и **Состоит в браке**. Для них рекомендуется поставить способ кодирования **Уникальные значения**. Порядок списка уникальных значений будет влиять на то, как будут коди-

роваться значения полей. Для поля **Пол** первое уникальное значение будет закодировано в 0 («женский»), второе – 1 («мужской»). Это означает, что при расчете кредитного рейтинга по уравнению логистической регрессии женщинам всегда будет начисляться 0 баллов, а мужчинам – какой-либо отличный от нуля балл.

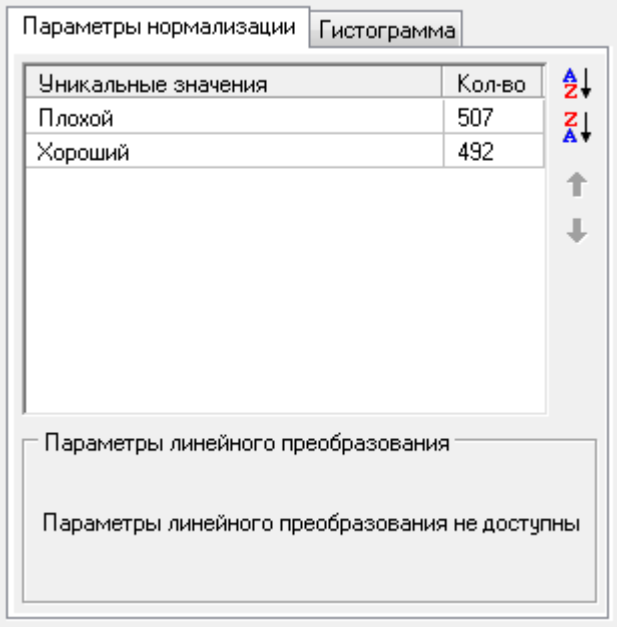


Рисунок 18.14 – Задание типов событий.

Аналогично для поля **Состоит в браке** зададим кодирование по уникальным значениям в следующем порядке: «Нет» (значение 0), «Да» (значение 1).

Для выходного поля **Класс заемщика** порядок сортировки уникальных значений (которых всегда два) определяет тип события: первое – отрицательное, второе – положительное (рисунок 18.14). В нашем случае - чем выше рейтинг, тем выше кредитоспособность, поэтому значение «Хороший» будет положительным исходом события, а «Плохой» – отрицательным.

После установки параметров нормализации продолжим идти по шагам **Мастера**. В следующем окне будет предложено настроить обучающие и тестовые множества (рис. 18.15).

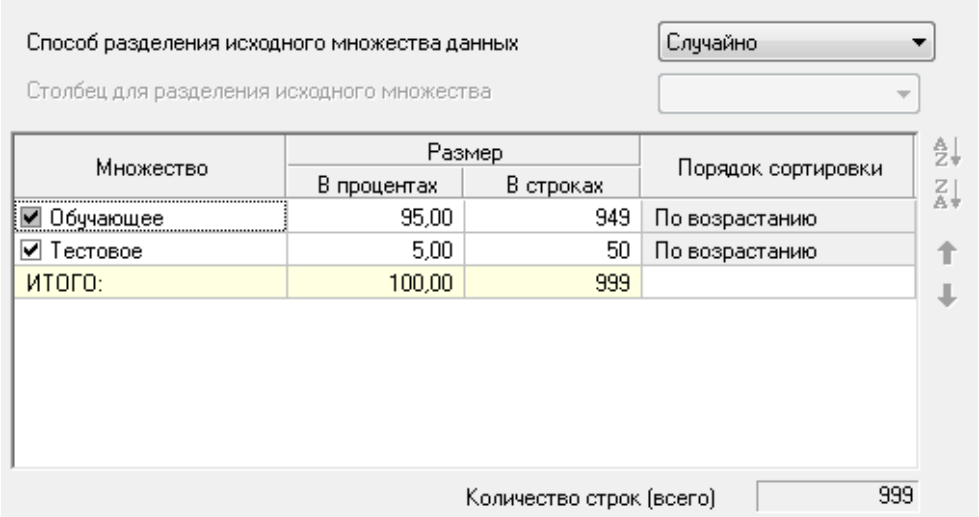


Рисунок 18.15 – Настройка разбиения набора данных

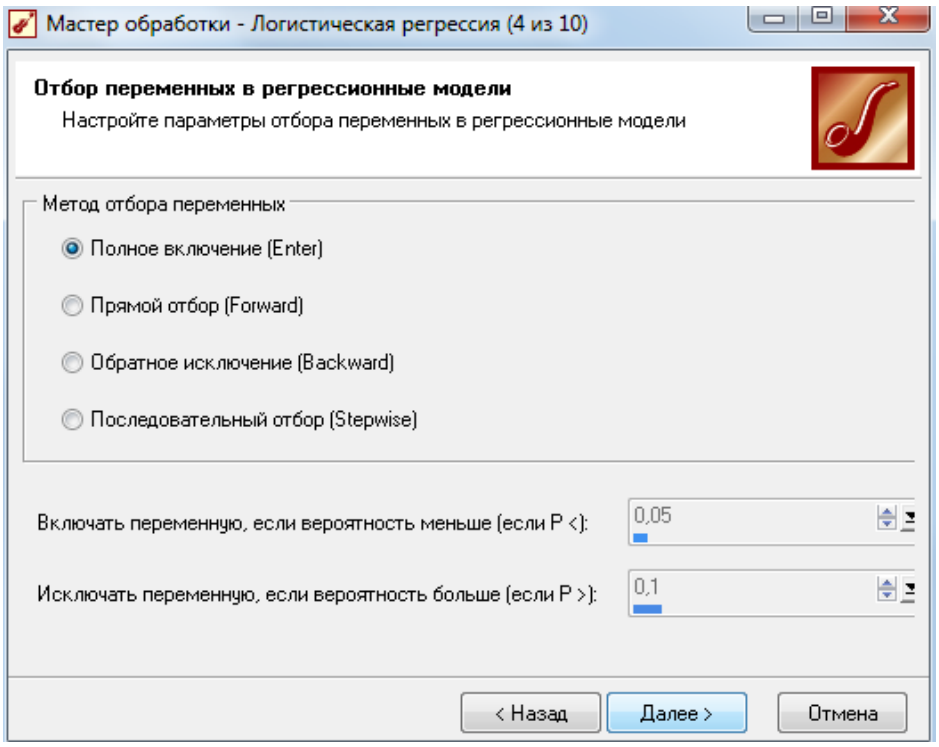


Рисунок 18.16 – Выбор метода отбора переменных

Четвертый шаг предлагает выбрать метод отбора переменных в логистической регрессии (рис.18.16). Пусть это будет – метод полного включения переменных.

Пятый шаг предлагает изменить параметры алгоритма логистической регрессии (рис. 18.17). По умолчанию предлагается порог классификации, равный 0.5. Оставим все без изменений.

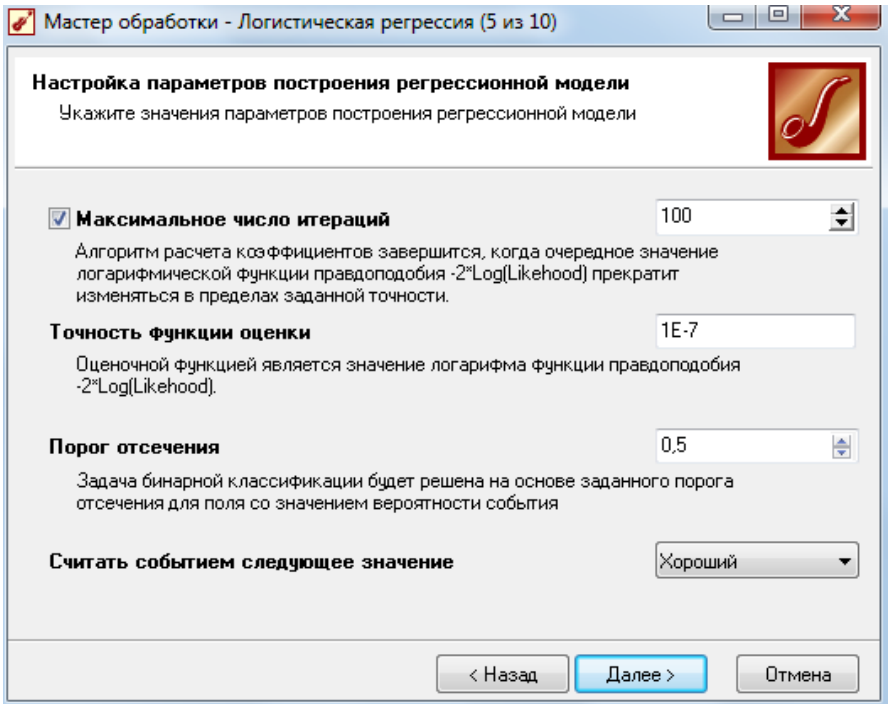


Рисунок 18.17 – Настройки алгоритма логистической регрессии

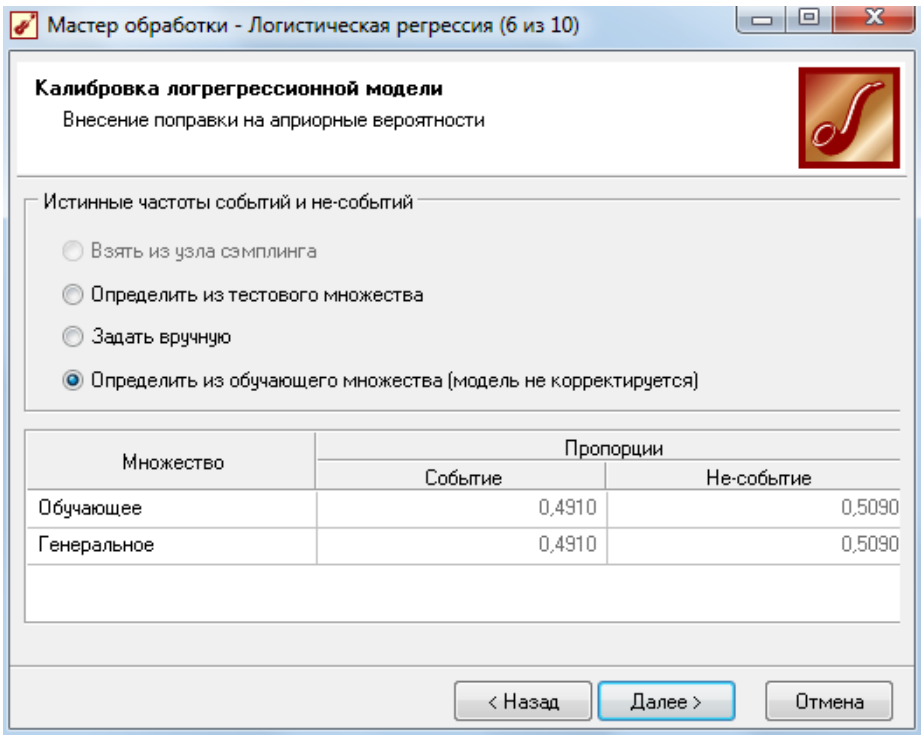


Рисунок 18.18 – Формирование обучающего множества

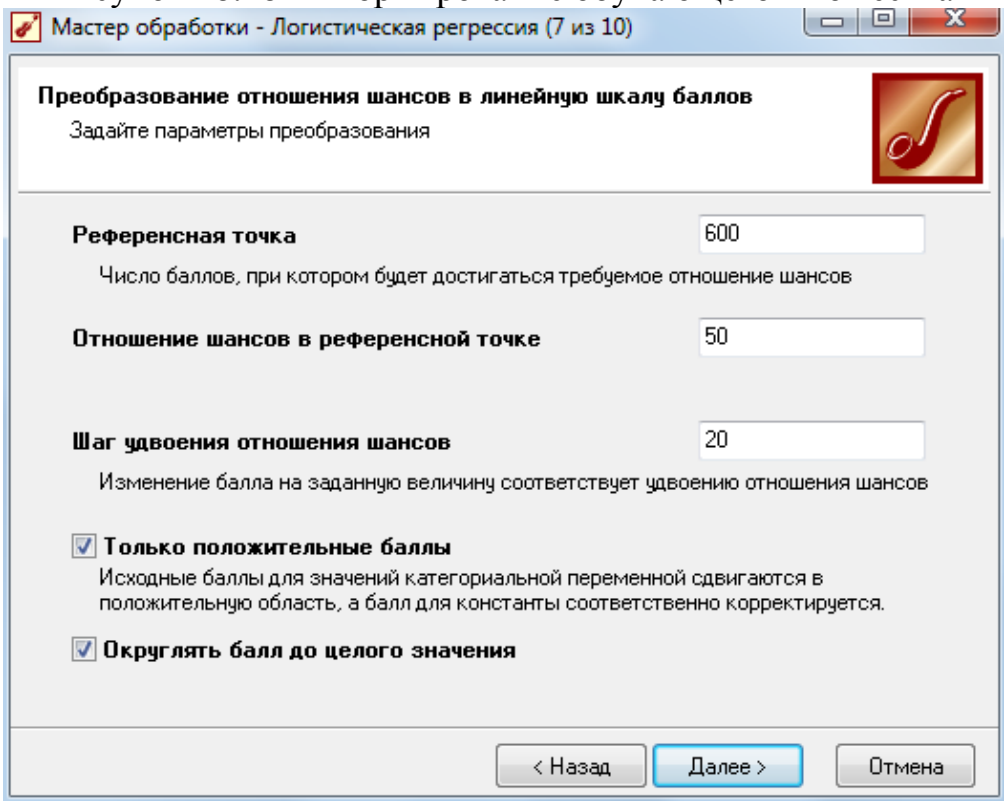


Рисунок 18.19. – Установление соответствия шансов и баллов

Нажав на кнопку Пуск на последнем шаге, будет построена модель и Мастер предложит выбрать визуализаторы узла (рисунок 18.20). Выберем следующие: **ROC-анализ**, **Коэффициенты регрессии**, **Что-если**, **Таблица сопряженности**, **Таблица**.

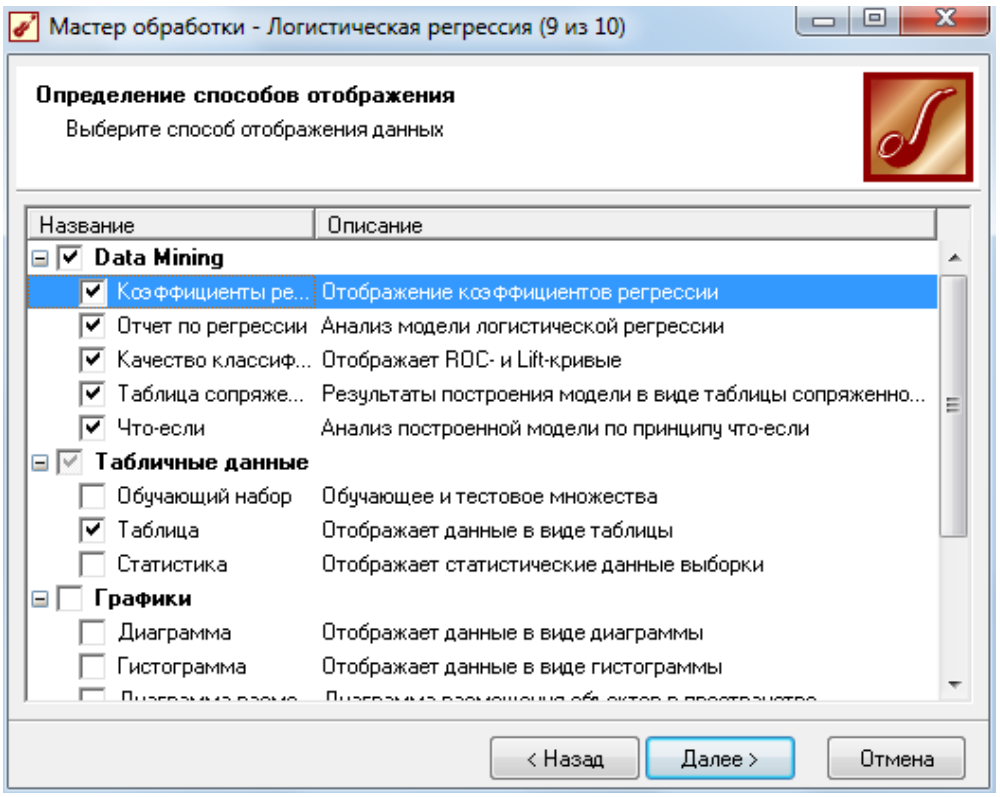


Рисунок 18.20 – Выбор визуализаторов узла

В визуализаторе **Таблица** видно, что после обработчика добавилось три новых колонки: **Класс заемщика_OUT, Класс заемщика_Вероятность события, Класс заемщика_Балл** (Рисунок 18.21). Рейтинг представляет собой рассчитанное значение *y* по уравнению логистической регрессии, а второе поле – принадлежность к тому или иному классу в зависимости от порога округления.

Класс заемщика	Класс заемщика_OUT	Класс заемщика_Вероятность события	Класс заемщика_Балл
Плохой	Плохой	0,308004511518965	376
Хороший	Хороший	0,922328790972217	339
Плохой	Плохой	0,0109343471828614	292
Хороший	Хороший	0,51432676634538	347
Хороший	Плохой	0,0111067977602539	266
Плохой	Плохой	0,0154087533051605	266
Плохой	Плохой	0,0550472774094755	292
Хороший	Хороший	0,689456484206642	292
Хороший	Хороший	0,999999998083161	402
Хороший	Хороший	0,999999964411651	339
Хороший	Хороший	0,513671937031936	339

Рисунок 18.21 – Таблица

Визуализатор **Коэффициенты регрессии** наглядно показывает рассчитанные коэффициенты логистической регрессии (рис. 18.22). Например, в нашем случае видно, что каждый дополнительный иждивенец уменьшает кредитный рейтинг заемщика на величину -1,91 (до логит-преобразования), а каждый дополнительный год стажа работы увеличивает рейтинг на 0,0011.

Атрибут	Коэффициент	Стандартная ошибка	Коэффициент Вальда	Значимость	Отношение шансов	Нижняя граница ДИ	Верхняя граница ДИ	Балл
9.0 <Константа>	-3,859775336							376
12 Возраст	-0,002587060...	0,02089851818	0,01532435159	0,9014803105	0,9974162833	0,9573864492	1,039119828	0
ab Пол								
женский	0							0
мужской	0,6320757758	0,2382719848	7,037077396	0,007983921...	1,881512126	1,179467052	3,001430073	18
ab Состоит в браке								
Да	0							0
Нет	0,2836525478	0,2323309605	1,490593427	0,2221244091	1,327971444	0,8422182157	2,09388508	8
12 Иждивенцы	-1,915155244	0,1859870322	106,0333882	7,250413116...	0,1473189604	0,1023159374	0,2121162807	-55
9.0 Доход	0,0007270435...	5,426110186E-5	179,5327536	6,129724538...	1,000727308	1,000620885	1,000833743	0
9.0 Опыт работы	0,0011831784...	0,02801766103	0,001783350642	0,9663155574	1,001183879	0,9476865857	1,057701116	0
9.0 Срок проживания	0,01350050742	0,01092649527	1,526646267	0,216616479	1,013592051	0,9921158212	1,035533174	0
9.0 Недвижимость	0,0115758417	0,005698759547	4,126145066	0,04222543542	1,011643101	1,000406356	1,02300606	0
9.0 Месячный платеж	-0,000893511...	8,696345659E-5	105,5666503	9,176019278...	0,9991068875	0,9989366059	0,9992771982	0

Рисунок 18.22 – Коэффициенты регрессии

Отчет по регрессии (рис.18.23) позволяет оценить адекватность логистической регрессии (R^2 Мак Фаддена, Хи-квадрат).

Визуализатор **ROC-анализ** выводит график ROC-кривой, на котором по умолчанию рисуется положение текущего порога отсечения и значения чувствительности и специфичности, показатель AUC и типы событий (рис. 18.24). Площадь под кривой равна 0,96, что говорит об очень хорошей предсказательной способности построенной модели.

Лог-регрессия "Финальная"

-2 Log Likelihood	R^2 МакФадден	Хи-квадрат	Число степеней свободы	AIC	AICc	Значимость	Метод отбора переменных
505,701	0,616	809,588	9	527,701	527,701	0,0000	Полное включение

Коэффициенты лог-регрессии

Фактор	Коэффициент	Стандартная ошибка	Коэффициент Вальда	Значимость	Отношение шансов	95% доверительный интервал отношения шансов	
						Минимум	Максимум
Константа	-3,85978						
Возраст	-0,00259	0,0209	0,0153	0,9015	0,99742	0,95739	1,03912
Пол = женский	0,0	0,0000			1,0	1,0	1,0
Пол = мужской	0,63208	0,2383	7,0371	0,0080	1,88151	1,17947	3,00143
Состоит в браке = Да	0,0	0,0000			1,0	1,0	1,0
Состоит в браке = Нет	0,28365	0,2323	1,4906	0,2221	1,32797	0,84222	2,09389
Иждивенцы	-1,91516	0,1860	106,0334	0,0000	0,14732	0,10232	0,21212
Доход	0,00073	0,0001	179,5328	0,0000	1,00073	1,00062	1,00083
Опыт работы	0,00118	0,0280	0,0018	0,9663	1,00118	0,94769	1,0577
Срок проживания	0,0135	0,0109	1,5266	0,2166	1,01359	0,99212	1,03553
Недвижимость	0,01158	0,0057	4,1261	0,0422	1,01164	1,00041	1,02301
Месячный платеж	-0,00089	0,0001	105,5667	0,0000	0,99911	0,99894	0,99928

Рисунок 18.23 – Отчет по регрессии регрессии

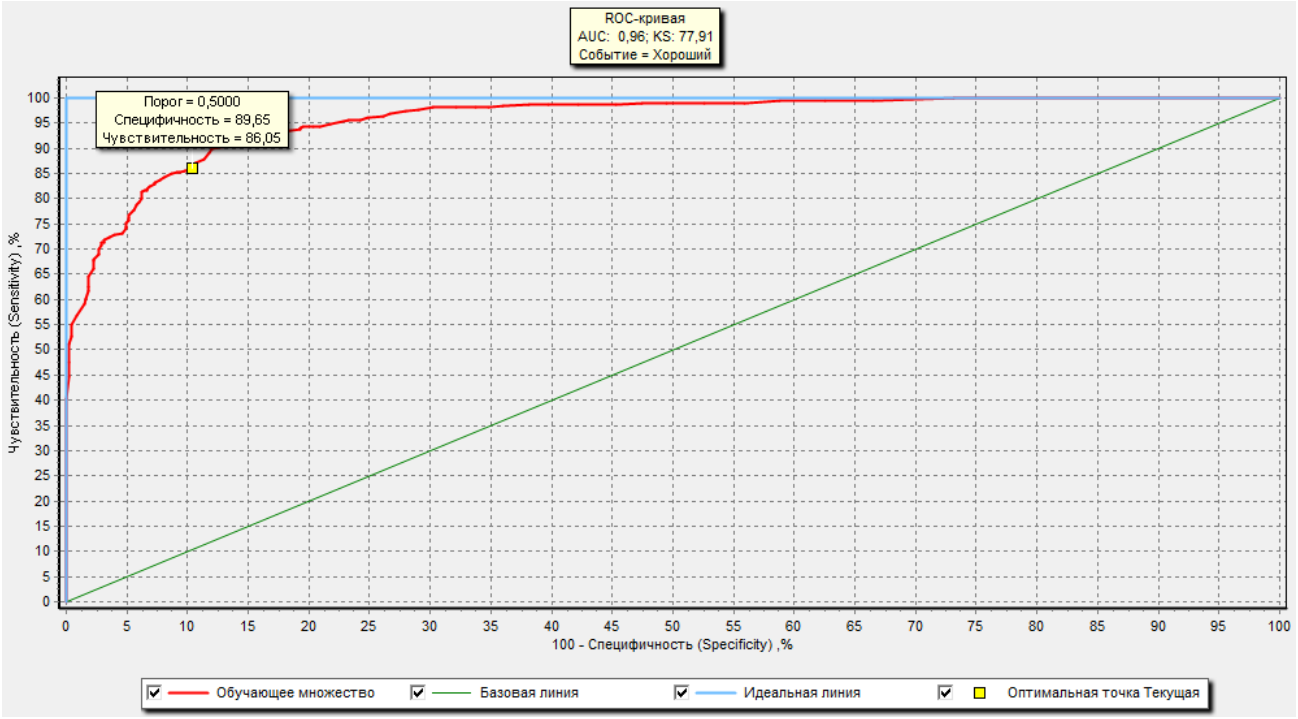


Рисунок 18.24 – График ROC-кривой скоринговой модели с порогом отсечения 0,5

Однако оптимальная точка для данной модели не равна 0,5. Максимальная суммарная чувствительность и специфичность достигается в точке 0,44 (для расчета и отображения оптимальной точки необходимо в меню кнопки  выбрать пункт **Максимум**. В этой оптимальной точке *Se*=88%, *Sp*=89,9%, что означает: 88% благонадежных заемщика будут выявлены классификатором. Специфичность равна 89,9%, следовательно, 10,1% недобросовестных заемщиков получат одобрение в выдаче кредита (кредитный риск).

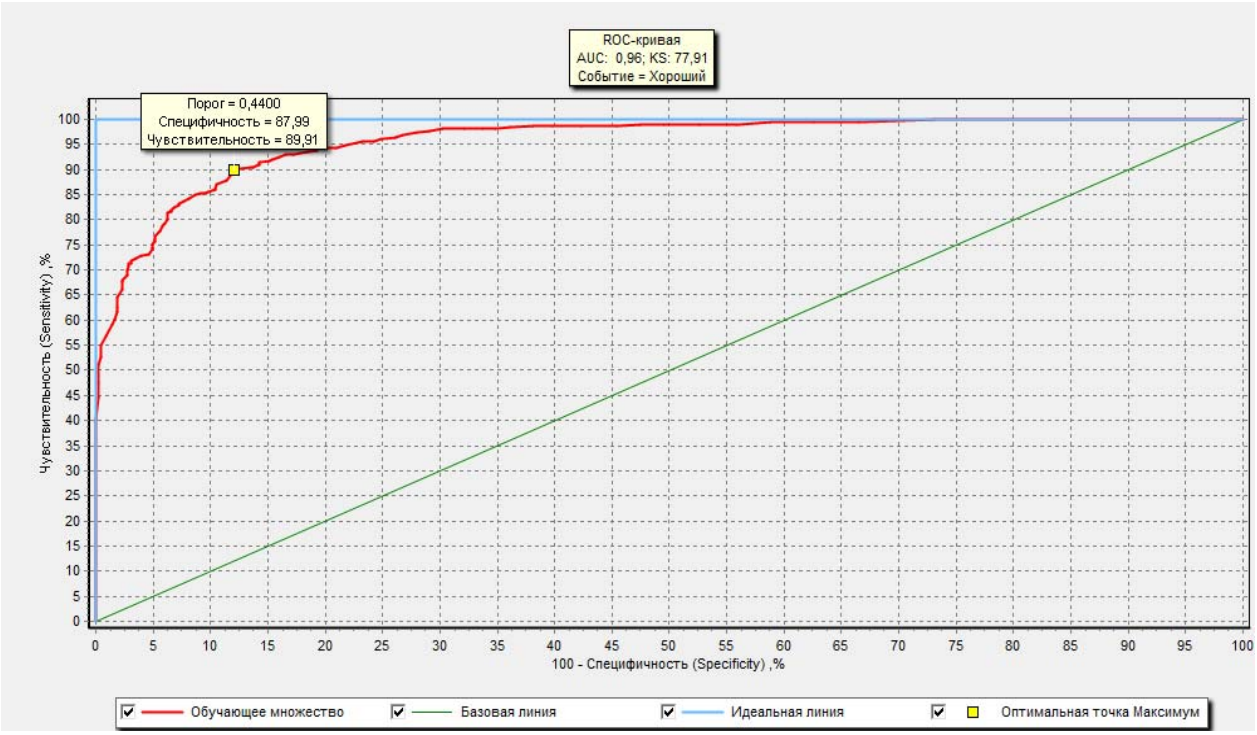


Рисунок 18.25 – График ROC-кривой скоринговой модели с порогом отсечения 0,44

Для установки нового порога отсечения, равного 0,44, необходимо перенастроить узел-обработчик логистической регрессии, нажав кнопку  .

В общем случае проецируя определения чувствительности и специфичности на скоринг (и учитывая то, что класс заемщика «Хороший» соответствует положительному исходу), можно заключить, что скоринговая модель с высокой специфичностью соответствует *консервативной кредитной политике* (чаще происходит отказ в выдаче кредита), а с высокой чувствительностью – *политике рискованных кредитов*. В первом случае минимизируется кредитный риск, связанный с потерями ссуды и процентов и дополнительными расходами на возвращение кредита, а во втором – коммерческий риск, связанный с упущенной выгодой. Это хорошо иллюстрирует визуализатор **Таблица сопряженности** (рис. 18.26). Она показывает результаты сравнения категориальных значений выходного поля исходной (обучающей или тестовой) выборки и категориальных значений выходного, рассчитанных с помощью модели с выбранным порогом отсечения (в данном случае – 0,44).

	Классифицировано		
Фактически	Плохой	Хороший	Итого
Плохой	425	58	483
Хороший	47	419	466
Итого	472	477	949

а

	Классифицировано		
Фактически	Плохой	Хороший	Итого
Плохой	22	2	24
Хороший	1	25	26
Итого	23	27	50

б

Рисунок 18.26 – Таблицы сопряженности: *а* – рабочая выборка, *б* – тестовая выборка

Из таблицы видно, что на обучающем множестве (Рисунок 18.26, *а*) модель реже отказывала в выдаче «хорошим» заемщикам (47 ошибочных случая) и чаще выдавала кредит «плохим» клиентам. Ошибка классификации составила 11%. На тестовом множестве (Рисунок 12.26, *б*) наблюдается примерно та же картина (ошибка классификации 6%). Если такая ситуация не устраивает, можно поднять порог отсечения и добиться того, чтобы модель чаще выдавала отрицательное решение.

Последний визуализатор **Что-если** позволяет исследовать, как будет вести себя построенная модель при подаче на ее вход тех или иных данных. Иначе говоря, проводится эксперимент, в котором, изменяя значения входных полей обучающей или рабочей выборки в нашем случае логистической регрессии, пользователь наблюдает за изменением значений на выходе.

Возможность анализа по принципу «Что-если» особенно ценна, поскольку позволяет исследовать правильность работы системы, достоверность полученных результатов, а также ее устойчивость. Под устойчивостью понимается то, насколько снижается достоверность полученных результатов при попадании на вход системы нетипичных данных – выбросов, пропусков данных и т.д. Такой анализ поз-

волит определить какую предварительную обработку данных нужно провести перед подачей на вход системы.

Визуализатор **Что-если** включает табличное и графическое представления, которые формируются одновременно (рис. 18.27).

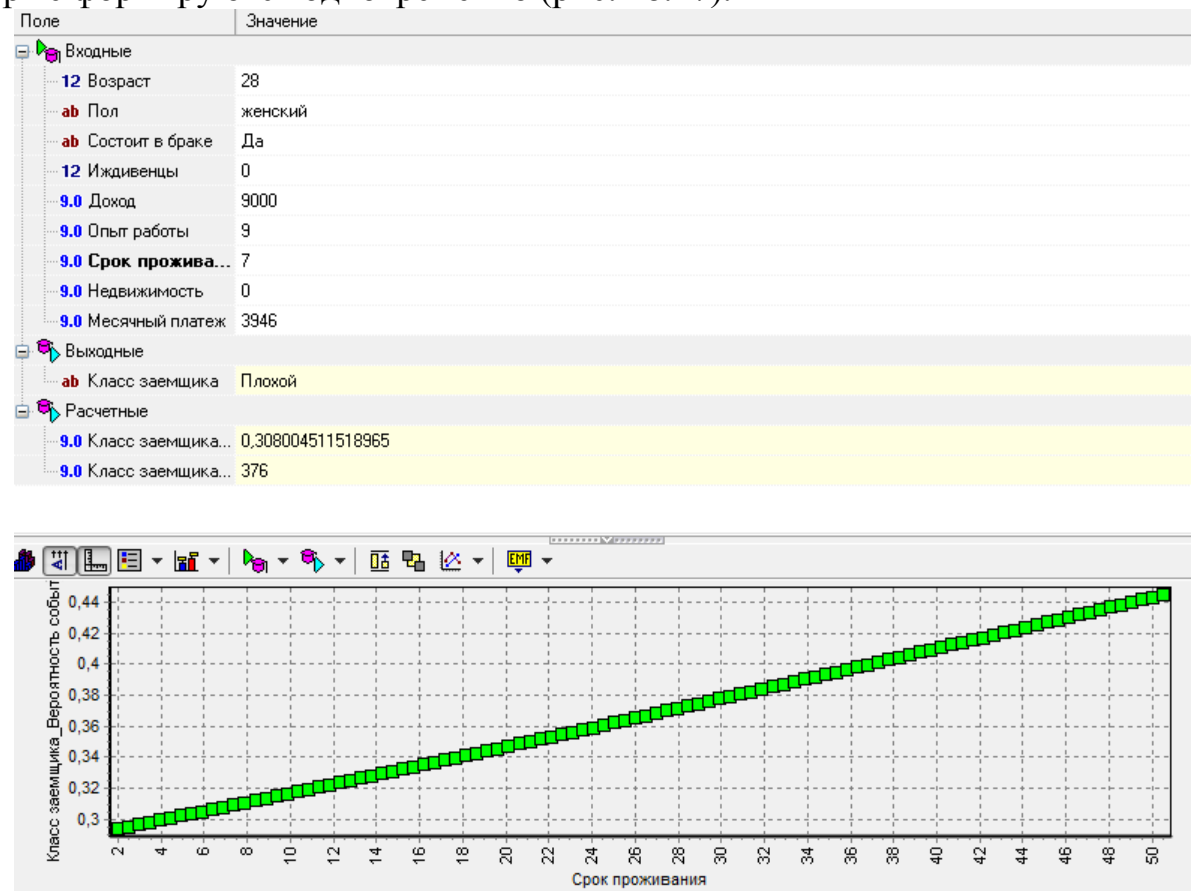


Рисунок 18.27 – Визуализатор «Что-если»

В табличном представлении в верхней части таблицы отображаются входные поля, а в нижней – выходные и расчетные. Изменяя значения входных полей, пользователь дает команду на выполнение расчета и наблюдает рассчитанные значения выходов логистической регрессии. Расчетные поля отличаются от выходных тем, что они не существуют в исходном наборе данных и были созданы в ходе обработки. Таким полями является **Рейтинг**.

В графическом представлении **Что-если** по горизонтальной оси диаграммы отображается весь диапазон значений текущего поля выборки, а по вертикальной – значения соответствующих выходов модели. По диаграмме **Что-если** видно, при каком значении входа изменяется значение на соответствующем выходе. Если, например, во всем диапазоне входных значений выходное значение для данного поля не изменялось, то диаграмма будет представлять собой горизонтальную прямую линию. В нашем случае установлена графическая зависимость изменения кредитного рейтинга конкретного клиента от срока его проживания (все остальные входы – константы). Видно, что с увеличением срока проживания рейтинг линейно растет.

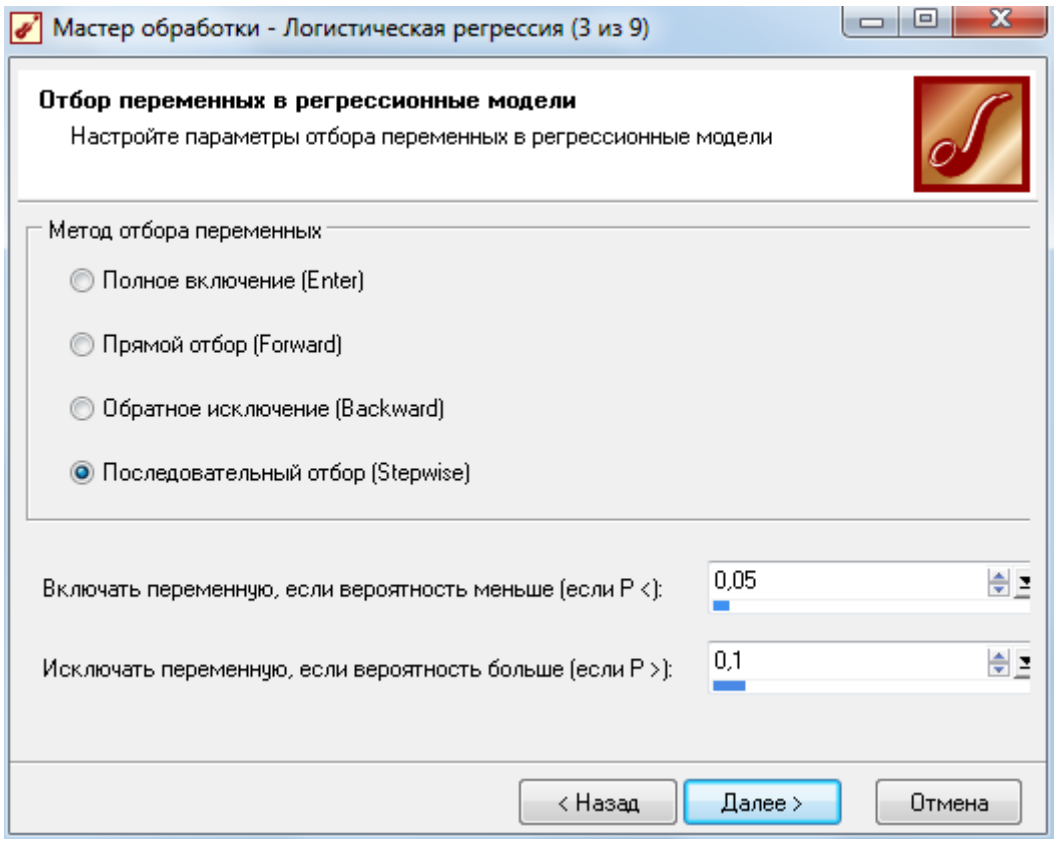


Рисунок 18.28 – Метод отбора переменных в регрессии

Замечание. Прямой отбор переменных в логистической регрессии показал, что часть коэффициентов при переменных (например, для имеющегося примера: возраст, опыт) не являются статистически значимыми, поэтому следует рекомендовать рассмотреть другие методы отбора переменных при построении логистической регресии (рис. 18.28) (например, обратное исключение или последовательный отбор) лучшую модель можно выбрать по наилучшей предсказательной силе, которая характеризуется площадью под ROC-кривой: AUC.

Теперь в этом же сценарии построим еще одну скоринговую модель, но уже на основе дерева решений. Добавим в ветку сценария одноименный обработчик **Дерево решений** (рис.18.29).

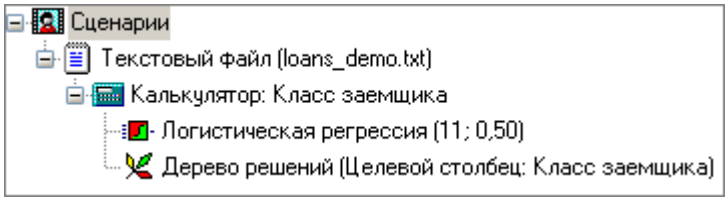


Рисунок 18.29 –Сценарий в Deductor

Откроется окно **Мастера обработки**. Первые два шага мастера аналогичны тем шагам, что делались ранее в обработчике **Логистическая регрессия**. На третьем шаге откроется окно выбора параметров алгоритма C4.5 (рис. 18.30).

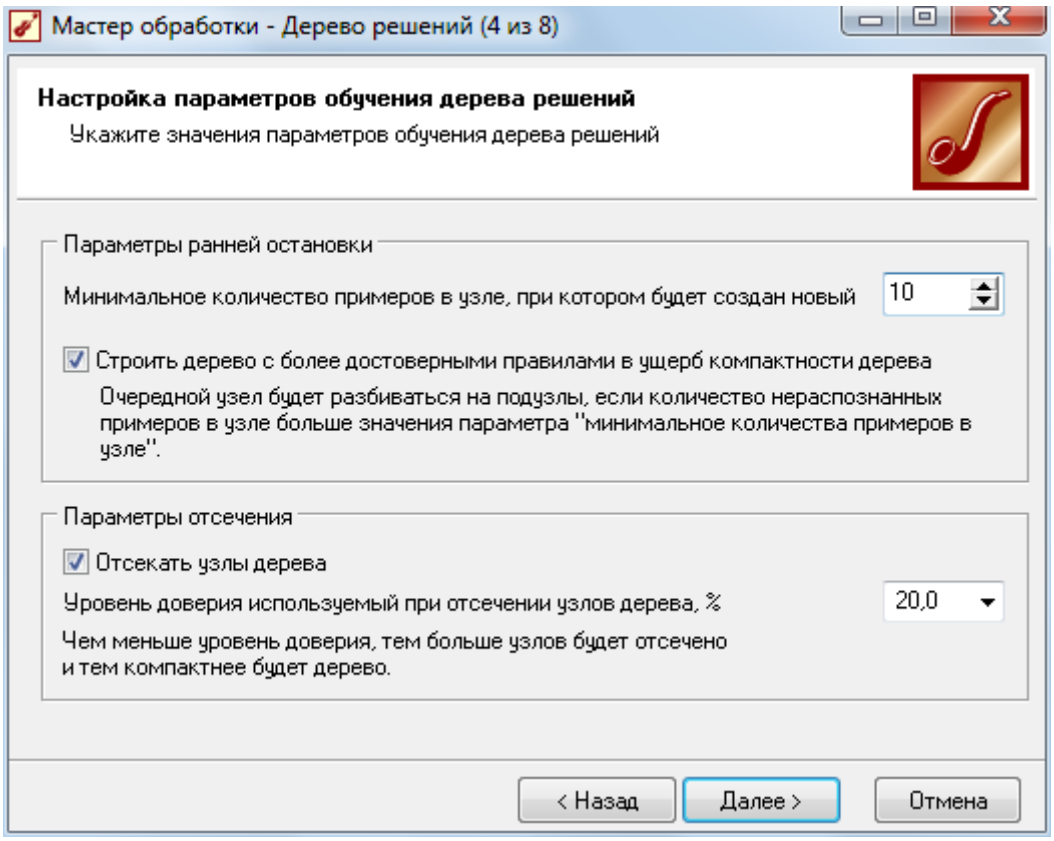


Рисунок 18.30 – Настройки алгоритма дерева решений

Оставим все настройки по умолчанию за исключением минимального количества примеров в узле, при котором будет создаваться новый. Этот параметр сделаем равным 1% от объема всей выборки. Меньшее значение может привести к появлению статистически недостоверных правил, бо́льшее – к почти полному отсутствию таковых («бедное» дерево решений). На следующем шаге построим дерево решений (кнопка **Пуск**) и снова выберем нужные нам визуализаторы: **Дерево решений**, **Правила**, **Значимость атрибутов**, **Что-если**, **Таблица сопряженности**, **Таблица**.

В результате работы алгоритма C4.5 было выявлено 18 правил, точность классификации на обучающей выборке составила 90,5%, на тестовой – 88%. Модель логистической регрессии с порогом отсечения 0.44 обеспечивает примерно такую же точность, а на тестовом множестве даже выше. Это означает, что между входами и выходами наблюдаются преимущественно линейные зависимости, и дерево решений не смогло до конца проявить свой потенциал, выражающийся в моделировании нелинейных связей.

Визуализатор **Дерево решений** позволяет просматривать правила в виде дерева, а также достоверность и поддержку каждого узла (рисунок 18.31). Те же самые правила в виде продукций «Если-то» можно просмотреть с помощью визуализатора **Правила**.

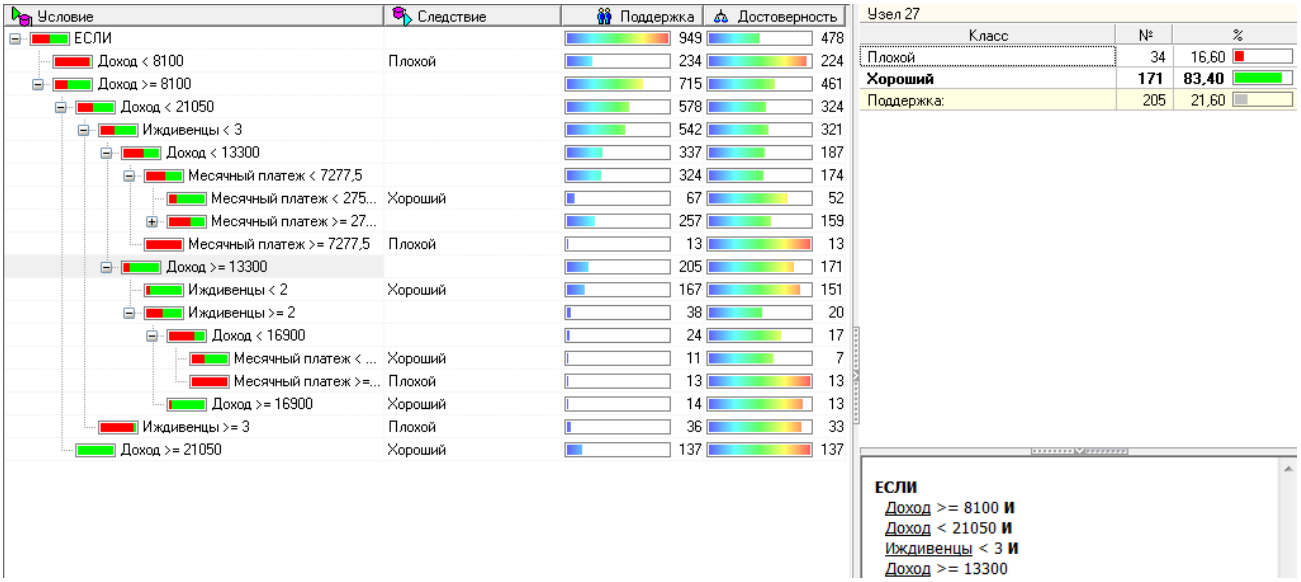


Рисунок 12.31 – Визуализатор «Дерево решений»

С помощью визуализатора **Значимость атрибутов** можно определить, на сколько сильно выходное поле зависит от каждого из входных факторов (рисунок 18.32). Визуализатор представляет из себя таблицу, состоящую из 3-х столбцов: «№», «Атрибут» и «Значимость, %». Каждая строчка таблицы содержит один из входных атрибутов, определенных при настройке назначения полей дерева решений. Каждому входному атрибуту соответствует значимость - степень зависимости выходного поля от этого атрибута. Параметр значимость тем больше, чем больший вклад вносит конкретный входной атрибут при классификации выходного поля. Фактически данный визуализатор показывает степень нелинейной зависимости между выходным и входными полями.

Целевой атрибут: Класс заемщика

№	Номер	Атрибут	Значимость, %
1	5	Доход	70,943
2	4	Иждивенцы	12,279
3	9	Месячный платеж	11,531
4	1	Возраст	2,226
5	7	Срок проживания	1,699
6	8	Недвижимость	1,323
7	2	Пол	0,000
8	6	Опыт работы	0,000
9	3	Состоит в браке	0,000

Рисунок 18.32 – Значимость атрибутов в дереве решений

Задание

Используя предложенные преподавателем кредитные истории, хранящиеся в текстовых файлах, выполните следующее:

1. Постройте скоринговую модель на основе логистической регрессии

2. Рассчитайте оптимальный скоринговый балл и балл, при котором достигается 90% чувствительность модели.

3. Постройте несколько моделей деревьев решений при различных настройках алгоритма. Постройте различные деревья решений для заемщиков, состоящих в

браке и не состоящих в браке. Выберите из них ту модель, которая чаще отказывает в выдаче кредита.

3. Сравните качество моделей друг с другом. Выработайте рекомендации по выбору моделей.

Вопросы для самоконтроля

- Что такое скоринг?
- Почему важно использовать скоринг в розничном кредитовании.
- Какие алгоритмы позволяют строить скоринговые модели?
- Как строится ROC-кривая?
- Что такое чувствительность и специфичность?
- Как подобрать оптимальный скоринговый балл?
- На каких идеях базируется алгоритм C4.5?
- В чем достоинства и недостатки логистической регрессии и деревьев решений применительно к скорингу?

Практическое занятие № 19

Ассоциативные правила

Цель занятия: изучить технику применения ассоциативных правил на примере задачи стимулирования розничных продаж

Теоретические сведения

Основные понятия

Аффинитивный анализ (*affinityanalysis*) - один из самых распространенных методов изучения массовых данных. Его название происходит от английского слова *affinity* (близость, сходство), а его цель - обнаружить ассоциации между событиями, то есть найти правила для количественного описания взаимной связи между двумя или более событиями, которые происходят совместно. Такие правила принято называть ассоциативными (*associationrules).*

С помощью данного метода решаются многие практические задачи, в том числе:выявление товаров, которые часто покупаются вместе (в одном наборе);определение доли клиентов, положительно относящихся к нововведениям в сфере обслуживания; создание профиля посетителя веб-ресурса; определение доли случаев, в которых новое лекарство вызывает нежелательные побочные эффект и т.д.

Базовым понятием в теории ассоциативных правил является *транзакция* - некоторое множество событий, происходящих совместно.

Типичным примером может служить приобретение клиентом некоторого набора товаров в супермаркете. Как правило, такой набор формируется не случайно; покупка одного товара влияет на вероятность приобретения других (увеличивает или уменьшает ее).

Эту связь и устанавливают ассоциативные правила; так, может быть обнаружено, что покупатель, купивший молоко, с вероятностью 75% купит также и хлеб.

Анализ рыночной корзины (*marketbasketanalysis)* - стандартная область применения аффинитивного анализа.

Современные кассовые аппараты в супермаркетах позволяют собирать обширную информацию о покупках, которая затем может сохраняться в базе данных и использоваться для поиска ассоциативных правил.

В табл. 19.1 приведен простейший пример, содержащий данные о 10 транзакциях, касающихся 13 видов продуктов. Хотя на практике чаще приходится иметь дело с тысячами и даже миллионами транзакций, в которые вовлечены десятки и сотни различных продуктов, этого будет достаточно для иллюстрации основных особенностей рассматриваемого метода.

Сопоставление приведенных данных показывает, что все четыре транзакции, в которых фигурирует салат, включают также и помидоры, и что четыре из семи транзакций, содержащих помидоры, также содержат и салат. Таким образом, салат и помидоры в большинстве случаев покупаются вместе. Ассоциативные правила и предназначены для описания таких совпадений.

Таблица 19.1 – Пример набора транзакций

Номер транзакции	Состав рыночной корзины
1	Сливы, салат, помидоры
2	Сельдерей, конфеты
3	Конфеты
4	Яблоки, морковь, помидоры, картофель, конфеты
5	Яблоки, апельсины, салат, конфеты, помидоры
6	Персики, апельсины, сельдерей, помидоры
7	Фасоль, салат, помидоры
8	Апельсины, салат, морковь, помидоры, конфеты
9	Яблоки, бананы, сливы, морковь, помидоры, лук, конфеты
10	Яблоки, картофель

Любое ассоциативное правило состоит из двух наборов предметов, один из которых называется *условием (antecedent*), другой - *следствием (consequent*), а также логической связи между ними (оператора «если - то»). Записывают его в виде *X→Y,* что означает «из *X* следует *Y*», или «если *X*, то *Y*»; примером может служить только что выявленная закономерность *{помидоры} → {салат}.*

Условие и следствие часто называют соответственно левосторонним *(LHS - left-handside)* и правосторонним *(RHS - right-handside)* компонентом ассоциативного правила.

Связь между условием и следствием характеризуется двумя показателями - *поддержкой* и *достоверностью,* обозначаемых соответственно*S (support*) и *C* (*confidence*). Если через *P* обозначить число транзакций, удовлетворяющих определенному условию, названные показатели для правила*A→B* можно рассчитать следующим образом:

S(A→B) = P(A∩B) = (Количество транзакций, содержащих A и *B*): (Общее количество транзакций)

C(A→B) = *P(A|B) = P(A∩B)/P(A)* = (Количество транзакций, содержащих *A* и *B)*: (Количество транзакций, содержащих *A)*

Показатель поддержки характеризует частоту, с которой интересующий нас набор *AB* встречается в общей совокупности данных, а показатель достоверности - частоту, с которой в этой совокупности соблюдается правило *A→B* («если *A,* то *B»),* Если поддержка и достоверность велики, это позволяет с достаточной степенью уверенности утверждать, что любая будущая транзакция, которая включает условие, будет также содержать и следствие.

Рассчитаем показатели поддержки и достоверности для ассоциации *{салат} → {помидоры}* из табл. 19.1. Поскольку количество транзакций, содержащих оба элемента, равно 4, а их общее число - 10, поддержка данной ассоциации составит:

S ({*салат*} → {*помидоры*}) = *4:10 = 0,4.*

Количество транзакций, содержащее только условие *{салат},* равно 4; следовательно, достоверность данной ассоциации

C ({*салат*} —» {*помидоры*}) = *4:4 = 1.*

Таким образом, все наборы покупок, содержащие салат, содержат также и помидоры, из чего следует, что данная ассоциация может рассматриваться как правило. Интуитивно это вполне объяснимо, поскольку оба продукта часто используются вместе для приготовления различных блюд.

Теперь рассмотрим другую ассоциацию: *{конфеты}→{помидоры}*. Эти продукты слабо совместимы в гастрономическом плане (тот, кто хочет сделать домашний салат, вряд ли станет покупать конфеты, а покупатель, желающий приобрести что-нибудь сладкое к чаю, скорее всего, не станет заодно покупать и помидоры). Поддержка данной ассоциации S = 4:10 = 0,4, а достоверность C = 4:7 = 0,57. Сравнительно низкая достоверность дает повод усомниться в том, что она является правилом.

При анализе предпочтение может отдаваться правилам, имеющим высокую поддержку или высокую достоверность, но чаще принимают во внимание лишь те ассоциации, по которым оба показателя достаточно велики. Правила, для которых значения поддержки или достоверности превышают некоторый порог, заданный пользователем, называются *сильными (strongrules)*.

Например, аналитика может интересовать, какие товары в супермаркете, покупаемые вместе, образуют ассоциации с минимальной поддержкой 20% и минимальной достоверностью 70%. С другой стороны, при анализе мошенничеств уровень поддержки может быть уменьшен до 1%, поскольку к этой категории относится лишь очень небольшая часть транзакций.

Значимость ассоциативных правил

Высокие уровни поддержки и достоверности сами по себе еще не свидетельствуют о значимости обнаруженной ассоциации. Например, если товар A встречался в 70 транзакциях из 100, а товар B – в 80, и в 50 случаях из 100 они оказываются в одном наборе, то ассоциация $A \rightarrow B$ не может считаться правилом, хотя в данном случае S = 0,5, а C = 0,5:0,7 = 0,71. Просто эти товары очень популярны и только поэтому часто встречаются в одной транзакции.

Если решения о покупке двух товаров независимы, естественно, говорить о каком-то правиле, их связывающем, не приходится. Из математической статистики известно, что если условие и следствие не зависят друг от друга, поддержка правила в целом будет примерно равна произведению поддержки только условия и поддержки только следствия. В данном случае $S(A)$ = 0,7, $S(B)$ = 0,8, а их произведение $S(A)S(B)$ = 0,56, то есть примерно совпадает с $S(A \rightarrow B)$ = 0,5. Таким образом, предположение о независимости решений о покупке товаров A и B достаточно обоснованно.

Фиктивные «правила», игнорирующие указанное обстоятельство, встречаются довольно часто. Например, если статистика дорожно-транспортных происшествий по Москве показывает, что из 100 аварий в 70 участвуют иномарки, то, на первый взгляд, это выглядит как правило: «если *{авария},* то *{иномарка}*». Однако если учесть, что московский парк легковых автомобилей на две трети состоит из иномарок, такое правило нельзя назвать значимым.

Таким образом, кроме поддержки и достоверности при поиске ассоциативных правил необходимо использовать показатели, отражающие степень независимости причины и следствия; самый простой из них – так называемый лифт

$$L(A \to B) = \frac{C(A \to B)}{S(B)} = \frac{P(A \cap B)}{P(A)P(B)}.$$

Лифт - это отношение частоты появления следствия в транзакциях, которые также содержат и условие, к частоте появления следствия в целом. Поэтому, если $L > 1$, более вероятно появление следствия в транзакциях, содержащих условие, чем во всех остальных. Можно сказать, что лифт является обобщенной мерой связи двух предметных наборов: при $L > 1$ связь положительная, при $L = 1$ она отсутствует, а при $L < 1$ - отрицательная.

Например, для ассоциации *{помидоры}-{салат}* из табл.19.1 получим

$$L(A \to B) = 0{,}4 : [0{,}7 \cdot 0{,}4] = 1{,}425.$$

Точно так же для ассоциации *{помидоры}→{конфеты}*

$$L(A \to B) = 0{,}4 : [0{,}7 \cdot 0{,}6] = 0{,}95 \approx 1.$$

Таким образом, в первом случае между элементами ассоциации обнаруживается положительная связь, а во втором случае какая- либо связь отсутствует.

Поиск ассоциативных правил

Простейший алгоритм поиска состоит в том, что для всех ассоциаций, которые могут быть построены на основе базы данных, определяется поддержка и достоверность, а затем отбираются те из них, для которых эти показатели превышают заданное пороговое значение. Однако в большинстве случаев такое элементарное решение неприемлемо, поскольку число ассоциаций, которое при этом придется анализировать, слишком велико.

Так, если выборка содержит всего 100 предметов, количество образуемых ими ассоциаций будет порядка 10^{31},а в реальных ситуациях (например, при анализе покупок в супермаркете) номенклатура учитываемых продуктов может достигать нескольких тысяч и более. Очевидно, что никаких вычислительных мощностей на такой расчет не хватит.

Поэтому на практике при поиске ассоциативных правил используют различные приемы, которые позволяют снизить пространство поиска до размеров, обеспечивающих приемлемые затраты машинного времени. Сейчас одним из наиболее распространенных является алгоритм*apriori* (Agrawal и Srikant, 1994), основанный на понятии *популярного набора (frequentitemset,* часто встречающийся предметный набор). Этот термин обозначает предметный набор, частота появления которого в общей совокупности транзакций превышает некоторый заранее заданный уровень.

Таким образом, алгоритм *apriori* включает два этапа:

1) поиск популярных наборов;

2) формулировка ассоциативных правил, удовлетворяющих заданным ограничениям по уровням поддержки и достоверности.

Пример: Стимулирование продаж в интернет-магазине

Построение набора правил

В Deductor для решения задач рассматриваемого типа применяется обработчик Ассоциативные правила, в котором реализован алгоритм*apriori.* На входе он запрашивает два поля: идентификатор транзакции и элемент транзакции. В качестве идентификатора может использоваться, например, номер чека или код клиента; в этом случае элементом будет наименование заказанного товара или услуги.

По завершении работы алгоритма формируется набор данных следующей структуры (табл. 19.2).

Таблица 19.2 – Поля результирующего набора данных

№	Имя	Метка	Тип	Описание
1	N	№	Целый	Номер ассоциативного правила
2	ANTECEDENT	Условие	Строковый	Условие ассоциативного правила (заключено в двойные кавычки)
3	CONSEQUENT	Следствие	Строковый	Следствие ассоциативного правила (заключено в двойные кавычки)
4	SUPPORTCOUNT	Поддержка, количество случаев	Целый	Число транзакций, удовлетворяющих данному правилу
5	SUPPORT	Поддержка, %	Вещественный	Поддержка ассоциативного правила в процентах
6	CONFIDENCE	Достоверность, %	Вещественный	Достоверность ассоциативного правила в процентах

Вся прочая информация, полученная в результате решения, доступна через специализированные визуализаторы **Правила, Популярные наборы, Дерево правил, Что-если**.

Рассмотрим конкретный пример из области розничной торговли. Компания AdventureWorkCycleRussia является дистрибьютором спортивных (серия *Sport),* горных (серия *Mountain)* и дорожных (серия *Road)* велосипедов и комплектующих к ним компании AdventureWorkCycle на территории России и стран СНГ. Офисы компании работают в шести городах России, а также на Украине и в Казахстане. В большинстве регионов компания работает через своих партнеров, центральный офис находится в Москве. У фирмы есть склад и собственная сборочная база.

Отдел маркетинга заинтересован в увеличении продаж через интернет-магазин, размещенный на web-сайте компании. Для этого важно знать, какие товары покупатели могут выбрать в дальнейшем в зависимости от того, что уже имеется в их корзинах. Такой прогноз позволит также оптимизировать структуру сайта - товары, часто покупаемые вместе, будут расположены по соседству на одной web-странице.

Для решения поставленной задачи отдел маркетинга предоставил данные о 5 тыс. чеков от предыдущих покупателей; соответствующая информация содержится в сценарии **cycles.ded**. Откроем Deductor Studio, загрузим в программу этот сценарий и создадим новый проект.

Импортируем данные из текстового файла **cycle_store.txt**; в этом наборе имеются два строковых поля (столбца):**ID** (код чека) и **ITEM**(наименование товара).

Рассмотрим решение задачи в Deductor по шагам. К узлу импорта добавим обработчик **Ассоциативные правила**, причем поле **ID** сделаем идентификатором транзакции, а **ITEM –** ее элементом (рис.19.1).

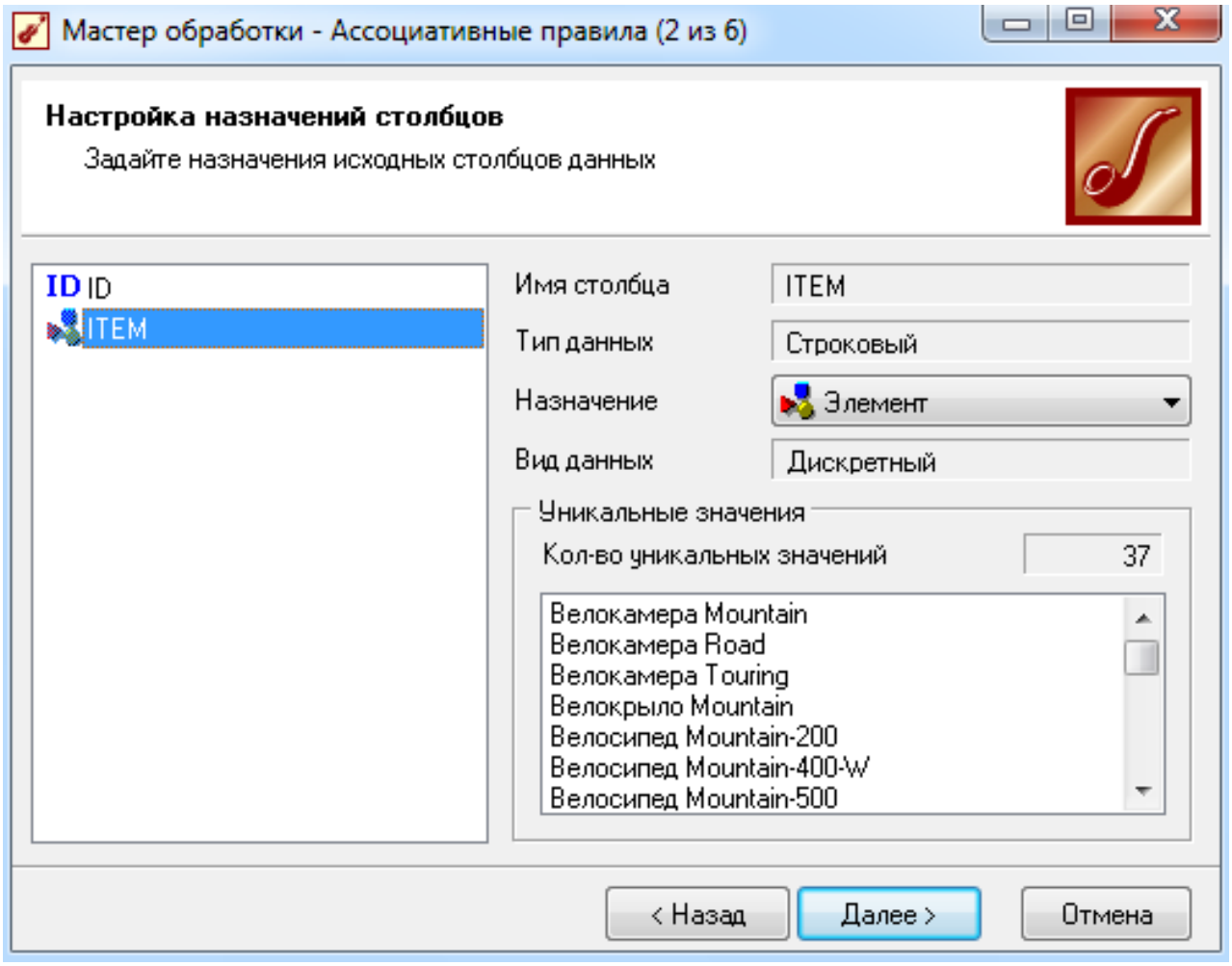


Рисунок 19.1 – Настройка назначений столбцов

Далее следует выбрать параметры построения правил, то есть, по сути, параметры работы алгоритма *a priori* (рис.19.2).

В данном окне можно указать пороговые уровни (максимальный и минимальный) поддержки и достоверности искомых правил, а также максимальную численность популярных наборов, которые программа будет рассматривать (параметр Максимальная мощность искомых часто встречающихся множеств).

Например, если в этом поле установить значение «4», генерация популярных наборов будет остановлена после получения множества 4-предметных наборов. Такое ограничение позволяет избежать появления длинных ассоциативных правил, которые с трудом поддаются содержательной интерпретации.

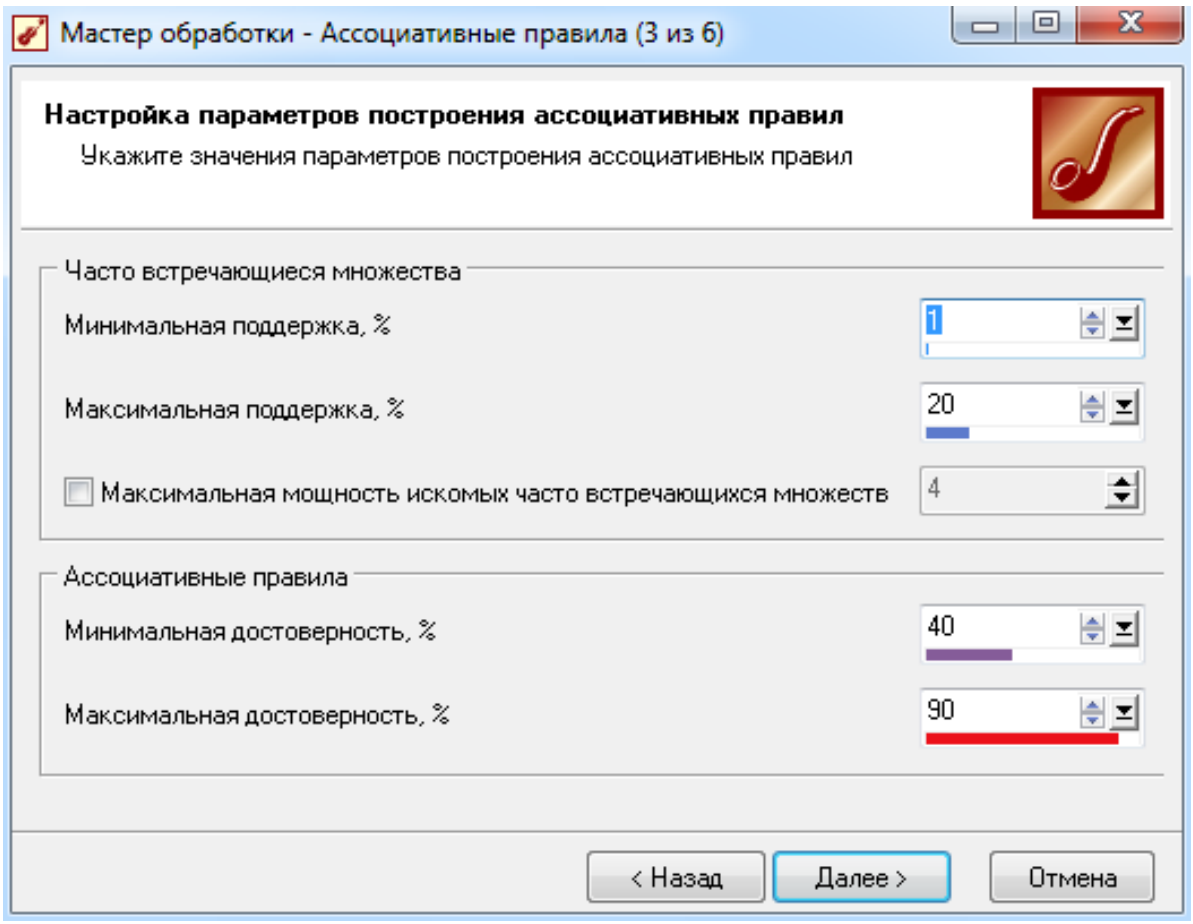


Рисунок 19.2 – Настройка параметров алгоритма

Оставим настройки, принятые по умолчанию, и щелкнем по кнопке**Далее**. Будет запущен алгоритм поиска ассоциативных правил, и по завершении его работы появится окно, содержащее следующую информацию (рисунок 19.3):

Кол-во множеств - число популярных наборов, удовлетворяющих заданным условиям минимальной поддержки и достоверности;

Кол-во правил - число сгенерированных программой ассоциативных правил.

В следующем окне следует выбрать способы представления результатов анализа; отметим все специализированные визуализаторы, а также визуализатор**Таблица** (рис. 19.4).

Все визуализаторы (кроме**Что-если**) позволяют более детально рассмотреть те или иные аспекты полученного решения; рассмотрим их подробнее.

На вкладке **Популярные наборы**, как и следует из ее названия, отображается множество найденных популярных предметных наборов в виде списка. Соответствующие кнопки позволяют: выбрать несколько вариантов сортировки списка, вызвать окно настройки фильтрации множеств.

Например, задав в фильтре минимальное значение поддержки 6% и отсортировав их по убыванию этого параметра, получим следующие 16 популярных наборов (рис.19.5).

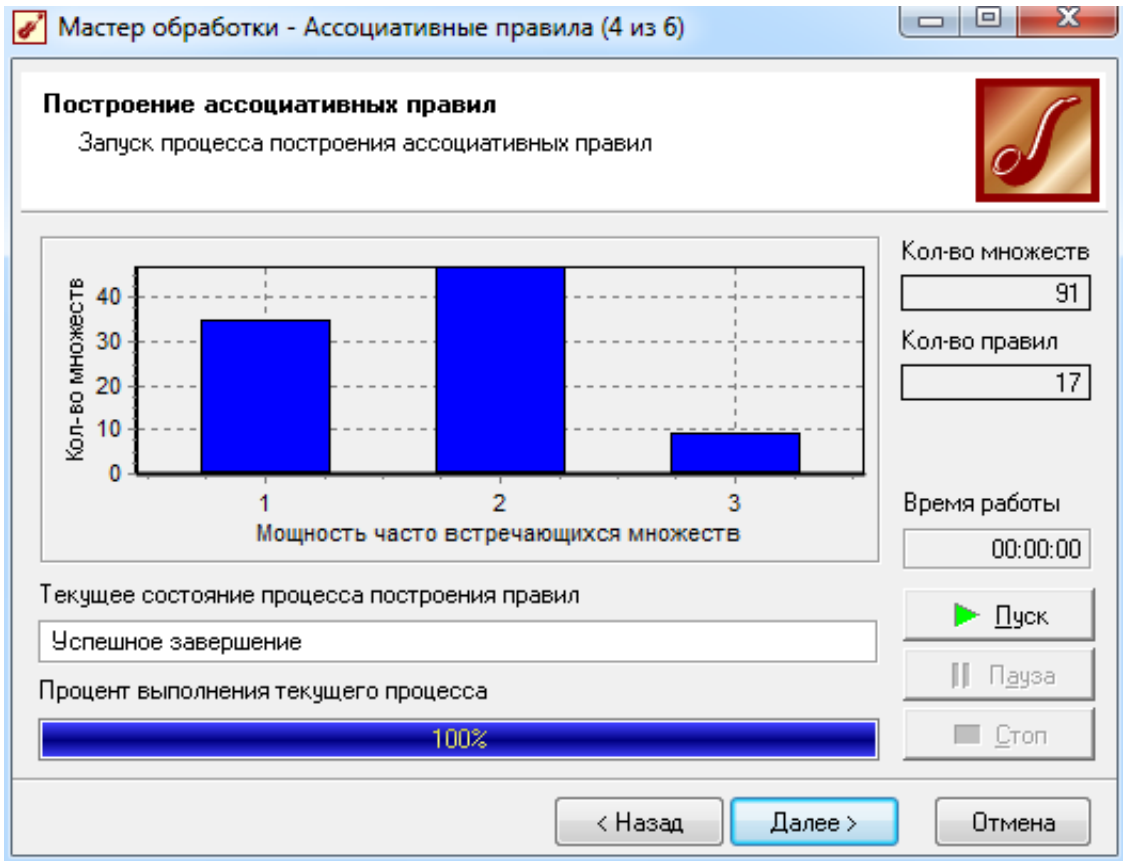


Рисунок 19.3 – Результаты поиска

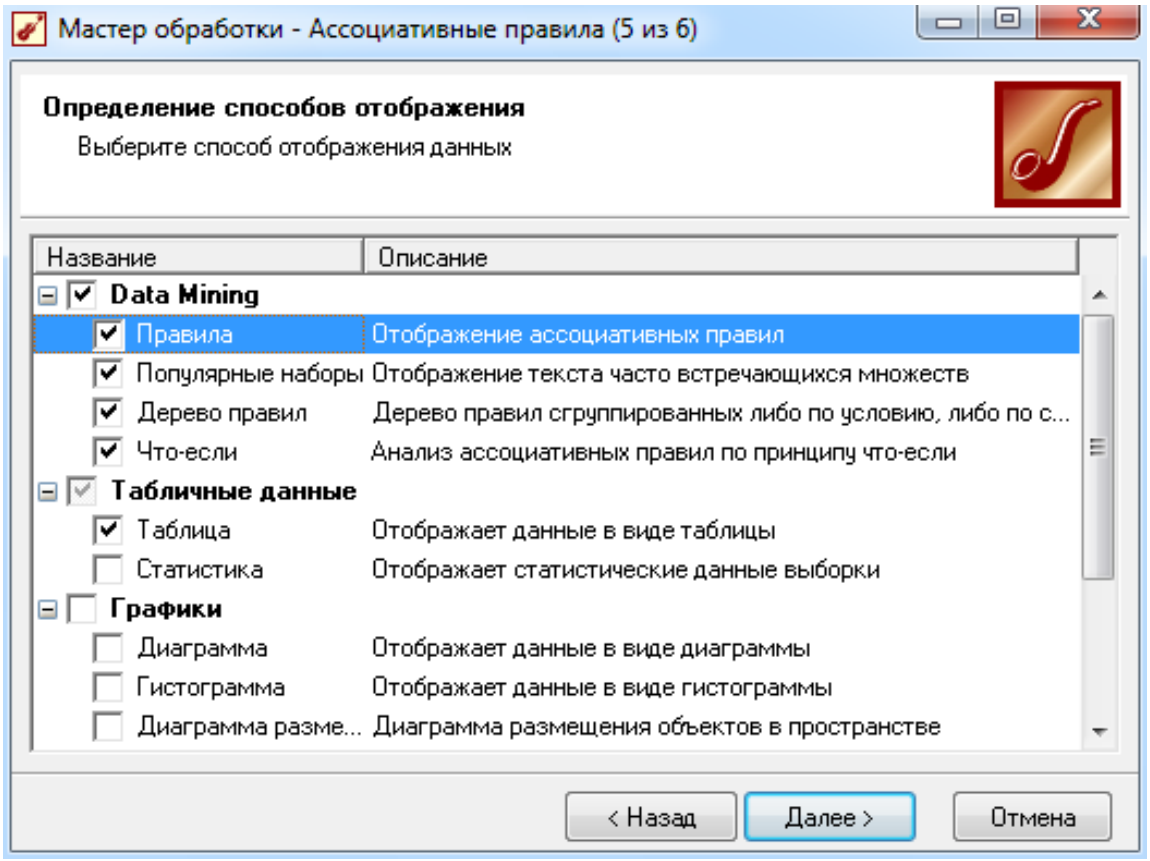


Рисунок 19.4 – Выбор средств визуализации

Множеств: 16 из 91 | Фильтр: Минимальная поддержка = 6,00

№	Номер множества	Элементы	Поддержка		Мощность
			Кол-во	%	
1	1	Велокамера Mountain	292	14,28	1
2	2	Велокамера Road	208	10,17	1
3	3	Велокамера Touring	144	7,04	1
4	4	Велокрыло Mountain	195	9,54	1
5	5	Велосипед Mountain-200	221	10,81	1
6	11	Велосипед Road-750	129	6,31	1
7	16	Втулка Logo Jersey	154	7,53	1
8	17	Держатель фляги Mountain	200	9,78	1
9	68	Держатель фляги Mountain Фляга	167	8,17	2
10	18	Держатель фляги Road	152	7,43	1
11	70	Держатель фляги Road Фляга	131	6,41	2
12	23	Пластыри для велокамеры	288	14,08	1
13	26	Тенниска фирменная	146	7,14	1
14	27	Фляга	392	19,17	1
15	28	Шапочка велосипедная	199	9,73	1
16	29	Шина HL Mountain	133	6,50	1

Рисунок 19.5 – Популярные наборы

На вкладке **Дерево правил** предлагается удобное средство отображения полученных ассоциативных правил; они выводятся в виде дерева, которое может строиться двумя способами: либо по условию, либо по следствию.

В первом случае на верхнем уровне располагаются узлы с условиями, на нижнем - узлы с соответствующими следствиями. Во втором случае порядок ассоциаций будет противоположным: из следствий «вырастают» ветви условий (рис.19.6).

Если выделить мышью любую из ветвей, в правой части окна будет выведен список правил, построенных по этому узлу, и для каждого из них указан уровень поддержки и достоверности.

Если дерево построено «по условию», оно приводится в верхней части списка, состоящего из обнаруженных следствий. Такие правила отвечают на вопрос: какие товары и с какой вероятностью будут куплены при заданном условии.

Если же дерево построено «по следствию», можно получить ответ на другой вопрос: какие товары должны быть куплены предварительно, чтобы ожидать этого

следствия (то есть покупки товара или товарного набора, который мы хотим продать).

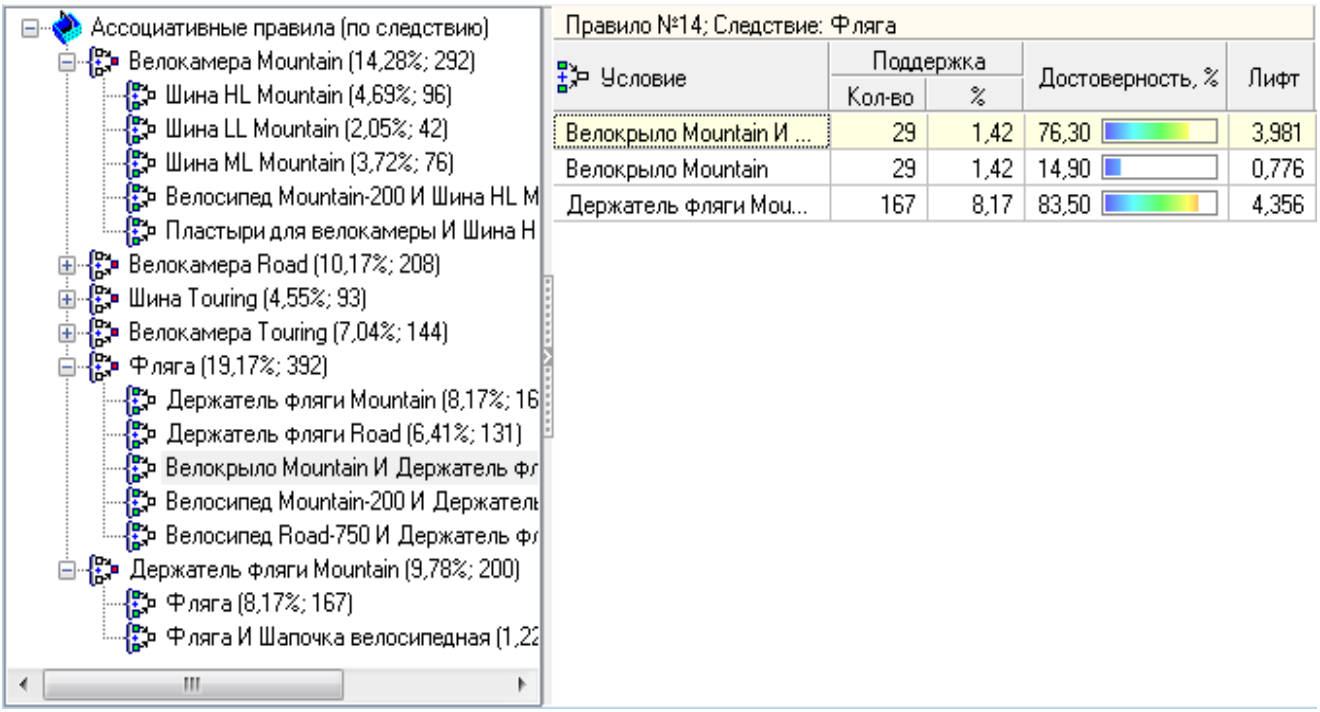


Рисунок 19.6 – Дерево ассоциативных правил

Интерпретация полученных правил

Сами по себе ассоциативные правила, полученные в результате работы некоторого алгоритма, еще непригодны к использованию. Их нужно правильно интерпретировать, то есть понять, какие из них представляют практический интерес, отражают реальные закономерности, а какие носят случайный характер или вообще являются артефактом.

Этот этап требует тщательной аналитической работы и глубокого понимания предметной области, в которой решается задача поиска ассоциаций.

Весь массив ассоциативных правил можно разделить на три группы:

полезные правила, содержащие новую информацию, которая может быть содержательно интерпретирована, имеющие ясную логику. Такие правила могут использоваться на практике для принятия эффективных решений;

тривиальные правила, отражающие действительность, легко объяснимые, но не дающие никакой новой информации (например, при изучении рыночных корзин самую высокую поддержку и достоверность покажут товары - лидеры продаж, что ясно и без всякого анализа). Практическая ценность таких правил близка к нулю;

непонятные правила, содержащие информацию, которую нельзя внятным образом объяснить. Они могут отражать как случайности выборки, так и глубоко скрытые взаимосвязи.

Напрямую их использовать невозможно, поскольку принимаемые на их основе решения, подобно «интуиции» биржевого игрока, не имеют четкого обоснования и могут привести к непредсказуемым последствиям.

В этих случаях по возможности следует провести дополнительный анализ выявленных закономерностей.

Изменяя верхний и нижний пределы поддержки и достоверности (см. рисунок 19.2), можно избавиться от тривиальных и статистически недостоверных закономерностей и увеличить долю полезных правил, генерируемых программой.

Оптимальные значения этих параметров очень сильно зависят от особенностей предметной области, поэтому какие-либо конкретные указания здесь невозможны. Тем не менее, существуют рекомендации общего порядка, которые могут оказаться полезными.

1. При большом значении параметра **Максимальная поддержка** программа будет формировать множество тривиальных правил, не содержащих никакой новой информации и не представляющих практического интереса. Поэтому не рекомендуется устанавливать его на уровне более 20%.

2. Хотя большинство практически ценных правил обнаруживается при невысоком значении порога поддержки, слишком низкий его уровень приводит к генерации статистически недостоверных зависимостей. Поэтому правила, которые кажутся интересными, но имеют низкую поддержку, нужно анализировать дополнительно, с учетом показателя лифта.

3. Как уже отмечалось, следует ограничивать параметр **Мощность часто встречающихся множеств**. Правила, в условие которых включено более 2-3 предметов, обычно очень трудно интерпретировать.

4. Уменьшение порога достоверности приводит к необоснованному увеличению количества правил, поэтому значение этого параметра не должно быть слишком низким. Кроме того, правило с достоверностью порядка 10%, даже если оно отражает реальные взаимосвязи, чаще всего не будет иметь никакого практического значения.

5. Правила с очень большой достоверностью (85-90% и более) также не имеют ценности в контексте решаемой задачи. Товары, входящие в следствие такого правила, покупатель, скорее всего, купит сам, без каких-либо усилий со стороны маркетинговых служб.

Вернемся к рассматриваемой задаче по стимулированию продаж в интернет-магазине (файл данных **cycle_store.txt**. содержит сведения о продажах товаров для велосипедного спорта). При настройках алгоритма, принятых по умолчанию, будет получено 18 правил (рис.19.7); рассмотрим их содержательную интерпретацию.

Например, третье правило *{Велокамера Mountain → Шина HL Mountain}* имеет уровень поддержки S = 4,7%, достоверности C = 72,2% и лифт L = 5,1. Напомним, в чем состоит смысл этих показателей.

1. Если покупатель решил приобрести что-либо в данном магазине, с вероятностью 4,7% это будет набор *Шина HL Mountain + Велокамера Mountain.*

2. Если клиент положил в корзину товар *велокамера Mountain,* то с вероятностью 72,2% он купит и *Шину HL Mountain;*

3. Клиент, купивший *Велокамеру Mountain,* в 5,1 раза чаще выберет *Шину HLMountain,* чем какой-либо другой товар.

Правил: 17 из 17 | Фильтр: Без фильтрации

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	Шина HL Mountain	Велокамера Mountain	96	4,69	72,18	5,055
2	2	Шина LL Mountain	Велокамера Mountain	42	2,05	48,28	3,381
3	3	Шина ML Mountain	Велокамера Mountain	76	3,72	69,72	4,883
4	4	Шина HL Road	Велокамера Road	53	2,59	59,55	5,855
5	5	Шина ML Road	Велокамера Road	53	2,59	58,89	5,790
6	6	Шина Road	Велокамера Road	50	2,44	47,17	4,638
7	7	Велокамера Touring	Шина Touring	79	3,86	54,86	12,064
8	8	Шина Touring	Велокамера Touring	79	3,86	84,95	12,064
9	9	Держатель фляги Mountai	Фляга	167	8,17	83,50	4,356
10	10	Фляга	Держатель фляги Mountai	167	8,17	42,60	4,356
11	11	Держатель фляги Road	Фляга	131	6,41	86,18	4,496
12	12	Велосипед Mountain-200 Шина HL Mountain	Велокамера Mountain	34	1,66	70,83	4,961
13	13	Пластыри для велокамерь Шина HL Mountain	Велокамера Mountain	28	1,37	71,79	5,028
14	14	Велокрыло Mountain Держатель фляги Mountai	Фляга	29	1,42	76,32	3,981
15	15	Велосипед Mountain-200 Держатель фляги Mountai	Фляга	56	2,74	84,85	4,426
16	16	Велосипед Road-750 Держатель фляги Road	Фляга	37	1,81	84,09	4,387
17	17	Фляга Шапочка велосипедная	Держатель фляги Mountai	25	1,22	46,30	4,734

Рисунок 19.7 – Интерпретация правил, полученных при настройках, принятых по умолчанию (Т- тривиальное, ? – непонятное правило)

шины, велокамеры и велосипеды часто встречаются в условиях и следствиях правил, это лидеры продаж магазина (см. табл. 19.3), поэтому и правила с ними имеют высокую достоверность (до 85%);

правила, входящие в группы {*Велокамера → Шина*} и {*Шина → Велокамера*}, тривиальны сами по себе: понятно, что эти запчасти обычно меняют одновременно;

правила типа {*Фляга → Держатель фляги*} тоже тривиальны, так как никому не нужна велосипедная фляга без возможности закрепить ее на раме;

наконец, правила типа {*Велосипед → Фляга*} хотя и тривиальны, но, возможно, имеют ценность; никогда не будет лишним при покупке велосипеда предложить флягу и держатель к ней.

Теперь рассмотрим правило {*Пластыри для велокамеры + Шина HL Mountain → Велокамера Mountain*}. Его условие непонятно: почему пластыри покупаются именно с шинами *Mountain,* ведь есть и другие шины? Возможно, это происходит из-за того, что велокамеры *Mountain* продаются чаще других камер (что, в свою очередь, объясняется популярностью велосипедов *Mountain).* Анализ популярных наборов подтверждает такую гипотезу.

№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт
				Кол-во	%		
1	1	Велокамера Mountain	Пластыри для велокамерь	77	3,77	26,37	1,872
2	2	Пластыри для велокамерь	Велокамера Mountain	77	3,77	26,74	1,872
3	3	Велокамера Mountain	Шина HL Mountain	96	4,69	32,88	5,055
4	4	Велокамера Mountain	Шина ML Mountain	76	3,72	26,03	4,883
5	5	Велокамера Road	Шина HL Road	53	2,59	25,48	5,855
6	6	Велокамера Road	Шина ML Road	53	2,59	25,48	5,790
7	7	Велокрыло Mountain	Велосипед Mountain-200	66	3,23	33,85	3,132
8	8	Велосипед Mountain-200	Велокрыло Mountain	66	3,23	29,86	3,132
9	9	Велосипед Mountain-200	Держатель фляги Mountai	66	3,23	29,86	3,054
10	10	Держатель фляги Mountai	Велосипед Mountain-200	66	3,23	33,00	3,054
11	11	Велосипед Mountain-200	Фляга	56	2,74	25,34	1,322
12	12	Шина HL Mountain	Велосипед Mountain-200	48	2,35	36,09	3,340
13	13	Велосипед Road-250	Держатель фляги Road	25	1,22	28,41	3,822
14	14	Велосипед Road-250	Фляга	23	1,12	26,14	1,363
15	15	Велосипед Road-250	Шина HL Road	28	1,37	31,82	7,311
16	16	Шина HL Road	Велосипед Road-250	28	1,37	31,46	7,311
17	17	Велосипед Road-750	Держатель фляги Road	44	2,15	34,11	4,589

Рисунок 19.8 – Правила, полученные при измененных настройках алгоритма

Обратим внимание на следующее обстоятельство: все правила, приведенные на рисунке 19.7, имеют уровень достоверности более 40%, и даже при достоверности 42-43% получаются тривиальные правила. Вероятно, имеет смысл сделать следующее:

запустить заново алгоритм *apriori* с интервалом допустимой достоверности от 25% до 40%;

не рассматривать правила, в следствиях и условиях которых содержатся велосипеды, шины и велокамеры (очевидные лидеры продаж).

После повторного «прогона» алгоритма получим пять новых правил (рисунок 19.8). Очевидно, их можно считать полезными: они нетривиальны, но вполне объяснимы и имеют достаточно высокий уровень достоверности.

ЗАДАНИЕ

Небольшая сеть из трех магазинов, продающих мелкие штучные товары, желает провести исследование связанных покупок. По мнению специалистов компании, знание того, какие товары покупаются совместно, поможет правильно расположить их на витринах. С этой целью были собраны все чеки за последние три месяца (около 20 тыс.). В них присутствуют 17 товаров (анальгетик, дезодорант, журнал, зубная паста, карандаши и др.). Таблица данных содержится в файле **transactions.txt** и включает два столбца:**Транзакция** и **Товар**.

Опираясь на имеющиеся данные, выполните следующие действия.

1. Решите задачу поиска ассоциаций в Deductor.

2. Выделите непонятные, на ваш взгляд, ассоциативные правила, а также правила, представляющие интерес. Сколько правил попало в эти категории?

3. Найдите правило, имеющее максимальный лифт.

4. Заказчика данного исследования интересует, какие товары покупают с поздравительной открыткой. Сколько таких товаров оказалось в выбранном перечне? Какая из ассоциаций представляет в этом плане наибольший интерес (имеет максимальный лифт)?

Вопросы для самоконтроля

- В чем состоит цель аффинитивного анализа и какие практические задачи решаются с его помощью?
- Дайте определение понятия «транзакция».
- Из каких элементов состоит ассоциативное правило?
- Как рассчитывается поддержка и достоверность ассоциативного правила и какое значение имеют эти показатели в процессе их поиска?
- В чем состоит смысл показателя «лифт»?
- Какие параметры необходимо указать при настройке алгоритма *apriori?*
- На какие группы подразделяются ассоциативные правила при их содержательной интерпретации?
- Перечислите рекомендации, которых следует придерживаться при анализе ассоциаций в программе Deductor и объясните их смысл.

Практическое занятие № 20

«Статистика поисковых запросов»[13]

Цель: ознакомится с возможностями методики получения статистики поисковых запросов при помощи сервисов предоставляемых компаниями Yandex, Google и Rambler, получить навыки анализа полученных статистических данных.

Теоретические сведения

Согласно теории искусственного интеллекта знания подразделяются на структурированные (кодифицированные) и неотделимые. Кодифицированные знания формализованы и достаточно легко могут храниться, копироваться и распространяться.

В свою очередь, неформальные знания неотделимы от человека, не оформлены, накапливаются через личный опыт, обучение в процессе деятельности, социальные взаимоотношения и т.д. Как следствие, они трудно поддаются количественному определению, хранению или передаче. Они выходят далеко за рамки технического прогресса и инноваций, находящих материальное воплощение в продуктах, услугах или процессах.

В свою очередь переносимые знания подразделяются на структурированные, слабоструктурированные и неструктурированные. В первом случае знания имеют определенную последовательность удобных для восприятия форм: схемы, таблицы, а также связи, которые позволяет облегчить их обработку и передачу.

Неструктурированные же знания являются более сложными для обработки. Слабоструктурированные знания имеют свойства и структурированных и неструктурированных.

Неструктурированная форма знания представляется в виде текста или же WEB-контента. Развитию последней способствовало появление глобальных каналов связи, в частности сети Интернет.

Рост объема доступных через Интернет данных, хранимых в слабо структурированном виде, способствовал появлению автоматических программных средств поиска информации и получения данных об использовании определенных ресурсов. Возник целый ряд интеллектуальных систем, основная задача которых состоит в эффективном извлечении знаний из Интернет.

Большинство систем мониторинга сети Интернет предоставляют возможность фильтрации и получения статистической информации о запросах пользователей. Подобные инструменты помогает определять количество обращений к разным файлам и серверам, адресам отдельных ресурсов.

Статистика запросов фактически представляет собой корпус языка, зачастую позволяющий проводить исследования, которые невозможно провести никаким другим способом.

Так, к примеру, подобного рода статистика является наиболее доступным источником современного языка, в отличие от анализа поисковых результатов, результаты которого могут лишь приблизительно говорить о текущем словоупотреб-

[13]Разработали магистранты ФПИ (2013г.): Кириченко Е.В., Сытников Д.А., Петухов А.В.

лении, в силу того, что в интернете сосуществуют тексты самой различной степени давности, в том числе и прошлого, и позапрошлого веков. Кроме того, корпус запросов к поисковой системе считается одним из наиболее репрезентативных источников живого языка

О том, как применять подобные инструменты и какую информацию мы можем получить, будет рассказано в данной лабораторной работе.

Описание и сравнение предоставляемых сервисов.

Прежде чем перейти к системам, позволяющим просмотреть статистику запросов пользователей, необходимо раскрыть понятие «статистика запросов».

Поисковый запрос – это информация, с помощью которой осуществляется поиск поисковой системой, такой как: Yandex, Google, Rambler. Как правило, поисковый запрос задаётся в виде фраз или слов. Бывают также запросы в виде изображений. Формат ключевых запросов зависит от типа информации для поиска и устройства конкретной поисковой системы

Ключевое слово – это слово, которое в совокупности с другими ключевыми словами, представляет текст сайта. Используется ключевое слово для поиска. Содержание текста, представленное ключевыми словами, анализируется лингвистическими и математическими методами. Например, анализ частоты появления слова в тексте.

Статистика запросов — информация об обращениях пользователей к поисковой системе по «ключевым словам».

Другими словами статистика запросов – это количество поисковых запросов пользователей по «ключевым словам» за определенный промежуток времени.

В большинстве случаев при работе с сервисом статистики имеется возможность фильтровать результаты по территориальному признаку, по языку в хронологическом порядке. При этом, обычно, сервис показывает не только данные об искомом запросе, но также и о словосочетаниях, синонимах и близких темах.

Рассмотрим некоторые преимущества и недостатки предоставляемых сервисов каждой из поисковых систем.

Яндекс Wordstat

Яндекс предоставляет доступ к своей статистике всем желающим в рамках системы по продаже рекламы Яндекс.Директ. Кроме стандартной информации о количестве запросов в месяц, а также словосочетаниях и близких темах, поисковик предоставляет возможность фильтровать результаты по регионам, городам, а также по месяцам.

Учитывая тот факт, что Яндекс является самой популярной в Рунете поисковой системой, подобная статистика является наиболее репрезентативной при оценке положения дел в Рунете.

В Яндекс Wordstat статистика запросов представляется в несколько упрощенном виде — объединяются все возможные словоформы (падежи, числа и т.п.), в большинстве случаев не учитываются предлоги, а так же вопросительные формы (например, «что такое» и тому подобное).

Правда, при помощи специальных операторов вы сможете добиться конкретизации статистики Яндекса именно по интересующей вас словоформе поискового запроса. Обычно для этого достаточно бывает заключить нужный поисковый запрос в кавычки. При этом учитываться будут только эти слова запроса, но в любой допустимой словоформ. Кроме того, вместе с кавычками можно будет дополнительно поставить восклицательные знаки перед каждым из слов, обязав тем самым статистику Яндекса учитывать только эти слова и только в выбранной вами словоформе.

Следует отметить, что в статистике поисковых запросов Яндекса приводятся не только производные от введенных вами слов (в левой колонке как раз будут показаны эти самые расширенные варианты запросов с добавлением других слов), но еще дополнительно в правой колонке будут показаны ассоциативные запросы, которые набирали те же самые пользователи в Яндексе вместе с введенными вами словами за одну и ту же сессию поиска.

Rambler статистика

Система статистики имеется и у Рамблера. Она менее репрезентативна в силу меньшей популярности поисковой системы, чем статистика Яндекса, но её преимуществом является более подробная информация.

К примеру, сервис выдает информацию о количестве запросов не только с заглавной страницы, но также и со всех остальных. Кроме того, статистика Рамблера позволяет использовать несложный язык запросов для уточнения или, наоборот, расширения результата.

Статистика Рамблер отличается от статистики запросов в Яндексе тем, что в ней не объединяются результаты для разных словоформ. Т.е. можно без дополнительных операторов получить статистику частотности запроса по словам в нужном падеже и требуемом числе.

Google Trends

Крупнейшая в мире поисковая система Google также предоставляет открытый доступ к своей статистике запросов. В отличие от двух предыдущих, количественная статистика доступна в формате csv.

Визуально статистика представляется лишь относительно – в видеграфика. Отчёты выделяются особой подробностью: например, кроме обычной статистики запросов пользователей, можно посмотреть степень конкуренции рекламодателей за конкретный поисковый запрос, просмотреть историю трафика для выбранных ключевых слов; предоставляется подсказка возможно полезных минус-слов.

В особом виде статистику отображают графики Google Trends. Сервис позволяет вводить до 5 разных запросов, изучать и сравнивать изменение интереса к ним в мире в виде графика за прошедшие 2-3 года.

Работа со статистикой поисковых запросов на примере запроса «ВТО»

Для проведения исследования поисковых запросов нам необходимо составить список интересующих нас запросов – семантический словарь.

Для того что бы в различных сервисах мы получали наиболее актуальную информацию конкретизируем условия нашего поиска. Поиску будем производить по всем регионам России (Центр, Северо-Запад, Поволжье, Юг, Сибирь, Дальний Восток, Северный Кавказ, Урал), период с 01 марта 2011 по 01 марта 2013 года.

В таблице 20.1 представлен используемый нами семантический словарь.

Таблица 20.1 – Используемый сематический словарь

Слово:
ВТО
Вступление +в ВТО
ВТО плюсы +минусы
ВТО +влияние
ВТО +пошлины

Далее переходим по ссылке **http://wordstat.yandex.ru/** для доступа к сервису Яндекс статистика.

Вводим в п. 1 (рис. 20.1) ключевое слово.

Выбираем регион, запросы которого будут анализироваться п. 2

Заполнить код зашиты автоматической регистрации п. 3 и нажать на кнопку «Подобрать».

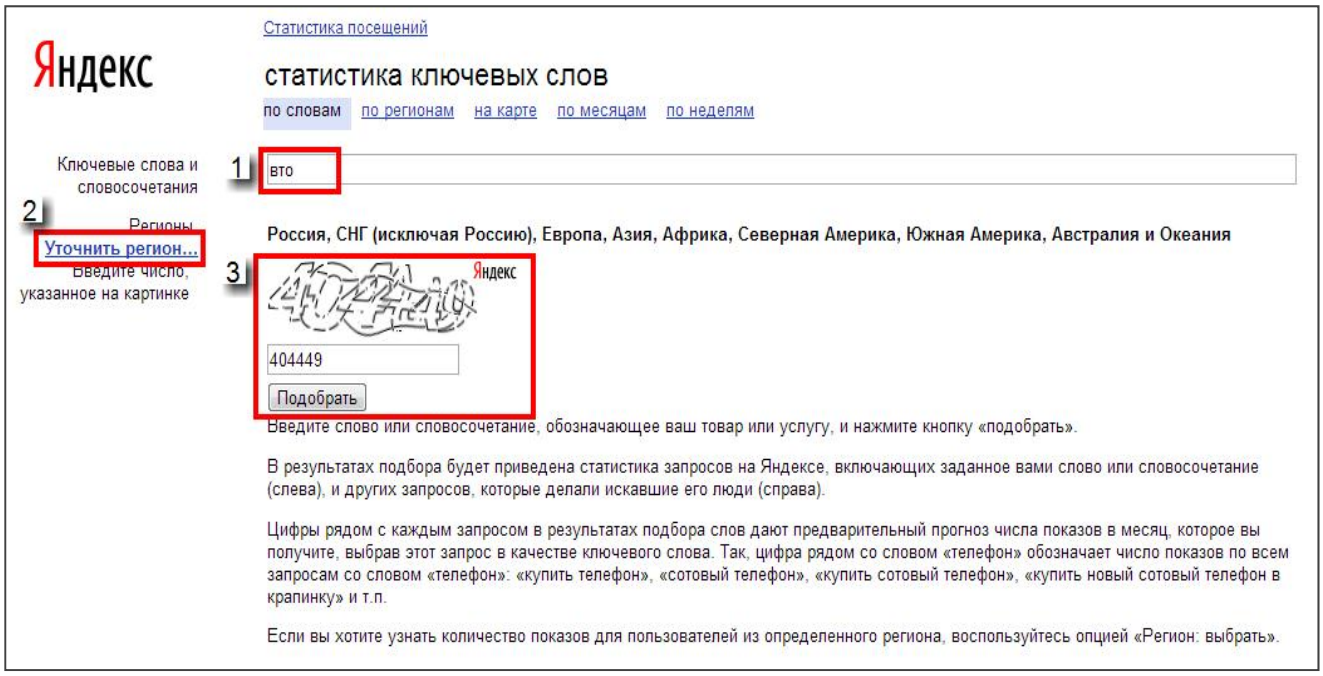


Рисунок 20.1 – Стартовая страница сервиса от Яндекс

Получаем таблицу с данными, в которой отображено количество пользователей, которые совершали указанный нами запрос в поисковой системе за месяц. Результаты указаны на рисунке 20.2. Кроме того, в таблице отражены словосочетания, которые пользователи искали вместе с интересующим нас запросом а также запросы, которые пользователи искали сразу же после интересующего нас запроса.

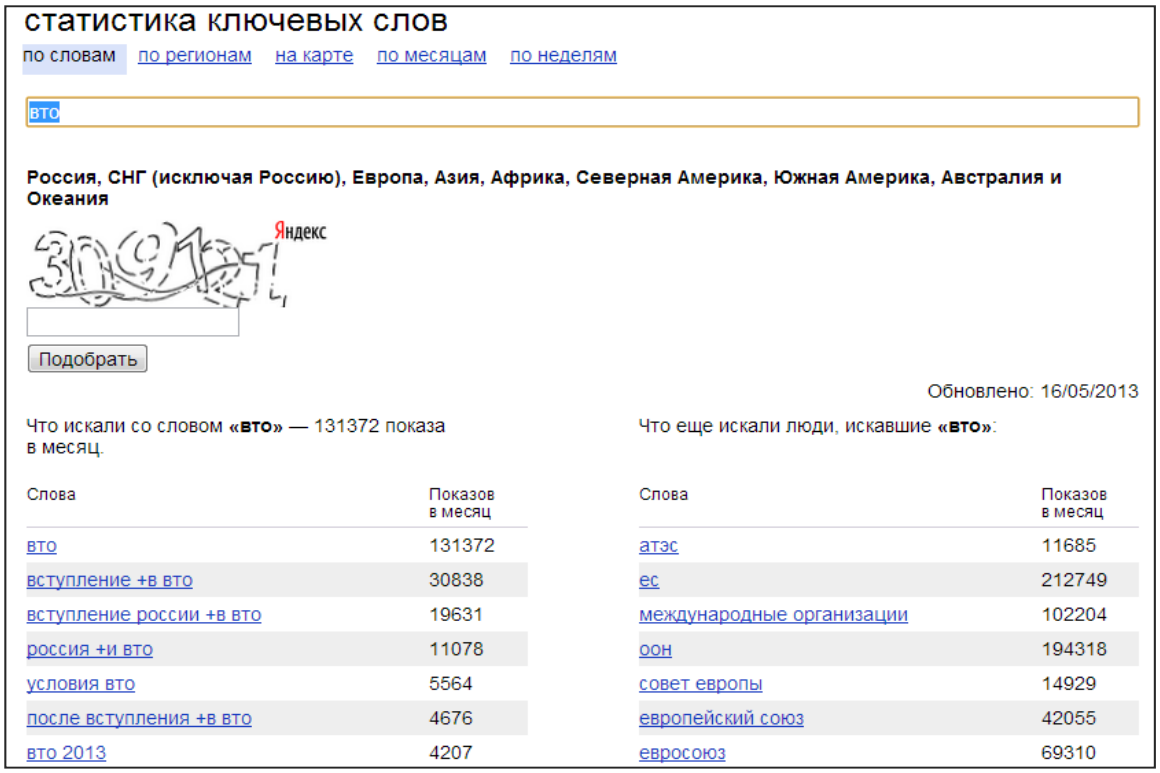


Рисунок 20.2 – Страница статистики ключевых слов в режиме отоброжения по словам

Для получения графика необходимо перейти в режим отображения информации «по месяцам» или «по неделе», при необходимости задать период анализа данных (рис. 20.3).

Рисунок 20.3 – Страница статистики ключевых слов в режиме отображения по месяцам

Запомним периоды пиковых значений.

Далее по очереди вбиваем запросы из составленного нами ранее семантического словаря и также запоминаем периоды пиковых значений.

По полученным данным можно отметить, что поведение всех графиков практически одинаково и в определенные промежутки времени, наблюдается как возрастание интересов, так и затухание. А именно в период с сентября 2011 года по всем запросам активность пользователей возросла и достигла своего максимума в марте 2011 года. Далее по всем запросам пошел спад и пользователи с практически постоянной частотой интересовались исследуемыми запросами. Следующий скачек активности россиян приходится на июль – август 2012 года.

Далее проведем анализ семантического словаря с помощью сервиса от компании Google.

Для этого перейдем по ссылке **http://www.google.ru/trends/**.

Примечание: для использования сервиса Google Тренды вам необходимо зарегистрироваться в сервисе Google +.

В появившемся окне необходимо ввести интересующий нас запрос и нажать кнопку «Показать».

Заполним параметры построения графика как на рисунке 20.4.

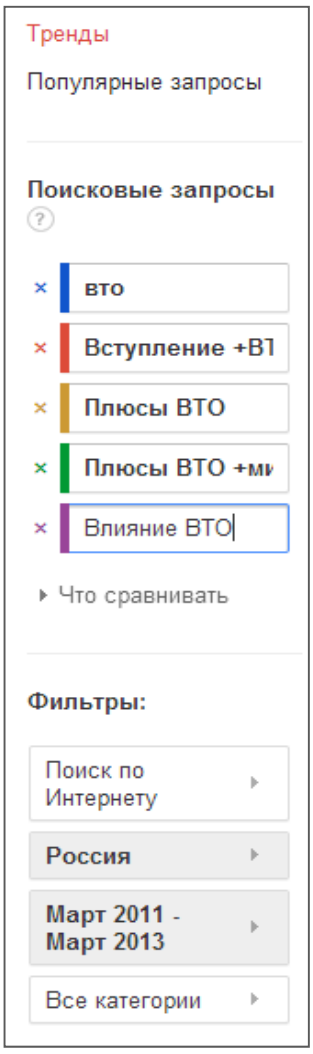


Рисунок 20.4 – Настройка параметров построения графика

В полученном графике, указанном на рисунке 20.5 запоминаем периоды дат пиковых значений.

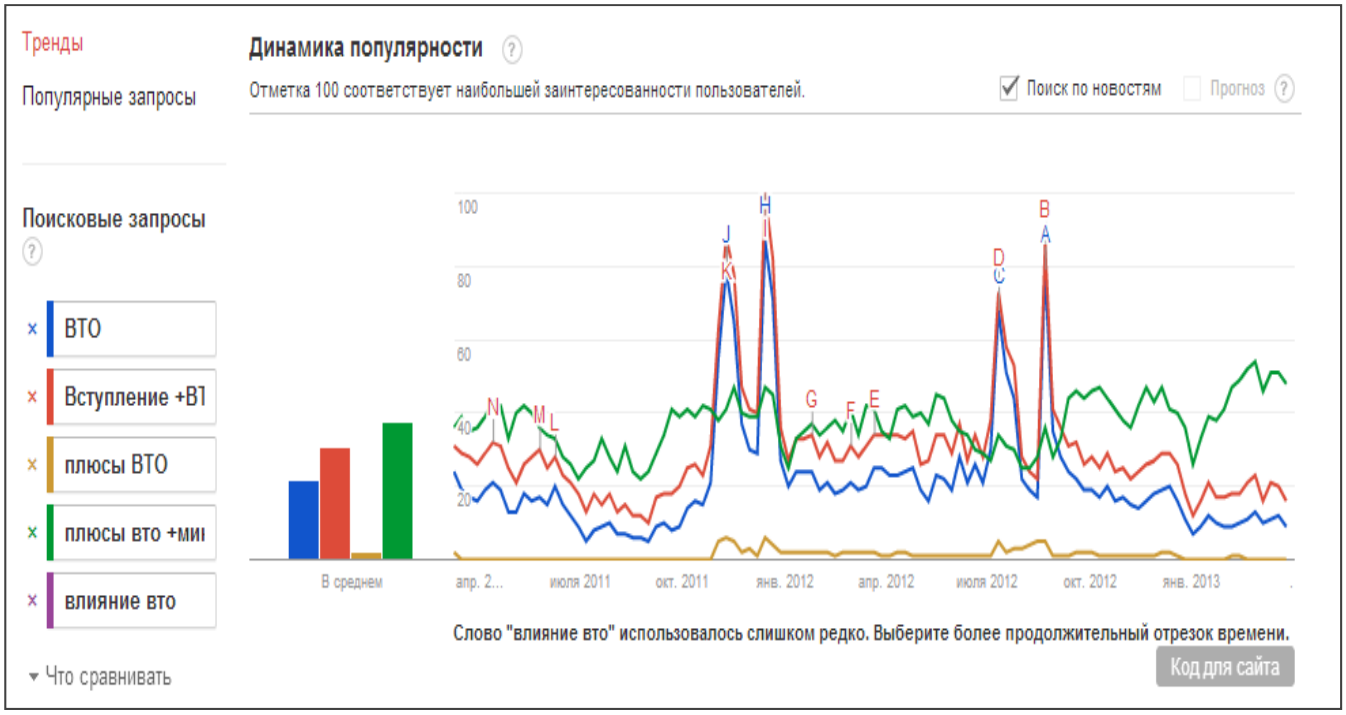


Рисунок 20.5 – Графическое представление количества поисковых запросов в определенный период времени по данным Google Тренды

Замечание: сервис от компании Google представляет данные в относительном виде. Числа на графике показывают долю запросов по ключевым словам в общем числе запросов, выполненных в Google за определенное время. Они являются не абсолютным выражением объема поисковых запросов, а относительным, в масштабе от 0 до 100. Каждая точка на графике соотносится с максимальным значением, т. е. 100.

Далее проведем анализ семантического словаря с помощью сервиса от компании Рамблер.

Для этого перейдем по ссылке **http://adstat.rambler.ru/wrds/**.

Заполним информацию на рисунке 20.6.

Далее поочередно выбираем даты начиная с наименьшей. И нажимаем на кнопку **«посчитать»**.

Внизу появится таблица с количеством запросов пользователей за указанный месяц.

Запоминайте полученные данные и отметьте периоды наибольшей активности пользователей.

Примечание: Рамблер ввел ограничение на количество запросов незарегистрированных пользователей. Для полноценного использования предоставляемого сервиса необходимо зарегистрировать учетную запись на сайте поисковой системы Рамблер.

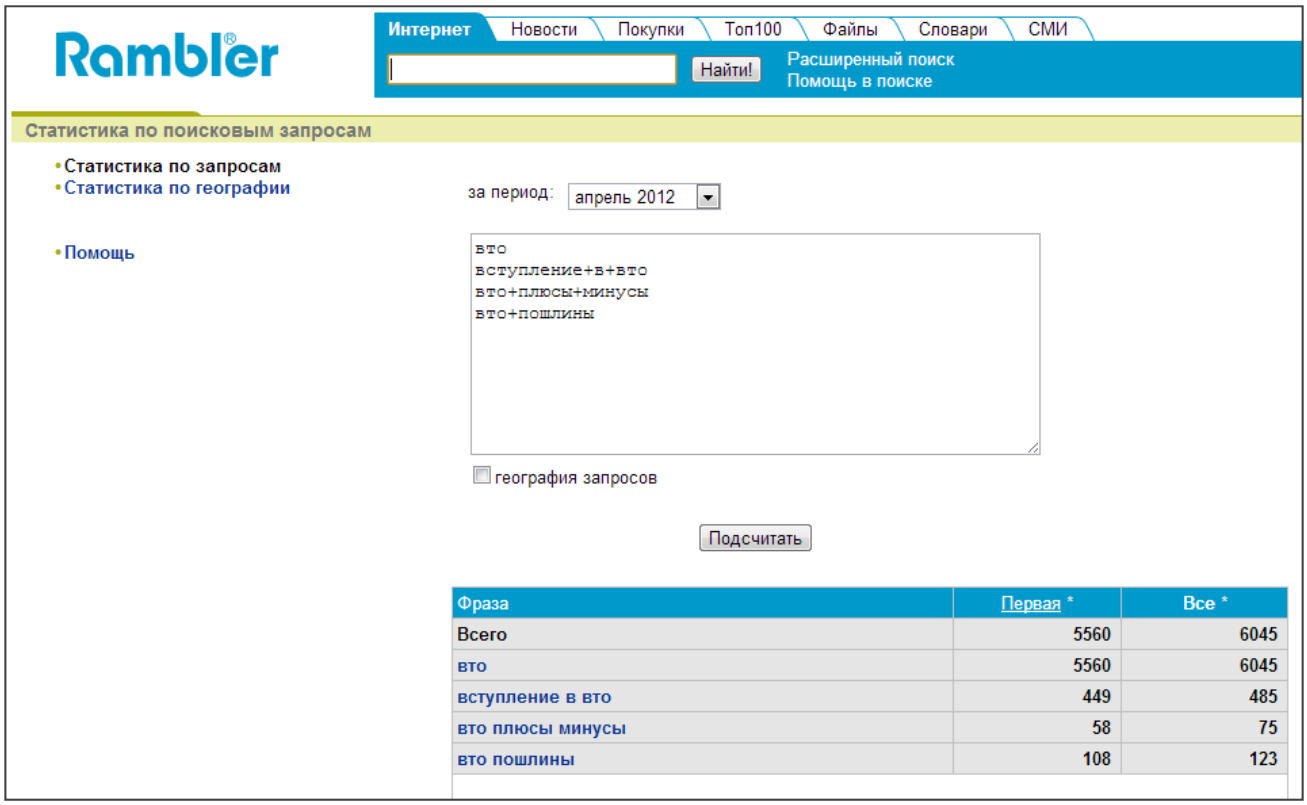


Рисунок 20.6 – Страница статистики ключевых слов сервиса «Рамблер статистика»

По полученным данным, так же как и в других системах, можно наблюдать одновременный рост и спад активности пользователей по всем запросам.

Проанализируем результаты, полученные с помощью всех рассматриваемых сервисов, и наложим их на сетку событий связанных со вступлением России в ВТО.

Для начала отметим наиболее важные события.

1. В течение осени 2011 г. были согласованы остававшиеся вопросы на переговорах с США.
2. Одновременно в течение осени 2011 г. в результате многомесячных неформальных российско-грузинских консультаций при посредничестве Швейцарии удалось выработать приемлемое для обеих сторон решение по контролю за передвижением гражданских грузов по территории Абхазии и Южной Осетии. В результате со стороны Грузии были сняты возражения по созыву формального заседания Рабочей группы.
3. 10 ноября 2011 г. переговоры о присоединении России к ВТО были завершены. Рабочая группа одобрила пакет документов о присоединении РФ к ВТО для внесения на рассмотрение Восьмой министерской конференции ВТО. Таким образом мандат Рабочей группы по присоединению России к ВТО был исчерпан , после чего она была распущена.
4. 16 декабря 2011 г. в ходе на 8-й Министерской конференции стран-членов ВТО в Женеве был одобрен пакет документов по присоединению России к ВТО. Пакет включал в себя: протокол о присоединении России к ВТО, содержащий Перечень тарифных уступок и перечень специфических обязательств по услугам; доклад Рабочей группы по присоединению РФ к ВТО.

5. В соответствии с правилами ВТО, России был предоставлен срок в 220 дней для ратификации пакета документов о присоединении к ВТО национальным парламентом.
6. 10 июля 2012 г. Государственная дума РФ 238 голосами против 208 и 1 воздержавшемся одобрила Протокол о присоединении России ко Всемирной торговой организации.
7. 18 июля 2012г. Совет Федерации РФ ратифицировал Протокол о присоединении России ко Всемирной торговой организации.
8. 21 июля 2012 г. Президент России В. Путин подписал федеральный закон "О ратификации Протокола о присоединении РФ к Марракешскому соглашению об учреждении Всемирной торговой организации от 15 апреля 1994 г."
9. 22 августа 2012 г. Российская Федерация официально стала 156-м членом Всемирной торговой организации.

Таким образом, мы четко прослеживаем рост активности пользователей в наиболее важные периоды государственной деятельности, связанной со вступлением России в ВТО.

Можно сделать вывод, что рассмотренные сервисы позволяют четко отразить интерес россиян по поводу вступления России во Всемирную Торговую Организацию. Следует отметить, что только совместное использование нескольких сервисов, позволяет отразить наиболее четкую картину.

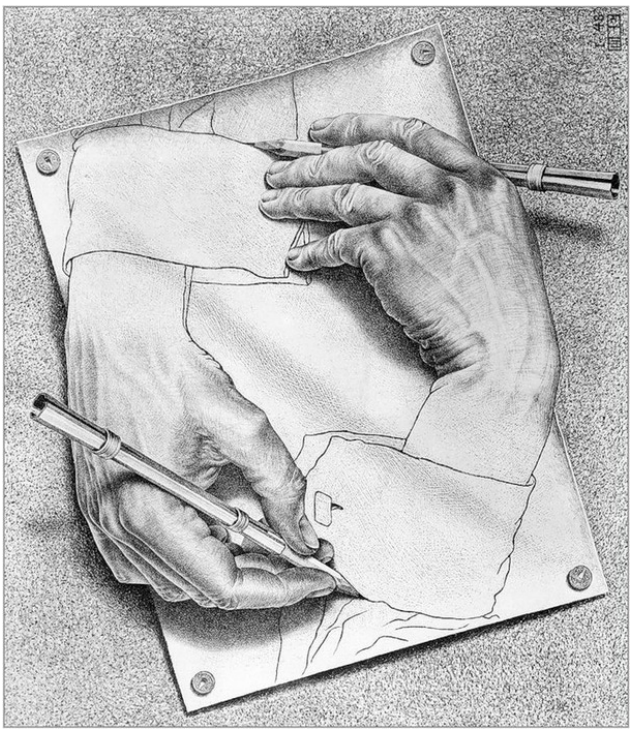
ЗАДАНИЯ

Проанализируйте и сопоставьте полученные данные с историческими событиями нижеуказанные запросы:

1. Олимпиада 2014.
2. Фукусима.
3. Чемпионат Европы 2012.
4. Владимир Путин.
5. Конец света.
6. Борис Березовский.
7. Уго Чавес.
8. Сирия.
9. Северная Корея.

Вопросы для самоконтроля

- Классификафия знаний в теории искуственного интелекта.
- Неструктурированная форма представления знаний.
- Понятия «статистика запросов».
- Преимущества и недостатки поисковых систем *Yandex, Google, Rambler.*



Часть V. МОДЕЛИ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

Практическое занятие № 21

Статистика нечисловых данных в экспертных оценках

Цель: ознакомится с практическим приложением статистики нечисловых данных в теории и практике экспеертных оценок.

Теоретические сведения.

Современная теория измерений и экспертные оценки. Какпроводить анализ собранных рабочей группой ответов экспертов? Для более углубленного рассмотрения проблем экспертных оценок понадобятся некоторые понятия *репрезентативной теории измерений*, служащей основой теории экспертных оценок, прежде всего той ее части, которая связана с анализом заключений экспертов, выраженных в качественном (а не в количественном) виде.

Как уже отмечалось, получаемые от экспертов мнения часто выражены в *порядковой шкале*. Другими словами, эксперт может сказать (и обосновать), что один тип продукции будет более привлекателен для потребителей, чем другой, один показатель качества продукции более важен, чем другой, первый технологический объект более опасен, чем второй, и т.д. Но эксперт не в состоянии обосновать, *во сколько раз* или *на сколько* более важен, соответственно, более опасен. Поэтому экспертов часто просят дать ранжировку (упорядочение) объектов экспертизы, т.е. расположить их в порядке возрастания (или, точнее, неубывания) интенсивности интересующей организаторов экспертизы характеристики.

Рассмотрим в качестве примера применения результатов теории измерений, связанных со средними величинами в порядковой шкале, один сюжет, связанный с ранжировками и рейтингами.

Методы средних баллов. В настоящее время распространены экспертные, маркетинговые, квалиметрические, социологические и иные опросы, в которых опрашиваемых просят выставить баллы объектам, изделиям, технологическим процессам, предприятиям, проектам, заявкам на выполнение научно-исследовательских работ, идеям, проблемам, программам, политикам и т.п. Затем рассчитывают средние баллы и рассматривают их *как интегральные (т.е. обобщенные, итоговые) оценки,* выставленные коллективом опрошенных экспертов. Какими формулами пользоваться для вычисления средних величин? Ведь существует очень много разных видов средних величин.

По традиции обычно применяют *среднее арифметическое*. Однако специалисты по теории измерений уже около 30 лет знают, что *такой способ некорректен*, поскольку баллы обычно измерены в *порядковой* шкале [38, глава 2.1.3]. Обоснованным является использование медиан в качестве средних баллов. Однако полностью игнорировать средние арифметические нецелесообразно из-за их привычности и распространенности. Поэтому *представляется рациональным использовать одновременно оба метода - и метод средних арифметических рангов (баллов), и методов медианных рангов. Т*акая рекомендация находится в согласии с общенаучной *концепцией устойчивости* [38, глава 1.4.7], рекомендующей применять различные методы для обработки одних и тех же данных с целью выделить выводы, получаемые одновременно при всех методах. Такие выводы, видимо, соответствуют реаль-

ной действительности, в то время как заключения, меняющиеся от метода к методу, зависят от субъективизма исследователя, выбирающего метод обработки исходных экспертных оценок.

Пример сравнения восьми проектов. Рассмотрим конкретный пример применения только что сформулированного подхода.

По заданию руководства фирмы анализировались восемь проектов, предлагаемых для включения в план стратегического развития фирмы. Они обозначены следующим образом: Д, Л, М-К, Б, Г-Б, Сол, Стеф, К (по фамилиям менеджеров, предложивших их для рассмотрения). Все проекты были направлены 12 экспертам, включенным в экспертную комиссию, организованную по решению Правления фирмы. В табл. 21.1 приведены ранги восьми проектов, присвоенные им каждым из 12 экспертов в соответствии с представлением экспертов о целесообразности включения проекта в стратегический план фирмы. При этом эксперт присваивает ранг 1 самому лучшему проекту, который обязательно надо реализовать. Ранг 2 получает от эксперта второй по привлекательности проект, ... , наконец, ранг 8 - наиболее сомнительный проект, который реализовывать стоит лишь в последнюю очередь.

Таблица 21.1 - Ранги 8 проектов по степени привлекательности для включения в план стратегического развития фирмы

№ эксперта	Д	Л	М-К	Б	Г-Б	Сол	Стеф	К
1	5	3	1	2	8	4	6	7
2	5	4	3	1	8	2	6	7
3	1	7	5	4	8	2	3	6
4	6	4	2,5	2,5	8	1	7	5
5	8	2	4	6	3	5	1	7
6	5	6	4	3	2	1	7	8
7	6	1	2	3	5	4	8	7
8	5	1	3	2	7	4	6	8
9	6	1	3	2	5	4	7	8
10	5	3	2	1	8	4	6	7
11	7	1	3	2	6	4	5	8
12	1	6	5	3	8	4	2	7

Примечание. Эксперт № 4 считает, что проекты М-К и Б равноценны, но уступают лишь одному проекту - проекту Сол. Поэтому проекты М-К и Б должны были бы стоять на втором и третьем местах и получить баллы 2 и 3. Поскольку они равноценны, то получают средний балл (2+3)/ 2 = 5/ 2 = 2,5.

Анализируя результаты работы экспертов (т.е. упомянутую таблицу), члены аналитической подразделения Рабочей группы, анализировавшие ответы экспертов по заданию Правления фирмы, были вынуждены констатировать, что полного согласия между экспертами нет, а потому данные, приведенные в таблице, следует подвергнуть тщательному математическому анализу.

Метод средних арифметических рангов. Сначала для получения группового мнения экспертов был применен метод средних арифметических рангов. Прежде всего была подсчитана сумма рангов, присвоенных проектам (см. табл. 21.1). Затем эта сумма была разделена на число экспертов, в результате рассчитан средний арифметический ранг (именно эта операция дала название методу). По средним рангам строится итоговая ранжировка (в другой терминологии - упорядочение), исходя из принципа - чем меньше средний ранг, чем лучше проект. Наименьший средний ранг, равный 2,625, у проекта Б, - следовательно, в итоговой ранжировке он получает ранг 1. Следующая по величине сумма, равная 3,125, у проекта М-К, - и он получает итоговый ранг 2. Проекты Л и Сол имеют одинаковые суммы (равные 3,25), значит, с точки зрения экспертов они равноценны (при рассматриваемом способе сведения вместе мнений экспертов), а потому они должны бы стоять на 3 и 4 местах и получают средний балл (3+4)/2 = 3,5. Дальнейшие результаты приведены в табл. 21.2.

Итак, ранжировка по суммам рангов (или, что в данном случае то же самое, по средним арифметическим рангам) имеет вид:

$$Б < М\text{-}К < \{Л, Сол\} < Д < Стеф < Г\text{-}Б < К \ . \ (1)$$

Здесь запись типа "А<Б" означает, что проект А предшествует проекту Б (т.е. проект А лучше проекта Б). Поскольку проекты Л и Сол получили одинаковую сумму баллов, то по рассматриваемому методу они эквивалентны, а потому объединены в группу (в фигурных скобках). В терминологии математической статистики ранжировка (1) имеет одну связь.

Таблица 21.2 - Результаты расчетов по методу средних арифметическихи методу медиан для данных, приведенных в таблице 21.1.

	Д	Л	М-К	Б	Г-Б	Сол	Стеф	К
Сумма рангов	60	39	37,5	31.5	76	39	64	85
Среднее арифметическое рангов	5	3,25	3,125	2,625	6,333	3,25	5,333	7,083
Итоговый ранг по среднему арифметическому	5	3,5	2	1	7	3,5	6	8
Медианы рангов	5	3	3	2,25	7,5	4	6	7
Итоговый ранг по медианам	5	2,5	2,5	1	8	4	6	7

Метод медиан рангов. Значит, наука сказала свое слово, итог расчетов - ранжировка (1), и на ее основе предстоит принимать решение? Так был поставлен вопрос при обсуждении полученных результатов на заседании Правления фирмы. Но тут наиболее знакомый с современной эконометрикой член Правления вспомнил, что ответы экспертов измерены в порядковой шкале, а потому для них неправомерно проводить усреднение методом средних арифметических. Надо использовать метод медиан.

Что это значит? Надо взять ответы экспертов, соответствующие одному из проектов, например, проекту Д. Это ранги 5, 5, 1, 6, 8, 5, 6, 5, 6, 5, 7, 1. Затем их надо расположить в порядке неубывания (проще было бы сказать – «в порядке возрастания», но поскольку некоторые ответы совпадают, то приходится использовать непривычный термин «неубывание»). Получим последовательность: 1, 1, 5, 5, 5, 5, 5, 6, 6, 6, 7, 8. На центральных местах - шестом и седьмом - стоят 5 и 5. Следовательно, медиана равна 5.

Медианы совокупностей из 12 рангов, соответствующих определенным проектам, приведены в предпоследней строке табл. 21.2. (При этом медианы вычислены по обычным правилам статистики - как среднее арифметическое центральных членов вариационного ряда.) Итоговое упорядочение комиссии экспертов по методу медиан приведено в последней строке таблицы. Ранжировка (т.е. упорядочение - итоговое мнение комиссии экспертов) по медианам имеет вид:

$$Б < \{М\text{-}К, Л\} < Сол < Д < Стеф < К < Г\text{-}Б \ .$$

Поскольку проекты Л и М-К имеют одинаковые медианы баллов, то по рассматриваемому методу ранжирования они эквивалентны, а потому объединены в группу (кластер), т.е. с точки зрения математической статистики ранжировка (4) имеет одну связь.

Сравнение ранжировок по методу средних арифметических и методу медиан. Сравнение ранжировок (1) и (2) показывает их близость (похожесть). Можно принять, что проекты М-К, Л, Сол упорядочены как М-К < Л < Сол, но из-за погрешностей экспертных оценок в одном методе признаны равноценными проекты Л и Сол (ранжировка (1)), а в другом - проекты М-К и Л (ранжировка (2)). Существенным является только расхождение, касающееся упорядочения проектов К и Г-Б: в ранжировке (3) Г-Б < К, а в ранжировке (4), наоборот, К < Г-Б. Однако эти проекты - наименее привлекательные из восьми рассматриваемых, и при выборе наиболее привлекательных проектов для дальнейшего обсуждения и использования на указанное расхождение можно не обращать внимания.

Рассмотренный пример демонстрирует сходство и различие ранжировок, полученных по методу средних арифметических рангов и по методу медиан, а также пользу от их совместного применения.

Метод согласования кластеризованных ранжировок. Проблема состоит в выделении общего нестрогого порядка из набора кластеризованных ранжировок (в другой терминологии - ранжировок со связями). Этот набор может отражать мнения нескольких экспертов или быть получен при обработке мнений экспертов различными методами. Рассмотрим *метод согласования кластеризованных ранжировок, позволяющий «загнать» противоречия внутрь специальным образом построенных кластеров (групп), в то время как упорядочение кластеров соответствует одновременно всем исходным упорядочениям.*

В различных прикладных областях возникает необходимость анализа нескольких кластеризованных ранжировок объектов. К таким областям относятся прежде всего инженерный бизнес, менеджмент, экономика, социология, экология, прогнозирование, научные и технические исследования и т.д., особенно те их разделы, что связаны с экспертными оценками (см., например, [8, 35]). В качестве объектов могут выступать образцы продукции, технологии, математические модели, проекты,

кандидаты на должность и др. Кластеризованные ранжировки могут быть получены как с помощью экспертов, так и объективным путем, например, при сопоставлении математических моделей с экспериментальными данными с помощью того или иного критерия качества. Описанный ниже метод был разработан в связи с проблемами химической безопасности биосферы и экологического страхования [35].

В настоящем пункте рассматривается метод построения кластеризованной ранжировки, согласованной (в раскрытом ниже смысле) со всеми рассматриваемыми кластеризованными ранжировками. При этом противоречия между отдельными исходными ранжировками оказываются заключенными внутри кластеров согласованной ранжировки. В результате упорядоченность кластеров отражает общее мнение экспертов, точнее, то общее, что содержится в исходных ранжировках**.**

В кластеры заключены объекты, по поводу которых некоторые из исходных ранжировок *противоречат* друг другу. Для их упорядочения необходимо провести новые исследования. Эти исследования могут быть как формально-математическими (например, вычисление медианы Кемени, упорядочения по средним рангам или по медианам и т.п.), так и требовать привлечения новой информации из соответствующей прикладной области, возможно, проведения дополнительных научных или прикладных работ.

Введем необходимые понятия, затем сформулируем алгоритм согласования кластеризованных ранжировок в общем виде и рассмотрим его свойства.

Пусть имеется конечное число объектов, которые мы для простоты изложения будем изображать натуральными числами 1,2,3,...,*k* и называть их совокупность «носителем». *Под кластеризованной ранжировкой, определенной на заданном носителе, понимаем следующую математическую конструкцию*. Пусть объекты разбиты на группы, которые будем называть кластерами. В кластере может быть и один элемент. Входящие в один кластер объекты будем заключать в фигурные скобки. Например, объекты 1,2,3,...,10 могут быть разбиты на 7 кластеров: {1}, {2,3}, {4}, {5,6,7}, {8}, {9}, {10}. В этом разбиении один кластер {5,6,7} содержит три элемента, другой - {2,3} - два, остальные пять - по одному элементу. Кластеры не имеют общих элементов, а объединение их (как множеств) есть все рассматриваемое множество объектов (весь носитель).

Вторая составляющая кластеризованной ранжировки - это строгий линейный порядок между кластерами. Задано, какой из них первый, какой второй, и т.д. Будем изображать упорядоченность с помощью знака < . При этом кластеры, состоящие из одного элемента, будем для простоты изображать без фигурных скобок. Тогда кластеризованную ранжировку на основе введенных выше кластеров можно изобразить так:

$$A = [1 < \{2,3\} < 4 < \{5,6,7\} < 8 < 9 < 10\ ] .$$

Конкретные кластеризованные ранжировки будем заключать в квадратные скобки. Если для простоты речи термин "кластер" применять только к кластеру не менее чем из 2-х элементов, то можно сказать, что в кластеризованную ранжировку *A* входят два кластера {2,3} и {5,6,7} и 5 отдельных элементов.

Введенная описанным образом кластеризованная ранжировка является бинарным отношением на носителе – множестве {1,2,3,...,10}. Его структура такова. Задано

отношение эквивалентности с 7-ю классами эквивалентности, а именно, {2,3}, {5,6,7}, а 5 классов остальные состоят из оставшихся 5 отдельных элементов. Затем введен строгий линейный порядок между классами эквивалентности.

Введенный математический объект известен в литературе как "ранжировка со связями" (М. Холлендер, Д. Вулф), "упорядочение" (Дж. Кемени, Дж. Снелл [10]), "квазисерия" (Б.Г. Миркин), "совершенный квазипорядок" (Ю.А. Шрейдер [36, с.127, 130]). Учитывая разнобой в терминологии, было признано полезным ввести собственный термин *"кластеризованная ранжировка",* поскольку в нем явным образом названы основные элементы изучаемого математического объекта - кластеры, рассматриваемые на этапе согласования ранжировок как классы эквивалентности, и ранжировка - строгий совершенный порядок между ними (в терминологии Ю.А.Шрейдера [36, гл.IV]).

Следующее важное понятие - *противоречивость*. Оно определяется для четверки - две кластеризованные ранжировки на одном и том же носителе и два различных объекта - элементы того же носителя. При этом два элемента из одного кластера будем связывать символом равенства = , как эквивалентные.

Пусть A и B - две кластеризованные ранжировки. *Пару объектов (a,b) назовем «противоречивой»относительно кластеризованных ранжировокA и B, если эти два элемента по-разному упорядочены в A и B, т.е. $a < b$ в A и $a > b$ в B (первый вариант противоречивости) либо $a > b$ в A и $a < b$ в B (второй вариант противоречивости).* Отметим, что в соответствии с этим определением пара объектов (a ,b), эквивалентная хотя бы в одной кластеризованной ранжировке, не может быть противоречивой: эквивалентность $a = b$ не образует "противоречия" ни с $a<b$, ни с $a>b$. Это свойство оказывается полезным при выделении противоречивых пар.

В качестве примера рассмотрим, кроме A, еще две кластеризованные ранжировки

$$B = [\{1,2\} < \{ 3,4, 5\} < 6 < 7 < 9 < \{8, 10\}],$$
$$C = [3 < \{1, 4\} < 2 < 6 < \{5, 7, 8\} < \{9, 10\}].$$

Совокупность противоречивых пар объектов для двух кластеризованных ранжировок A и B назовем «ядром противоречий»и обозначим S(A,B). Для рассмотренных выше в качестве примеров трех кластеризованных ранжировок A, B и C, определенных на одном и том же носителе {1, 2, 3,..., 10}, имеем

$$S(A,B) = [(8, 9)],\ S(A,C) = [(1, 3), (2,4)],$$
$$S(B,C) = [(1, 3), (2, 3), (2, 4), (5, 6), (8,9)].$$

Как при ручном, так и при программном нахождении ядра можно в поисках противоречивых пар просматривать пары (1,2), (1,3), (1,4), ... , (1,k), затем (2,3), (2,4), ..., (2,k), потом (3,4), ..., (3, k), и т.д., вплоть до последней пары (k-1, k).

Пользуясь понятиями дискретной математики, «ядро противоречий» можно изобразить *графом* с вершинами в точках носителя. При этом *противоречивые пары задают ребра этого графа.* Граф для $S(A,B)$ имеет только одно ребро (одна связная компонента более чем из одной точки). Граф для $S(A,C)$ - 2 ребра (две связные компоненты более чем из одной точки). Граф для $S(B,C)$ - 5 ребер (три связные компоненты более чем из одной точки, а именно, {1, 2 , 3, 4}, {5, 6} и {8, 9}).

Каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать матрицей $||x(a,b)||$ из 0 и 1 порядка k x k. При этом $x(a,b) = 1$ тогда и

только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b,a) = 0$, а во втором $x(b,a) = 1$. При этом хотя бы одно из чисел $x(a,b)$ и $x(b,a)$ равно 1. Из определения противоречивости пары (a, b) вытекает, что для нахождения всех таких пар достаточно поэлементно перемножить две матрицы $||x(a,b)||$ и $||y(a,b)||$, соответствующие двум кластеризованным ранжировкам, и отобрать те и только те пары, для которых $x(a,b)y(a,b) = x(b,a)y(b,a)=0$.

Алгоритм согласования некоторого числа (двух или более) кластеризованных ранжировок состоят из трех этапов. На первом *выделяются противоречивые пары* объектов во всех парах кластеризованных ранжировок. На втором формируются кластеры итоговой кластеризованной ранжировки (т.е. классы эквивалентности - *связные компоненты графов*, соответствующих объединению попарных ядер противоречий). На третьем этапе эти *кластеры* (*классы эквивалентности) упорядочиваются*. Для установления порядка между кластерами произвольно выбирается один объект из первого кластера и второй - из второго, порядок между кластерами устанавливается такой же, какой имеет быть между выбранными объектами в любой из рассматриваемых кластеризованных ранжировок. (Если в одной из исходных кластеризованных ранжировок имеет быть равенство, а в другой – неравенство, то при построении итоговой кластеризованной ранжировки используется неравенство.)

Корректность подобного упорядочивания, т.е. его независимость от выбора той или иной пары объектов, вытекает из соответствующих теорем, доказанных в работе [35].

Два объекта из разных кластеров согласующей кластеризованной ранжировки могут оказаться эквивалентными в одной из исходных кластеризованных ранжировок (т.е. находиться в одном кластере). В таком случае надо рассмотреть упорядоченность этих объектов в какой-либо другой из исходных кластеризованных ранжировок. Если же во всех исходных кластеризованных ранжировках два рассматриваемых объекта находились в одном кластере, то естественно считать (и это является уточнением к этапу 3 алгоритма), что они находятся в одном кластере и в согласующей кластеризованной ранжировке.

Результат согласования кластеризованных ранжировок A, B, C,... обозначим $f(A, B, C,...)$. Тогда

$$f(A, B) = [1<2<3<4<5<6<7<\{8, 9\}<10],$$
$$f(A, C) = [\{1,3\}<\{2, 4\}<6<\{5,7\}<8<9<10],$$
$$f(B, C) = [\{1,2,3,4\}<\{5,6\}<7<\{8,9\}<10],$$
$$f(A, B, C) = f(B, C) = [\{1,2,3,4\} <\{5,6\}<7<\{8, 9\}<10].$$

Итак, в случае $f(A, B)$ дополнительного изучения с целью упорядочения требуют только объекты 8 и 9. В случае f(A, C) кластер {5,7} появился не потому, что относительно объектов 5 и 7 имеется противоречие, а потому, что в обеих исходных ранжировках эти объекты не различаются. В случае $f(B, C)$ четыре объекта с номерами 1, 2, 3, 4 объединились в один кластер, т.е. кластеризованные ранжировки оказались настолько противоречивыми, что процедура согласования не позволила провести достаточно полную декомпозицию задачи нахождения итогового мнения экспертов.

Обсудим некоторые свойства алгоритмов согласования.

1. Пусть $D = f(A, B, C,...)$. Если $a<b$ в согласующей кластеризованной ранжировке D, то $a<b$ или $a=b$ в каждой из исходных ранжировок A, B, C, ..., причем хотя бы в одной из них справедливо строгое неравенство.

2. Построение согласующих кластеризованных ранжировок может осуществляться поэтапно. Вчастности,

$$f(A,B,C) = f(f(A,B), f(A,C), f(B,C)).$$

Ясно, что ядро противоречий для набора кластеризованных ранжировок является объединением таких ядер для всех пар рассматриваемых ранжировок.

3. Построение согласующих кластеризованных ранжировок нацелено на выделение общего упорядочения в исходных кластеризованных ранжировках. Однако при этом некоторые общие свойства исходных кластеризованных ранжировок могут теряться. Так, при согласовании ранжировок B и C, рассмотренных выше, противоречия в упорядочении элементов 1 и 2 не было - в ранжировке B эти объекты входили в один кластер, т.е. 1 = 2, в то время как 1<2 в кластеризованной ранжировке C. Значит, при их отдельном рассмотрении можно принять упорядочение 1<2. Однако в $f(B,C)$ они попали в один кластер, т.е. возможность их упорядочения исчезла. Это связано с поведением объекта 3, который "перескочил" в C на первое место и "увлек с собой в противоречие" пару (1, 2), образовав противоречивые пары и с 1, и с 2. Другими словами, связная компонента графа, соответствующего ядру противоречий, сама по себе не всегда является полным графом. Недостающие ребра при этом соответствуют парам типа (1, 2), которые сами по себе не являются противоречивыми, но "увлекаются в противоречие" другими парами.

4. Необходимость согласования кластеризованных ранжировок возникает, в частности, при разработке методики применения экспертных оценок в задачах экологического страхования и химической безопасности биосферы. Как уже говорилось, популярным является метод упорядочения по средним рангам, в котором итоговая ранжировка строится на основе средних арифметических рангов, выставленных отдельными экспертами [8, 37]. Однако из теории измерений известно (см. главу 2.1), что более обоснованным является использование не средних арифметических, а медиан. Вместе с тем метод средних рангов весьма известен и широко применяется, так что просто отбросить его нецелесообразно. Поэтому было принято решение об одновременном применении обеих методов. Реализация этого решения потребовала разработки методики согласования двух указанных кластеризованных ранжировок.

5. Область применения рассматриваемого метода не ограничивается экспертными оценками. Он может быть использован, например, для сравнения качества математических моделей процесса испарения жидкости. Имелись данные экспериментов и результаты расчетов по 8 математическим моделям. Сравнивать модели можно по различным критериям качества. Например, по сумме модулей относительных отклонений расчетных и экспериментальных значений. Можно действовать и по другому. В каждой экспериментальной точке упорядочить модели по качеству, а потом получить единые оценки методами средних рангов и медиан. Использовались и иные методы. Затем применялись методы согласования кластеризованных ранжировок, полученных различными способами. В результате оказалось возможным упорядочить модели по качеству и использовать это упорядочение при

разработке банка математических моделей, используемого в задачах химической безопасности биосферы.

6. Рассматриваемый метод согласования кластеризованных ранжировок построен в соответствии с *методологией теории устойчивости*, согласно которой результат обработки данных, инвариантный относительно метода обработки, соответствует реальности, а результат расчетов, зависящий от метода обработки, отражает субъективизм исследователя, а не объективные соотношения.

Основные математические задачи анализа экспертных оценок. Ясно, что при анализе мнений экспертов можно применять самые разнообразные статистические методы, описывать их - значит описывать практически всю прикладную статистику. Тем не менее можно выделить основные широко используемые в настоящее время методы математической обработки экспертных оценок - это проверка согласованности мнений экспертов (или классификация экспертов, если нет согласованности) и усреднение мнений экспертов внутри согласованной группы.

Поскольку ответы экспертов во многих процедурах экспертного опроса - не числа, а такие объекты нечисловой природы, как градации качественных признаков, ранжировки, разбиения, результаты парных сравнений, нечеткие предпочтения и т.д., то для их анализа оказываются полезными методы статистики нечисловых данных.

Почему ответы экспертов часто носят нечисловой характер? Наиболее общий ответ состоит в том, что люди не мыслят числами. В мышлении человека используются образы, слова, но не числа. Поэтому требовать от эксперта ответ в форме чисел - значит насиловать его разум. Даже в экономике менеджеры и предприниматели, принимая решения, лишь частично опираются на численные расчеты. Это видно из условного (т.е. определяемого произвольно принятыми соглашениями, обычно оформленными в виде нормативных актов и инструкций) характера балансовой прибыли, амортизационных отчислений и других экономических показателей. Поэтому фраза типа «фирма стремится к максимизации прибыли» не может иметь строго определенного смысла. Достаточно спросить: «Максимизация прибыли - за какой период?» И сразу станет ясно, что степень оптимальности принимаемых решений зависит от горизонта планирования (на экономико-математическом уровне этот сюжет рассмотрен в монографии [1]).

Эксперт может сравнить два объекта, сказать, какой из двух лучше (метод парных сравнений), дать им оценки типа "хороший", "приемлемый", "плохой", упорядочить несколько объектов по привлекательности, но обычно не может ответить, во сколько раз или на сколько один объект лучше другого. Другими словами, ответы эксперта обычно измерены в порядковой шкале, или являются ранжировками, результатами парных сравнений и другими объектами нечисловой природы, но не числами. *Распространенное заблуждение состоит в том, что ответы экспертов стараются рассматривать как числа, занимаются "оцифровкой" их мнений, приписывая этим мнениям численные значения - баллы, которые потом обрабатывают с помощью методов прикладной статистики как результаты обычных физико-технических измерений.* В случае произвольности "оцифровки" выводы, полученные в результате подобной обработки данных, могут не иметь отношения к реальности. В связи с "оцифровкой" уместно вспомнить классическую притчу о человеке, который ищет потерянные ключи под фонарем, хотя потерял их в кустах.

На вопрос, почему он так делает, отвечает: "Под фонарем светлее". Это, конечно, верно. Но, к сожалению, весьма малы шансы найти потерянные ключи под фонарем. Так и с "оцифровкой" нечисловых данных. Она дает возможность имитации научной деятельности, но не возможность найти истину.

Проверка согласованности мнений экспертов и классификация экспертных мнений. Ясно, что мнения разных экспертов различаются. Важно понять, насколько велико это различие. Если мало - усреднение мнений экспертов позволит выделить то общее, что есть у всех экспертов, отбросив случайные отклонения в ту или иную сторону. Если велико - усреднение является чисто формальной процедурой. Так, если представить себе, что ответы экспертов равномерно покрывают поверхность бублика, то формальное усреднение укажет на центр дырки от бублика, а такого мнения не придерживается ни один эксперт. Из сказанного ясна важность проблемы проверки согласованности мнений экспертов.

Разработан ряд методов такой проверки. Статистические методы проверки согласованности зависят от математической природы ответов экспертов. Соответствующие статистические теории весьма трудны, если эти ответы - ранжировки или разбиения, и достаточно просты, если ответы - результаты независимых парных сравнений. Отсюда вытекает рекомендация по организации экспертного опроса: не старайтесь сразу получить от эксперта ранжировку или разбиение, ему трудно это сделать, да и имеющиеся математические методы не позволяют далеко продвинуться в анализе подобных данных.

Например, рекомендуют проверять согласованность ранжировок с помощью коэффициента ранговой конкордации Кендалла-Смита. Но давайте вспомним, какая статистическая модель при этом используется. Проверяется нулевая гипотеза, согласно которой ранжировки независимы и равномерно распределены на множестве всех ранжировок. Если эта гипотеза принимается, то конечно, ни о какой согласованности мнений экспертов говорить нельзя. А если отклоняется? Тоже нельзя. Например, может быть два (или больше) центра, около которых группируются ответы экспертов. Нулевая гипотеза отклоняется. Но разве можно говорить о согласованности?

Эксперту гораздо легче на каждом шагу сравнивать только два объекта. Пусть он занимается парными сравнениями. *Непараметрическая теория парных сравнений (теория люсианов)позволяет решать более сложные задачи, чем статистика ранжировок или разбиений.* В частности, вместо гипотезы равномерного распределения можно рассматривать гипотезу однородности, т.е. вместо совпадения всех распределений с одним фиксированным (равномерным) можно проверять лишь совпадение распределений мнений экспертов между собой, что естественно трактовать как согласованность их мнений. Таким образом, удается избавиться от неестественного предположения равномерности.

При отсутствии согласованности экспертов естественно разбить их на группы сходных по мнению. Это можно сделать различными методами статистики объектов нечисловой природы, относящимися к кластер-анализу, предварительно введя метрику в пространство мнений экспертов. Идея американского математика Джона Кемени об аксиоматическом введении метрик нашла многочисленных продолжателей. Однако методы кластер-анализа обычно являются эвристическими. В

частности, обычно невозможно с позиций статистической теории строго обосновать "законность" объединения двух кластеров в один. Имеется важное исключение - *для независимых парных сравнений (люсианов) разработаны методы, позволяющие проверять возможность объединения кластеров как статистическую гипотезу*. Это - еще один аргумент за то, чтобы рассматривать теорию люсианов как ядро математических методов экспертных оценок.

Нахождение итогового мнения комиссии экспертов. Пусть мнения комиссии экспертов или какой-то ее части признаны согласованными. Каково же итоговое (среднее, общее) мнение комиссии? Согласно идее Джона Кемени следует найти среднее мнение как решение *оптимизационной задачи*. А именно, надо минимизировать суммарное расстояние от кандидата в средние до мнений экспертов. Найденное таким способом среднее мнение называют "медианой Кемени".

Математическая сложность состоит в том, что мнения экспертов лежат в некотором пространстве объектов нечисловой природы. Общая теория подобного усреднения рассмотрена выше [38, глава 2.1.5]. В частности, показано, что в силу закона больших чисел (в пространствах произвольной природы) среднее мнение при увеличении числа экспертов (чьи мнения независимы и одинаково распределены) приближается к некоторому пределу, который, как известно, является *математическим ожиданием* (случайного элемента, имеющего то же распределение, что и ответы экспертов).

В конкретных пространствах нечисловых мнений экспертов вычисление медианы Кемени может быть достаточно сложным делом. Кроме свойств пространства, велика роль конкретных метрик. Так, в пространстве ранжировок при использовании метрики, связанной с коэффициентом ранговой корреляции Кендалла, необходимо проводить достаточно сложные расчеты, в то время как применение показателя различия на основе коэффициента ранговой корреляции Спирмена приводит к упорядочению по средним рангам.

Бинарные отношения и расстояние Кемени. Как известно, бинарное отношение A на конечном множестве $Q = \{q_1, q_2,..., q_k\}$ - это подмножество *декартова квадрата* $Q^2 = \{(q_m, q_n), m,n = 1, 2, \ldots, k\}$. При этом пара (q_m, q_n) входит в A тогда и только тогда, когда между q_m и q_n имеется рассматриваемое отношение.

Напомним, что каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать квадратной матрицей $\|x(a,b)\|$ из 0 и 1 порядка $k \times k$. При этом $x(a, b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b, a) = 0$, а во втором $x(b, a) = 1$. При этом хотя бы одно из чисел $x(a, b))$ и $x(b, a)$ равно 1.

В экспертных методах используют, в частности, такие бинарные отношения, как ранжировки (упорядочения, или разбиения на группы, между которыми имеется строгий порядок), отношения эквивалентности, толерантности (отношения сходства). Как следует из сказанного выше, каждое бинарное отношение A можно описать матрицей $\| a(i ,j) \|$ из 0 и 1, причем $a(i ,j) = 1$ тогда и только тогда, когда q_i и q_j находятся в отношении A, и $a(i ,j) = 0$ в противном случае.

Определение.*Расстоянием Кемени между бинарными отношениями A и B, описываемыми матрицами* $\| a(i ,j) \|$ и $\| b(i ,j) \|$ *соответственно, называется число*

$$d(A, B) = \sum \left| a(i ,j) - b(i ,j) \right|,$$

где суммирование производится по всем i, j от 1 *до k, т.е. расстояние Кемени между бинарными отношениями равно сумме модулей разностей элементов, стоящих на одних и тех же местах в соответствующих им матрицах.*

Легко видеть, что расстояние Кемени - это число несовпадающих элементов в матрицах $||a(i,j)||$ и $||b(i,j)||$.

Расстояние Кемени основано на некоторой системе аксиом. Эта система аксиом и вывод из нее формулы для расстояния Кемени между упорядочениями содержится в книге [10]. Она сыграла большую роль в развитии в нашей стране такого научного направления, как анализ нечисловой информации (см. историю вопроса в монографиях [1, 8]). В дальнейшем под влиянием работ Дж. Кемени были предложены различные системы аксиом для получения расстояний в тех или иных нужных для социально-экономических, технических, медицинских и иных исследований пространствах [38, глава 1.1.6].

Медиана Кемени и законы больших чисел. С помощью расстояния Кемени находят итоговое мнение комиссии экспертов. Пусть A_1 , A_2 , A_3 ,..., A_p - ответы p экспертов, представленные в виде бинарных отношений. Для их усреднения используют *медиану Кемени*

$$\text{Argmin} \sum d(A_i,A) ,$$

где Arg min - то или те значения A, при которых достигает минимума указанная сумма расстояний Кемени от ответов экспертов до текущей переменной A, по которой и проводится минимизация. Такимобразом,

$$\sum d(A_i,A) = d(A_1,A) + d(A_2,A) + d(A_3,A) + \ldots + d(A_p,A).$$

Кроме медианы Кемени, используют *среднее по Кемени***,** в котором вместо $d(A_i, A)$ стоит $d^2(A_i, A)$.

Медиана Кемени - частный случай определения эмпирического среднего в пространствах нечисловой природы. Для нее справедлив закон больших чисел, т.е. эмпирическое среднее приближается при росте числа составляющих (т.е. p - числа слагаемых в сумме), к теоретическому среднему:

$$\text{Arg min} \sum d(A_i,A) \rightarrow \text{Arg min } Md(A_1, A) .$$

Здесь M - символ математического ожидания. Предполагается, что ответы p экспертов $A_1, A_2, A_3, \ldots, A_p$ есть основания рассматривать как независимые одинаково распределенные случайные элементы (т.е. как случайную выборку) в соответствующем пространстве произвольной природы, например, в пространстве упорядочений или отношений эквивалентности. Систематически эмпирические и теоретические средние и соответствующие различные варианты законов больших чисел рассмотрены в [38, глава 2.1.5].

Законы больших чисел показывают, во-первых, что медиана Кемени обладает *устойчивостью* по отношению к незначительному изменению состава экспертной комиссии; во-вторых, при увеличении числа экспертов она *приближается к некоторому пределу.* Его естественно рассматривать как *истинное мнение* экспертов, от которого каждый из них несколько отклонялся по случайным причинам.

Вычисление медианы Кемени - задача целочисленного программирования. Для ее нахождения используется различные алгоритмы дискретной математики, в частности, основанные на методе ветвей и границ. Применяют также алгоритмы, осно-

ванные на идее случайного поиска, поскольку для каждого бинарного отношения нетрудно найти множество его соседей.

Рассмотрим упрощенный пример вычисления медианы Кемени. Пусть дана квадратная матрица (порядка 9) попарных расстояний для множества бинарных отношений из 9 элементов A_1, A_2, A_3,..., A_9 (см. табл.21.3). Пусть требуется найти в этом множестве *медиану* для множества из 5 элементов $\{A_2, A_4, A_5, A_8, A_9\}$.

Таблица 21.3 - Матрица попарных расстояний

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1	0	2	13	1	7	4	10	3	11
A_2	2	0	5	6	1	3	2	5	1
A_3	13	5	0	2	2	7	6	5	7
A_4	1	6	2	0	5	4	3	8	8
A_5	7	1	2	5	0	10	1	3	7
A_6	4	3	7	4	10	0	2	1	5
A_7	10	2	6	3	1	2	0	6	3
A_8	3	5	5	8	3	1	6	0	9
A_9	11	1	7	8	7	5	3	9	0

В соответствии с определением медианы Кемени следует ввести в рассмотрение функцию

$$C(A) = \sum d(A_i, A) = d(A_2,A)+d(A_4,A)+d(A_5,A)+d(A_8,A)+d(A_9,A),$$

рассчитать ее значения для всех A_1, A_2, A_3, ..., A_9 и выбрать наименьшее. Проведем расчеты:

$$C(A_1) = d(A_2, A_1)+d(A_4, A_1)+d(A_5, A_1)+d(A_8, A_1)+d(A_9, A_1) = \\ = 2 + 1 +7 +3 +11 = 24,$$

$$C(A_2) = d(A_2, A_2)+d(A_4, A_2)+d(A_5, A_2)+d(A_8, A_2)+d(A_9, A_2) = \\ = 0 + 6 + 1 + 5 + 1 = 13,$$

$$C(A_3) = d(A_2, A_3)+d(A_4, A_3)+d(A_5, A_3)+d(A_8, A_3)+d(A_9, A_3) = \\ = 5 + 2 + 2 + 5 +7 = 21,$$

$$C(A_4) = d(A_2, A_4)+d(A_4, A_4)+d(A_5, A_4)+d(A_8, A_4)+d(A_9, A_4) = \\ = 6 + 0 + 5 + 8 + 8 = 27,$$

$$C(A_5) = d(A_2, A_5)+d(A_4, A_5)+d(A_5, A_5)+d(A_8, A_5)+d(A_9, A_5) = \\ = 1 + 5 + 0 +3 + 7 = 16,$$

$$C(A_6) = d(A_2, A_6)+d(A_4, A_6)+d(A_5, A_6)+d(A_8, A_6)+d(A_9, A_6) = \\ = 3 + 4 + 10 + 1 + 5 = 23,$$

$$C(A_7) = d(A_2, A_7)+d(A_4, A_7)+d(A_5, A_7)+d(A_8, A_7)+d(A_9, A_7) = \\ = 2 + 3 +1 + 6 + 3 = 15,$$

$$C(A_8) = d(A_2, A_8)+d(A_4, A_8)+d(A_5, A_8)+d(A_8, A_8)+d(A_9, A_8) = \\ = 5 + 8 + 3 + 0 +9 = 25,$$

$$C(A_9) = d(A_2, A_9)+d(A_4, A_9)+d(A_5, A_9)+d(A_8, A_9)+d(A_9, A_9) = \\ = 1 + 8 + 7 + 9 + 0 = 25.$$

Из всех вычисленных сумм наименьшая равна 13, и достигается она при $A=A_2$, следовательно, медиана Кемени - это множество $\{A_2\}$, состоящее из одного элемента A_2.

ЗАДАНИЯ

1. В табл. 21.4 приведены упорядочения 7 инвестиционных проектов, представленные 7 экспертами.

Таблица 21.4 – Упорядочения проектов экспертами

Эксперты	Упорядочения
1	1 < {2,3} < 4 < 5 < {6,7}
2	{1,3} < 4 < 2< 5< 7 < 6
3	1 < 4 < 2 < 3 < 6 < 5 < 7
4	1 < {2, 4} < 3 < 5 < 7 <6
5	2 < 3 < 4 < 5 <1 <6 <7
6	1 < 3 < 2 < 5 < 6 < 7 < 4
7	1 < 5 < 3 < 4 < 2 < 6 < 7

Найдите:

а) итоговое упорядочение по средним арифметическим рангам;

б) итоговое упорядочение по медианам рангов;

в) кластеризованную ранжировку, согласующую эти два упорядочения.

2. Выпишите матрицу из 0 и 1, соответствующую бинарному отношению (кластеризованной ранжировке) 5 < {1, 3} < 4 < 2 < {6, 7}.

3. Найдите расстояние Кемени между бинарными отношениями - упорядочениями A = [3< 2 <1< {4,5}] и B = [1 < {2 ,3} < 4 < 5].

4. Дана квадратная матрица (порядка 9) попарных расстояний (мер различия) для множества бинарных отношений из 9 элементов A_1 , A_2 , A_3 ,..., A_9 (табл.21.5). Найдите в этом множестве медиану для множества из 5 элементов $\{A_2, A_3, A_5, A_6, A_9\}$.

Таблица 21.5 – Попарные расстояния между бинарными отношениями

0	5	3	6	7	4	10	3	11
5	0	5	6	10	3	2	5	7
3	5	0	8	2	7	6	5	7
6	6	8	0	5	4	3	8	8
7	10	2	5	0	10	8	3	7
4	3	7	4	10	0	2	3	5
10	2	6	3	8	2	0	6	3
3	5	5	8	3	3	6	0	9
11	7	7	8	7	5	3	9	0

Вопросы для самоконтроля

- Как случайные толерантности используются в теории нечетких толерантностей?
- В теории люсианов выведите из общего вида несмещенной оценки многочлена от p по результатам m независимых испытаний Бернулли с вероятностью успеха p в каждом (формула (12)) несмещенную оценку в случае $f(p) = 2p(1 - p)$ (формула (13)).
- Как можно проводить кластерный анализ совокупности нечетких множеств?
- Чем метод средних арифметических рангов отличает от метода медиан рангов?
- Почему необходимо согласование кластеризованных ранжировок и как оно проводится?
- В чем состоит проблема согласованности ответов экспертов?
- Как бинарные отношения используются в экспертизах?
- Как бинарные отношения описываются матрицами из 0 и 1?
- Что такое расстояние Кемени и медиана Кемени?
- Чем закон больших чисел для медианы Кемени отличается от "классического" закона больших чисел, известного в статистике?

Практическое занятие № 22

Математическое представление когнитивных моделей в виде графов

Цель. Дать представдение о представлении знаний эксперов в виде когнитивных карт**.**

Теоретические сведения.

Когнитивное моделирование. Понятие «когнитивное моделирование» имеет различное толкование в зависимости от цели и объекта когнитивных исследований; содержание действий когнитивного моделирования зависит от субъекта и объекта исследований. Когнитивное моделирование автоматизирует часть функций процессов познания, поэтому оно с успехом может применяться во многих областях.

Когнитивное моделирование сложных систем включает в себя построение когнитивных моделей (в первую очередь – когнитивных карт) с использованием различных когнитивных технологий их разработки, экспертный анализ моделей, в том числе с помощью формальных математических методов, предсказание (прогнозирование) развития ситуаций, разработку возможных сценариев развития системы, разработку управленческих решений и ряд других действий, направленных на совершенствование модели, а также на выбор лучшего сценария развития системы.

Когнитивное моделирование предназначено для структуризации, анализа и принятия управленческих решений в сложных и неопределенных ситуациях (геополитических, внутриполитических, военных и т.п.), при отсутствии количественной или статистической информации о происходящих процессах в таких ситуациях.

Методология когнитивного моделирования, предназначенная для исследования сложных систем, для анализа и принятия решений в плохо определенных ситуациях, была предложена Р.Аксельродом[14], известным американским социологом и политологом, который развил аппарат когнитивных карт для анализа и прогнозирования решений, принимаемых политиками. Основой методологии является моделирование субъективных представлений экспертов о ситуации, использующее методологию структуризации ситуации, методы анализа ситуации и основывается на моделировании субъективных представлений экспертов о ситуации в виде знакового орграфа (когнитивной карты), являющегося моделью знаний эксперта. В работе Р.Аксельрода знаковый орграф обозначается как $G = (F, W)$, где F – множество факторов ситуации, W – множество причинно-следственных отношений между факторами ситуации.

Когнитивное компьютерное моделирование. Для когнитивно-информационной поддержки постановки и решения новых научных проблем особую актуальность (начиная с 90-х годов XX столетия) приобрели разработки в области интерактивной машинной графики (ИМГ). ИМГ уже стало практически обязательным атрибутом всех современных систем моделирования и может быть ис-

[14]Axelrod R. The Structure of Decision: Cognitive Maps of Political Elites. – Princeton. UniversityPress, 1976

пользовано для максимальной активизации творческого и познавательного потенциала человеческого интеллекта. Одним из важнейших шагов на этом пути стала разработка концепции когнитивной (т.е. способствующей познанию) *компьютерной графики*, основной задачей которой является "создание таких моделей представления знаний (когнитивных моделей) в которых была бы возможность однообразными средствами представлять как объекты, характерные для алгебраического мышления, так и образы-картины, с которыми оперирует геометрическое мышление"[15].

Важнейшим атрибутом когнитивного компьютерного моделирования являются *когнитограммы*, т.е. специальным образом организованная визуализация моделей, данных и результатов моделирования, ориентированная на максимальную активизацию образно-интуитивных механизмов мышления. Можно выделить три категории когнитограмм: искусственные (абстрактные), естественные (в той или иной степени имитирующие реальную или виртуальную визуальную обстановку) и комбинированные или совмещенные, объединяющие свойства первых двух категорий. Самыми распространенными в настоящее время являются *искусственные когнитограммы*, наиболее простым вариантом реализации которых можно считать использование традиционной визуализации результатов в виде разнообразных графиков и диаграмм различной размерности. В настоящее время разрабатываются визуальные словари, например[16], в которых для каждого термина составляется *графическая интерактивная карты термина – когнитограмма*, содержащей все его концептуальные связи в рамках данной словарной базы.

Когнитивная структуризация предметной области – это выявление будущих целевых и нежелательных состояний объекта управления и наиболее существенных (базисных) факторов управления и внешней среды, влияющих на переход объекта в эти состояния, а также установление на качественном уровне причинно-следственных связей между ними, с учетом взаимовлияния факторов друг на друга

Когнитивная структуризация (*cognitive mapping*) или концептуализация состоит из: разработки структуры полученных знаний о предметной области (определяется список основных понятий о предметной области), выявлении отношений между понятиями, определении связи данной предметной области с окружающим миром.

Цель когнитивной структуризации состоит в формировании и уточнении гипотезы о функционировании исследуемого объекта, рассматриваемого как сложная система, которая состоит из отдельных, но взаимосвязанных между собою элементов и подсистем.

Когнитивная (познавательно-целевая) структуризация знаний об исследуемом объекте и внешней для него среды может производиться на основе PEST-анализа и SWOT-анализа и других технологий.

Когнитивная технология синтезирует системный и когнитивный подходы, является универсальным научным инструментарием понимания поведения сложных систем: экономических, социальных, политических, социотехнических и др.

---

[15] Краткий словарь когнитивных терминов / Под общей редакциейЕ.С. Кубряковой. – М.: Филол. ф-т МГУ им. М.В. Ломоносова, 1997. – 245 с.

[16] «Визуальныйсловарик» http://vslovar.org.ru/ ; «VisualThesaurus» http://www.visualthesaurus.com/index.jsp

Когнитивная структуризация является удобным инструментом исследования слабоструктурированных проблем сложных систем, способствует лучшему их пониманию, а также выявлению противоречий и качественному анализу систем.

Основные определения, представленные выше, сгруппированы в табл. 22.1.

Таблица 22.1 - Определения в области когнитивной науки

№	Понятие	Определение
1	Когнитивная наука, когнитивистика	Это целостная междисциплинарная область, совокупность наук о сознании, о приобретении, хранении, преобразовании и использовании знания, о переработке информации.
2	Когнитология(инженерия знаний)	Это область междисциплинарных исследований, посвященных проблемам структуризации знаний экспертов в конкретной предметной области и использующих интуицию, опыт, ассоциативность мышления, догадки эксперта; когнитология обеспечивает процесс передачи ЭВМ знаний экспертов; важнейшей задачей когнитологии является задача обеспечения технологического синтеза интеллектуальных возможностей человека и ЭВМ, разработка интерактивных систем и визуализация информации, создание систем поддержки принятия решений.
3	Когнитивный подход (познавательный)	Это решение традиционных для определенной науки проблем, но методами, учитывающими когнитивные аспекты; это исследования самых разных предметностей с учетом человеческого фактора. Принципиальной ценностью в когнитивных исследованияхявляется объединения точного, естественного и гуманитарного знания.
4	Когнитивный стиль	Это совокупность критериев выбора предпочтений при решении задач и познании мира, специфическая для каждого человека. Это система средств и индивидуальных приемов, к которым прибегает человек для организации своейпознавательной деятельности.
5	Метафора в когнитивных исследованиях	Это видение одного объекта через другой; в когнитивных процессах сложные непосредственно ненаблюдаемые мыслительные пространства соотносятся через метафору с более простыми, хорошо знакомыми мыслительными пространствами.
6	Технология когнитивного (познавательно-целевого) моделирования	Это совокупность приемов, способов, методов извлечения и порождения новых знаний, а также обеспечивающие этот процесс специальные программные средства. Определяющим для любых когнитивных технологий является репрезентация и визуализация идей-концепций-концептов и их связей. Различные технологии когнитивного моделирования используются для построения моделей, которые по своему назначению могут быть отнесены к двум большим классам: • модели человеческого интеллекта, • модели сложных систем.
7	Визуализация	Это прием, с помощью которого можно наглядно представить взаимосвязи в сложной системе. Результатом такой визуализации является когнитивная модель (карта) системы (системы понятий, концептов, объектов сложных систем и .п.) в виде графа.
8	Когнитивный инструментарий	В узком смысле - этопрограммные средства для репрезентации знаний, построения когнитивных моделей и последующих исследований на них.
9	Когнитивное моделирование	Имеет различное толкование в зависимости от цели и объекта когнитивных исследований; содержание действий когнитивного моделирования зависит от субъекта и объекта исследований. Это: построение когнитивных моделей (в первую очередь – когнитивных карт) с использованием различных когнитивных технологий их разработки, экспертный анализ моделей, в том числе с помощью формальных математических методов, прогнозирование развития ситуаций на моделях, разработка и обоснование решений; автоматизирует часть функций процессов познания. Когнитивное моделирование сложных систем предназначено для структуриза-

		ции, анализа и принятия управленческих решений в сложных и неопределенных ситуациях.

Продолжение таблицы 22.1

10	Методология когнитивного моделирования	Это (с общенаучных позиций) логическая организация деятельности исследователя, состоящая в определении цели и предмета исследования, подходов и ориентиров в проведении исследования, выборе средств и методов, определяющих наилучший результат. Методология когнитивного моделирования различается в зависимости от того, какая и для чего строится когнитивная модель. Для сложных систем - это совокупность методов структуризации знаний о ситуациях в системе, методы анализа ситуации; это моделирование субъективных представлений экспертов о ситуации в виде знакового орграфа. В настоящее время методология когнитивного моделирования дополняется новыми задачами и методами их решения.
11	Когнитивное компьютерное моделирование	Это использование когнитивной (т.е. способствующей познанию) компьютерной графики для визуализации представлений исследователя о проблемной ситуации. Важнейшим атрибутом когнитивного компьютерного моделирования являются когнитограммы, т.е. специальным образом организованная визуализация моделей, данных и результатов моделирования, ориентированная на максимальную активизацию образно-интуитивных механизмов мышления.
12	Когнитивная структуризация или концептуализация	Это: разработка структуры полученных знаний о предметной области (определяется список основных понятий о предметной области), выявление отношений между понятиями, определение связи данной предметной области с окружающим миром. Цель когнитивной структуризации состоит в формировании и уточнении гипотезы о функционировании исследуемого объекта, рассматриваемого как сложная система, которая состоит из отдельных, но взаимосвязанных между собою элементов и подсистем.
13	Когнитивная модель	Существуют различные определения когнитивной модели. С позиций когнитивной психологии, которые являются общими в когнитивной науке - это метафора, основанная на наблюдениях и выводах, сделанных из этих наблюдений, и описывающих, как обнаруживается, хранится и используется информация.
14	Когнитивная карта	Одна из форм когнитивной модели. Это схема, визуальное отображениесубъектом (исследователем, экспертом, индивидом) его представления о системе связей (отношений, влияний, взаимодействий) между концептами (объектами, сущностями, понятиями, факторами), составляемая с определенной целью в рамках того или иного направления когнитивной науки. Это структура знаний, это графическое представление причинных связей между понятиями, факторами, показателями, взаимодействующими системами и их блоками. С формальных позиций – это знаковый ориентированный граф, вершины которого отображают сущности (объекты, понятия, факторы, переменные) предметной области, дуги – отношения между вершинами.

Типология когнитивных моделей

Чаще всего в качестве основания для классификации моделей берется вид языка, на котором они формулируются:

•содержательная модель формулируется на естественном языке, такие модели чаще всего используются в гуманитарной сфере;

•формальная модель воплощается с помощью одного или нескольких формальных языков (например, языков математических теорий или языков программирования).

В процессе построения, изучения и совершенствования содержательной модели когнитивная модель непрерывно модифицируется и усложняется. В гуманитарных науках цикл моделирования на этом обычно и заканчивается, но в некоторых случаях модель удается формализовать до такой степени, что становится возможным построение и изучение формальной модели объекта. Любая модель является, в конечном счете, моделью объекта, фрагмента реальности.

Когнитивные карты имеют не только визуальное, но и математическое обоснование. Это четкие и нечеткие графы (нечеткие когнитивные карты).

Граф оказывается подходящей моделью для представления отношений между экономическими объектами (предприятия, организации, средства и факторы производства, элементы социальной сферы, характеризуемые как объект, в котором сосредоточена или на который направлена экономическая деятельность, и представляющие определенную сторону экономических отношений), между субъектами социальных процессов (например, людьми, группами людей), между подсистемами социально-экономических систем, между другими концептами, сущностями и т.п.

В книге американского математика Фреда Стивена Робертса (о применении теории графов и комбинаторики в задачах моделирования в социальных науках и биологии) встречается название «когнитивная карта» для знакового ориентированного графа(Когнитивная карта, построенная на основе обсуждений в Британском Комитете по делам Востока в 1918 г. британской политики в Персии)[17]: «знаковый орграф, изображенный на рисунке 22.1 является когнитивной картой или переводом в знаковый орграф мнений реального лица, принимающего решения».

Этот специальный орграф выражал точку зрения одного из членов Комитета Марлинга. Дуги представляют причинные связи, показывающие влияние одной понятийной переменной на другую. Р. Аксельрод (Axelrod, 1971г.) заметил, что такие когнитивные карты стремятся стать сбалансированными и не содержат контуров. Он считает, что отсутствие контуров характеризует тенденцию лица, принимающего решения, пропускать важные обратные связи.

На Рисунок 9 приведено еще одно изображение когнитивной карты, иллюстрирующее разнообразие исследуемых проблем сложных систем, а также средств визуализации их в виде когнитивных карт.

Когнитивная карта на рисунке 22.1 интересна еще и тем, что ее дуги имеют веса, т.е. это взвешенный граф.

Для введения формального аппарата теории графов, воспользуемся определениями Ф. Робертса, данными в его книге[17]:

«Знаковый граф (знаковый орграф) – *это граф, в котором вершины соответствуют членам группы; из вершины V_i в вершинуV_j проводится дуга, если наблюдается отчетливо выраженное отношение V_iк V_j, причем дуга $e_{ij} = (V_i, V_j)$ имеет знак плюс (+), если V_i«симпатизирует» V_j и знак минус (-) в противном случае».*

[17] *Робертс Ф.С.* Дискретные математические модели с приложениями к социальным, биологическим и экологическим задачам. – М.: Наука, 1986. С.175,176.

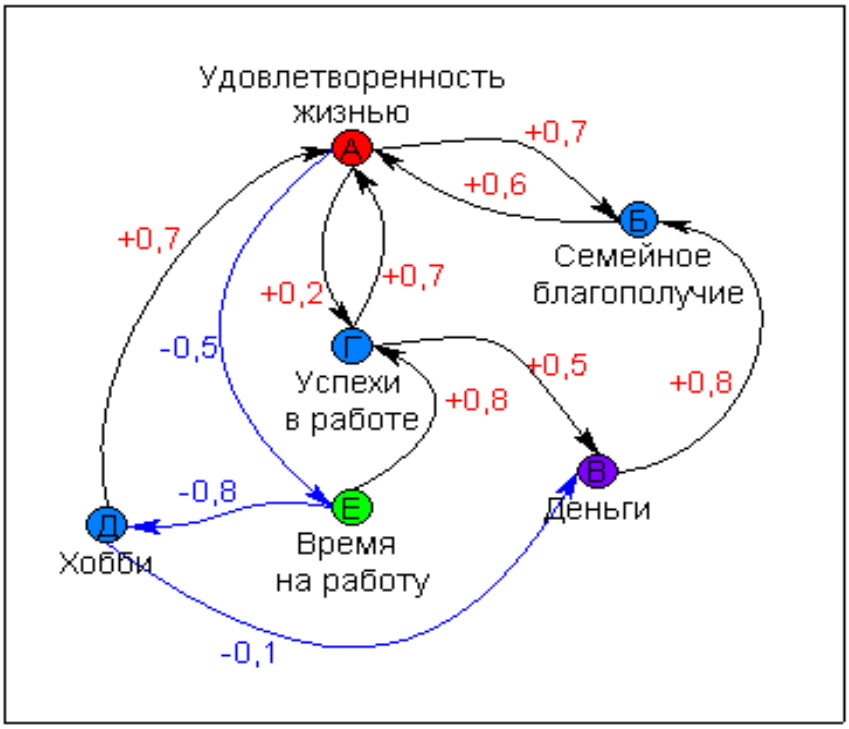


Рисунок 22.1 - Когнитивная карта «Удовлетворенность жизнью» (см. сноску 18)

Понятие «знаковый орграф» может иметь разнообразные приложения, поэтому дуги и знаки интерпретируются по разному в зависимости от изучаемой сложной системы. Кроме того, теоретические исследования сложных систем развивается в рамках более сложной модели, нежели знаковый орграф – в рамках взвешенного орграфа, *в котором каждой дуге* e_{ij} *приписано действительное число (вес)* w_{ij}.

Рисунки 22.2 выполнены с помощью программной системы ПСКМ[18]. Сплошные линии дуг соответствуют $w_{ij} = +1$, штрихпунктирные - $w_{ij} = -1$. Знак «+» может быть интерпретирован как «положительные (отрицательные) изменения в вершине v_iприводят к положительным (отрицательным) изменениям в вершине v_j», т.е. это однонаправленные изменения;знак «-» - как «положительные (отрицательные) изменения в вершине v_iприводят к отрицательным (положительным) изменениям в вершине v_j» - разнонаправленные изменения.

Встречные стрелки отображают взаимовлияние вершин, цикл графа; такое отношение симметрично.

Большинство понятий орграфов применимо и к взвешенным орграфам. Это понятия: *путь, простой путь, полупуть, контур, цикл, полуконтур; сильная, слабая, односторонняя связность*, «*знак пути, замкнутого пути, контура*».

Знак пути, цепи, замкнутого пути, замкнутой цепи, контура цикла и т.д. *определяется как произведение знаков входящих в них дуг*.

[18] *Горелова Г.В. Захарова Е.Н., Радченко С.А.*Исследование слабоструктурированных проблем социально-экономических систем: когнитивный подход. - Ростов н/Д: Изд-во РГУ, 2006. - 332с.

Очевидно, что путь, цикл и т.п. имеют знак «-», если число отрицательных дуг, содержащихся в них нечетно, в противном случае они имеют знак «+». Так, для графа «Ромео и Джульетта» путь $V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow V_1$ является отрицательным, а цикл $V_1 \rightarrow V_2 \rightarrow V_1$ – положительным.

При математическом моделировании сложных систем перед исследователем возникает проблема нахождения компромисса между точностью результатов моделирования и возможностью получения точной и подробной информации для построения модели.

В такой ситуации знаковые и взвешенные орграфы пригодны для разработки «простых» математических моделей и при анализе результатов, получаемых при минимальной информации.

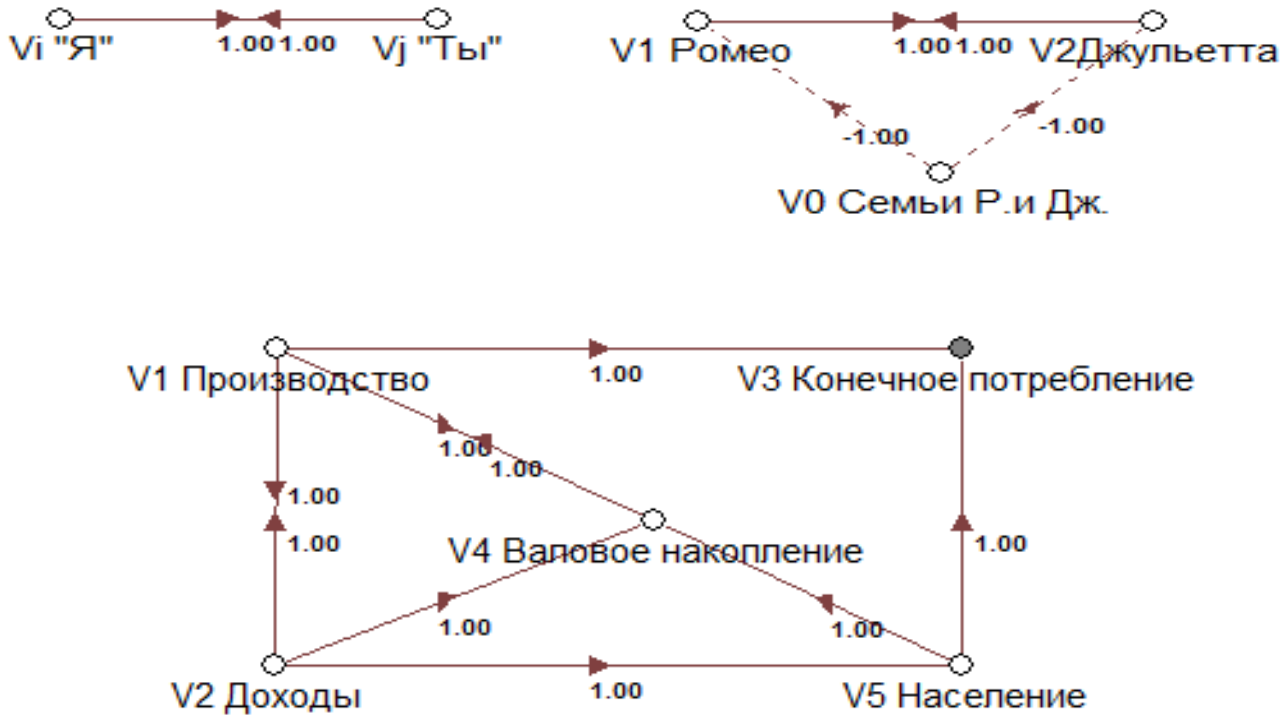


Рисунок 22.2 - Знаковые орграфы с именами вершин и весами дуг $w_{ij} = +1$ и $w_{ij} = -1$

Приведем еще два примера из [Roberts, с. 161, 162] – рисунки 22.3 и 22.4, интересных с исторической точки зрения, как одни из первых когнитивных карт, но не потерявших актуальности и сейчас.

На рисунке 22.2 контур $V_1 \rightarrow V_3 \rightarrow V_5 \rightarrow V_6 \rightarrow V_1$ противодействует отклонению в вершине V_1. Увеличение любой переменной в этом контуре в итоге приводит через другие вершины к уменьшению данной переменной и наоборот (Интерпретация: чем больше людей в городе, тем больше отходов, чем больше отходов, тем больше бактерий, чем больше бактерий, тем больше заболеваемость, чем больше заболеваемость, тем меньше людей и т.п.). Это контур отрицательной обратной связи. Контур $V_1 \rightarrow V_2 \rightarrow V_4 \rightarrow V_1$ является контуром, усиливающим отклонение, т.е. контуром положительной обратной связи.

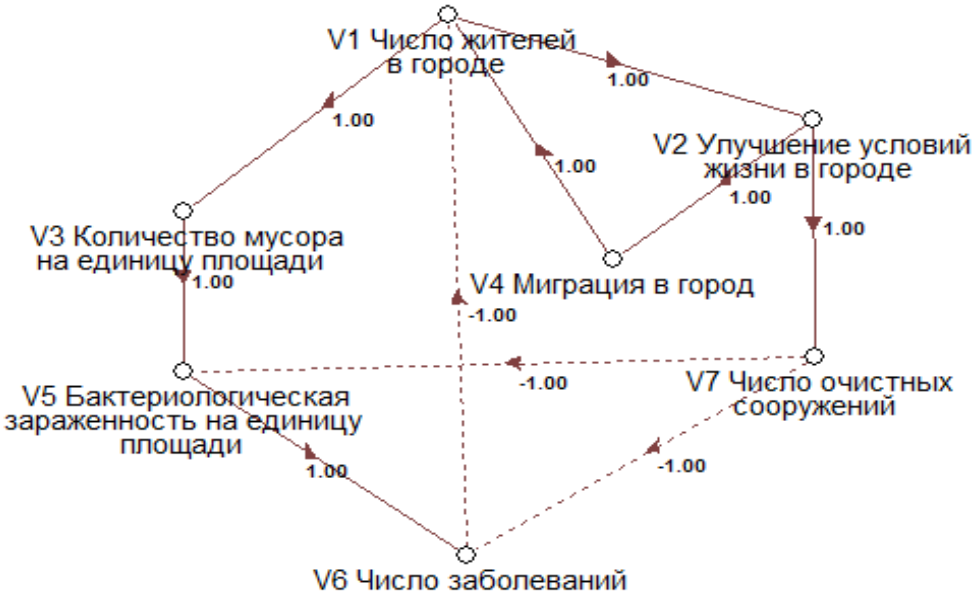


Рисунок 22.3 - Знаковый орграф для анализа проблем удаления твердых отходов (Maryama [1963, с. 176])

Утверждение Маруямы: Контур усиливает отклонение тогда и только тогда, когда он содержит четное число отрицательных дуг (усиление отклонений происходит и при отсутствии в контуре отрицательных дуг).

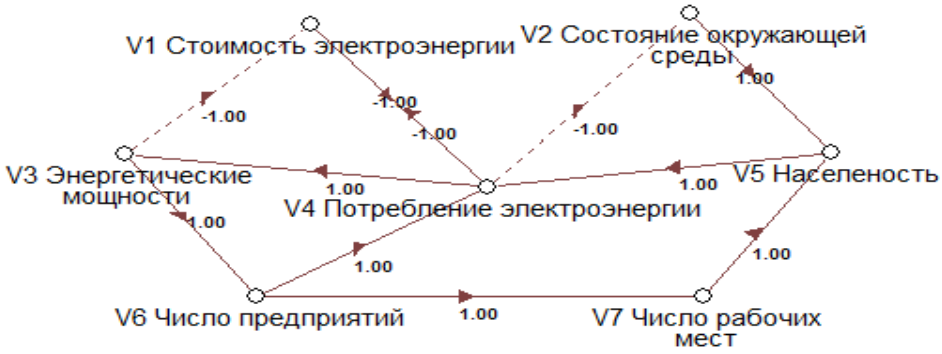


Рисунок 22.4 - Знаковый орграф для анализа проблем потребления электроэнергии в конкретном регионе (Roberts [1971], с.162)

Схема рисунке 22.4 содержит небольшое число вершин и связей для удобства предварительного анализа. Более тщательное анализ проблемы потребления электроэнергии потребует, по словам Робертса, значительно большее число переменных и более тонкие методы для их выбора. При этом возникает проблема объединения мнений экспертов.

Для решения проблем, обозначенных в примерах рисунках 22.3 и 22.4, недостаточно только построить граф той или иной сложности и проанализировать цепочки его связей (пути) и циклы, необходим более глубокий анализ его структуры,

свойств устойчивости (неустойчивости), анализ влияния изменений параметров вершин на другие вершины, анализ чувствительности[19].

ЗАДАНИЕ

Построить и проанализировать когнитивную карту по заданию преподавателя.

Вопросы для самоконтроля

- Определить понятия: когнитивная наука, когнитология и объяснить их отличие
- В чем состоит отличие по целям и методам когнитивных исследований в области человеческого интеллекта и в области исследования сложных систем?
- Определить понятия: когнитивная модель, когнитивная карта. В чем их сходство и отличие?
- Определите понятие и приведите пример когнитивной карты
- В чем отличие между разными типами когнитивных моделей?
- В чем заключается преимущество когнитивных моделей?
- Определить понятия: технология когнитивного моделирования, когнитивный инструментарий, когнитивное моделирование. В чем их сходство и отличие?
- В чем состоит когнитивная методология исследования сложных систем?

[19] *Горелова Г.В. Захарова Е.Н., Радченко С.А.*Исследование слабоструктурированных проблем социально-экономических систем: когнитивный подход. - Ростов н/Д: Изд-во РГУ, 2006. - 332с.

Практическое занятие № 23

Выбор и концептуальное описание предметной области задачи принятия решений

Цель работы: Овладение методикой выделения и концептуального описания предметной области задачи; построение на основе этого описании концептуальной модели данной предметной области.

Описание работы

Данноепрактическое занятие соответствует этапу системного анализа задачи, с которого начинается проектирование любой экспертной системы. Это очень важный этап, на который следует обратить особое внимание. Он включает в себя процессы идентификации и концептуализации задачи, т.е. выделения соответствующей предметной области (ПО), сбора необходимой информации, консультаций с экспертами и начальной формализации задачи в виде концептуальной модели предметной области.

Для проведения системного анализа и построения концептуальной модели предлагается специальная ***методика***, определяющая последовательность выделения и описания необходимых понятий, а также система приобретения знаний ***"Помощник эксперта"***.

Методика позволяет направить процесс анализа задачи так, чтобы наиболее рациональным образом выделить и организовать необходимые знания. Это дает возможность правильно определить их объем и степень детализации, а также избежать нехватки важной информации или наличия лишних сведений, что обеспечивает целостность и обоснованность приобретаемых знаний, существенно повышая качественный уровень проводимого системного анализа.

"Помощник эксперта" осуществляет инструментальную поддержку данной работы на двух этапах: визуализации концептуальных моделей элементарных и производных решений, в качестве которых выступают возможные действия ЛПР; построения понятийно – объектной модели предметной области (ПОМ).

Методика системного анализа задачи и построения концептуальной модели предметной области приведены ниже.

Руководство пользователя по системе "Помощник эксперта" распространяется вместе с самой системой.

Теоретические основы построения концептуальной модели предметной области задачи принятия решений (КМПО)

Концептуальная модель предметной области предполагает описание данных и процессов задачи, а также построение ее пространства состояний, на семиотическом уровне, т.е. на уровне, когда объекты предметной области и происходящие в ней процессы представляются их знаковыми или лексическими эквивалентами - понятиями, а затем раскрываются объемы и содержания этих понятий. Анализ понятий данной ПО, связанных с решаемой задачей, позволяет выделить характер-

ные признаки и свойства соответствующих объектов, необходимых для реализации требуемых задачей процессов.

Предметная область задачи представляется в виде совокупности следующих множеств: X, C, R, G, F , где:

$X = \{x_1, x_2, \ldots x_n\}$ - множество имен объектов (предметов и сущностей внешнего мира), с которыми мы имеем дело при решении задачи;

$C = \{c_1, c_2, \ldots c_m\}$ - множество имен свойств объектов из множества X (характерных признаков этих объектов). Каждый объект из множества X получает свое содержание в виде совокупности необходимых для решения данной задачи свойств, т.е.:

$x_i = (c_j, c_k, \ldots c_z)$, где для каждого свойстваопределяются области значений: $c_j = (c_{j1}, c_{j2}, \ldots, c_{jp})$, …, $c_z = (c_{z1}, c_{z2}, \ldots, c_{zq})$;

$R = (r_1, r_2, \ldots r_n)$ - множество имен отношений, в которые могут вступать объекты моделируемой ПО;

$G = (g_1, g_2, \ldots g_k)$ - множество имен действий (операций), которые допустимы над этими объектами путем изменение значений их свойств и отношений между ними.

Состояние ПО $S_{ПО}$ представляет собой совокупность всех *фактов*, определяющих состояние каждого объекта. Под *фактами* понимаются означенные свойства и отношения. Состояние ПО изменяется под влиянием действий из множества G и определяется в фиксированный момент t_i времени следующим образом:

$$S_{ПО}(t_i) = \{ X(t_i), C(t_i), R(t_i) \} \qquad (23.1)$$

Совокупность всех возможных состояний образует *пространство состояний* данной ПО. Каждому действию из множества G соответствует *состояние-предусловие*, т.е. состояние ПО, к которому применимо данное действие. В результате выполнения этого действия изменяются значения некоторых элементов множеств C и R, связанных с применяемым действием, т.е. формируются новые факты. Получившееся новое состояние является *состоянием-постусловием* данного действия. Таким образом, каждое действие представляет собой функциональное отображение одного (или нескольких) состояний в другое (другие). В зависимости от описываемого выражением (23.1) предусловия на момент $t_i \in [t_{нач}, t_{кон}]$ выбирается то действие, которое необходимо для решения задачи. Обозначим через F множество взаимоотображений пространства состояний и множества действий:

$$F: (S_{ПО}(t) \leftrightarrow G)$$

или

$$F: (\{ X(t), C(t), R(t) \} \leftrightarrow G). \qquad (23.2)$$

Задача состоит в том, чтобы перевести ПО из начального состояния - $S_н$ в некоторое заданное, определяемое как целевое - $S_ц$. Процесс решения задачи, таким образом, заключается в том, чтобы определить цепочку действий, последовательное применение которых к начальному состоянию ПО переводит ее в целевое состояние. Схема решения выражается формулой:

$$Z = (S_н \xrightarrow{G} S_ц) \qquad (23.3)$$

А целевое состояние выражением:

$$S_ц = g_i\,(g_n(g_m(\ldots\ldots g_k(\,S_н\,))))$$

Последовательность (g_k, …, g_m, g_n, g_i) и представляет собой алгоритм решения задачи, поиском которого занимается экспертная система.

Методика анализа задачи и построения концептуальной модели предметной области

Предлагаемая методика системного анализа задачи и построения КМПО разработана на основе методов ситуационного управления и заключается в анализе по определенному алгоритму так называемых ***структур действий***, отражающих процессы рассматриваемой ПО и функции разрабатываемой интеллектуальной системы (ЭС) (рис. 23.1). В основе методики лежит понимание концептуальной модели ПО, как самой первой ступени формализации знаний, связанных с задачами в данной ПО, требующими принятия решений. Знания в КМПО представляются в виде определенной системы понятий, связанных между собой различными отношениями, – так называемого, **понятийногопространства** ПО. Анализ и раскрытие **содержания** этих понятий и является тем способом приобретения и организации знаний, который позволяет получить достаточно полную и, в то же время, не избыточную модель, необходимую для дальнейшей формализации с целью создания базы знаний и проведения логического вывода в данной ПО.

Как было сказано выше, понятия в КМПО по своему типу делятся на 4 основные группы:

- понятия-объекты, характеризующие объекты, явления и события данной ПрО;
- понятия-свойства, характеризующие различные черты, признаки, особенности и свойства объектов, а также их возможные состояния;
- понятия-отношения, характеризующие разного рода взаимосвязи и взаимозависимости между объектами;
- понятия-действия, характеризующие различные процессы, протекающие в ПО в разные моменты времени.

Содержание понятий каждого типа раскрывается по-разному. Под **содержанием понятия-объекта** (рис. 23.1) понимается:

- во-первых, совокупность его свойств и характерных признаков, важных для описания объекта с точки зрения решаемой задачи,
- во-вторых, набор различных состояний, в которых объект может находиться в процессе выполнения действий,
- в-третьих, множество отношений данного объекта с другими объектами ПО, которые существуют постоянно или могут возникать во время исполнения моделируемых процессов,

- в-четвертых, описание иерархической взаимосвязи данного понятия-объекта с другими понятиями-объектами задачи, определяющее степень общности его описания.

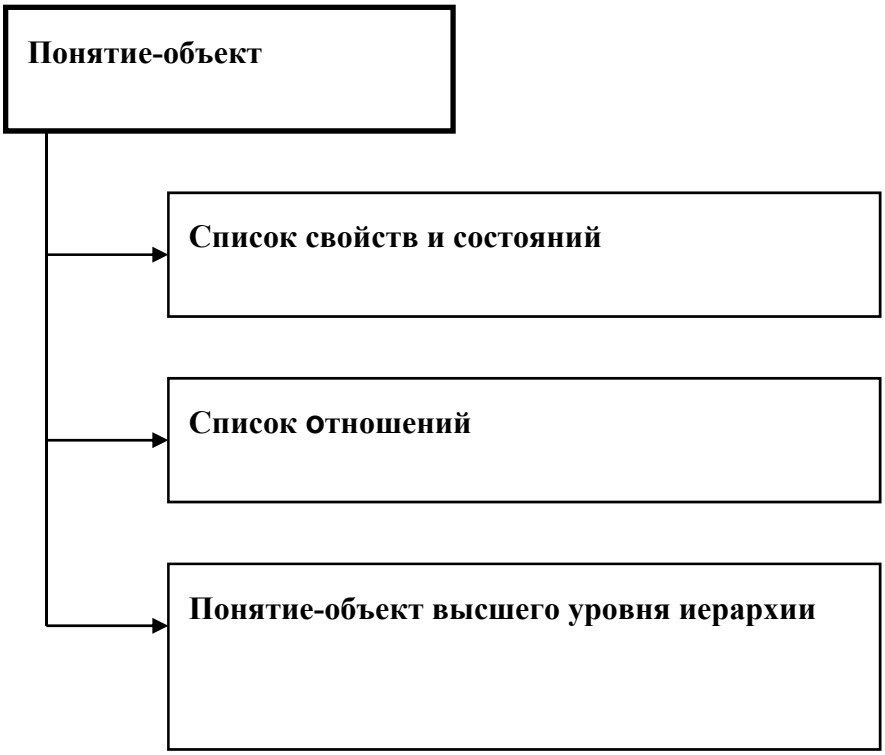


Рисунок 23.1 – Содержание понятия-объекта

Рисунок 23.2 – Содержание понятия-свойства

Содержание понятия-свойства (рис.23.2) раскрывается через область его значений, их тип и текущее значение свойства, а **содержание понятия-отношения** – через совокупность объектов, с которыми это отношение возможно, и значение его (логическая величина) для этих объектов в текущий момент времени. Оба эти понятия могут также содержать описание зависимости своих значений от значений других свойств и отношений объектов ПО.

Понятие-действие является основным и наиболее сложным понятием КМПО. Его **содержание** (рис.23.3) раскрывается через описание всех объектов, которые каким-либо образом в нем участвуют, и действий, которые необходимо выполнить для реализации данного действия. Содержание понятия-действия включает в себя следующее:

- указание на **субъект** действия, т.е. объект, который может выполнять данное действие;
- указание на **объект** действия, т.е. объект, на изменение свойств и/или отношений которых направлено данное действие;
- указание на **компоненты** действия, т.е. другие объекты ПО, от значения свойств и/или отношений которых зависит совершение данного действия или значения свойств и/или отношений которых также изменяются в результате выполнения данного действия;
- описание условий совершения действия и его результатов в виде совокупности значений свойств и/или отношений связанных с ним объектов, т.е. ситуаций **предусловия** и **постусловия** данного действия;
- описание действий, которые необходимо выполнить для создания условий совершения данного действия, т.е. **поддействий**, или действий нижнего уровня, данного действия.

Рисунок 23.3 – Содержание понятия-действия

На базе основных формируются производные понятия, служащие для более полного описания процессов и решений данной ПО. К таким понятиям, в первую очередь, относятся понятия-факты и понятия-ситуации.

Под **содержанием понятия-факта** понимается конкретное значение или некая область значений одного свойства или отношения объекта. Совокупность всех фактов, связанных с одним объектом, описывает ***состояние*** этого объекта в данный момент времени. Описание состояний всех объектов из КМПО в конкретный момент времени представляет собой описание всей ***ситуации*** в ПО в этот момент, а описание состояний какой-то части объектов отображает определенный ***фрагмент*** этой ***ситуации***. Состояние одного объекта тоже рассматривается как фрагмент ситуации. Набор соответствующих понятий-фактов составляет **содержание понятия-ситуации**.

Таким образом, общее понятийное пространство задачи представляет собой совокупность нескольких подпространств, таких как:

- **пространство объектов *X*** - набор понятий-объектов, описанных через понятия-свойства и понятия-отношения;
- **пространство действий *G*** - набор понятий-действий, возможных в данной ПО;
- **пространство ситуаций (ситуационное пространство) *S*** - набор понятий-ситуаций, описывающих состояния различных объектов в разные моменты времени.

Понятия-факты и понятия-ситуации формируются на основе описания понятий-объектов, поэтому можно сказать, что ситуационное пространство порождается пространством объектов.

В задачах принятия решений основным объектом исследования является его подпространство –

- **пространство проблемных ситуаций S^P**- совокупность понятий-ситуаций, формирующихся в результате возникновения событий - инцидентов, нарушающих нормальное течение моделируемых процессов. Такие ситуации требуют специального анализа и принятия решений по их устранению.

Решение – это определенная последовательность действий, снимающая проблемную ситуацию и превращающая ее в ситуацию, соответствующую нормальному течению процесса. Таким образом, на основе пространства элементарных действий формируется еще одно пространство –

- **пространство решений *U*** - совокупность понятий-решений, описывающих последовательности действий, необходимых для устранения проблемных ситуаций в данной задаче. В **содержание понятия-решения** входит набор ситуаций, к которым применимо данное решение, и набор составляющих его действий.

Согласно данной методике, каждое действие описывается через имя и набор объектов: объект, над которым совершается действие, субъект, который его выполняет, и компоненты – другие объекты, которые в нем участвуют. Эти объекты определяются множеством свойств и отношений, из которых формируются усло-

вия и результаты данного действия. Все это вместе составляет **концептуальную структуру действия**, которая изображена на рисунке 23.4.

Таким образом, работая по данной методике, студент должен последовательно проанализировать процессы, происходящие в выбранной им ПО, т.е. все действия, связанные с принятием ею различных решений. Каждое ***действие*** описывается через имя и набор объектов: ***объект***, над которым совершается действие, ***субъект***, который его выполняет, и ***компоненты*** – другие объекты, которые в нем участвуют. В результате формируется определенное **понятийное пространство** соответствующей ПО, содержащее **пространство объектов** и **пространство действий**.

Пространство объектов представляет собой описание объектов задачи через их ***свойства*** и ***отношения***, в которые они могут вступать между собой. **Пространство действий** отражает содержание всех действий, выполняемых объектами ПО. Под ***содержанием действия*** при этом понимаются те изменения в значениях свойств и отношений объектов, которые происходят в результате выполнения данного действия.

Пространство объектов порождает **пространство состояний** ПО, а пространство действий – **пространство решений**, возможных в данной задаче.

Полученное понятийное пространство в дальнейшем может служить основой для более глубокого исследования ПО, моделирования происходящих в ней процессов, а также для проведения логического вывода и оценки качества принимаемых решений посредством анализа их воздействия на ПО. В настоящем лабораторном практикуме оно используется для построения продукционной базы знанийразрабатываемой ИС.

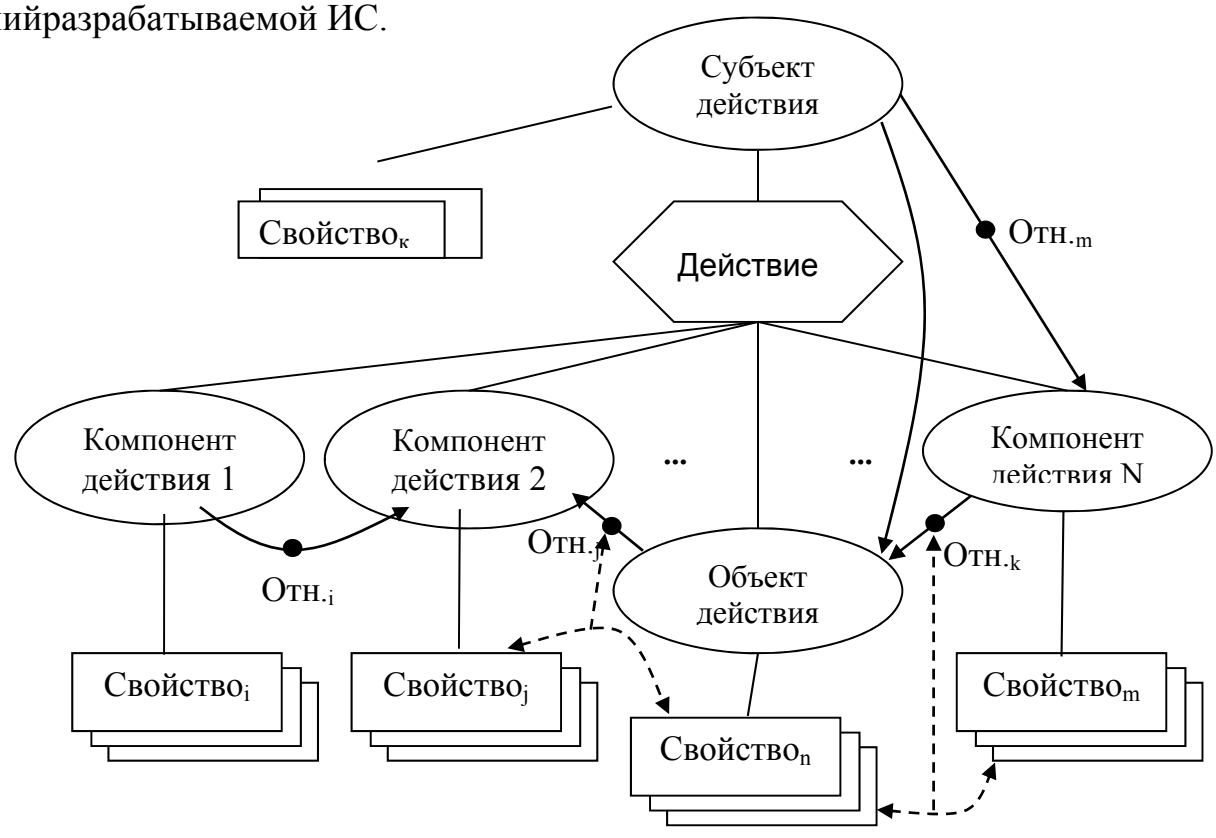


Рисунок 23.4 - Концептуальная модель действия

Таким образом, методика анализа структуры действий состоит из следующих шагов:

1. Определить действие (процесс), соответствующее постановке задачи (в общем случае таких действий может быть несколько, тогда они анализируются последовательно).
2. Проанализировать его содержание и условия выполнения, на основе этого определить структуру данного действия (т.е. выделить субъект действия, объект действия, возможные компоненты действия), свойства соответствующих объектов и отношения между ними.
3. Проанализировать условия выполнения данного действия и на основе этого определить возможные поддействия данного действия.
4. Последовательно анализировать поддействия до получения конечных (элементарных) действий.
5. Добавить новые свойства и отношения в модель, если понятия в структурах действий могут быть различной степени общности.
6. Добавить новые свойства и отношения, если объем каких-либо понятий-объектов больше единицы.

Порядок выполнения работы

В данной работе студент должен выполнить следующее:

1. Согласовать с преподавателем задачу для выполнения лабораторных работ.
2. На основании приведенного в Приложении 1 описания методики системного анализа, лекционного материала и руководства пользователя по системе «Помощник Эксперта» построить КМПО выбранной задачи.
3. Сохранить полученную модель в виде файла требуемого формата.

Результатом выполнения работы является:

- графические образы концептуальных моделей элементарных и производных решений;
- концептуальная модель ПО поставленной задачи;
- формализация этой модели в виде понятийных структур, порождаемых методикой системного анализа.

Примеры выполнения лабораторных работ

Задача об обезьяне и бананах

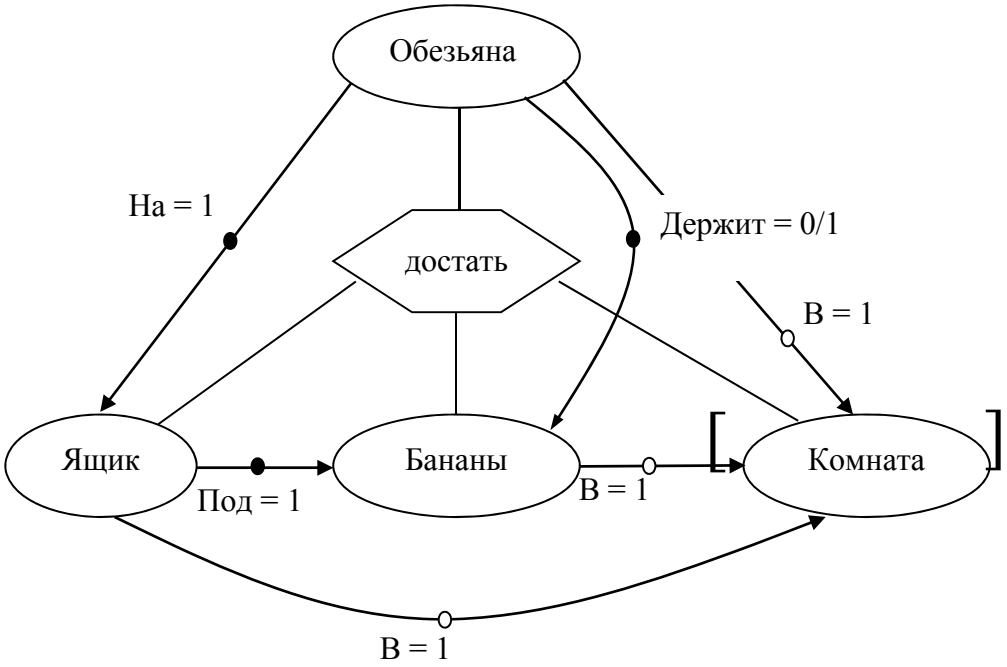
1. Системный анализ предметной области задачи и построение ее концептуальной модели

На этапе постановки задачи известно, что в комнате находятся обезьяна, ящик и бананы, причем бананы подвешены на недосягаемой с пола высоте. Спрашивается, какую последовательность действий должна совершить обезьяна, чтобы достать бананы?

Из условия задачи сразу видно как называется моделируемый процесс, кто его субъект, на какой объект он направлен, какие компоненты требуются для его выполнения.

Обратите внимание, что в данном случае выражение «достать бананы» может описывать как саму задачу, алгоритм решения которой подразумевает выполнение нескольких действий, так и действие, выполняющееся в этом алгоритме последним, и приводящее к целевой ситуации.

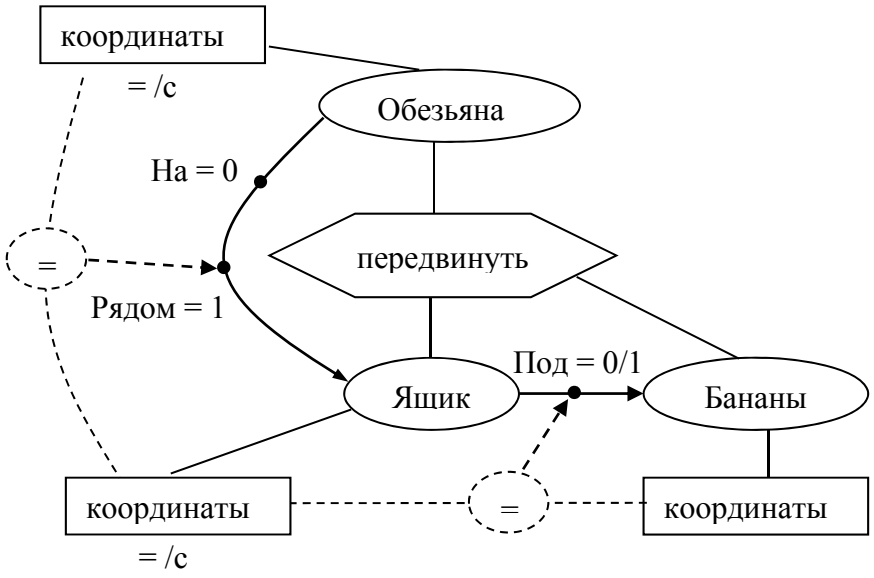
Построим структуру действия «Обезьяна достать Бананы».

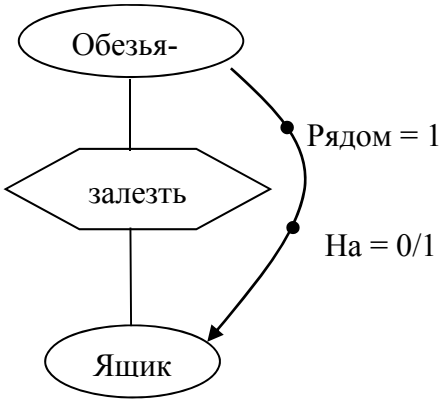


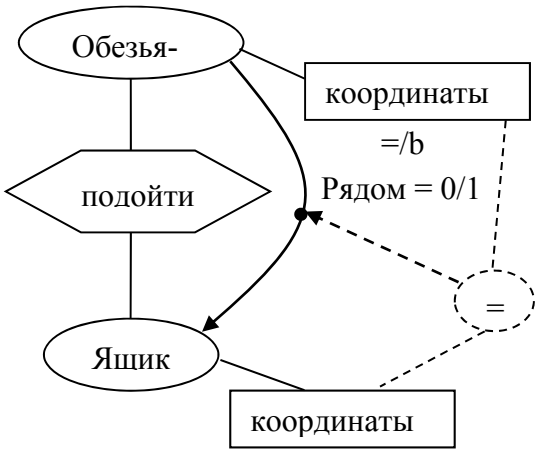
На данной структуре в формализованной форме отражена ситуация, когда обезьяна может достать бананы (т.е. взять их) и ситуация-результат этого действия. Обезьяна может взять бананы, когда она находится на ящике, т.е отношение На (О, Я) = 1, который стоит под бананами, т.е. отношение Под (Б, Я) = 1 и еще не держит бананы (Держит (О,Б)=0.

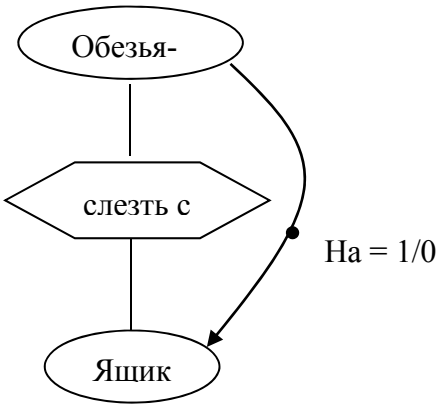
Эти факты отражены на структуре оценкой истинности соответствующих отношений: «На», «Под», «Держит». В результате реализации действия возникает новая ситуация, в которой отношение «Держит» становится равным истине. Также на приведенной структуре показан объект «Комната» и выделены отношения «В» между всеми остальными объектами и комнатой. Объект «Комната» и отношения «В» раскрывают контекст моделируемого процесса, и незримо присутствуют при реализации всех действий.

Поэтому, если помнить об их присутствии, то от них можно абстрагироваться. Факт наличия неизменяемых объектов, свойств и отношений отражается на структурах действий специальными символами: объекты берутся в квадратные скобки, отношения выделяются пустым кружочком, свойства обозначаются прямоугольником с двойной рамкой. По аналогии с действием «достать» раскроем структуры остальных действий.

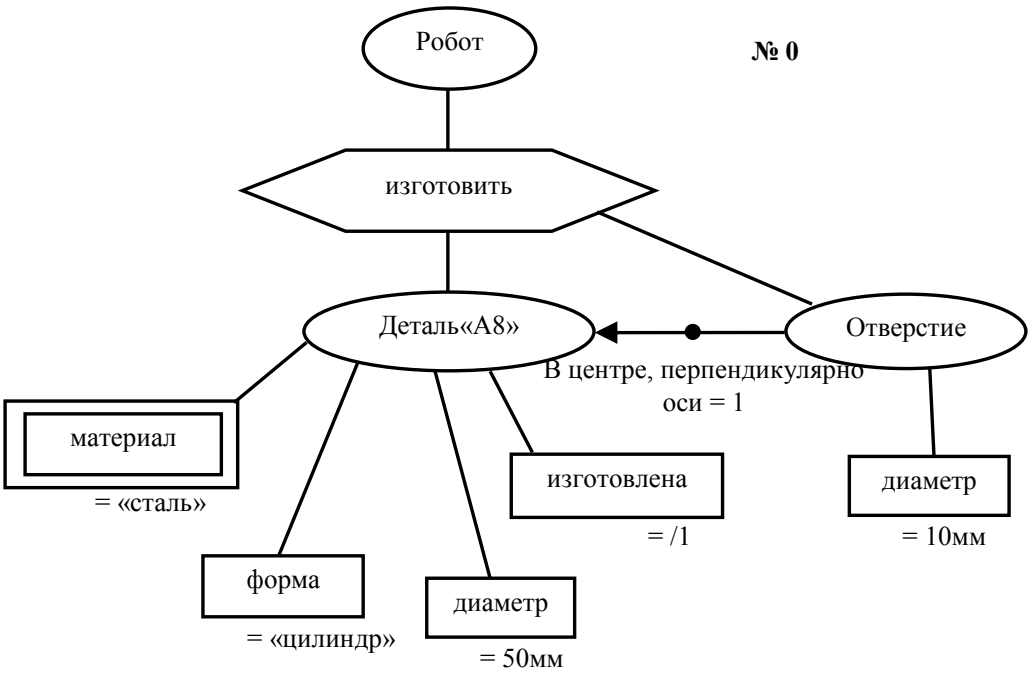

координаты
= /c
Обезьяна
На = 0
=
передвинуть
Рядом = 1
Под = 0/1
Ящик
Бананы
координаты
=
координаты
= /c


Обезья-
Рядом = 1
залезть
На = 0/1
Ящик


Обезья-
координаты
=/b
подойти
Рядом = 0/1
=
Ящик
координаты



Последнее действие добавлено только для иллюстрации. Точно также могут быть добавлены новые действия, расширяющие исходную модель. Например, такие: подойти к дивану, сесть на диван, съесть бананы. Можно еще более усложнить задачу, например, так: добавить еще одну обезьяну, не фиксировать точку, где находятся бананы, подвесить еще связку яблок и т.п. Раскрывая структуры действий, мы увидим, как КМПО будет пополняться новыми объектами, свойствами и отношениями.



На следующем шаге строиться понятийно-объектная модель (ПОМ) предметной области задачи с помощью той же программы – ПЭ.

Задача об управлении роботом

Рассмотрим процесс выполнения лабораторных работ на примере разработки прототипа экспертной системы ситуационного управления роботом (ЭССУР), ко-

торый должен изготавливать определенные детали. Предполагается, что робот воспринимает внешний мир посредством датчиков, информация с которых агрегируется в переменные, соответствующие понятиям предметной области, и описывающиетекущую ситуацию. Далее эта ситуация обрабатывается механизмом логического вывода ЭССУР на основе базы знаний с целью выработки необходимых управляющих воздействий, которые сообщаются эффекторам робота.

Данный пример приводится исключительно в иллюстративных целях и не претендует на адекватное описание технологического процесса, поскольку в действительности это описание будет зависеть от множества конкретных производственных условий, применяемого оборудования, от качества используемых источников знаний – экспертов, инструкций и т.д., а также от профессионализма разработчиков ЭССУР.

Кроме того, проиллюстрируем на этом примере работу инженера когнитолога, т.е. специалиста по разработке прикладных экспертных систем, с экспертом, т.е. токарем.

Системный анализ и построение концептуальной модели предметной области

Представим, что в качестве эксперта выступает токарь, который ранее изготавливал рассматриваемые детали, и мы интервьюируем эксперта по нашей методике. Предположим также, что единственное, что инженер – когнитолог предварительно знает об анализируемой предметной области, это то, что речь идет о некотором технологическом процессе.

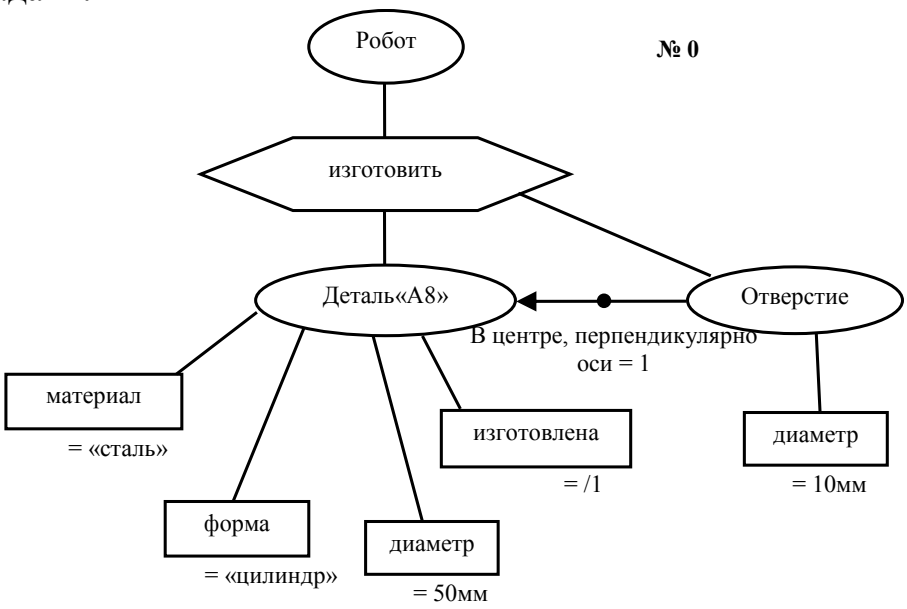
Введем обозначения: **ИК** – инженер-когнитолог; **Т** – токарь; *курсивом* – будем обозначать комментарии, поясняющие цель вопросов когнитолога на различных этапах системного анализа, ***жирным курсивом*** будем выделять имена выявленных понятий.

Процесс диалога с экспертом мог бы быть примерно следующим.

1. Построение структуры действия, соответствующего постановке задачи.
1.1. Определение целевой ситуации, т.е. результата технологического процесса.

ИК: На изготовление чего направлен рассматриваемый технологический процесс?
Т: В результате должна быть изготовлена деталь «А-8».
1.2. Выявление признаков достижения целевой ситуации.
ИК:Опишите, что собой представляет деталь «А-8»?
Т: Это стальной цилиндр, диаметром 50 мм., в центре которого перпендикулярно оси вращения просверлено отверстие диаметром 10мм.
1.3. Выделение понятий и построение структуры действия.
В приведенном выше диалоге мы выяснили, что речь идет об изготовлении ***детали «А-8»*** и ***токарь*** приходит к выводу, что эта деталь ***изготовлена***, если ***материалом*** детали является ***сталь***, если по ***форме*** деталь представляет собой ***цилиндр диаметром 50 мм.,в центре, перпендикулярно оси*** его вращения расположено ***отверстиедиаметром 10 мм.***

К такому же выводу должна прийти и ИСУР, которая должна принимать решения на уровне эксперта-токаря. Поэтому в качестве субъекта принимаемых решений будем подразумевать робота, как носителя ИСУР.
Выделенные понятия образуют следующую структуру действия, соответствующую постановке задачи.



Обратите внимание, что эксперт не произносил слов «материал», «форма», «изготовлена». Они возникли в результате анализа приведенного выше диалога инженером - когнитологом. Эксперт сказал «это стальной цилиндр», а ИК определил взаимосвязи понятий «сталь», «цилиндр» и «Деталь «А-8»». В результате анализа ИК выяснил, что «сталь» характеризует «материал», из которого изготавливается деталь, а «цилиндр» описывает ее «форму».

Поэтому «материал» и «форма» стали понятиями-свойствами, а «сталь» и «цилиндр» их значениями.
Целевое свойство «изготовлена» было введено когнитологом потому, что технологический процесс подразумевает изготовление детали и этот процесс будет завершен, когда деталь будет изготовлена. Деталь может быть либо уже изготовлена, либо еще нет, поэтому данное свойство является логическим.

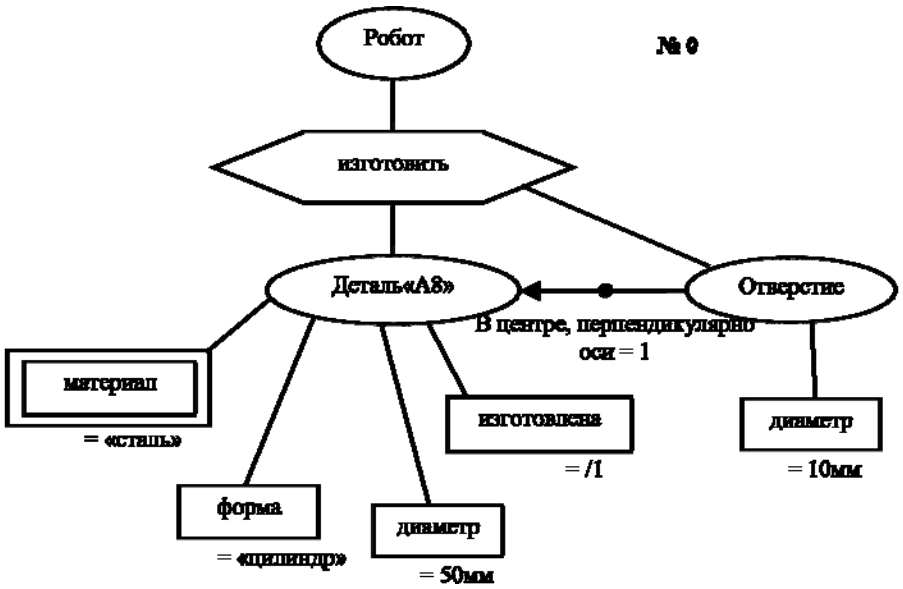
В структуре, соответствующей постановке задачи, сформулировано название всего моделируемого процесса - «Робот изготовить деталь», указан формальный признак завершения процесса - «деталь изготовлена», а также раскрыто представление об изготовленной детали через ее объективные признаки.

В данном случае целевую структуру можно понимать как умозаключение, к которому приходит токарь, когда наблюдает все характерные признаки готовности детали: «Деталь А-8 изготовлена, ЕСЛИ материал детали - сталь, форма детали – цилиндр, диаметр детали – 50мм., отверстие находится в центре перпендикулярно оси детали, диаметр отверстия – 10мм». Поскольку структура соответствует постановке задачи, то ей присвоен нулевой уровень.
2. Анализ условий действия «Робот изготовить Деталь «А-8»», выявление его поддействий.

Теперь отправной точкой для дальнейшего анализа являются выявленные и формализованные ранее условия действия. Относительно каждого условия мы должны выяснить способ его образования. Таких способов может быть три:
- условие формируется другими действиями, выполняющимися раньше рассматриваемого;
- условие статично и неизменно в течение всего анализируемого технологического процесса, т.е. раскрывает контекст протекания процесса;
- условие динамично, но неизменно в течение всего анализируемого технологического процесса, т.е. вариант реализации процесса определяется этим условием, задаваемым изначально;
- условие определяется исходными данными.

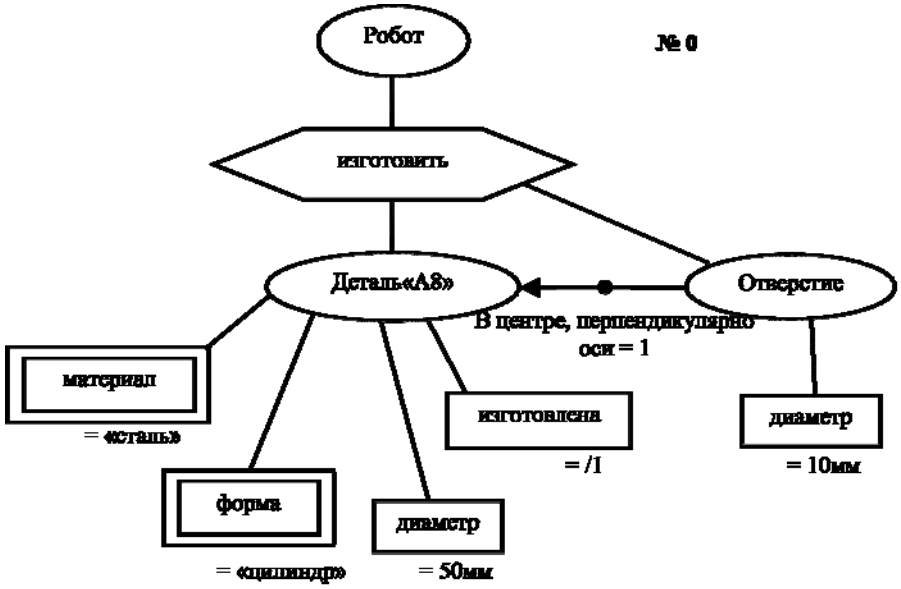
2.1. Анализ условия «материал детали - сталь»
ИК: Уточните, всегда ли деталь «А-8» изготавливается из стали?
Т: Да, всегда.

Таким образом, данное условие статично и неизменно, его анализ не выводит нас на новые действия, поэтому, подразумевая в дальнейшем, что деталь изготавливается из стали, от данного условия можно абстрагироваться. Отмечаем этот факт на структуре, выделяя свойство «материал» двойной рамкой.



2.2. Анализ условия «форма детали - цилиндр»
ИК: Каким образом формируется цилиндрическая форма детали?
Т: Путем обтачивания на токарно-фрезерном станке цилиндрических заготовок до получения нужного диаметра.

Выяснилось, что цилиндрическую форму детали также можно считать статичным и неизменным свойством, а диаметр детали меняется в результате реализации действия ***обточить***. Отобразим неизменность формы на структуре нулевого уровня.



2.3. Анализ условий «Отверстие в центре, перпендикулярно оси Детали» и «Диаметр отверстия – 10 мм.»

ИК: Каким образом отверстие диаметром 10 мм. получается в центре детали, перпендикулярно ее оси?

Т: Оно просверливается сверлом соответствующего диаметра.

ИК: А что значит «соответствующего диаметра»?

Т: Меньше диаметра отверстия на 0,2 мм.

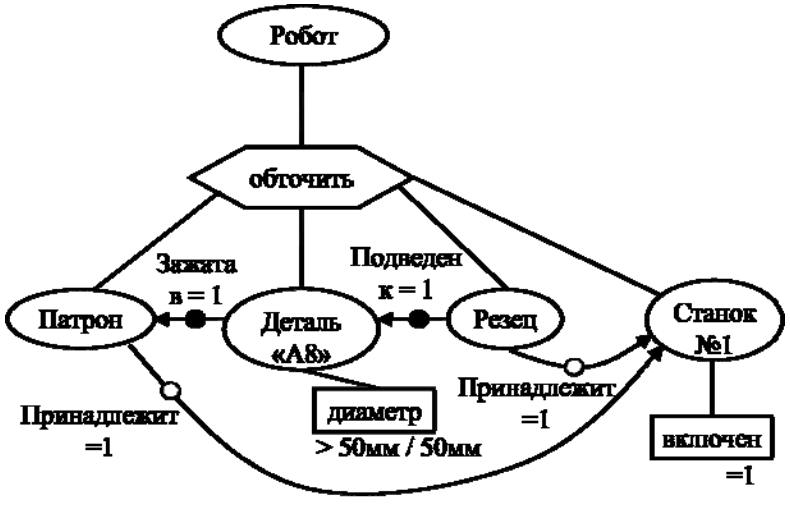
Таким образом, анализ условий действия «Робот изготовить Деталь» позволил выйти на действия следующего уровня – «Робот ***обточить*** Деталь» и «Робот ***просверлить*** Деталь». Проанализируем теперь их.

3. Построение структуры действия «Робот обточить Деталь»

ИК: Каким образом происходит обтачивание детали?

Т: Деталь ***зажимается впатронестанка №1***, ***резец подводится к*** детали, ***станок №1 включается***.

Обратите внимание, что данный ответ эксперта позволяет нам как сразу формализовать условия свершения действия «Робот обточить Деталь», так и выявить его поддействия – «Робот зажать Деталь», «Робот подвести Резец», «Робот включить Станок №1».



Структура читается следующим образом: «Если диаметр детали больше 50 мм, и деталь зажата в патроне станка №1, и резец подведен к детали, то деталь может быть обточена до диаметра 50мм»

4. Построение структуры действия «Робот просверлить Деталь»

ИК: Что нужно для того, чтобы просверлить деталь сверлом?

Т: Деталь необходимо предварительно обточить, далее она фиксируется держателем от возможного проворачивания в патроне, сверло позиционируется над местом сверления.

ЗАДАНИЯ

Каждый студент выбирает или придумывает сам среду принятия решений для построения ее модели в разных типах моделей представления знаний. Когда среда принятия решений выбирается самостоятельно, студент обязательно должен согласовать тему с преподавателем. При возникновении затруднений можно взять тему из предлагаемых вариантов.

В качестве заданий на проект могут фигурировать:

- процессы (или их фрагменты) различных видов деятельности, в том числе технологические;
- проблемные ситуации, возникающие в процессе выполнения какой-либо деятельности, требующие своего разрешения;
- задачи осуществления выбора одной или нескольких из множества допустимых альтернатив с учетом нескольких критериев.

Первый тип тем характеризуется тем, что в моделируемом процессе можно выделить ряд действий, которые должны быть выполнены в определенной последовательности. Задача состоит в том, чтобы построить модель описания пространства состояний всех необходимых действий, в котором система будет самостоятельно выстраивать пути достижения требуемых целевых состояний.

Второй тип тем предполагает наличие проблемной ситуации и возможность выделения множества альтернативных выходов (действий) из данной проблемы. Модель такой ПО должна содержать описание этих действий и критерии предпочтения альтернатив в зависимости от исходных начальных состояний. В качестве основных видов деятельности для выделения проблемных ситуаций предлагаются муниципальное управление и сфера малого предпринимательства.

В темах третьего типа моделируется принятие решений о выполнении одного действия (или о его невыполнении) на заданном множестве альтернативных вариантов выбора.

1. Примеры тем – процессов:

1.1. Диагностика кассового аппарата;

1.2. Регистрация патента;
1.3. Ремонт комнаты;
1.4. Поиск обрыва в сети;
1.5. Установка системы сигнализации для защиты здания (или еще ч-л);
1.6. Сборка компьютера;
1.7. Приготовление борща;
1.8. Сборка дома из деталей детского конструктора;
1.9. Установка машины – сервера;
1.10. Замена велосипедной камеры;
1.11. Сдача экзамена на вождение автомобиля;
1.12. Прием заявки на обслуживание абонента;
1.13. Парковка автомобиля;
1.14. Экспедирование груза;
1.15. Диагностика состояния и заправка катриджа;
1.16. Ремонт копировального аппарата;
1.17. Диагностика и ремонт принтера;
1.18. Процесс электросварки;
1.19. Диагностика неисправностей холодильника;
1.20. Заправка бензобака автомобиля;
1.21. Установка машины в гараж;
1.22. Сертификация товара;
1.23. Долив бутылок в автоматизированном комплексе;
1.24. Диагностика неисправностей блока питания ПК;
1.25. Управление лифтом;
1.26. Обслуживание абонента в библиотеке;
1.27. Сортировка поездов на ж\д станции;
1.28. Разграничение прав пользователей сети;
1.29. Диагностика вируса в ПК;
1.30. Управление освещением дома;
1.31. Посадка цветка;
1.32. Мойка автомобиля;
1.33. Диагностика травмы;
1.34. Пересадка цветка;
1.35. Диспетчерское сопровождение самолета;
1.36. Игра «Крестики – нолики»;
1.37. Игра «Морской бой»;
1.38. Модернизация ПК;
1.39. Игра в бильярд (ситуации из 3-4 шаров);
1.40. Установка связи между абонентами и станцией;
1.41. Диагностика ОС;
1.42. Обновление библиотеки из Интернета;
1.43. Поиск информации в Интернете по запросу;
1.44. Поиск свободного места на складе;
1.45. Уход за домашним цветком;
1.46. Распределение пакетов в ЛВС;

1.47. Управление автомобилем при приближении к перекрестку;
1.48. Управление движением на Т-образном перекрестке.
1.49. Чрезвычайная ситуация в районе (городе);
1.50. Сопровождение заказа менеджером;
1.51. Управление ЛВС.

2. Примеры тем – проблемных ситуаций при управлении городом или предприятием.

2.1. Управление городом. Проблемная ситуация: нехватка учителей в школах;
2.2. Управление городом. Проблемная ситуация: нехватка жилья для населения;
2.3. Управление городом. Проблемная ситуация: молодежная наркомания;
2.4. Управление городом. Проблемная ситуация: загромождение улиц парковкой автомобилей;
2.5. Управление городом. Проблемная ситуация: низкий уровень доходов у населения;
2.6. Управление городом. Проблемная ситуация: плохое качество продуктов на рынках;
2.7. Управление городом. Проблемная ситуация: повышенная криминогенность;
2.8. Управление городом. Проблемная ситуация: много праздношатающейся молодежи;
2.9. Управление городом. Проблемная ситуация: плохое качество работы ЖКХ;
2.10. Управление городом. Проблемная ситуация: плохое экологическое состояние;
2.11. Управление городом. Проблемная ситуация: грязные улицы;
2.12. Управление городом. Проблемная ситуация: недостаточное количество услуг населению (по видам);
2.13. Управление городом. Проблемная ситуация: очень маленький бюджет;
2.14. Управление городом. Проблемная ситуация: высокая безработица; и т.п.
2.15. Управление предприятием. Проблемная ситуация: плохая покупаемость оборудования фирмы;
2.16. Управление предприятием. Проблемная ситуация: сокращение прибыли фирмы;
2.17. Управление предприятием. Проблемная ситуация: плохие отношения с налоговой инспекцией;
2.18. Управление предприятием. Проблемная ситуация: мало клиентов на фирме;
2.19. Управление предприятием. Проблемная ситуация: недовольство поставщиком товаров;

2.20. Управление предприятием. Проблемная ситуация: нет заказов на оргтехнику;
2.21. Управление предприятием. Проблемная ситуация: плохая трудовая дисциплина на фирме;
2.22. Управление предприятием. Проблемная ситуация: не работает сетевое окружение;
2.23. Управление предприятием. Проблемная ситуация: недостаточная реклама товара;
2.24. Управление предприятием. Проблемная ситуация: не запускается программа;
2.25. Управление предприятием. Проблемная ситуация: необходимость привлечения персонала на работу;
2.26. Управление предприятием. Проблемная ситуация: появилась сильная конкуренция;
2.27. Управление предприятием. Проблемная ситуация: плохое качество связи в сети; и т.п.

3. Примеры тем – выбор альтернативы

3.1. Выбор двигателя для автомобиля;
3.2. Выбор подарка на день рождения;
3.3. Выбор и покупка телевизора;
3.4. Выбор варианта соединения линии связи;
3.5. Подбор комплектации ПК под заданные условия;
3.6. Выбор фирмы для поставки оборудования;
3.7. Выбор мебельного гарнитура;
3.8. Выбор маршрута движения от дома до института;
3.9. Выбор программного обеспечения для Web – сервера;
3.10.Подбор программного обеспечения для АСУТП;
3.11.Покупка автомашины;
3.12.Закупка ВТ;
3.13.Подбор тура для клиента;
3.14.Выбор работы;
3.15.Выбор и покупка квартиры;
3.16.Выбор ассортимента товара в магазин;
3.17.Подбор кандидата на вакансию;
3.18.Выбор проекта ЛВС;
3.19.Выбор направления предпринимательской деятельности;
3.20.Выбор Интернет – провайдера;
3.21.Выбор дачного дома;
3.22.Выбор игрушки для ребенка;
3.23.Выбор аппарата сотовой связи;
3.24.Выбор и покупка монитора;
3.25.Подбор варианта ПК;
3.26.Выбор резины для автомобиля; и т.п.

Вопросы для самоконтроля

- Понятия о концептуальной модели предметной области.
- Содержание понятия обьекта.
- Содержание понятия свойство.
- Содержание понятия действия.

Практическая работа № 24

Байесовские сети[20]

Цель работы – освоить принципы работы с байесовскими сетями как одного из вариантов создания экспертной системы в программе *Netica*.

Теоретические сведения

Теорема умножения предполагает, что

$P(AB) = P(A/B)P(B) = P(B/A)P(A)$

Отсюда

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

это известная формула Байеса.

Используя теоретико-множественный смысл события можно записать

$$B = (B \cap A) \cup (B \cap \overline{A})$$

тогда в силу независимости

$$P(B) = P(B \cap A) + P\ (B\ \cap \bar{A})$$

или

$$P(B) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A})$$

Теперь формулу Байеса можно переписать в виде

$$P(A/B) = \frac{P(B/A)P(A)}{P(B/A)P(A) + P\left(B/\overline{A}\right)P(\overline{A})}$$

В современных информационных технологиях эта формула используется как основа для управления неопределенностью – «делать выводы вперед и назад».

Основной современного логического вывода считается «если..., то...» правила. В нашем случае «если событие *H* истинно, то событие *E* будет наблюдаться с вероятностью *P*», а если событие *E* уже произошло, то какова вероятность истинности *H*?

H – событие, заключающееся в том, что данная гипотеза верна

E – событие, заключающееся в том, что наступило определенное доказательство (свидетельство), которое может подтвердить правильность указанной гипотезы.

Формула Байеса примет вид

$$P(H/E) = \frac{P(E/H)P(H)}{P(E/H)P(H) + P(E/\overline{H})P(\overline{H})}$$

[20]Разработано по материалам: Хабаров С. Экспертные системы.: Уч. пос.- ФЭУ, С-Пб. ЛТА

Она устанавливает связь гипотезы H и свидетельства E и в то же время свидетельства с неподтвержденной гипотезой. $P(H)$ – априорная вероятность гипотезы, известная до наступления события E.

Предполагается, что $P(H)$ и $P(E/H)$ находятся опытным или экспериментальным путем.

Формулу Байеса обобщают на случай множества гипотез ($H_1, H_2, \ldots H_m$) и множества свидетельств ($E_1, E_2, \ldots E_n$).

Вероятности каждой из гипотез можно определить по формуле

$$P(H_i / E_1E_2 \ldots E_n) = \frac{P(E_1E_2 \ldots E_n/ H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1E_2 \ldots E_n/ H_k)P(H_k)}$$

$i = 1, m$

Сложность формулы заключается в необходимости знать все условные вероятности знаменателя, поэтому часто делается довольно сильное предположение о независимости свидетельств (подход называют наивный Байес – naïveBayes). Тогда формула приобретает вид

$$P(H_i / E_1E_2 \ldots E_n) = \frac{P(E_1E_2 \ldots E_n/ H_i)P(H_i)}{\sum_{k=1}^{m} P(E_1/ H_i)P(E_2/ H_i) \ldots P(E_n/ H_k)P(H_k)}$$

Пример1. Имеется три взаимно независимые фирмы. Гипотезы

H_1 – «средняя надежность фирмы»

H_2 – «высокая надежность фирмы»

H_3 – «низкая надежность фирмы».

Имеется два условно независимых свидетельства, подтверждающих в разной степени исходные гипотезы.

$P(i)$	1	2	3
$P(H_1)$	0,6	0,4	0,1
$P(E_1/H_1)$	0,3	0,7	0,2
$P(E_2/H_2)$	0,6	0,8	0,0

Условно независимые свидетельства, поддерживающие исходные гипотезы:

E_1 – «наличие прибыли у фирмы»,

E_2 – «своевременный расчет с бюджетом».

Новые факты, получаемые в процессе сбора, будут повышать или понижать вероятности гипотез.

Пусть с вероятностью 1 наступило событие E_2, тогда апостериорные вероятности для гипотез согласно формуле Байеса для одного свидетельства:

$$P(H_i / E_2) = \frac{P(E_2/ H_i)P(H_i)}{\sum_{k=1}^{3} P(E_2/ H_k)P(H_k)}$$

$i = 1, 2, 3.$

Имеем $P(H_1/\,E_2) = \frac{0,6\cdot 0,6}{0,6\cdot 0,6+0,4\cdot 0,8+0,1\cdot 0,0} = \frac{0,36}{0,68} = 0,53$

$$P(H_2\,/\,E_2) = \frac{0,4\cdot 0,8}{0,6*0,6+0,4\cdot 0,8+0,1\cdot 0,0} = \frac{0,32}{0,68} = 0,47$$

$$P(H_3\,/\,E_2) = \frac{0,1*0,0}{0,6\cdot 0,6+0,4\cdot 0,8+0,1\cdot 0,0} = 0$$

После того как событие E_2 произошло доверие к гипотезе H_1 и H_3 понизилось, а доверие к H_2 возросло. Если есть факты, подтверждающие и событие E_1, и событие E_2, то при условии их независимости формула Байеса будет выглядеть в следующем виде:

$$P(H_i\,/\,E_1E_2) = \frac{P(E_1/\,H_i\,)P(E_2/\,H_i\,)P(H_i\,)}{\sum_{k=1}^{3} P(E_1/\,H_k\,)P(E_2/\,H_k\,)P(H_k)}$$

$i = 1, 2, 3.$

Таким образом

$$P(H_1\,/\,E_1E_2) = \frac{0,3\cdot 0,6\cdot 0,6}{0,3\cdot 0,6\cdot 0,6+0,7\cdot 0,8\cdot 0,4+0,2\cdot 0,0\cdot 0,1} = \frac{0,108}{0,332} = 0,325$$

$$P(H_2\,/\,E_1E_2) = \frac{0,7\cdot 0,8\cdot 0,4}{0,3\cdot 0,6\cdot 0,6+0,7\cdot 0,8\cdot 0,4+0,2\cdot 0,0\cdot 0,1} = \frac{0,224}{0,332} = 0,675$$

$$P(H_3\,/\,E_1E_2) = \frac{0,2\cdot 0,0\cdot 0,1}{0,3\cdot 0,6\cdot 0,6+0,7\cdot 0,8\cdot 0,4+0,2\cdot 0,0\cdot 0,1} = 0.$$

После получения свидетельств E_1 и E_2 осталось только две гипотеза H_1 и H_2. при этом H_2 более вероятно, чем H_1.

Данный пример иллюстрирует процесс распространения вероятностей по элементам экспертной системы (ЭС), основанной на Байесовских сетях, при поступлении в нее новых свидетельств. Можно показать что последовательное поступление свидетельств приводит к результатам аналогичным применению формулы Байеса для одновременно поступающих свидетельств.

Байесовская сеть доверия (БСД) – это направленный ациклический граф со следующими свойствами:

- каждая вершина – событие, описывающееся случайной величиной,
- вершины, связанные с «родительскими» определяются таблицей или функцией условных вероятностей,
- вероятности состояний вершин без «родителей» являются безусловными.

Таким образом, в байесовских сетях доверия вершины – случайные величины, дуги – вероятностные зависимости, определяющиеся таблицей или функцией условных вероятностей.

Следует отметить, что деревья, основанные на «если-то» правилах при наличии неоднозначных прецедентов автоматически переходят в байесовские сети.

Пример 2. *Построение байесовской сети доверия в программе* Hugin.

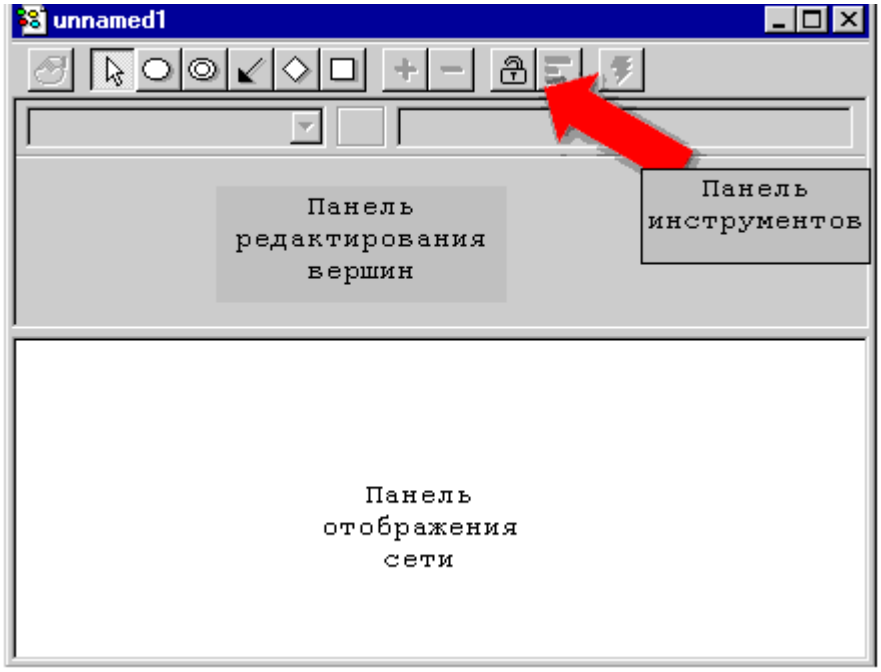


Рисунок 24.1 – Построение сети

После запуска HUGIN, откроется окно системы HUGIN, которое содержит панель меню, панель инструментов *(tool bar pane*), а также панель редактирования вершин (*node edit pane*) и панель графического отображения сети (*network pane*). Новая пустая сеть доверия получает имя *"unnamed1"* и автоматически открывается в окне сети (*network window*).

После запуска системы HUGIN (рис.24.1) она автоматически устанавливается в режим редактирования (*edit mode*), что позволяет немедленно приступить к построению новой БСД и определению состояния ее вершин.

Другим важным режимом работы системы является режим выполнения *(run mode)*, который позволяет использовать БСД для получения требуемых результатов.

Добавление новых вершин в проектируемую БСД

Построение БСД начинается с определения отдельных вершин, входящих в проектируемую БСД. При этом для создания вершин с дискретными состояниями, определения их свойств и задания причинно-следственных связей между вершинами используются, приведенные на Рисунок24.2, инструменты интерфейса системы, которые активируются путем нажатия левой кнопки мышки на их пиктограммах.

- вершины с дескретными состояниями *[discrete chance tool]*

- свойствавершин *[node properties tool]*

- добавление связей *[Link tool]*

Рисунок 24.2 – Инструменты интерфейса

Для построения рассматриваемого примера БСД, первое, что необходимо сделать – это создать вершину ***reliability*** (“надежность”). Для этого необходимо:

- Установить режим добавления вершин с дискретными состояниями, выбрав соответствующую пиктограмму в панели инструментов (рис. 24.2).
- Щелкнуть мышкой в любом месте панели отображения сети (рис. 24.1), где предполагается размещения добавляемой вершины.

После щелчка в панели отображения сети должна отобразиться в виде овала вершина, название которой по умолчанию будет **C1**. В соответствии с нашей моделью, необходимо этой вершине присвоить название ***reliability***. Для этого:

- Выделите курсором мыши нужную вершину.
- Установите режим определения свойств вершины [*Node Properties*], щелкнув на соответствующей пиктограмме панели инструментов (рис. 24.2).
- Измените содержимое полей **Name** = ***reliability*** и **Label** = **Надежность**.
- Нажмите "*OK*".

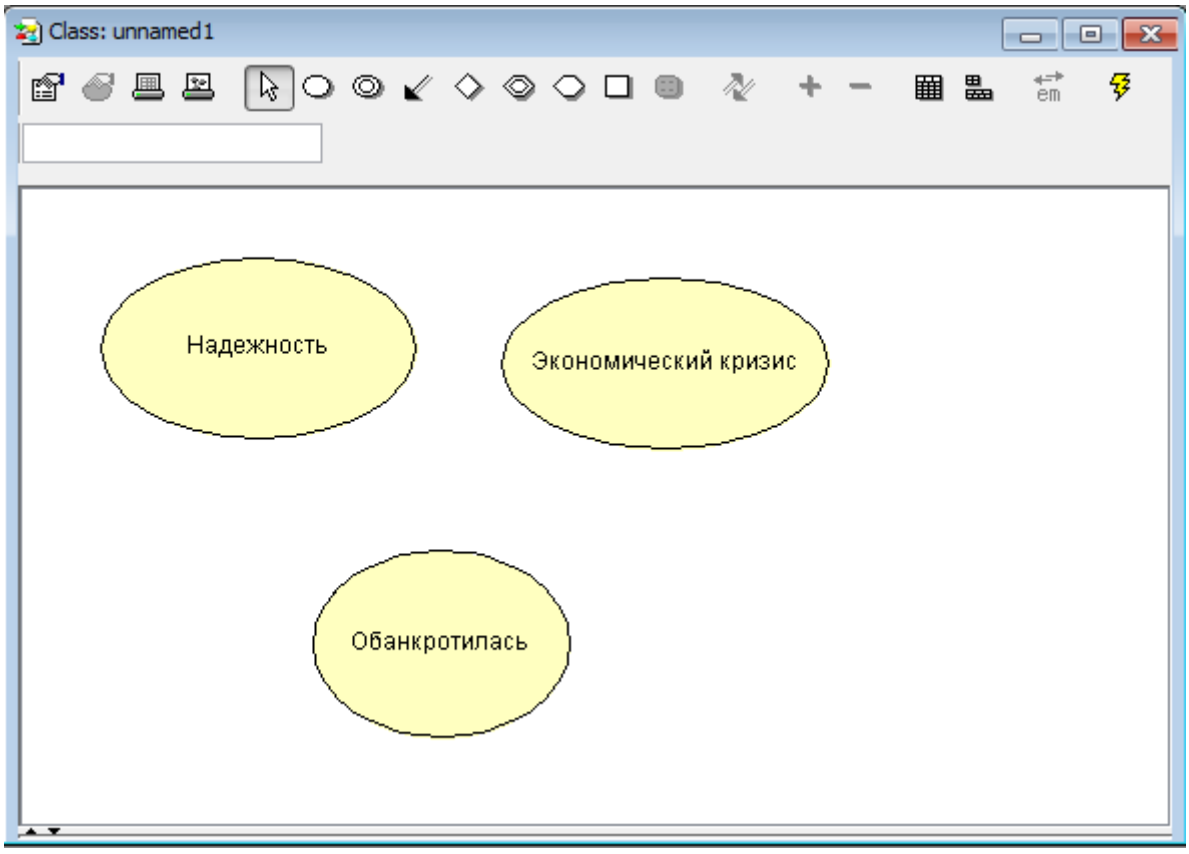


Рисунок 24.3 – Проектируемая БСД

Name - это внутреннее имя вершины, пока **Label** является названием вершины. Если поле **Label** не задано (как это было до того момента, пока мы не изменили поле **Label**), то его значение соответствует внутреннему имени. Внутреннее имя может содержать символы 'a' - 'z' и 'A' - 'Z', цифры '0' - '9', а также знак подчеркивания '_'.

Вершины ***crisis*** (“Экономический кризис”) и ***bankrot*** (“Обанкротилась”) добавляются в БСД таким же образом. Можно добавлять вершины не нажимая каж-

дый раз на пиктограмму вершин с дискретными состояниями, а однажды выбрав режим добавления и удерживая клавишу *SHIFT* щелкнуть мышью в окне отображения сети столько раз, сколько вершин с необходимо добавить в БСД.

Установление причинно-следственных связей между вершинами проектируемой БСД

На текущий момент проектируемая БСД имеет вид, похожий на тот, что приведен на Рисунок24.3. Следующий этап проектирования БСД состоит в установлении причинно - следственных связей между событиями. Эти связи в модели БСД отображаются в виде стрелок, соединяющих между собой вершины БСД. Для добавления стрелок от вершины ***Reliability*** к вершины ***Bankrot*** и от ***Crisis*** к ***Bankrot***, необходимо сделать следующее:

- Нажмите иконку добавления связей (рисунок 341).
- Протяните мышью стрелку от ***Reliabilit*** к ***Bankrot*** , нажав левую кнопку и удерживая нажатой клавишу *SHIFT*.
- Протяните мышью стрелку от ***Crisis*** к ***Bankrot*** нажав левую клавишу.

Теперь Вы имеете полное качественное представление, подобное изображенному на рисунке 24.4.

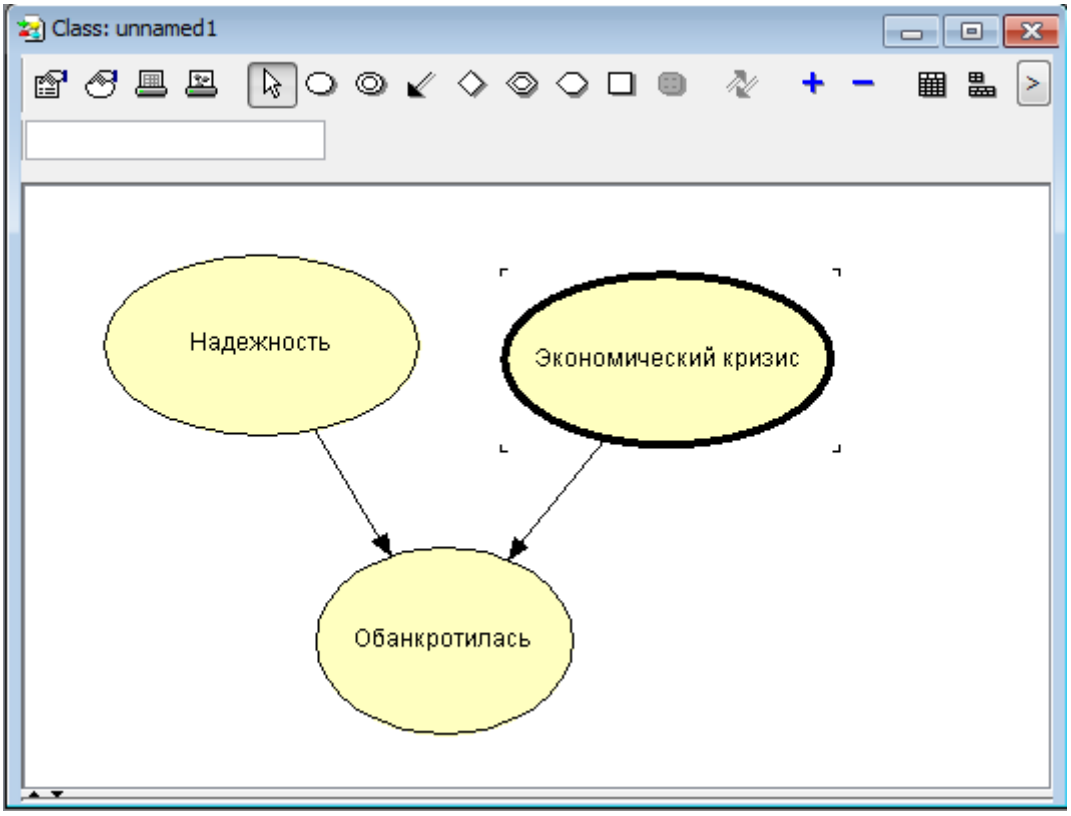


Рисунок 24.4 – Полное качественное представление

Следующий шаг - это задание состояний и таблиц условных вероятностей для каждой вершины.

Определение всех возможных состояний каждой из вершин БСД

Ранее, каждую из вершин мы определили так, что каждая из них может находиться в одном из двух состояний: вершина ***Reliabilit*** в состояниях ‘affected’ и ‘not affected’, вершина ***Crisis*** в состояниях ‘higt’ и ‘not higt’, а вершина ***Bankrot*** в со-

стояниях - ‘bankrot’ и ‘notbankrot’. Как задаются состояния каждой из вершин БСД в системе HUGIN рассмотрим на примере определения состояний вершины ***Reliabilit***. Для этого необходимо:

Рисунок 24.5

- Сделать активной вершину ***Reliabilit***, выбрав ее в раскрывающемся списке на панели инструментов или дважды щелкнув на ней мышью.
- Нажать пиктограмму добавления состояний на панели инструментов (рис. 24.5).
- Щелкнув мышью, перевести курсор в ТУВ на поле, содержащем текст **State 0** и ввести в него текст ‘affected’, нажать кнопку Rename. Затем **State 1** заменить на ‘notaffected’, задавая второе состояние вершины.

Самостоятельно проделайте аналогичные операции для вершины ***Crisis*** и ***Bankrot***.

Задание значений таблиц условных вероятностей каждой из вершин БСД

Следующий шаг проектирования БСД - это корректное задание значений ТУВ каждой из вершин.

Для перехода в режим редактирования таблиц условных вероятностей, необходимо щелкнуть правой кнопкой мыши на каждой из вершин и выбрать пункт ***OpenTables.***

Обратите внимание, что при определении вершины ***Bankrot***, таблица ее условных вероятностей будет отличаться от ТУВ вершин ***Reliabilit*** и ***Crisis***. Это связано с тем, что вершина Bankrot имеет родительские вершины, которых Reliabilit и Crisis не имеют

По умолчанию система HUGIN для всех вершин устанавливает равномерное распределение. Для нашего примера значения ТУВ были заданы табл. 24.1, 24.2, 24.3 и в новых обозначениях будут иметь вид:

Таблица 24.1

affected	0.1
not affected	0.9

Таблица 24.2

high	0.1
not high	0.9

Таблица 24.3

crisis	affected		not affected	
reliability	high	not high	high	not high
bankrupot	0.95	0.85	0.9	0.02
not bankrot	0.05	0.15	0.1	0.98

Процесс заполнения таблиц условных вероятностей рассмотрим на примере заполнения ТУВ для вершины ***Reliabilit***:

- Выберите вершину ***Reliabilit***.
- Щелкните на поле содержащем ***Reliabilit*** = ’**Reliabilit’**.
- Введите значение 0.1, соответствующее табл.4.
- Щелкните на поле содержащем ***Reliabilit*** = ’**not’**.
- Введите значение 0.9, соответствующее табл.4.

Таким же образом задайте значения для ***Crisis*** и ***Bankrot***. Когда будите вводить значения для ***Bankrot***, будьте внимательны, чтобы значения соответствовали тем, что даны в табл.24.3

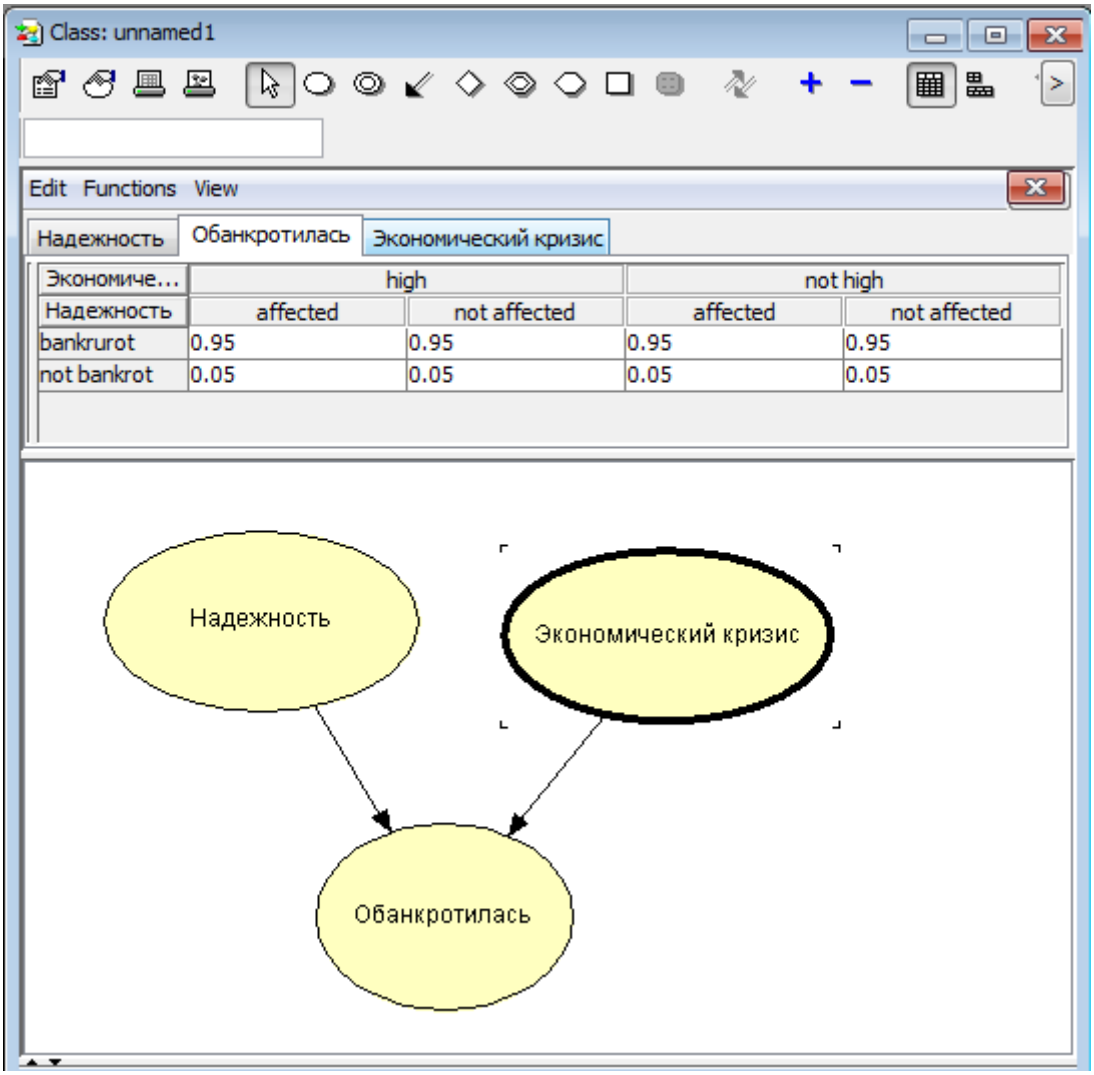


Рисунок 24.6 – Окно отображения сети

После задания таблиц условных вероятностей для всех вершин байесовской сети доверия, окно отображения сети будет выглядеть так, как показано на рисунке 24.6.

На этом конструирование проектируемой БСД заканчивается.

Теперь необходимо сохранить результаты работы. Это можно сделать следующим образом:

- Выберите "*Save As*" в меню "*File*".

•Введите имя файла (например *apple.hkb* или *lab1.hkb*).
•Нажмите "*Save*".

Компилирование спроектированной БСД и работа с ней

Рисунок 24.7

После того как БСД спроектирована и определена в системе HUGIN, ее необходимо скомпилировать и посмотреть как она работает. Для этого необходимо перевести систему в режим вычислений. С этой целью щелкните мышкой по пиктограмме режима вычислений [*run mode tool button*] в панели инструментов (рис. 24.7). Если Вы четко следовали описанию лабораторной работы, то ошибок компиляции не будет. Компиляция такой малой БСД, как наша, закончится быстро и система перейдет в режим вычислений [*"run" mode*].

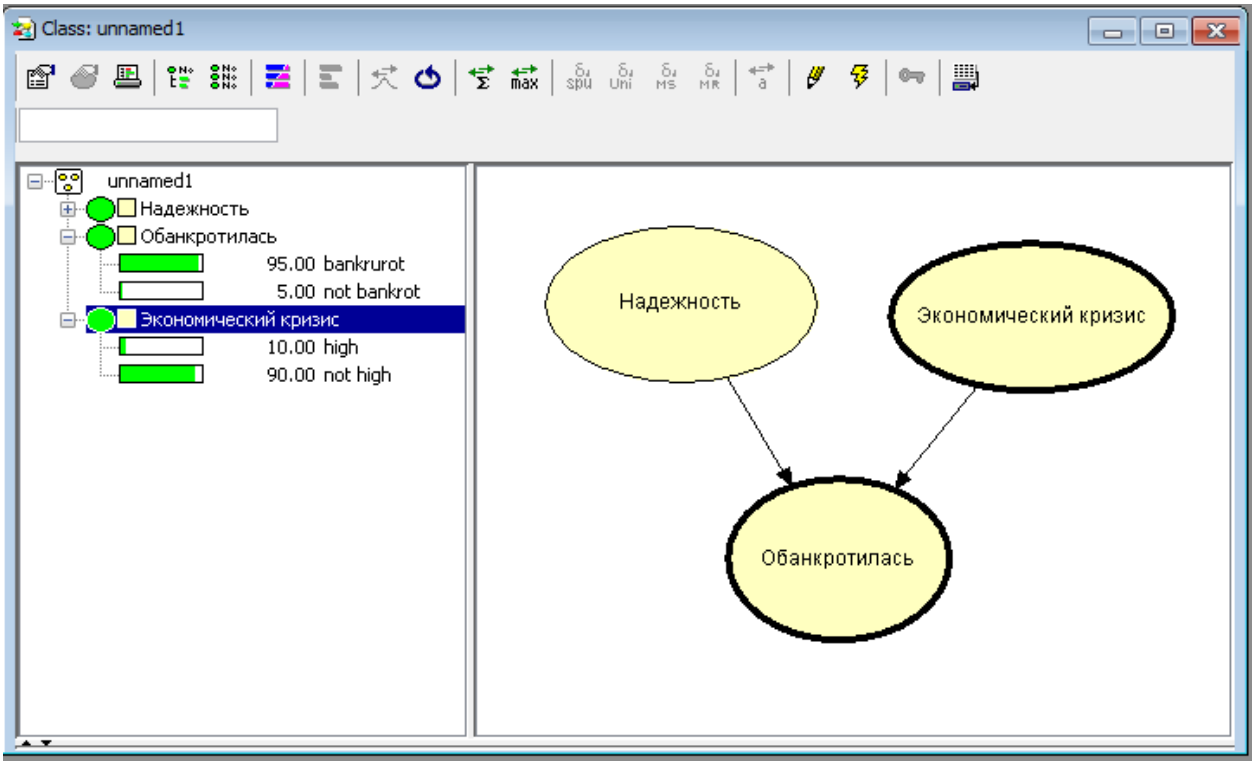


Рисунок 24.8 – Окно сети

В этом режиме окно сети поделено на две вертикальные секции (рис. 24.8). Левая секция представляет собой панель списка вершин [*node list pane*], а правая - панель отображения сети [*network pane*].

В левой части окна отображаются вершины БСД, все возможные состояния каждой из вершин, а также вероятности пребывания вершины в каждом из состояний. в секции списка вершин [*node list pane*]. Посмотреть эти вероятности можно дважды щелкнув на имени вершины в списке вершин. Теперь раскроем вершины ***Bankrot*** и ***Reliabilit***. Для этого дважды щелкните на имени вершины ***Bankrot***, а затем дважды щелкните на ***Reliabilit***. Вы также можете просмотреть вероятности

пребывания всех вершин во всех возможных состояниях, нажав пиктограмму раскрытия списка вершин [*expand node list tool*], вторую слева на панели инструментов.

4. Распространение вероятностей в БСД при поступлении новых свидетельств

Теперь, представим себе, что Вы хотите использовать БСД для нахождения вероятности банкротства фирмы, имея информацию о экономическом кризисе. Другими словами в базу знаний нашей экспертной системы добавляется новый факт банкротстве фирмы и нас интересует результат вывода ЭС о кризисе и о надежносте. Реализации такого типа запроса к экспертной системе на основе БСД выполняется следующим образом:

Рисунок 24.9

- Раскройте список всех вершин, нажав [*expand node list tool*].
- Введите факт банкротства дважды щелкнув на состоянии '**bankrot'** для вершины ***Bankrot***.
- Нажмите пиктограмму распространения (рис. 24.9) на панели инструментов для распространения данного факта на всю БСД.
- Посмотрите вероятность пребывания события ***Reliabilit*** в состоянии '**Reliabilit'**. Результаты распространения вероятностей по байесовской сети доверия приведены на рисунке 24.10.

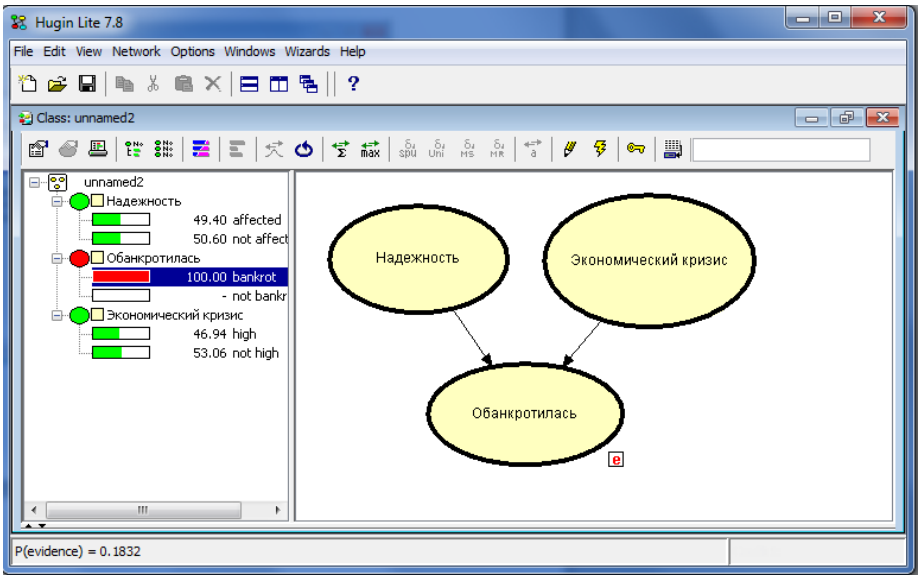


Рисунок 24.10 – Результаты распространения вероятностей

В примере – *пусть известно, что фирма обанкротилась (вероятностью 1 или на 100%)*. После этого можно узнать вероятность того, что повлиял экономический кризис. Для приведенных выше данных, опираясь на результаты вывода путем распространения сумм по БСД, можно показать, что кризис повлиял с вероятностью 0,469, а надежность фирмы с вероятностью 0,494.

Можно предположить, что на банкротство фирмы с вероятностью 0,531 не повлияет кризис и с вероятностью 0,506 не повлияет надежность фирмы. Однако этот вывод является преждевременным. Для нахождения наиболее вероятной комбинации состояния вершин в программе Hugin вместо распространения сумм нужно использовать распространение максимумов. Каждое из состояний вершин, имеющее значение 100,00 будет принадлежать к наиболее вероятной комбинации состояний. В рассматриваемом примере наиболее вероятно, что кризис не повлияет, а надежность фирмы будет высокая.

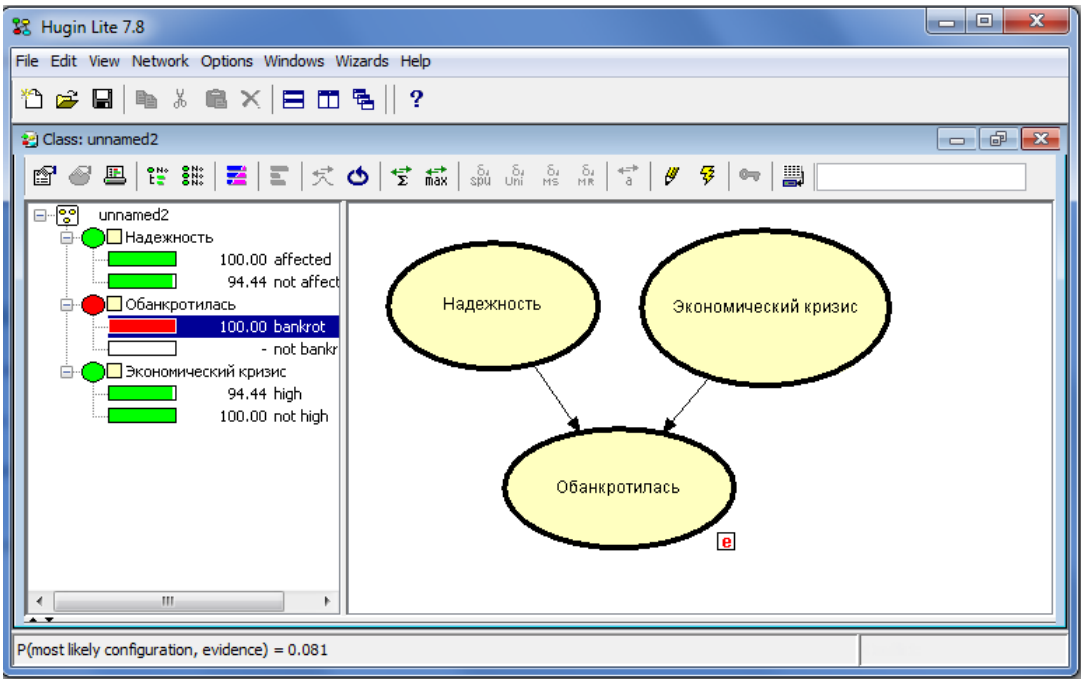


Рисунок 24.11 – Вероятная ситуация 1

Расчет вероятности комбинаций состояний

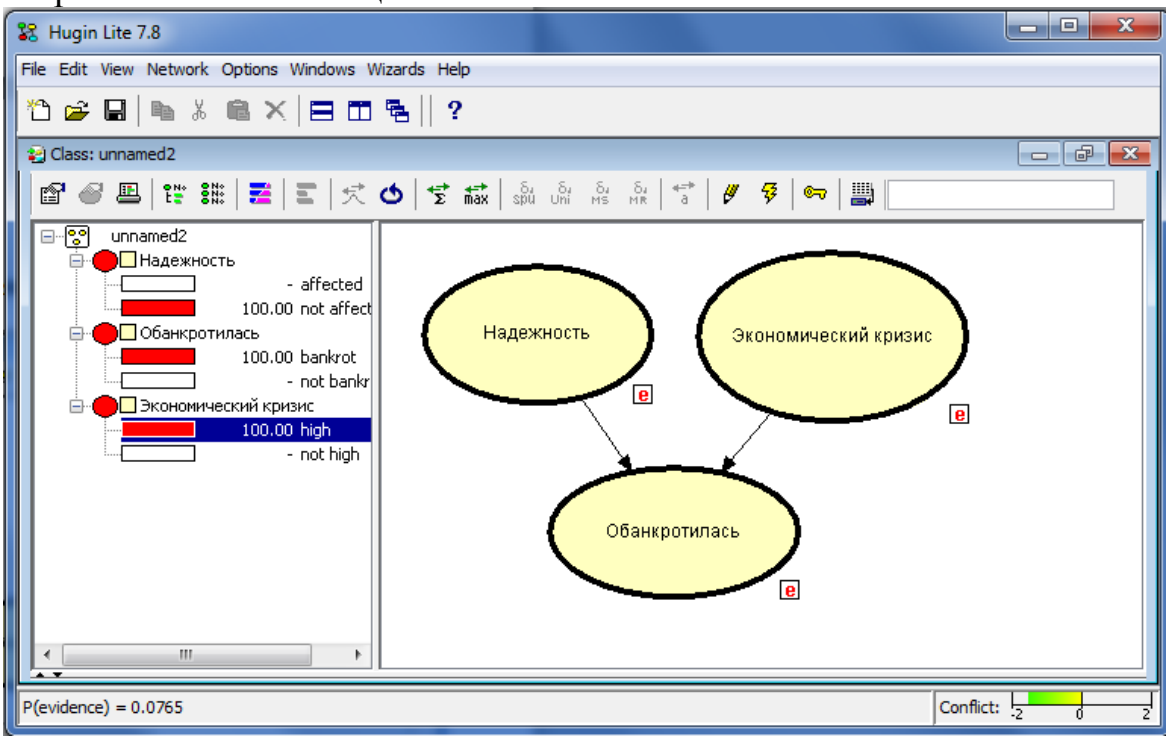


Рисунок 24.12 – Вероятная ситуация 2

Рассмотрим наиболее вероятную комбинацию состояний, полученных при наличии факта, что фирма обанкротилась.

P(reliability = «high», crisis = «not affected» / bankrupt= «bankrupt»)

Используемформулу$P(A|B) = P(AB)/P(B)$.Внашихобозначениях

P(reliability = «high», crisis = «not affected» / bankrupt)=

=P(reliability = «high», crisis = «not affected», bankrupt= «bankrupt»)/ P(bankrupt= «bankrupt»)=0,0765/0,1832=0,417.

Итак, наиболее вероятная ситуация, что кризис не повлиял и надежность фирмы была высокой имеет значение вероятности 0,417 (рис. 24.11-24.12). При разработке систем принятия решений и экспертных систем используют диаграммы влияния, к БСД добавляют вершины решения - decisions (прямоугольники) и вершины пользы – utility (ромбы) (рис. 24.13).

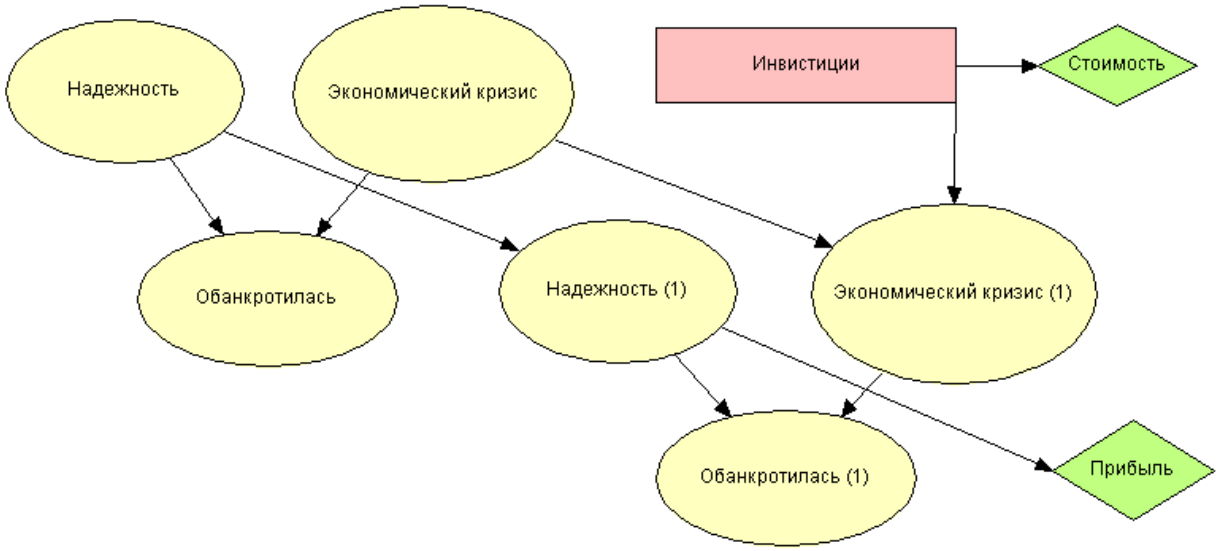


Рисунок 24.13 – Диаграммы влияния

Предположим, что мы для повышения надежности фирмы инвестируем 8000 д.е. или 0 д.е. и таблицы условных вероятностей для дополнительных переменных reliability(1) и crisis(2) (рис. 24.13), полученные на основе обработки знаний экспертов имеют вид, представленный на рисунке 24.14.

investment	investment		not	
reliability	high	not high	high	not high
high	0.2	0.99	0.99	0.02
not high	0.8	0.01	0.01	0.98

crisis	affected	not affected
affected	0.6	0.05
not affected	0.4	0.95

Рисунок 24.14 – Таблицы условных вероятностей

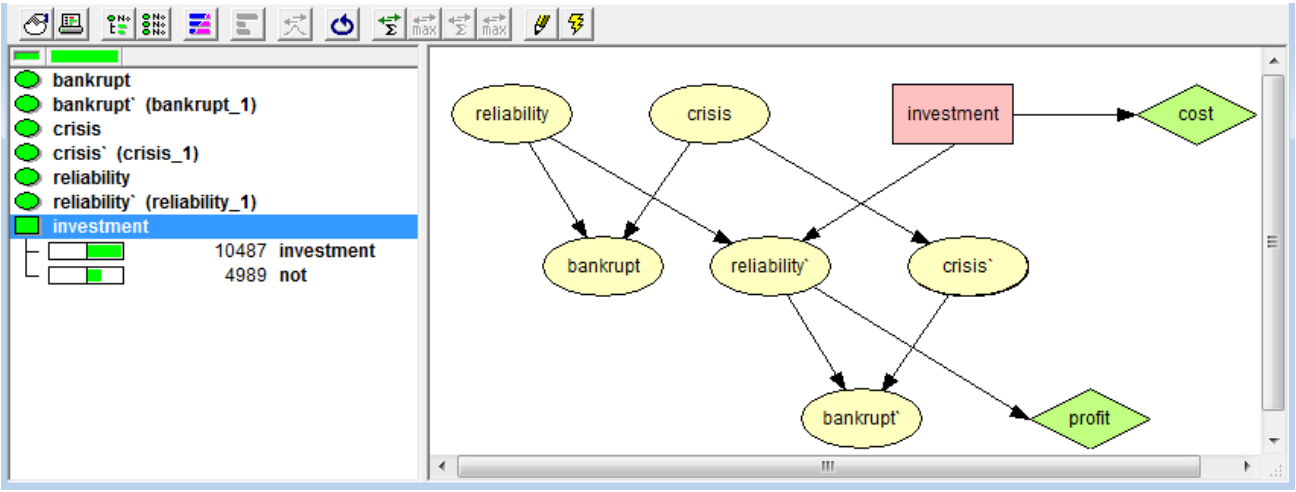


Рисунок 24.15 – Полезность инвестиций

Тогда полезность от инвестиций составляет 10487 д.е. и 4989 д.е. в противном случае (рис. 24.15). То есть инвестиции более предпочтительны.

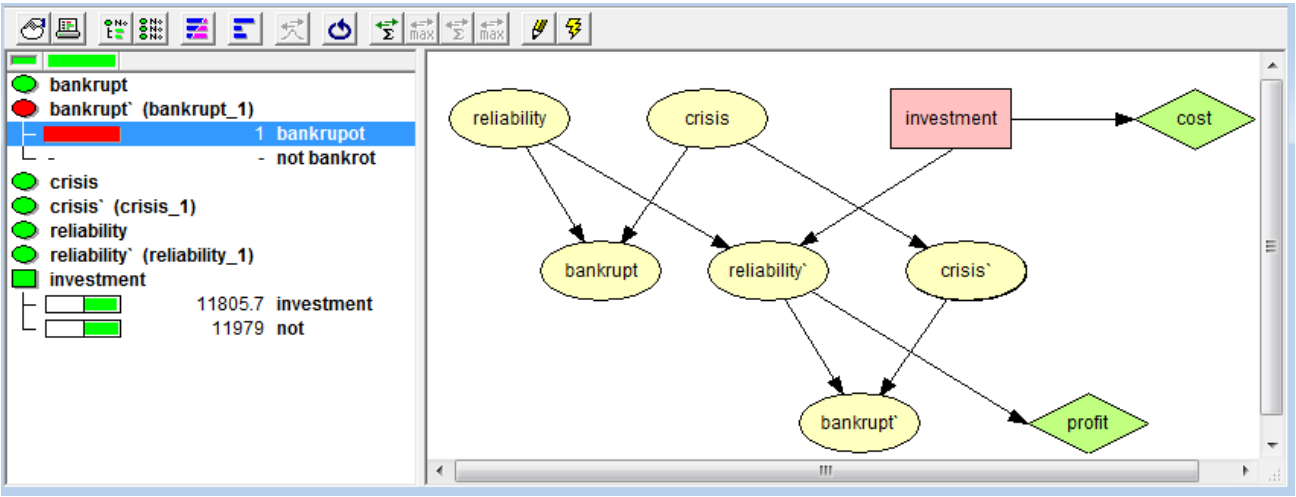


Рисунок 24.16 – Результативность инвестиций если фирма банкрот

Если получено свидетельство о том, что фирма банкрот, то результативность инвестиций 11805,7 д.е., а их отсутствия 11979 д.е., то есть инвестировать нет смысла (Рисунок24.16).

Сегодня байесовские сети доверия имеют широкое применение при разработке экспертных систем и систем принятия решений для оценки рисков в различных областях деятельности: медицине, финансах, коммерции и т.д.

ЗАДАНИЕ

Используя предложенные преподавателем (или собственные) вероятностные модели предметной области постройте байесовскую сеть доверия.

Практическое занятие № 25

Системно-когнитивный анализ изображений (обобщение, абстрагирование, классификация и идентификация)

Цель: рассмотреть применение системно-когнитивного анализа, его математической модели – системной теории информации и программного инструментария – системы «Эйдос» для синтеза обобщенных изображений классов, их абстрагирования, классификации обобщенных изображений (кластеры и конструкты) сравнения конкретных изображений с обобщенными образами (идентификация)

Теоретические сведения.

Универсальная когнитивная аналитическая система "Эйдос-X++" является современным инструментарием системно-когнитивного анализа, разработана в универсальной постановке, не зависящей от предметной области, и обеспечивает:

- формализацию предметной области;
- многопараметрическую типизацию, синтез, повышение качества и верификацию 3 статистических моделей и 7 моделей знаний предметной области;
- распознавание (системную идентификацию и прогнозирование);
- поддержку принятия решений и исследование модели, в т.ч.: дивизивную и агломеративную когнитивную кластеризацию, конструктивный и СК-анализ моделей: семантические и нейронные сети, когнитивные диаграммы, классические и интегральные когнитивные карты.

Есть в системе и ряд других новых возможностей. Переосмыслена иерархическая структура системы, учтен значительный опыт проведения научных исследований и преподавания ряда дисциплин с применением системы «Эйдос» и систем окружения.

Управление – это достижение цели путем принятия и реализации решений об определенных действиях, способствующих достижению этой цели. Цели управления обычно заключаются в том, чтобы определенная система, которая называется объектом управления, находилась в определенном целевом (желаемом) состоянии или эволюционировала по определенному заранее известному или неизвестному сценарию.

Действия, способствующие достижению цели, называются управляющими воздействиями. Решения об управляющих воздействиях принимаются управляющей системой. Управляющее воздействие вырабатывается управляющей системой на основе модели объекта управления и информации обратной связи о его состоянии и условиях окружающей среды.

Это означает, что система управления сложными динамичными объектами должна быть интеллектуальной, т.к. именно системы этого класса позволяют проводить обучение, адаптацию или настройку модели объекта управления за счет накопления и анализа информации о поведении этого объекта при различных сочетаниях действующих на него факторов. Таким образом, решив первую проблему, т.е. разработав технологию создания модели сложного объекта управления, мы этим самым создаем основные предпосылки и для решения и второй проблемы, т.к.

для этого достаточно применить эту технологию непосредственно в цикле управления.

Непосредственно на основе матрицы сопряженности (абсолютных частот) или с использованием матрицы условных и безусловных процентных распределений с использованием количественных мер знаний можно получить 7 различных моделей знаний, приведенных в таблице 25.1:

Таблица 25.1 – Различные аналитические формы частных критериев знаний в системе «Эйдос-х++»

Наименование модели знаний и частный критерий	Выражение для частного критерия	
	через относительные частоты	через абсолютные частоты
INF1, частный критерий: количество знаний по А.Харкевичу, 1-й вариант расчета вероятностей: Nj – суммарное количество признаков по j-му классу. Вероятность того, что если у объекта j-го класса обнаружен признак, то это i-й признак	$I_{ij} = \Psi \times Log_2 \frac{P_{ij}}{P_i}$	$I_{ij} = \Psi \times Log_2 \frac{N_{ij} N}{N_i N_j}$
INF2, частный критерий: количество знаний по А.Харкевичу, 2-й вариант расчета вероятностей: Nj – суммарное количество объектов по j-му классу. Вероятность того, что если предъявлен объект j-го класса, то у него будет обнаружен i-й признак.	$I_{ij} = \Psi \times Log_2 \frac{P_{ij}}{P_i}$	$I_{ij} = \Psi \times Log_2 \frac{N_{ij} N}{N_i N_j}$
INF3, частный критерий: Хи-квадрат: разности между фактическими и теоретически ожидаемыми абсолютными частотами	---	$I_{ij} = N_{ij} - \frac{N_i N_j}{N}$
INF4, частный критерий: ROI - Return On Investment, 1-й вариант расчета вероятностей: Nj – суммарное количество признаков по j-му классу	$I_{ij} = \frac{P_{ij}}{P_i} - 1 = \frac{P_{ij} - P_i}{P_i}$	$I_{ij} = \frac{N_{ij} N}{N_i N_j} - 1$
INF5, частный критерий: ROI - Return On Investment, 2-й вариант расчета вероятностей: Nj – суммарное количество объектов по j-му классу	$I_{ij} = \frac{P_{ij}}{P_i} - 1 = \frac{P_{ij} - P_i}{P_i}$	$I_{ij} = \frac{N_{ij} N}{N_i N_j} - 1$
INF6, частный критерий: разность условной и безусловной вероятностей, 1-й вариант расчета вероятностей: Nj – суммарное количество признаков по j-му классу	$I_{ij} = P_{ij} - P_i$	$I_{ij} = \frac{N_{ij}}{N_j} - \frac{N_i}{N}$
INF7, частный критерий: разность условной и безусловной вероятностей, 2-й вариант расчета вероятностей: Nj – суммарное количество объектов по j-му классу	$I_{ij} = P_{ij} - P_i$	$I_{ij} = \frac{N_{ij}}{N_j} - \frac{N_i}{N}$

Система «ЭйдосХ++» обеспечивает синтез и верификацию всех этих моделей знаний. При этом верификация (оценка достоверности) модели может осуществляться как с использованием всей обучающей выборки в качестве распознаваемой, так и с использованием различных ее подмножеств на основе бутстрепного подхода. Диалог режима синтеза модели и ее верификации приведен на рисунке 25.1:

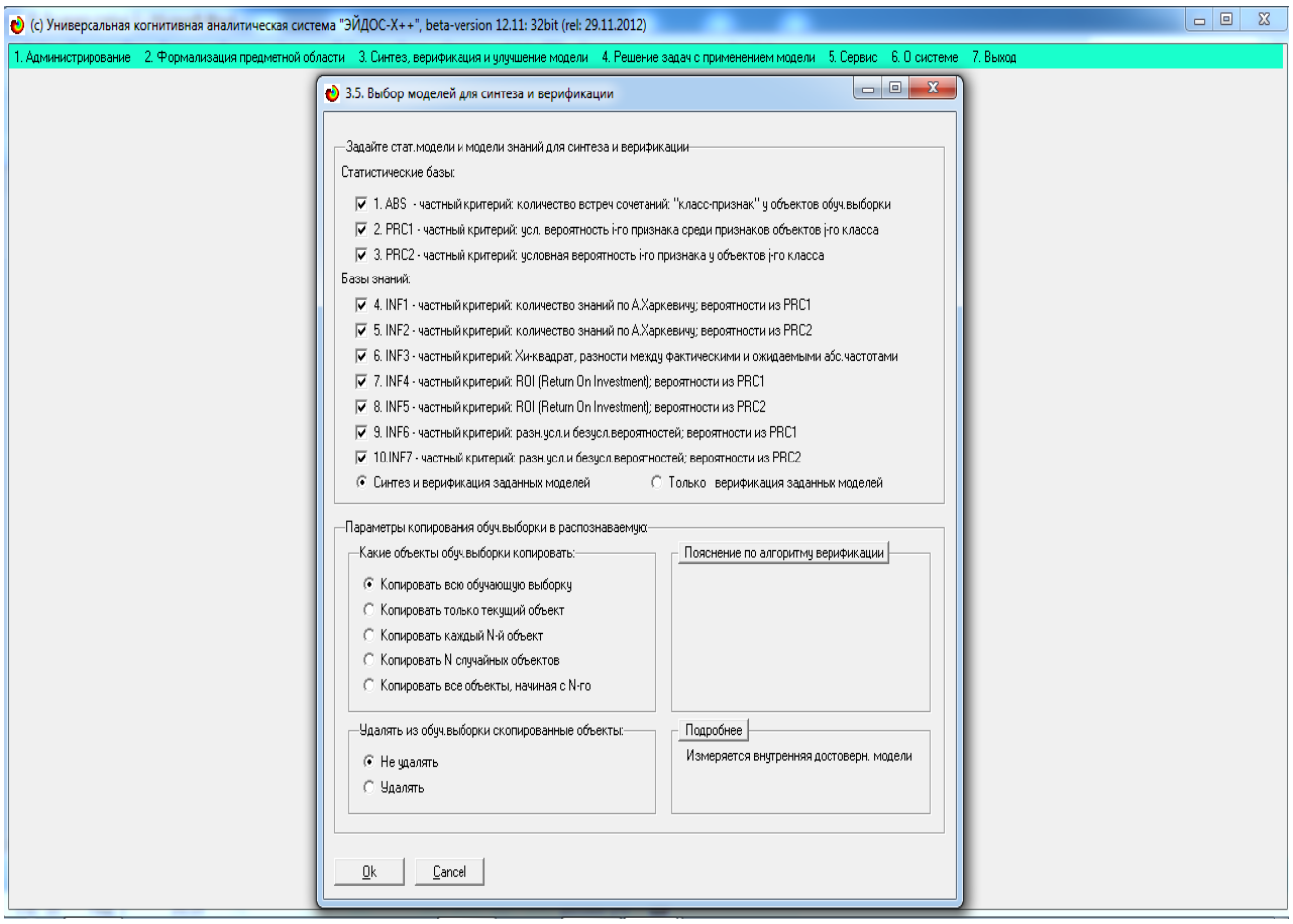


Рисунок 25.1 – Диалог режима синтеза модели и ее верификации в Системе «ЭйдосX++»

Количественные значения коэффициентов Iij таблицы 1 являются знаниями о том, что "объект перейдет в j-е состояние" если "на объект действует i-е значение фактора". Когда количество знаний Iij > 0 – i–й фактор способствует переходу объекта управления в j-е состояние, когда Iij < 0 – препятствует этому переходу, когда же Iij = 0 – никак не влияет на это.

В векторе i-го фактора (строка матрицы знаний) отображается, какое количество знаний о переходе объекта управления в каждое из будущих состояний содержится в том факте, что данное значение фактора действует.

В векторе j-го состояния класса (столбец матрицы знаний) отображается, какое количество знаний о переходе объекта управления в соответствующее состояние содержится в каждом из значений факторов, представленных в модели.

Схема обработки данных и их преобразования в информацию и знания в системе Эйдос-X++ представлена на рисунке 25.2.

Все задачи идентификации, прогнозирования, приятия решений и исследования предметной области решаются в системе Эйдос-X++ на основе моделей знаний, хотя для этого могут использоваться и статистические модели.

Поэтому там, где возможности статистических систем заканчиваются, работа системы Эйдос-X++ только начинается.

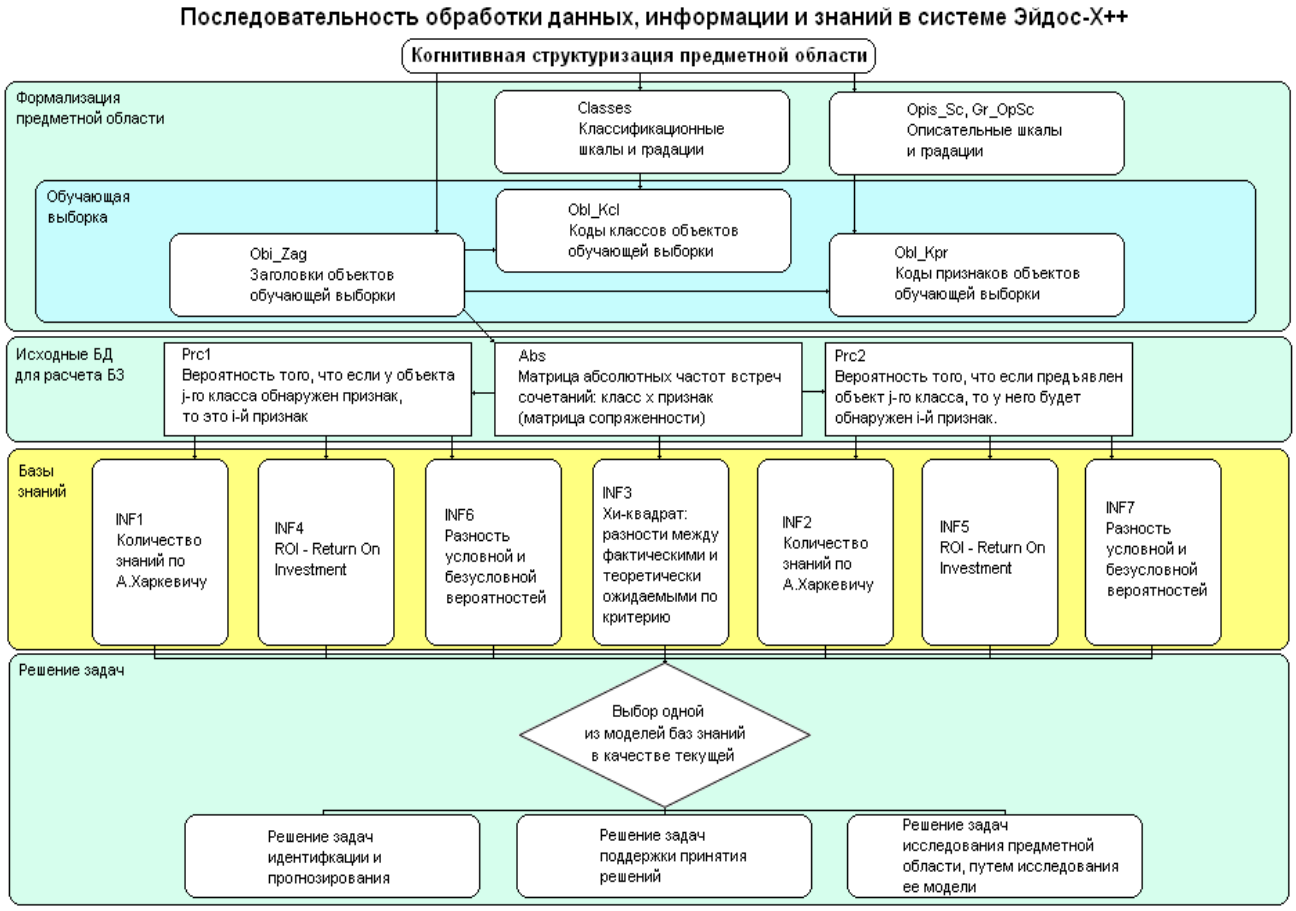


Рисунок 25.2 – Схема обработки данных и их преобразования в информацию и знания в системе Эйдос-Х++

Таким образом, модель системы Эйдос-Х++ позволяет рассчитать какое количество знаний содержится в любом факте о наступлении любого события в любой предметной области, причем для этого не требуется повторности этих фактов. Если же эти повторности осуществляются и при этом наблюдается некоторая вариабельность значений факторов, обуславливающих наступление тех или иных событий, то модель обеспечивает многопараметрическую типизацию, т.е. синтез обобщенных образов классов или категорий наступающих событий с количественной оценкой силы и направления влияния на их наступление различных значений факторов. *Причем эти факторы могут быть различной природы (физические, экономические, социальные, психологические, организационные и другие), как количественными, так и качественными и измеряться в различных единицах измерения и обрабатываться в одной модели знаний сопоставимо друг с другом за счет того, что для любых значений факторов в модели оценивается количество знаний, которое в них содержится о наступлении событий, переходе объекта управления в определенные состояния или просто о его принадлежности к тем или иным классам.*

Рассмотрим поведение объекта управления при воздействии на него не одного, а целой системы значений факторов:

$$I_j = f(\vec{I}_{ij}). \qquad (25.1)$$

В теории принятия решений скалярная функция Ij векторного аргумента называется интегральным критерием. Основная проблема состоит в выборе такого аналитического вида функции интегрального критерия, который обеспечил бы эффективное решение задач, решаемых управляющей системой САУ и АСУ.

Учитывая, что частные критерии (таблица) имеют смысл количества знаний, а знания, как и информация, является аддитивной функцией, предлагается ввести интегральный критерий, как аддитивную функцию от частных критериев в виде:

$$I_j = (\vec{I}_{ij}, \vec{L}_i). \qquad (25.2)$$

В выражении (2) круглыми скобками обозначено скалярное произведение, т.е. свертка. В координатной форме это выражение имеет вид:

$$I_j = \sum_{i=1}^{M} I_{ij} L_i, \qquad (25.3)$$

,

где:

$\vec{I}_{ij} = \{I_{ij}\}$ – вектор j-го класса-состояния объекта управления;

$\vec{L}_i = \{L_i\}$ – вектор состояния предметной области (объекта управления), включающий все виды факторов, характеризующих объект управления, возможные управляющие воздействия и окружающую среду (массив-локатор), т.е. Li=n, если i-й признак встречается у объекта n раз.

Таким образом, предложенный интегральный критерий представляет собой суммарное количество знаний, содержащихся в системе значений факторов различной природы (т.е. факторах, характеризующих объект управления, управляющее воздействие и окружающую среду) о переходе объекта управления в то или иное будущее состояние.

В многокритериальной постановке задача прогнозирования состояния объекта управления, при оказании на него заданного многофакторного управляющего воздействия Ij, сводится к максимизации интегрального критерия:

$$j^* = \arg\max_{j \in J}((\vec{I}_{ij}, \vec{L}_i)), \qquad (25.4)$$

т.е. к выбору такого состояния объекта управления, для которого интегральный критерий максимален.

Результат прогнозирования поведения объекта управления, описанного данной системой факторов, представляет собой список его возможных будущих состояний, в котором они расположены в порядке убывания суммарного количества знаний о переходе объекта управления в каждое из них.

Задача принятия решения о выборе наиболее эффективного управляющего воздействия является обратной задачей по отношению к задаче максимизации интегрального критерия (идентификации и прогнозирования), т.е. вместо того, чтобы по набору факторов прогнозировать будущее состояние объекта, наоборот, по заданному (целевому) состоянию объекта определяется такой набор факторов, который с наибольшей эффективностью перевел бы объект управления в это состояние.

Предлагается обобщение фундаментальной леммы Неймана-Пирсона, основанное на косвенном учете корреляций между знаниями в векторе состояний при

использовании средних по векторам. Соответственно, вместо простой суммы количеств информации предлагается использовать корреляцию между векторами состояния и объекта управления, которая количественно измеряет степень сходства этих векторов:

$$I_j = \frac{1}{\sigma_j \sigma_l A} \sum_{i=1}^{M} \left(I_{ij} - \bar{I}_j\right)\left(L_i - \bar{L}\right), \qquad (25.5)$$

где:

$\bar{I}_j$ – средняя информативность по вектору класса;

$\bar{L}$ – среднее по вектору идентифицируемой ситуации (объекта).

σ_j – среднеквадратичное отклонение информативностей вектора класса;

σ_l – среднеквадратичное отклонение по вектору распознаваемого объекта.

Выражение (25.5) получается непосредственно из (25.3) после замены координат перемножаемых векторов их стандартизированными значениями. Необходимо отметить, что выражение для интегрального критерия сходства (25.5) по своей математической форме является корреляцией двух векторов, координатами которых являются частные критерии знаний (поэтому в системе «Эйдос-Х++» этот интегральный критерий называется «Смысловой или семантический резонанс знаний», а критерий (25.3) – «Сумма знаний»).

Таким образом, в системе «Эйдос-Х++» возможна оценка достоверности 7 моделей знаний, а также 3 статистических моделей, с использованием двух интегральных критериев сходства конкретного образа идентифицируемого объекта с обобщенным образом класса:

- «Резонанс знаний».
- «Сумма знаний».

При этом система генерирует несколько различных форм по достоверности моделей с этими интегральными критериями:

1. Обобщающая форма по достоверности моделей при разных интегральных критериях.
2. Обобщающий статистический анализ результатов идентификации по моделям
3. Достоверность идентификации классов в различных моделях
4. Распределение уровней сходства верно и ошибочно идентифицированных и не идентифицированных объектов в различных моделях.
5. Детальный статистический анализ результатов идентификации в различных моделях по классам

Объем статьи не позволяет привести конкретные примеры этих форм, и здесь можно лишь отметить, что многочисленные численные эксперименты подтвердили возможность обоснованно выбрать на их основе наиболее достоверную модель в каждом конкретном случае. Это означает, что в системе «Эйдос-Х++» после синтеза модели мы имеем возможность не сразу применять ее для решения различных задач, а предварительно обосновано выбрать наиболее достоверную модель и уже затем использовать ее для решения конкретных задач.

Кроме того в системе «Эйдос-X++» реализуется возможность идентификации объекта с каждым классом именно в той модели и с тем интегральным критерием, при которых была наиболее высокая достоверность идентификации. Этот алгоритм идентификации был впервые разработан и реализован совместно с А.П.Труневым в системе «Эйдос-астра» (25.5) и продемонстрировал повышение вероятности верной идентификации и верной не идентификации около 20%. С приведенными монографиями можно ознакомиться на сайте автора системы «Эйдос».

Выводы. Система«Эйдос» за многие годы применения хорошо показала себя при проведении научных исследований в различных предметных областях и занятий по ряду научных дисциплин, связанных с искусственным интеллектом, представлениями знаний и управлению знаниями. Однако в процессе эксплуатации системы были выявлены и некоторые недостатки, ограничивающие перспективы применения системы. Создана качественно новая версия системы (система Эйдос-X++), в которой преодолены ограничения и недостатки предыдущей версии и реализованы новые важные идеи по ее развитию и применению в качестве программного инструментария системно-когнитивного анализа (СК-анализ).

Несомненный научный и практический интерес представляет *синтез обобщенных изображений* на основе ряда конкретных (задача-1). При этом в результате обобщения выясняется *ценность* признаков изображений для их дифференциации, а также *степень характерности* тех или иных признаков для конкретных изображений. Это позволяет без ущерба для адекватности модели *удалить* из нее малоценные признаки, т.е. осуществить *синтез абстрактных изображений* (задача-2), что обеспечивает в последующем сокращение затрат различных видов ресурсов на сбор и обработку графической информации. Над обобщенными изображениями возможны операции классификации, объединения наиболее сходных из них в кластеры и формирования систем наиболее различных кластеров, т.е. конструктов (задача-3). Можно также количественно сравнивать конкретные изображения с обобщенными, т.е. идентифицировать эти конкретные изображения (задача-4).

Рассмотрим решение этих задач в системно-когнитивном анализе (СК-анализ) [1, 2, 3, 4, 5]. При этом в качестве примера для решения перечисленных задач выберем изображения десятичных цифр, используемые для написания почтовых индексов (рис. 25.3):

Рисунок 25.3 – Образец написания цифр почтового индекса (ГОСТ Р 51506-99)[21].

Выбор данного примера связан прежде всего с его простотой, но тем ни менее он позволяет рассмотреть все методы и инструменты СК-анализа, которые имеют общее значение и могут быть применены для решения поставленных задач при обработке более сложных изображений.

[21] http://www.prodtp.ru/index.php?act=recipes&CODE=03&id=51

Для решения 1-й задачи прежде всего необходимо разбить исходные изображения на элементы, из которых они составлены и закодировать как сами эти элементы, так и изображения с использованием справочников элементов.

Рассмотрение самих способов выделения и кодирования элементов изображений не входит в задачи данной статьи. Отметим лишь, что на наш взгляд эти методы могут быть основаны на метрических и топологических свойствах изображений и оба этих подхода могут базироваться на выделении элементов на основе высоких градиентов (скорости *изменения*, т.е. 1-й производной) цвета, его насыщенности, яркости и кривизны распределения этих параметров в пространстве.

В ряде конкретных случаев эта работа уже проведена. Например, в МВД существуют, специальные атласы для разработки фотороботов лиц, содержащие изображения таких элементов лиц как носы, губы, брови, глаза, лбы, морщины, залысины, прически, усы, подбородки, бороды, уши и т.д. и т.п.

Очевидно, что рассматриваемый нами шрифт специально сконструирован специалистами таким образом, чтобы было вполне очевидно из каких элементов состоят цифры, т.к. они выделены яркостью и высокой кривизной на границах элементов (четко выраженные «углы»). Для нас несущественно, какие именно коды будут иметь эти элементы, поэтому пронумеруем их так, как проще (рис. 25.4).

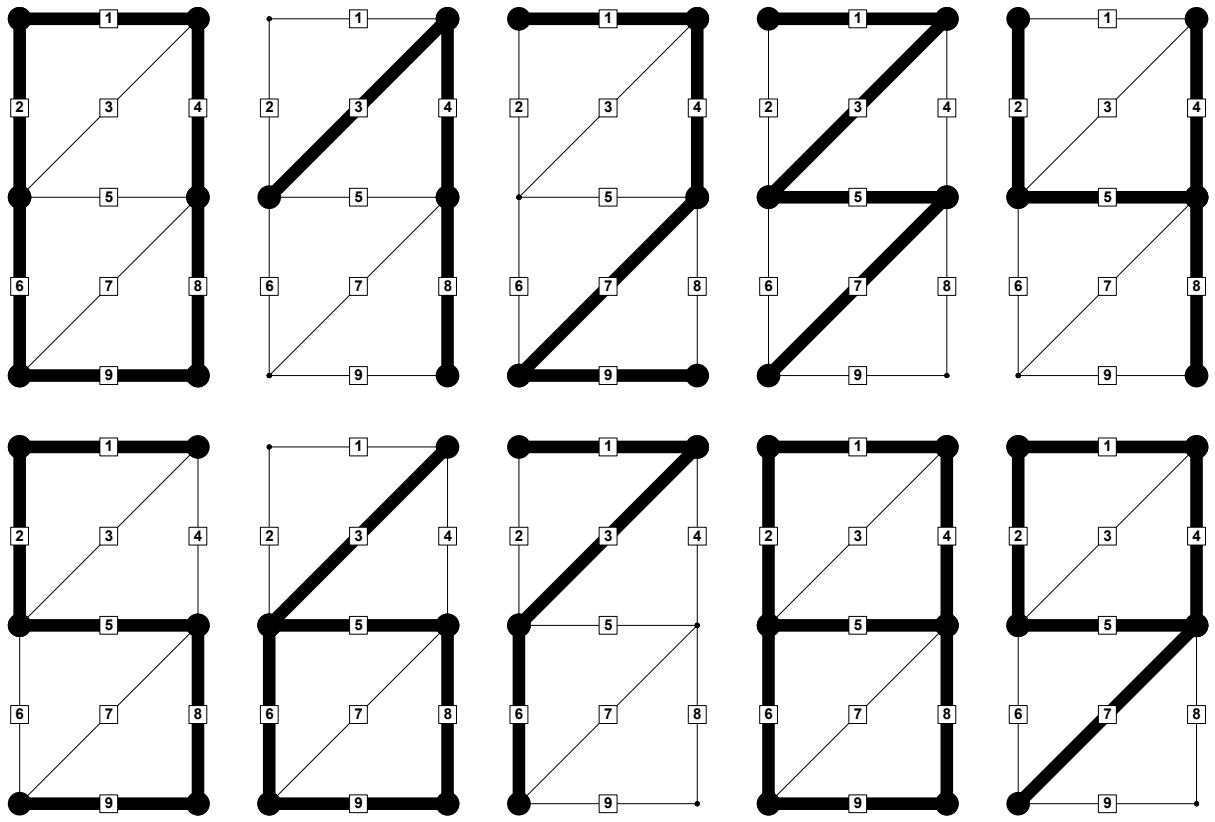


Рисунок 25.4 – Принцип кодирования изображений десятичных цифр, состоящих из пронумерованных элементов

С использованием системы кодирования элементов цифр, представленной на рисунке 25.3, сформируем следующий справочник элементов.

Таблица 25.2 – Справочник элементов (признаков) цифр

KOD	NAME
1	Элемент 1
2	Элемент 2
3	Элемент 3
4	Элемент 4
5	Элемент 5
6	Элемент 6
7	Элемент 7
8	Элемент 8
9	Элемент 9

Сами цифры кодируются в виде последовательности кодов элементов, из которых они состоят, при этом порядок кодов несущественен, т.к. код каждого элемента однозначно определяет его вид и положение в матрице символа (таблица 25.2).

Таблица 25.3 – Кодирование изображений десятичных цифр

Цифра	Признаки (элементы) цифр						
0	1	2	4	6	8	9	
1	3	4	8				
2	1	4	7	9			
3	1	3	5	7			
4	2	4	5	8			
5	1	2	5	8	9		
6	3	5	6	8	9		
7	1	3	6				
8	1	2	4	5	6	8	9
9	1	2	4	5	7		

В качестве справочника классов, формируемых на основе примеров конкретных изображений цифр, закодированных в таблице 1, примем как сами цифры, так и два обобщенных класса: «Четные» и «Не четные» (таблица 25.4):

Таблица 25.4 – Сравочник классов

KOD	NAME
1	Цифра 0
2	Цифра 1
3	Цифра 2
4	Цифра 3
5	Цифра 4
6	Цифра 5
7	Цифра 6
8	Цифра 7
9	Цифра 8
10	Цифра 9
11	Четная цифра
12	Не четная цифра

С использованием справочника классов (таблица 25.3) и закодированных изображений цифр (таблица 25.2) формируется обучающая выборка (таблица 25.5):

Таблица 25.5 – Обучающая выборка

Код	Наименование объекта	Коды классов		Коды признаков						
1	Цифра 0	1	11	1	2	4	6	8	9	
2	Цифра 1	2	12	3	4	8				
3	Цифра 2	3	11	1	4	7	9			
4	Цифра 3	4	12	1	3	5	7			
5	Цифра 4	5	11	2	4	5	8			
6	Цифра 5	6	12	1	2	5	8	9		
7	Цифра 6	7	11	3	5	6	8	9		
8	Цифра 7	8	12	1	3	6				
9	Цифра 8	9	11	1	2	4	5	6	8	9
10	Цифра 9	10	12	1	2	4	5	7		

В систему «Эйдос», являющуюся инструментарием СК-анализа, вводятся справочникb классов, признаков и обучающая выборка (таблицы 25.3, 25.1 и 25.5), а затем осуществляется синтез семантической информационной модели (СИМ-2).

В результате формируются матрица абсолютных и относительных частот, а также матрица знаний, содержащая информацию о том, какая цифра предъявлена, если установлено, что в нее входит *i-й* элемент (таблицы 25.6, 25.7, 25.8):

Таблица 25.6 – Матрица абсолютных частот

Код элемента	Наименование элемента	Наименования и коды классов												Всего:
		Цифра 0	Цифра 1	Цифра 2	Цифра 3	Цифра 4	Цифра 5	Цифра 6	Цифра 7	Цифра 8	Цифра 9	Четные	Не четные	
		1	2	3	4	5	6	7	8	9	10	11	12	
1	Элемент 1	1	0	1	1	0	1	0	1	1	1	3	4	14
2	Элемент 2	1	0	0	0	1	1	0	0	1	1	3	2	10
3	Элемент 3	0	1	0	1	0	0	1	1	0	0	1	3	8
4	Элемент 4	1	1	1	0	1	0	0	0	1	1	4	2	12
5	Элемент 5	0	0	0	1	1	1	1	0	1	1	3	3	12
6	Элемент 6	1	0	0	0	0	0	1	1	1	0	3	1	8
7	Элемент 7	0	0	1	1	0	0	0	0	0	1	1	2	6
8	Элемент 8	1	1	0	0	1	1	1	0	1	0	4	2	12
9	Элемент 9	1	0	1	0	0	1	1	0	1	0	4	1	10
	Всего:	1	1	1	1	1	1	1	1	1	1	5	5	20

Таблица 25.7 – Матрица условных вероятностей (%, СИМ-2)

Код эл.	Наименование элемента	Наименования и коды классов: Цифра 0	Цифра 1	Цифра 2	Цифра 3	Цифра 4	Цифра 5	Цифра 6	Цифра 7	Цифра 8	Цифра 9	Четные	Не четные	Безусловная вероятность
		1	2	3	4	5	6	7	8	9	10	11	12	
1	Элемент 1	100	0	100	100	0	100	0	100	100	100	60	80	70
2	Элемент 2	100	0	0	0	100	100	0	0	100	100	60	40	50
3	Элемент 3	0	100	0	100	0	0	100	100	0	0	20	60	40
4	Элемент 4	100	100	100	0	100	0	0	0	100	100	80	40	60
5	Элемент 5	0	0	0	100	100	100	100	0	100	100	60	60	60
6	Элемент 6	100	0	0	0	0	0	100	100	100	0	60	20	40
7	Элемент 7	0	0	100	100	0	0	0	0	0	100	20	40	30
8	Элемент 8	100	100	0	0	100	100	100	0	100	0	80	40	60
9	Элемент 9	100	0	100	0	0	100	100	0	100	0	80	20	50
	Всего:	1	1	1	1	1	1	1	1	1	1	5	5	20

Таблица 25.8 – Матрица знаний (Бит×100, СИМ-2)

Код эл.	Наименование элемента	Наименования и коды классов: Цифра 0	Цифра 1	Цифра 2	Цифра 3	Цифра 4	Цифра 5	Цифра 6	Цифра 7	Цифра 8	Цифра 9	Четные	Не четные	Ср.кв.откл.
		1	2	3	4	5	6	7	8	9	10	11	12	
1	Элемент 1	52		52	52		52		52	52	52	-22	19	28
2	Элемент 2	100				100	100			100	100	26	-32	53
3	Элемент 3		132		132			132	132			-100	59	76
4	Элемент 4	74	74	74		74				74	74	42	-59	45
5	Элемент 5				74	74	74	74		74	74			39
6	Элемент 6	132						132	132	132		59	-100	76
7	Элемент 7			174	174						174	-59	42	82
8	Элемент 8	74	74			74	74	74		74		42	-59	45
9	Элемент 9	100		100			100	100		100		68	-132	72
	Ср.кв.откл.	50	50	62	66	43	45	58	58	44	60	57	65	

Из матрицы знаний (таблица 25.8) видно, что различные элементы несут различное количество информации о том, что предъявлена некоторая цифра, т.е. для каждой цифры одни элементы более характерны, а другие менее характерны.

Например, для цифры «0» наиболее характерен 6-й элемент, за ним идут 2-й и 9-й элементы, а затем 4-й и 8-й, и самым нехарактерным для этой цифры является 1-й элемент.

Подобная информация представляется в системе «Эйдос» во многих различных формах, в частности в виде таблиц информационных портретов классов и рисунков нелокальных нейронов (таблица 25.9).

Предлагается *визуализировать в изображениях значимость элементов путем их отображения цветом и толщиной линии, соответствующими значимости:*

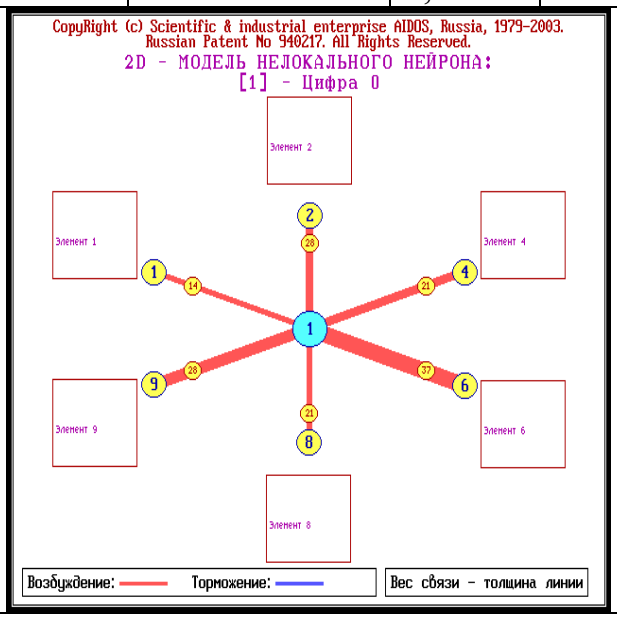
– черный или красный цвет означает положительное количество информации;

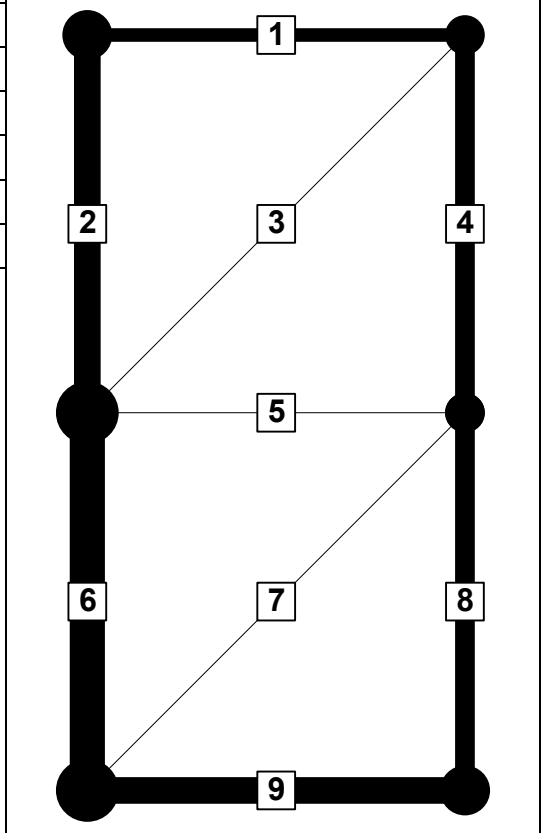
– синий цвет означает отрицательное количество информации;

– толщина линии соответствует модулю количества информации.

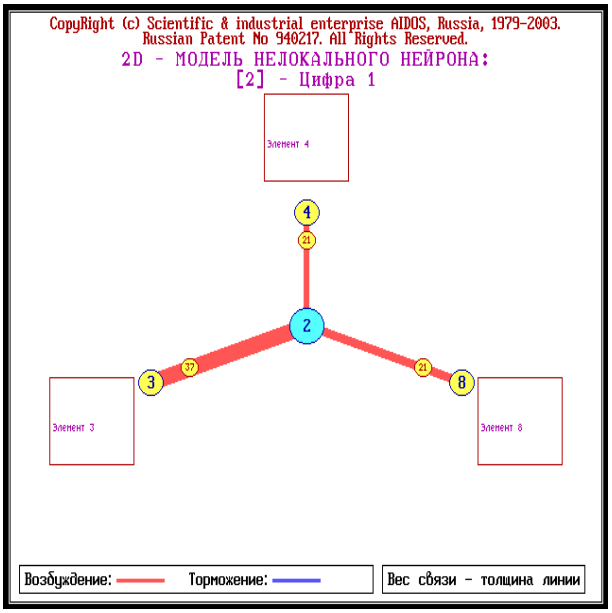
Таблица 25.9 – Информационные портреты классов, нелокальные нейроны и изображения цифр с отображением значимости элементов цветом и толщиной линии

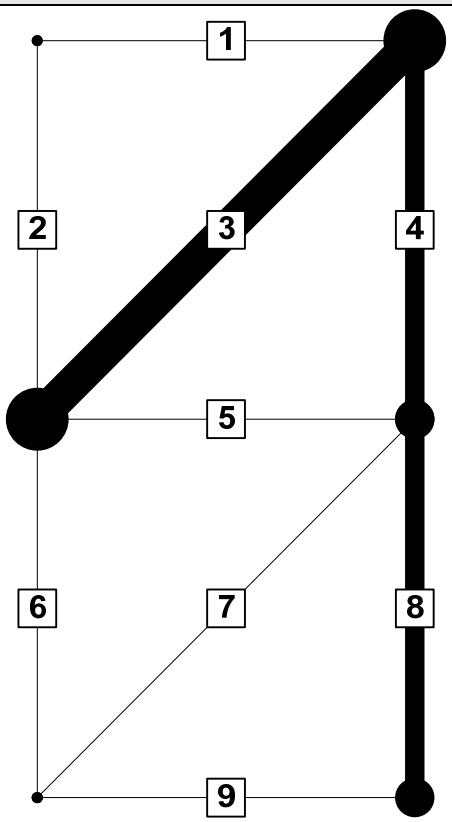
№	Код элемента	**ЦИФРА 0**		
		Наименование	INFBIT	INFPROC
1	6	Элемент 6	1,32193	36,87
2	2	Элемент 2	1,00000	27,89
3	9	Элемент 9	1,00000	27,89
4	4	Элемент 4	0,73697	20,56
5	8	Элемент 8	0,73697	20,56
6	1	Элемент 1	0,51457	14,35



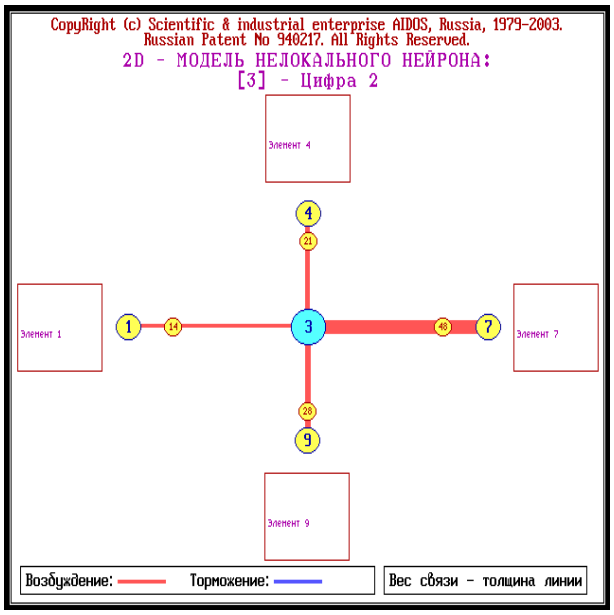


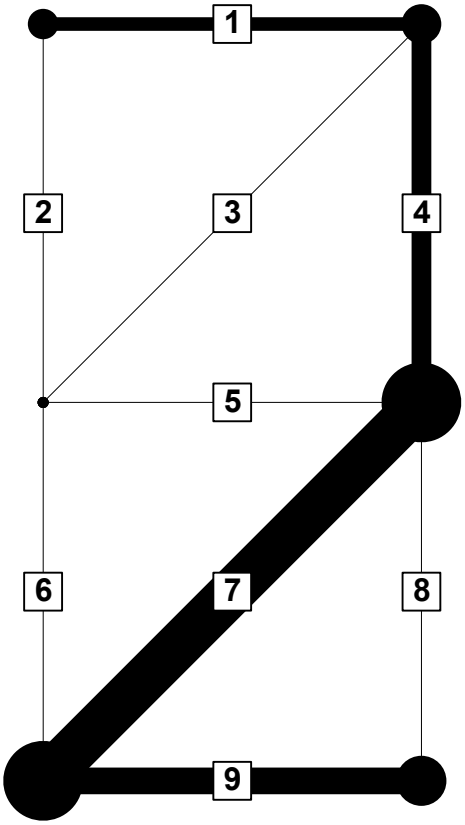
№	Код элемента	ЦИФРА 1		
		Наименование	INFBIT	INFPROC
1	3	Элемент 3	1,32193	36,87
2	4	Элемент 4	0,73697	20,56
3	8	Элемент 8	0,73697	20,56



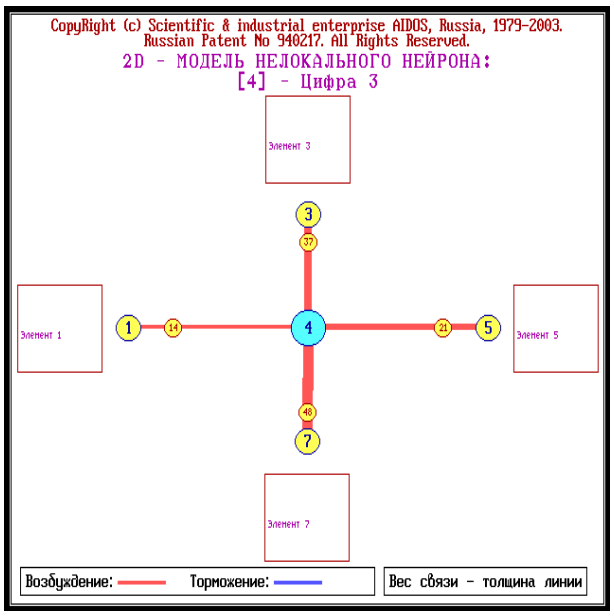


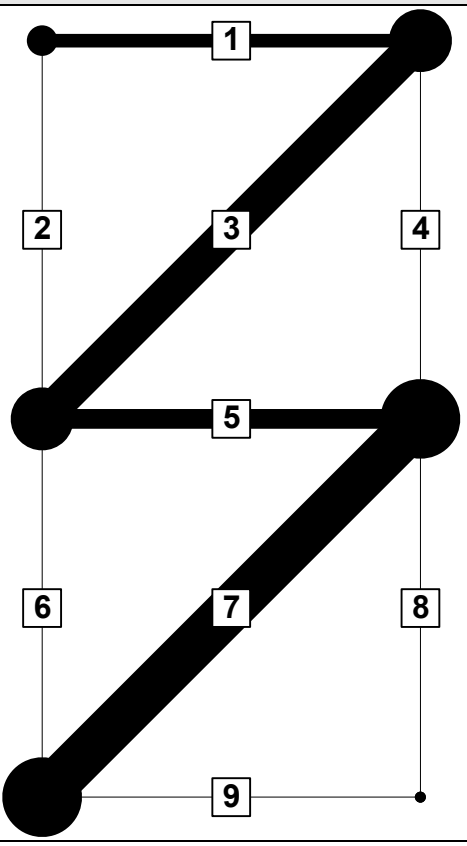
№	Код элемента	ЦИФРА 2		
		Наименование	INFBIT	INFPROC
1	7	Элемент 7	1,73697	48,45
2	9	Элемент 9	1,00000	27,89
3	4	Элемент 4	0,73697	20,56
4	1	Элемент 1	0,51457	14,35



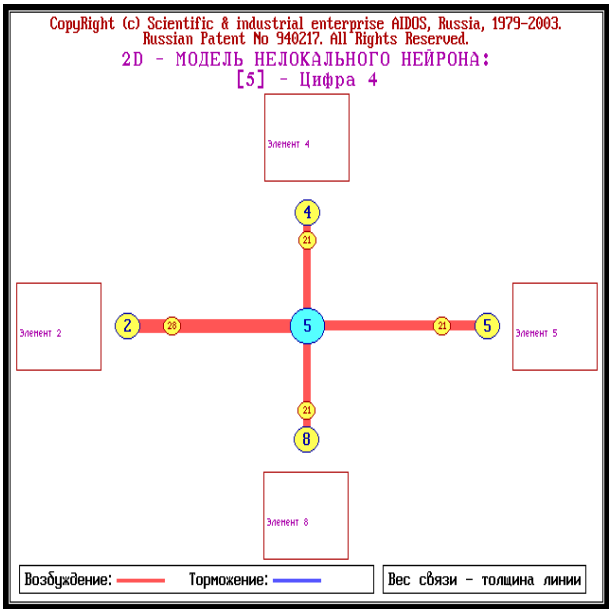


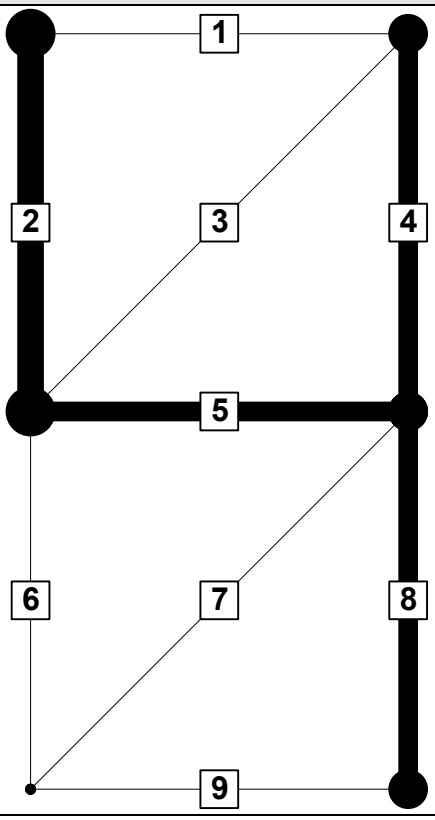
№	Код элемента	ЦИФРА 3		
		Наименование	INFBIT	INFPROC
1	7	Элемент 7	1,73697	48,45
2	3	Элемент 3	1,32193	36,87
3	5	Элемент 5	0,73697	20,56
4	1	Элемент 1	0,51457	14,35



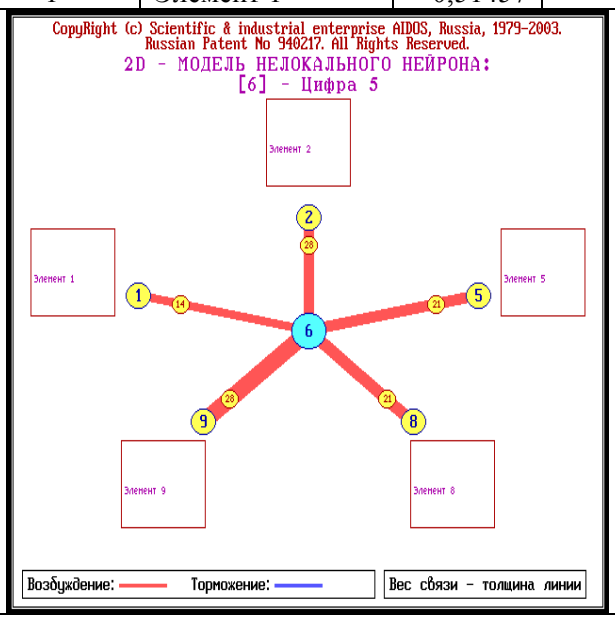


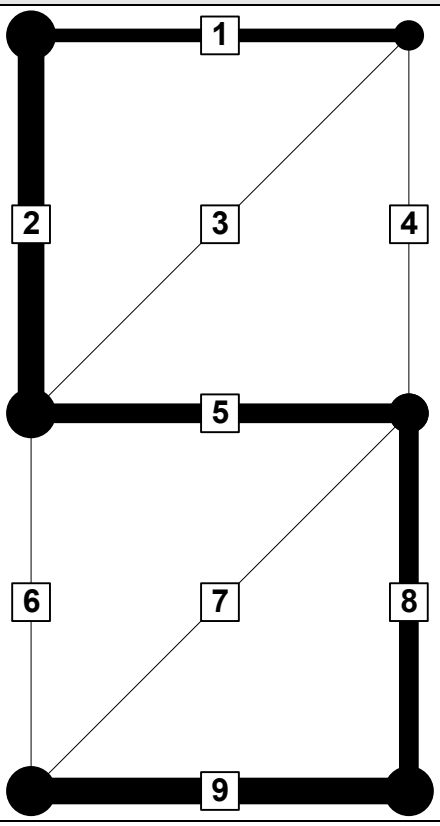
№	Код элемента	ЦИФРА 4		
		Наименование	INFBIT	INFPROC
1	2	Элемент 2	1,00000	27,89
2	4	Элемент 4	0,73697	20,56
3	5	Элемент 5	0,73697	20,56
4	8	Элемент 8	0,73697	20,56



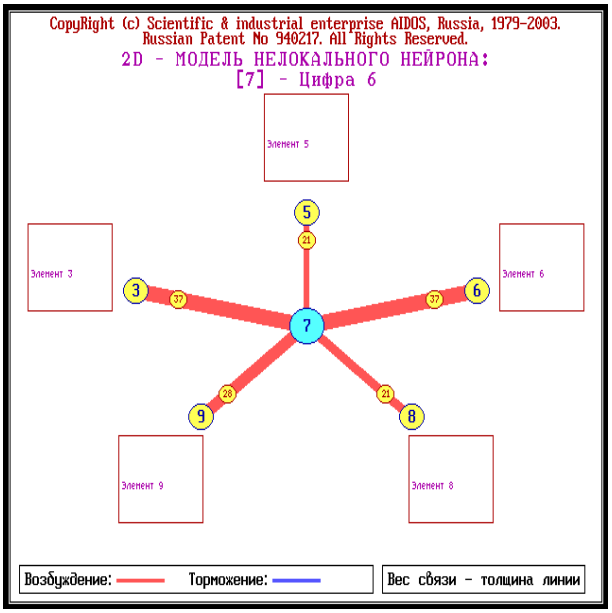


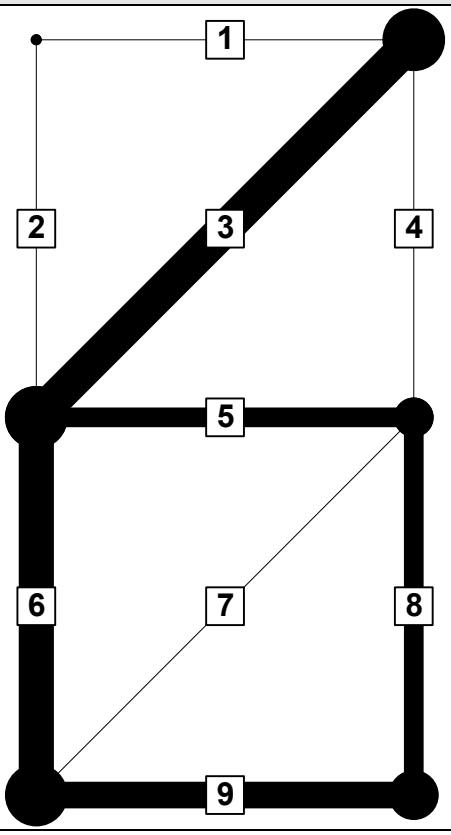
№	Код элемента	ЦИФРА 5		
		Наименование	INFBIT	INFPROC
1	2	Элемент 2	1,00000	27,89
2	9	Элемент 9	1,00000	27,89
3	5	Элемент 5	0,73697	20,56
4	8	Элемент 8	0,73697	20,56
5	1	Элемент 1	0,51457	14,35



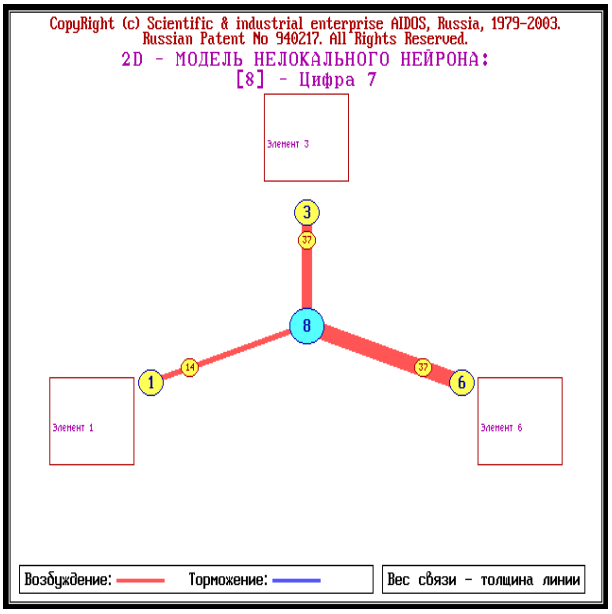


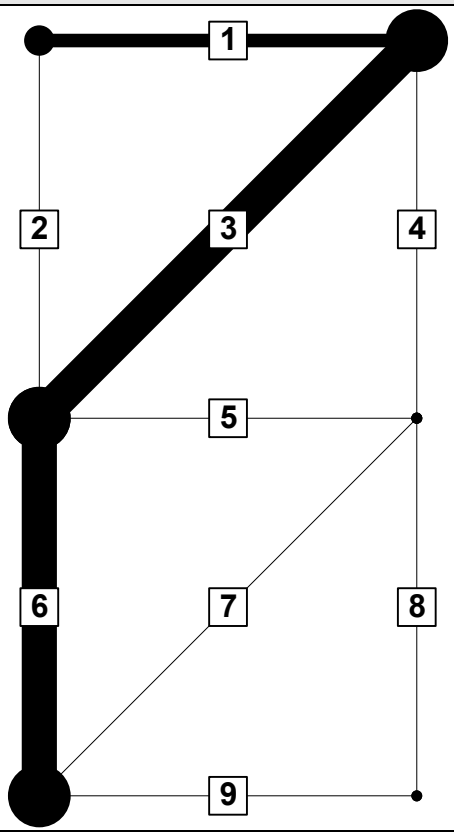
№	Код элемента	ЦИФРА 6		
		Наименование	INFBIT	INFPROC
1	3	Элемент 3	1,32193	36,87
2	6	Элемент 6	1,32193	36,87
3	9	Элемент 9	1,00000	27,89
4	5	Элемент 5	0,73697	20,56
5	8	Элемент 8	0,73697	20,56



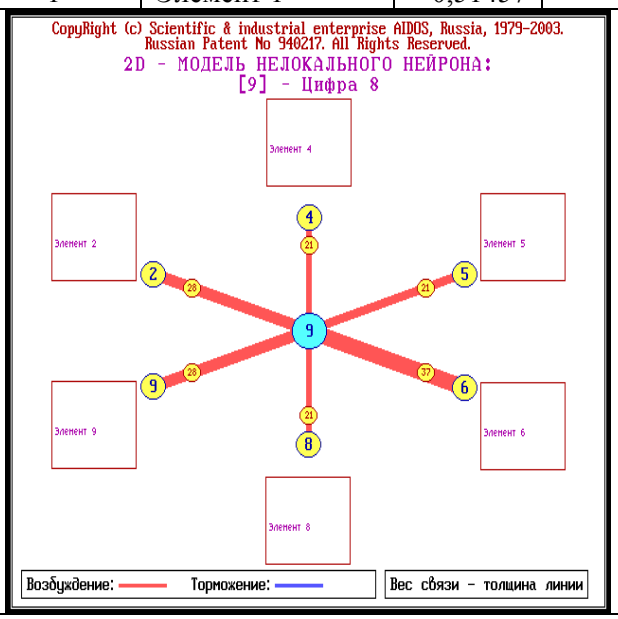


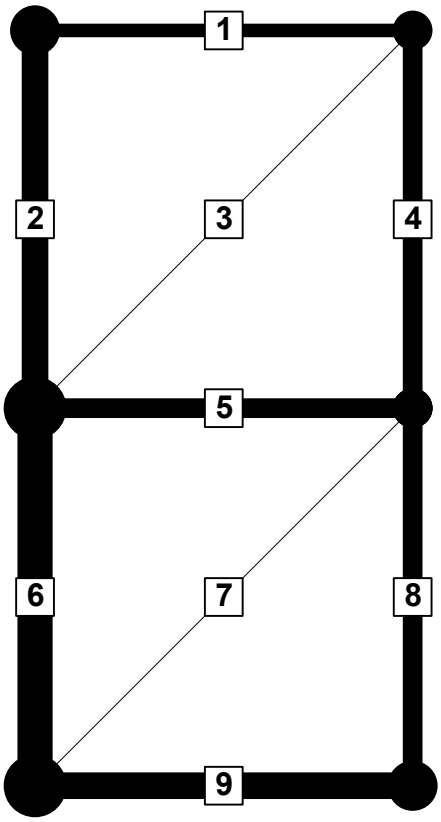
№	Код элемента	ЦИФРА 7		
		Наименование	INFBIT	INFPROC
1	3	Элемент 3	1,32193	36,87
2	6	Элемент 6	1,32193	36,87
3	1	Элемент 1	0,51457	14,35



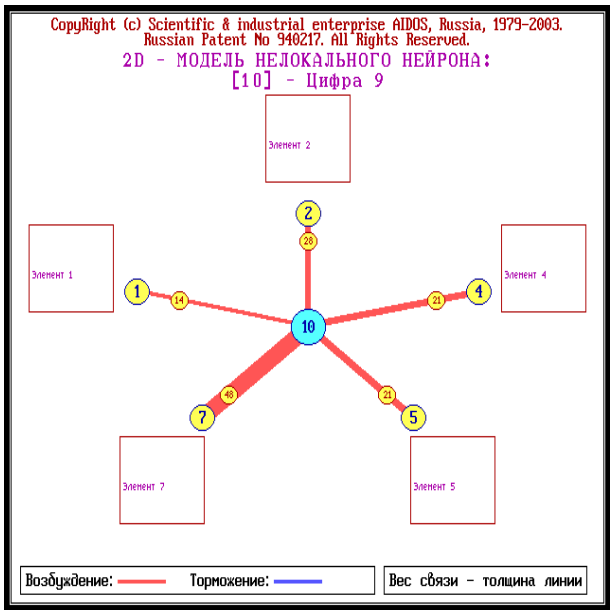


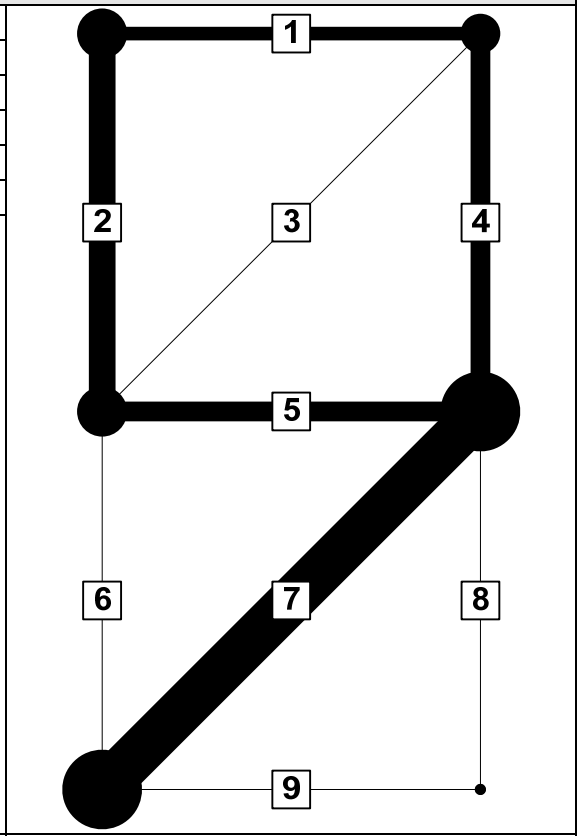
№	Код элемента	ЦИФРА 8		
		Наименование	INFBIT	INFPROC
1	6	Элемент 6	1,32193	36,87
2	2	Элемент 2	1,00000	27,89
3	9	Элемент 9	1,00000	27,89
4	4	Элемент 4	0,73697	20,56
5	5	Элемент 5	0,73697	20,56
6	8	Элемент 8	0,73697	20,56
7	1	Элемент 1	0,51457	14,35



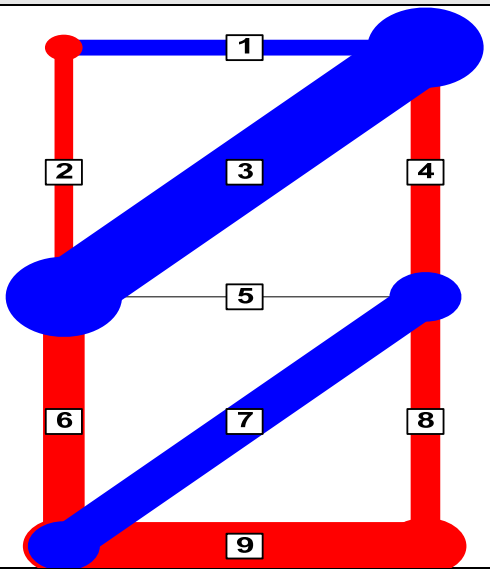


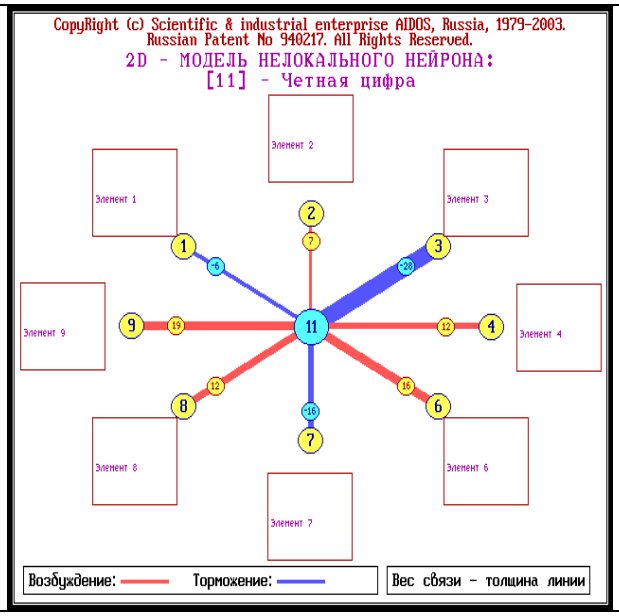
№	Код элемента	ЦИФРА 9		
		Наименование	INFBIT	INFPROC
1	7	Элемент 7	1,73697	48,45
2	2	Элемент 2	1,00000	27,89
3	4	Элемент 4	0,73697	20,56
4	5	Элемент 5	0,73697	20,56
5	1	Элемент 1	0,51457	14,35



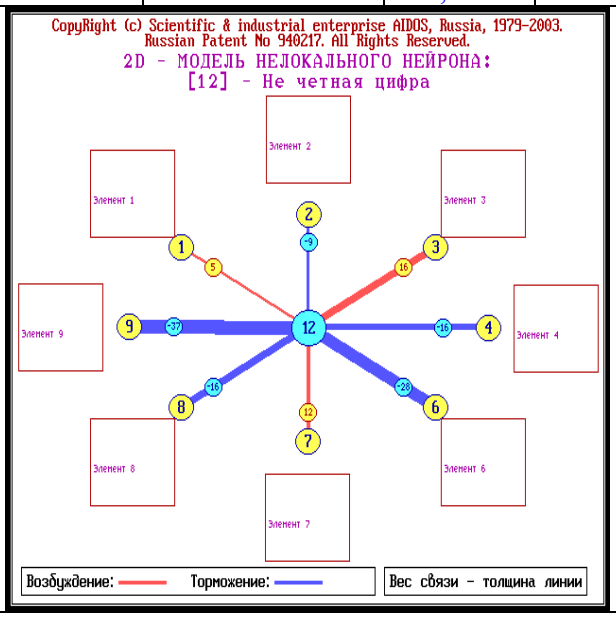


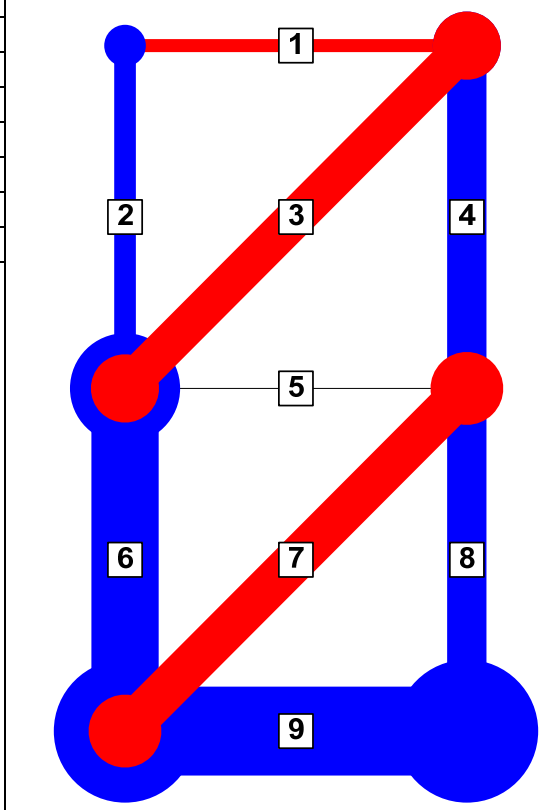
№	Код элемента	КЛАСС "ЧЕТНЫЕ"		
		Наименование	INFBIT	INFPROC
1	9	Элемент 9	0,67807	18,91
2	6	Элемент 6	0,58496	16,32
3	4	Элемент 4	0,41504	11,58
4	8	Элемент 8	0,41504	11,58
5	2	Элемент 2	0,26303	7,34
6	1	Элемент 1	-0,22239	-6,20
7	7	Элемент 7	-0,58496	-16,32
8	3	Элемент 3	-1,00000	-27,89





№	Код элемента	КЛАСС "НЕ ЧЕТНЫЕ" Наименование	INFBIT	INFPROC
1	3	Элемент 3	0,58496	16,32
2	7	Элемент 7	0,41504	11,58
3	1	Элемент 1	0,19265	5,37
4	2	Элемент 2	-0,32193	-8,98
5	4	Элемент 4	-0,58496	-16,32
6	8	Элемент 8	-0,58496	-16,32
7	6	Элемент 6	-1,00000	-27,89
8	9	Элемент 9	-1,32193	-36,87





Приведем изображения десятичных цифр, используемые для написания почтовых индексов, с указанием значимости элементов, из которых они состоят, с помощью толщины линии или условно говоря ее «нажима» (рис. 25.5).

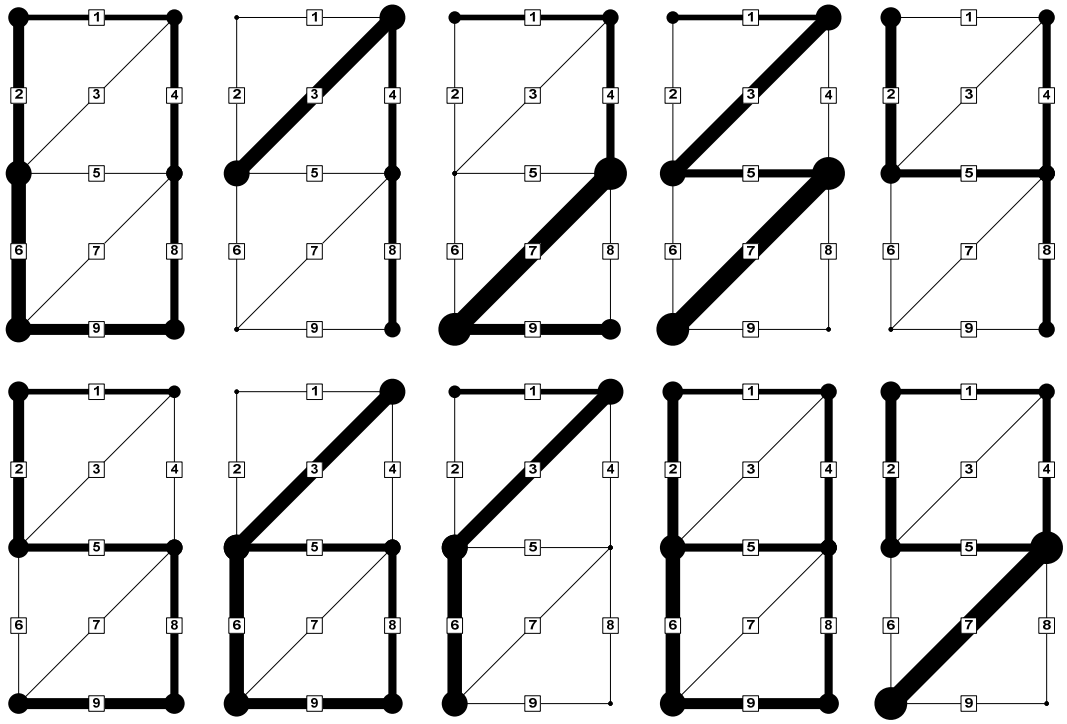


Рисунок 25.5 – Изображения десятичных цифр, используемые для написания почтовых индексов, с указанием значимости элементов, из которых они состоят, с помощью толщины линии

Подобные изображения, в которых наиболее характерные их элементы каким-то образом выделены, причем пропорционально степени их характерности, принято называть ***шаржами***. Обычно в искусстве выделение осуществляется с помощью размера, но можно это делать и с помощью *яркости* соответствующих элементов изображения, т.е. по сути ***силы нажима*** или ***толщины линии***, как это и сделано в данной статье. Представляет интерес на наш взгляд и несколько неожиданный результат, состоящий в том, что оказывается в исследованном шрифте вертикальные и горизонтальные элементы цифр ассоциированы, содержат больше информации о принадлежности к четным цифрам, а наклонные – с нечетным.

Приведенная технология обеспечивает синтез *шаржей* любых типов изображений, например, таких как папиллярные узоры, радужная оболочка глаза, почерк и фотороботы лиц, сгруппированных по их полу, возрасту, профессии, социальному статусу, уровню образования, степени успешности тех или иных видов деятельности, например учебным или профессиональным достижениям по различным дисциплинам и циклам дисциплин, риску невозврата кредита, риску совершения ДТП и сумме страховых выплат КАСКО и ОСАГО и т.п. и т.д. Аналогичным образом можно обобщать изображения автомобилей, различных видов растений и пород животных, а также любые другие изображения, например предназначенные для восприятия элементы компьютерного интерфейса и дорожные знаки.

Обратимся к таблице 25.10, в которой проранжируем элементы в порядке убывания среднеквадратичного отклонения содержащегося в них количества информации о принадлежности изображений, включающих эти элементы, к классам

цифр, а также к классам «четные» и «не четные». В результате получим (таблица 25.10):

Из таблицы 25.10 видно, что 7-й элемент практически в 3 раза более значим, чем 1-й. Если бы элементов было больше, то малозначимые элементы без особого ущерба для адекватности модели вполне можно было бы удалить из нее. Точно также из образов цифр без особого ущерба можно удалить малозначимые элементы. Эта операция и называется ***абстрагированием***.

Таблица 25.10 – Список графических элементов, ранжированный в порядке значимости

№	Наименование	Значимость
1	Элемент 7	82
2	Элемент 3	76
3	Элемент 6	76
4	Элемент 9	72
5	Элемент 2	53
6	Элемент 4	45
7	Элемент 8	45
8	Элемент 5	39
9	Элемент 1	28

Сравнение друг с другом обобщенных образов классов (цифр), т.е. классификация, осуществляется системой «Эйдос». В результате формируется матрица сходства классов, т.е. изображений цифр (таблица 25.11):

Таблица 25.11 – Матрица сходства классов

KOD	1	2	3	4	5	6	7	8	9	10	11	12
1	100,0	-28,7	-22,3	-86,8	11,2	25,8	11,8	6,5	86,9	-41,2	83,9	-85,4
2	-28,7	100,0	-28,4	15,6	5,2	-39,5	30,6	33,6	-45,0	-38,8	-39,2	28,4
3	-22,3	-28,4	100,0	41,9	-41,3	-22,9	-44,4	-39,9	-39,7	58,3	-13,9	7,2
4	-86,8	15,6	41,9	100,0	-43,6	-46,8	-5,2	16,4	-88,8	49,1	-90,4	83,4
5	11,2	5,2	-41,3	-43,6	100,0	43,3	-37,2	-56,8	30,8	19,3	34,7	-11,2
6	25,8	-39,5	-22,9	-46,8	43,3	100,0	-8,8	-55,4	42,4	-13,4	44,1	-35,5
7	11,8	30,6	-44,4	-5,2	-37,2	-8,8	100,0	63,8	19,2	-77,2	4,7	-27,2
8	6,5	33,6	-39,9	16,4	-56,8	-55,4	63,8	100,0	-5,4	-50,4	-32,8	15,8
9	86,9	-45,0	-39,7	-88,8	30,8	42,4	19,2	-5,4	100,0	-38,6	91,4	-85,9
10	-41,2	-38,8	58,3	49,1	19,3	-13,4	-77,2	-50,4	-38,6	100,0	-31,2	46,6
11	83,9	-39,2	-13,9	-90,4	34,7	44,1	4,7	-32,8	91,4	-31,2	100,0	-94,4
12	-85,4	28,4	7,2	83,4	-11,2	-35,5	-27,2	15,8	-85,9	46,6	-94,4	100,0

Фрагменты это матрицы сходства могут быть отображены средствами системы «Эйдос» в форме семантических сетей (рисунок 25.6):

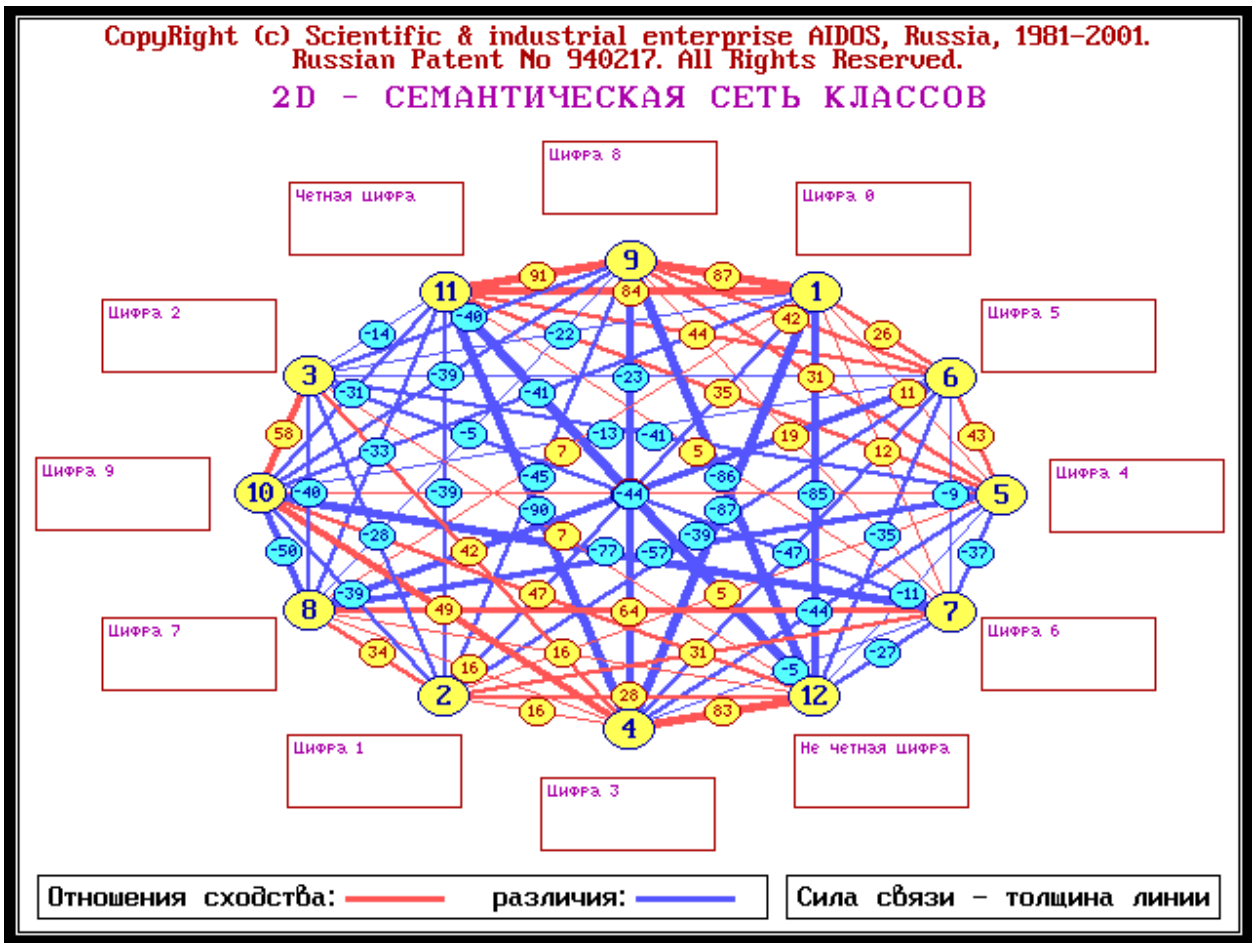


Рисунок 25.6 – Семантическая сеть классов, отображающая в наглядной графический форме степень сходства и различия обобщенных образов классов (изображений)

Из рисунка 25.6 видно, что изображения цифр группируются в кластерах вокруг обобщенных образов четных и нечетных цифр, представляющих собой полюса конструкта (таблица 25.12):

Таблица 25.11 – КОНСТРУКТ «Не четная цифра – четная цифра»

№	Код класса	Наименование класса	Уровень Сходства (%)
1	12	Не четная цифра	100,00
2	4	Цифра 3	83,43
3	10	Цифра 9	46,59
4	2	Цифра 1	28,38
5	8	Цифра 7	15,84
6	3	Цифра 2	7,17
7	5	Цифра 4	-11,21
8	7	Цифра 6	-27,23
9	6	Цифра 5	-35,49
10	1	Цифра 0	-85,39
11	9	Цифра 8	-85,91
12	11	Четная цифра	-94,44

Обращает на себя внимание, что цифра «2» больше похожа на обобщенный образ нечетных цифр, а цифра «5» – на обобщенный образ четных цифр.

На основе проведенного анализа можно даже сказать, что причина этого в том, что в изображении цифры «2» наиболее характерным является 7-й элемент, характерный именно для нечетных цифр, а в состав изображения цифры «5» входят элементы 2, 8 и 9, характерные для четных цифр.

Идентификация конкретных изображений цифр с их обобщенными образами осуществляется следующим образом:

– подсчитывается какое суммарное количество информации содержится в системе элементов данной конкретной цифры о ее принадлежности к каждому из обобщенных образов классов, сформированных в модели;

– классы ранжируются в порядке убывания суммарного количества информации о принадлежности к ним, содержащегося в систем признаков конкретной цифры;

– считается, что цифра относится к тому классу, о принадлежности к которому в ее системе признаков содержится максимальное количество информации.

Идентификация осуществляется в 4-й подсистеме системы «Эйдос». В результате формируются экранные формы карточек результатов идентификации и соответствующие выходные формы (рисунок 25.7):

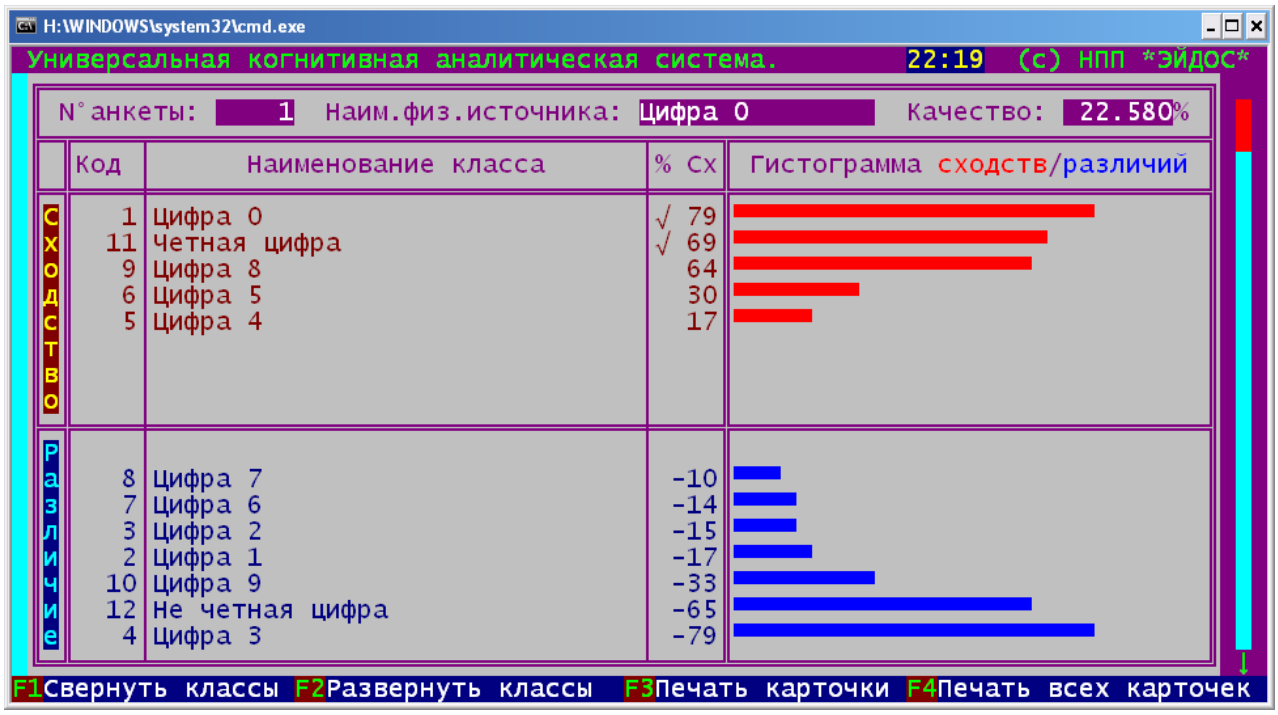


Рисунок 25.7 – Экранная форма карточки идентификации цифры «0»

На рисунке 25.8 приведена экранная форма результатов идентификации изображений конкретных цифр с обобщенным образом «Не четная цифра».

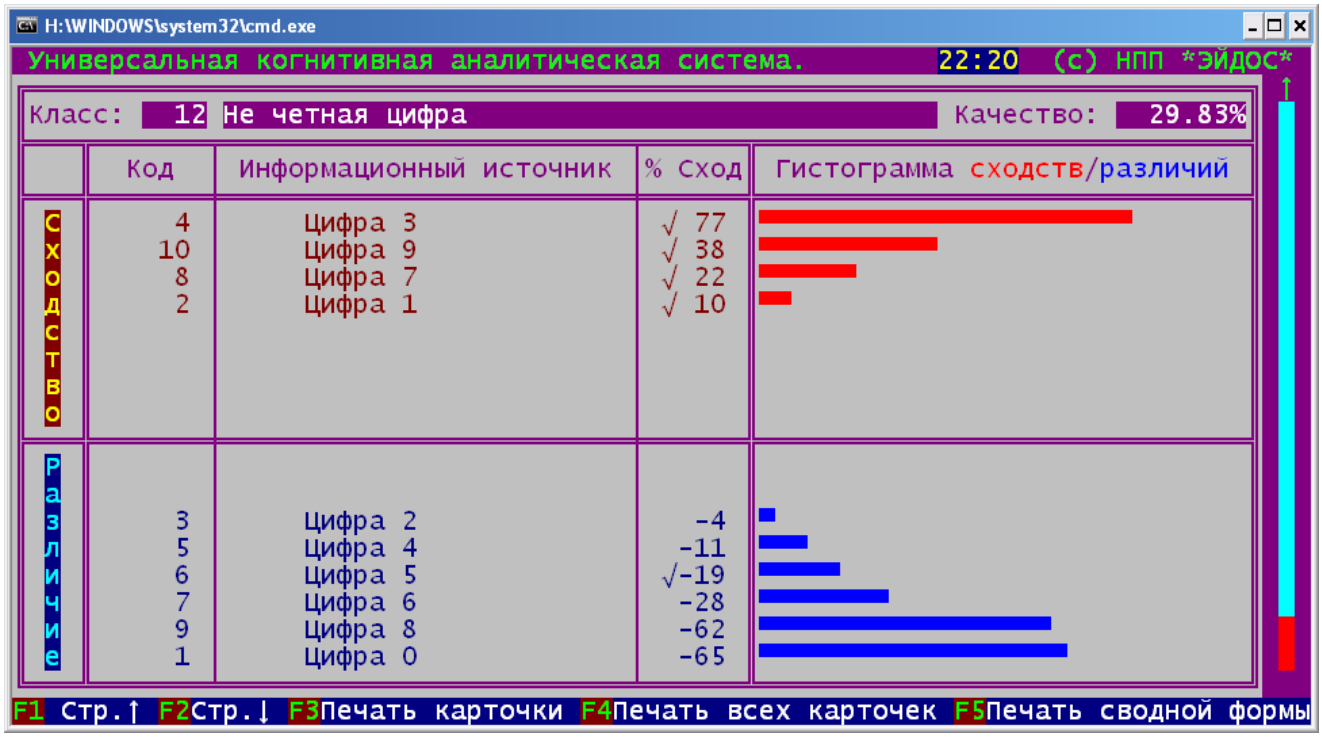


Рисунок 25.8 – Экранная форма результатов идентификации изображений конкретных цифр с обобщенным образом «Не четная цифра»

Обобщение результатов идентификации позволяет измерить степень достоверности созданной модели (таблица 25.13).

Таким образом, рассмотрено применение системно-когнитивного анализа, его математической модели – системной теории информации (СТИ) и программного инструментария – системы «Эйдос» для синтеза обобщенных изображений классов, их абстрагирования, классификации обобщенных изображений (кластеры и конструкты) сравнения конкретных изображений с обобщенными образами (идентификация).

Рассмотрение проведено на конкретном примере изображений цифр шрифта, используемого для написания почтовых индексов, но примененные при этом методы являются универсальными и могут быть использованы и при обработке других самых разнообразных изображений.

Необходимо также отметить, что в рассмотренном примере все цифры верно отнесены к классам, к которым они относятся, причем каждая цифра является наиболее похожей из всех именно на класс, к которому она действительно принадлежит, т.е. ошибка неидентификации равна нулю, однако к ним отнесены и другие цифры, которые к ним не относятся, т.е. *есть ошибка ложной идентификации*.

Учитывая это можно предположить, что актуальной является задача *разработки таких шрифтов и других знаков, предназначенных для визуального восприятия (например элементов компьютерного интерфейса, дорожных знаков и т.п.), начертания которых минимизировали бы ошибки как 1-го, так и 2-го рода, т.е. были бы* ***наиболее легко и безошибочно воспринимаемы*** *как человеком, так и системами машинного зрения. Это имеет значение как для минимизации затрат различных видов вычислительных и других ресурсов и времени, так и для улучше-*

ния восприятия затухающих текстов, чтения в условиях зашумленности или погрешностей самого аппарата восприятия.

Таблица 25.13 – Выходная форма по результатам измерения достоверности модели

ИЗМЕРЕНИЕ АДЕКВАТНОСТИ (ДИФФЕРЕНЦИАЛЬНОЙ И ИНТЕГРАЛЬНОЙ ВАЛИДНОСТИ) СЕМАНТИЧЕСКОЙ ИНФОРМАЦИОННОЙ МОДЕЛИ

Всего физических анкет: 10 (100% для п.15)
Всего логических анкет: 20

4. Средняя достоверность идентификации логических анкет с учетом сходства : 23.473%
5. Среднее сходство логических анкет, правильно отнесенных к классу : 12.487%
6. Среднее сходство логических анкет, ошибочно не отнесенных к классу : 0.474%
7. Среднее сходство логических анкет, ошибочно отнесенных к классу : 4.874%
8. Среднее сходство логических анкет, правильно не отнесенных к классу : 16.333%

9. Средняя достоверность идентификации логических анкет с учетом кол-ва : 54.000%
10. Среднее количество физич-х анкет, действительно относящихся к классу: 3.000 (100% для п.11 и п.12)
Среднее количество физич-х анкет, действительно не относящихся к классу: 7.000 (100% для п.13 и п.14)
Всего физических анкет: 10.000 (100% для п.15)
11. Среднее количество и % лог-их анкет, правильно отнесенных к классу: 2.750, т.е. 91.667%
12. Среднее количество и % лог-их анкет, ошибочно не отнесенных к классу: 0.250, т.е. 8.333% (Ошибка 1-го рода)
13. Среднее количество и % лог-их анкет, ошибочно отнесенных к классу: 2.050, т.е. 29.286% (Ошибка 2-го рода)
14. Среднее количество и % лог-их анкет, правильно не отнесенных к классу: 4.950, т.е. 70.714%

15. Средневзвешенная вероятность случайного угадывания принадлежности объекта к классу (%): 30.000
16. Средневзвешенная эффективность применения модели по сравнению со случ. угадыванием (раз): 5.900
17. Обобщенная достоверность модели (Д1+Д2)/2: 81.190%. Обобщенная ошибка (Е1+Е2)/2: 18.810%

21-02-09 22:23:03 г.Краснодар

N п/п	Код класса	Наименование класса	Достов. идентиф. лог.анк. с уч.количества эвр.крит	Кол-во лог.анк. дейст-но относящихся к классу	Количество логических анкет правильно или ошибочно отнесенных или не отнесенных к классу				Вероятн. случайного угадывания (%) =NLA/NFA	Эффектив модели по срав. со случ. угадыв. (раз)
					Правиль. отнесен.	Ошибочно не отнес	Ошибочно отнесен.	Правиль. не отнес		
1	2	3	9	10	11	12	13	14	15	16
1	1	Цифра 0	0.0	1	1	0	5	4	10.000	10.000
2	2	Цифра 1	20.0	1	1	0	4	5	10.000	10.000
3	3	Цифра 2	60.0	1	1	0	2	7	10.000	10.000
4	4	Цифра 3	40.0	1	1	0	3	6	10.000	10.000
5	5	Цифра 4	0.0	1	1	0	5	4	10.000	10.000
6	6	Цифра 5	0.0	1	1	0	5	4	10.000	10.000
7	7	Цифра 6	60.0	1	1	0	2	7	10.000	10.000
8	8	Цифра 7	40.0	1	1	0	3	6	10.000	10.000
9	9	Цифра 8	20.0	1	1	0	4	5	10.000	10.000
10	10	Цифра 9	40.0	1	1	0	3	6	10.000	10.000
11	11	Четная цифра	80.0	5	5	0	1	4	50.000	2.000
12	12	Не четная цифра	80.0	5	4	1	0	5	50.000	1.600
		Ср.взв.значения	54.0	3.0	2.8	0.3	2.1	5.0	30.000	5.900

Универсальная когнитивная аналитическая система НПП *ЭЙДОС*

ФОРМУЛЫ РАСЧЕТА ПОКАЗАТЕЛЕЙ ДИФФЕРЕНЦИАЛЬНОЙ ВАЛИДНОСТИ (ПО КЛАССАМ):
C04[k] = C05[k] - C06[k] - C07[k] + C08[k]
C09[k] = (C11[k] - C12[k] - C13[k] + C14[k]) / (C11[k] + C12[k] + C13[k] + C14[k]) * 100
C10[k] = C11[k] + C12[k]
C15[k] = C10[k] / NFiz * 100
C16[k] = C09[k] / C15[k]
где k - класс (соответствует строке)
где NFiz - суммарное количество физических анкет (объектов) в распознаваемой выборке

ФОРМУЛЫ РАСЧЕТА ПОКАЗАТЕЛЕЙ ИНТЕГРАЛЬНОЙ ВАЛИДНОСТИ (СРЕДНЕВЗВЕШЕННОЕ ПО ВСЕМ КЛАССАМ):
Ci = СУММА_по_k(Ci[k] * C10[k]) / NLog
где i = (4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16)
где NLog = СУММА_по_k(C10[k]) - суммарное количество логических анкет в распознаваемой выборке
ПРИМЕЧАНИЕ: учтены только результаты идентификации с модулем сходства не менее: 0→

Обоснованной выглядит гипотеза, что шрифт, в котором наиболее информативные элементы специально выделены, например яркостью или толщиной элементов как в представленном на рисунке 3, является более легким для восприятия, чем исходный шрифт с одинаковой толщиной графических элементов, представленный на рисунке 1. Однако экспериментальная проверка этой гипотезы требует проведения инженерно-психологических исследований и не входит в задачу данной статьи.

По-видимому, предложенный в статье подход может быть применен и для разработки эффективных *рекламных* стимульных материалов, а также в Ψ-технологиях[22].

Практическое занятие № 26

Системно-структурный синтез

[22] См., например, И. В. Смирнов: http://www.psycor.ru/main.php?smir

Цель работы: ознакомиться с элементами теории системно-структурного проектирования.

Теоретические сведения.

Направление системных исследований, ориентированное на синтез структур в системах различного вида деятельности человека, развивается на протяжении всего периода становления системных исследований. Работы в области синтеза систем базировались на различных подходах. Подход, который условно называли: *целевым* (см. гл. 1), можно считать синтезом структур от целей («сверху») к конечной структуре внизу. При применении подхода, называемого в разных изданиях *терминальным, лингвистическим, морфологическим,* синтез начинался от анализа пространства состояний, от элементов «снизу» вверх, к способам и принципам построения.

Исследования в этой области начинались на базе *математической логики* и были посвящены синтезу автоматов и схем путем введения правил взаимодействия логических элементов и минимизации структур на основе логических законов и теорем[23]. В настоящее время такого рода работы продолжают развиваться на базе *дискретной математики*.

Сфера приложений расширилась: синтезируются аналоговые и цифровые блоки электронных устройств[24], управляющие системы некоторых классов[25], одни и те же алгоритмы используются при проектировании структур электронных систем, человеко-машинных систем, проектирования баз знаний[26], разработки методик решения учебных задач.

Первым наиболее развитым и востребованным на практике не только для технических систем был ***кибернетический подход Л.А. Растригина***[27], основанный на идее *целевого* подхода.

Основную идею этого подхода составляет в терминах автора (двух стадийная схема принятия решений при управлении:

$$F_t \rightarrow Z^* \rightarrow U^*.$$

[23] ***Поспелов Д.А.*** Логические методы анализа и синтеза схем / Д. А. Поспелов. – М.: Энергия, 1968. – 328 с.

[24] ***Захаров В.К.***Электронные устройства автоматики и телемеханики: учеб. – 3-е изд. / В. К. Захаров, Ю. И. Лыпарь. – Л.: Энергоатомиздат, 1984. – 432 с.; ***Лыпарь Ю.И.***Структурный синтез электронных цепей / Ю. И. Лыпарь.. – Л.: Изд-во ЛПИ, 1982. – 84 с.; ***Лыпарь Ю.И.***Автоматизация проектирования избирательных усилителей и генераторов / Ю. И. Лыпарь.. – Л.:Изд-во Леннингр. ун-та, 1983. – 144 с.; ***Лыпарь Ю.И.***Системная теория структурного синтеза электронных схем / Ю. И. Лыпарь //Вычислительная техника, автоматика и радиоэлектроника: сб. трудов. – СПб: Изд-во СПбГПУ, 2002. – С. 120–127.

[25] ***Теория*** систем и методы системного анализа в управлении и связи. – М.: Радио и связь, 1983. – 248 с.

[26] ***Лыпарь Ю. И.*** База знаний для систем проектирования и обучения / Ю. И. Лыпарь // Тезисы докладов V межд. конф.: Региональная информатика. – 96. – СПб. 1996. – С. 251–252.

[27] ***Растригин Л. А.*** Современные принципы управления сложными объектами / Л. А. Растригин. – М.: Радио и связь, 1980. – 228 с.

На первой стадии $F_t \rightarrow Z^*$ определяется цель Z^* управления:

$$Z^* = \varphi_1(A_t, X), \qquad (26.1)$$

где φ_1 – алгоритм синтеза цели Z^* по потребностям A_t, и состоянию X среды. Формулируя цель, субъект как бы переводит свои потребности на язык состояния объекта Z^*: $Y \rightarrow Y^*_X$, что позволяет ему передать процедуру реализации управления U^*_X другому лицу (или даже автомату).

На второй стадии $Z^* \rightarrow U^*$ определяется управление U^*_X, реализация которого обеспечивает достижение цели Z^*:

$$U^*_X = \varphi_2(Z^*, X), \qquad (26.2)$$

где φ_2 – алгоритм управления. Этот алгоритм изучает кибернетика как наука об управлении.

Рисунок 26.1 – Последовательность из 8 этапов

Применительно к техническим системам получают формальные алгоритмы. Применительно к социально-экономическим объектам задачу первой стадии Л. А. Растригин предлагает решать на интуитивном уровне, а для второй стадии предлагает последовательность из 8 этапов (рис. 26.1).

Методы выполнения этапов зависят от конкретной задачи. Могут выполняться не все этапы, представленные на рисунке 26.1.

Идея организации процесса принятия решений в системах управления, предложенная ***Л. А. Растригиным***, остается актуальной и в настоящее время. Вместе с тем необходимо отметить, что в этом подходе задача формального синтеза структур объекта и системы управления не ставится, а решается параметрическая задача определения параметров модели системы управления.

Развивались и ведутся исследования по применению концепций системно-структурного синтеза, базирующихся на подходе «снизу»: структурно-функциональный подход ***А. С. Казарновского***[1]; номинально-структурный подход ***А. С. Лукьянченко***[2]. Они ориентированы на обеспечение полноты отображения *элементов* и *связей* системы путем различных вариантов формирования *структур* с помощью комбинаторных и морфологических приемов.

Однако для всех этих подходов характерно отсутствие даже постановки задачи о том, что делать со структурами, если их число будет огромно, как из них выбрать эффективные. Надо иметь в виду, что для реальных систем комбинаторным перебором порождаются структуры, образующие множество огромной мощности Например, если на всех аспектах проектирования и каждом их этапе рассматривать только девять вариантов, то мощность множества порождаемых структур превышает $4{,}8\times10^{101}$.В это число входят изоморфные структуры (их на два – три порядка больше неизоморфных), а также структуры, не ведущие к цели. Эта задача относится к так называемым *NP* полным, т. е. не решаемым простым перебором.

Поэтому путь проектирования снизу вверх от элементов системы к структуре объекта отдается на интуицию разработчика.

Однако потребитель нового объекта может описать его весьма упрощенно: сигналы на входе и выходе, некоторые условия работы, время разработки т вероятную стоимость выполнения проекта. В то же время результат должен быть хорошего качества, надежным и долговечным при малой стоимости и малом времени проектирования.

Поскольку на этом этапе еще нет структуры объекта и тем более численных значений параметров его элементов, то решать задачу структурного анализа невозможно. Более того, даже если известно множество структур объекта, выполнять параметрический синтез их всех и уже после этого сравнивать структуры между собой, чтобы убедиться, что ни одна из них не выполняет заданной функции выбора, экономически нецелесообразно.

Предложенная автором данного раздела теория системного синтеза структур обоснована в ряде работ [3]. Она позволяет на основе целей, свойств, которыми должна обладать проектируемая система, и ограничений осуществить синтез множества структур потенциально способных реализовать поставленную цель и свойства на всем множестве возможных решений (на множестве универсум *Un*).

---

[1] ***Казарновский А. С.***Структурно-функциональная модель сложного производственного объекта / А. С. Казарновский, Г. Ф. Ененко // Управляющие системы и машины, 1974, № 5. С. 3–7; ***Казарновский А. С.*** Совершенствование организационных структур промышленных предприятий: Вопросы методологии / А. С. Казарновский, П. А. Перлов, В. Т. Радченко. – Киев: Наукова думка, 1981. – 310 с.

[2] ***Лукьянченко А. С.*** Анализ и факторизация коммуникационных структур / А. С. Лукьянченко // Техника средств связи, 1979. – Сер. АСУ. – Вып. 1. – С. 59–72.

[3] ***ЛЫПАРЬ Ю.И.***Системное Проектирование. Функциональный и Структурный Аспекты. // Спб. Кибернетика и Информатика. Сборник Научных Трудов К 50-Летию Секции «Кибернетика» Дома Учёных Им. Горького РАН, Санкт-Петербург, Изд-во:Политехник. 2006 г.С. 217-238 ; ***Лыпарь Ю. И.*** Структурный синтез электронных цепей / Ю. И. Лыпарь. – Л.: Изд-во ЛПИ, 1982. – 84 с.; ***Лыпарь Ю. И.*** Автоматизация проектирования избирательных усилителей и генераторов / Ю. И. Лыпарь. – Л.: Изд-во Леннингр. ун-та, 1983. – 144 с.; ***Лыпарь Ю. И.*** Теория системного структурного синтеза / Ю. И. Лыпарь. // Труды Междунар. научно-практич. конф.: Системный анализ в проектировании и управлении. – СПб.: Изд-во СПбГТУ 2001. – С. 43–45.

Процедура проектирования (сверху – вниз) разбита на семь этапов (рис. 26.2). Принятые обозначения поясняются в процессе изложения подхода.

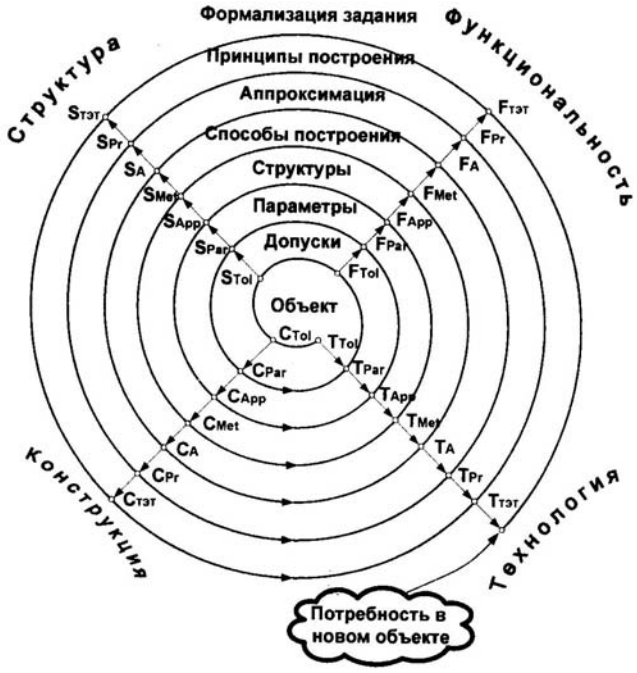


Рисунок 26.2 – Метамодель системно структурного проектирования объектов

На первом этапе осуществляется построение функций выбора, выделяющих с помощью лингвистических переменных и численных характеристик из множества *Un*подмножество определенного класса структур, потенциально способных выполнить задание (удовлетворить потребность потребителей объектов). Кроме того, лицо, принимающее решение (ЛПР) на функциональном аспекте (*F*), совместно с ЛПР по технологии(*T*)определяют возможность реализации объекта на существующей технологической базе за заданные время и стоимость.При положительном решении выполняется переход к проектированию на структурном (*S*), конструкторском (*C*)и технологическом (*T*)аспектах.Этап завершается формированием функции выбора для первого этапа проектирования с учётом всех аспектов.

На последующих этапах осуществляется построение функций выбора, помогающих для следующего этапа отсекать из этого класса те структуры, которыене удовлетворяют требованиям функционирования, изготовления и эксплуатации проектируемой системы в условиях ограничений и взаимодействия с окружающей средой.

Оставшиеся после отсечения решения неразличимы и названы эффективными для конкретного этапа.

Здесь важно подчеркнуть, что предлагаемые «алгоритмы синтеза не требуют указания куда идти, но указывают, куда не надо идти» (В. А. Трапезников), т. е. не требуется рассматривать для каждого уровня иерархии все элементы множества решений.

ЛПР на каждом из этапов выбирает одно из решений, полученного подмножества, основываясь на своём опыте. Вначале каждого *i*-го этапа проектирования

на функциональном аспекте создаются функции выбора $F_{\text{ТЭТ}\,i}$, которые помимо исходных содержат дополнительно сформулированные требования по результатам анализа решений выполненного этапа по всем аспектам.

Таким образом, отсекаются от исходного *NP* полного множества с помощью лингвистических переменных и части численных данных решения, не удовлетворяющие исходным и дополнительным требованиям. Число и последовательность этапов обусловлены необходимостью получения всех необходимых данных для решения задач очередного этапа. Принципиальный отказ от нахождения только одного «оптимального» решения обусловлен невозможностью учесть на i – м этапе проектирования все нюансы ограничений и требований последующих аспектов и этапов. Например, структура может быть отвергнута на конструкторском или технологическом аспекте. Поэтому задачи всех аспектов решаются с небольшим сдвигом по времени параллельно.

Чтобы вновь возникшие данные или ограничения можно было оперативно учесть, введены контуры обратной связи на всех аспектах и этапах. Возврат на i-ый этап сопровождается добавлением на функциональном аспекте в функции выбора предыдущего аспекта нового ограничения или требования, что просто сузит область решений этого этапа и упростит задачу выбора ЛПР.

Если же вновь возникшее требование не может быть реализовано на множестве решений i-го этапа, то по контуру обратной связи оно может быть решено на соответствующем аспекте этого же этапа или же на этапе i-1, i-2,..,1.

Заметим, что эта технология существенно сокращает время проектирования, что особенно важно в областях с высоким уровнем конкуренции, например, для электронных систем.

Обратите внимание на особенность подхода: проектирование идет от функций выбора с малым числом критериев к большему их числу, что прямо противоположно традиционным методам оптимизации.

Если бы процесс проектирования был нацелен на выработку самого лучшего решения, то в результате задача синтеза множества эффективных структур вообще не могла быть решена в рамках такого подхода.

Как показали исследования качества электронных систем, оно определяется, прежде всего, их структурой. На конструкторском и технологическом аспектах невозможно улучшить качество системы с плохой структурой. Более того, на последующих аспектах качество системы может только ухудшится из-за внесения различных дополнительных элементов и связей, отсутствовавших в решениях структурного аспекта.

Рассмотрим этапы проектирования (рис. 26.2). Прежде чем проектировать объект необходимо по возможности формализовать исходные требования $F_{\text{ТЭТ}}$, сформулированные потребителями и ЛПР на каждом аспекте проектирования. На первом и последующих этапах функционального аспекта $F_{\text{ТЭТ}}$ формируются для остальных аспектов многоцелевые функции выбора, которые образуют семи этапное сито с изменяющимися от этапа к этапу размерами и формой ячеек для выделения из множества универсум *Un* возможных вариантов объектов только удовлетворяющих ТЭТ.

Первый этап – синтез целей и их моделей, формализация свойств и ограничений $F_{ТЭТ}$; *второй этап* – синтез принципов построения S_{Pr}; *третий* – аппроксимация S_A (создание идеального облика (обликов), плана, характеристик предмета проектирования); *четвертый* – синтез способов построения S_{Met}; *пятый* – синтез структуры S_{App}; *шестой* и *седьмой* – соответственно синтез параметров S_{Par} и допусков на них S_{Tol}.

Необходимо отметить, что из-за различия языков описания, а также от зависимости формализации и уточнения целей предшествующих задач от последующих в метамодели введены контуры обратной связи на всех этапах и аспектах. Это увеличивает число ограничений в ТЭТ и решает проблему перехода от *NP* полной задачи к линейной.

Все этапы проектирования в литературе часто называют просто синтезом без уточнения предмета синтеза, из-за чего иногда возникают недоразумения. На этапах 1–5 решаются задачи синтеза структур, а на двух последних – осуществляют синтез параметров.

Третий, шестой и седьмой этапы проектирования совпадают по целям с этапами 3 – 8 подхода Растригина имеют развитый математический аппарат и решаются достаточно успешно при решении технических задач.

Заметим, что совершенствованию именно этих методов посвящается большинство публикаций по синтезу. Остальные этапы по сложности значительно превосходят вышеупомянутые и относят к разряду изобретательских: синтез оригинальной структуры, нового способа и принципа является основанием для патентования соответственно устройства и способа. Третий этап для художественных и дизайнерских задач пошива одежды также относится к изобретательским, хотя основа в них достаточно технична.

Формулировка и формализация целей в настоящее время ближе к искусству, чем к алгоритмизируемым шагам, хотя и здесь можно сослаться на работы в которых описаны подходы и методики, позволяющие с большим или меньшим успехом решать эти задачи в разных областях человеческой деятельности.

Представим формально процесс проектирования в виде отображения П, имеющего область определения на множестве значений ТЭТ, а область значений имеет во множестве структур $K^*_{\text{р}}$, во множестве значений параметров X^* их элементов допустимых по ТЭТ, и во множестве допусков d^*_{tol} на технологический разброс параметров X^* элементов.

Излагаемая ниже процедура проектирования имеет общий характер и применима для проектирования электронных устройств, систем управления, портфеля ценных бумаг, проектирования системы «оператор – ЭВМ», пошива одежды, построения художественных картин, разработки методик лечения больных, создания баз знаний, методик обучения.

Представим формально процесс проектирования в виде отображения П, имеющего область определения на множестве значений ТЭТ. Отображение П представим композицией промежуточных отображений

$$\Pi_S = S_{Tol} \circ S_{Par} \circ S_{App} \circ S_{Met} \circ S_A \circ S_{Pr} \circ F_{\text{ТЭТ}}, \qquad (26.3)$$

где $F_{\text{ТЭТ}}$ – потребность в новом объекте, выраженная через технологические, технические, эксплуатационные, экономические, экологические и эргономические

требования; S_{Pr} – синтез (выбор) множества принципов построения; S_A– синтез функций, аппроксимирующих характеристики объекта; S_{Met} – синтез множества способов построения (*Met* от *methods*); S_{App} – синтез множества структур устройства для каждого способа (*A p p* от appliance); S_{Par} – синтез параметров элементов для выбранной ЛПР структуры;S_{Tol} – синтез допусков (*tolerance*) на параметры элементов; ∘–символ композиции.

Начинают процесс проектирования с выполнения отображения $F_{\text{ТЭТ}}$, которое описывает процесс постепенной формализации ТЭТ для всех последующих этапов, делая ТЭТ, все более детальными.

$$F_{\text{ТЭТ}} : \text{ТЭТ} \rightarrow F(as);\ F(as) = (F_S, F_C, F_T), \tag{26.4}$$

где $F(as)$ – совокупность функций выбора для каждого аспекта проектирования. Все они содержит функциональные соотношения, условия и ограничения в задачах выбора решений.

Начинают процесс проектирования с функционального аспекта, преобразуя ТЭТ в функциональные соотношения и разделяя их по аспектам. В результате после каждого *i*–го этапа проектирования выполняется формирование новой функции выбора $F_{\text{ТЭТ}(i+1)}$, с большим числом критериев в функциях (6.43) и на меньшей мощности множества решений.

Таким образом, процесс поэтапной формализации выполняется для всех последующих этапов, увеличивая мощность множества ТЭТ, делая его все более детальным:

$$F_{\text{ТЭТ}\,i} : \text{ТЭТ}_i \rightarrow F_{Efi},\ i = (1, 2, \ldots, 7), \tag{26.5}$$

где F_{Efi} – *i*-ая функция выбора в решении задач выбора на *i* -ом этапе. В результате выполнения отображения (6.44) будут формироваться функции выбора для каждого из аспектов, а в целом образуется следующая последовательность этапов:

$$\begin{aligned} \Pi_S &= S_{Tol} \circ S_{Par} \circ S_{App} \circ S_{Met} \circ S_A \circ S_{Pr} \circ F_{\text{ТЭТ}}, \\ \Pi_C &= C_{Tol} \circ C_{Par} \circ C_{App} \circ C_{Met} \circ C_A \circ C_{Pr} \circ C_{\text{ТЭТ}}, \\ \Pi_T &= T_{Tol} \circ T_{Par} \circ T_{App} \circ T_{Met} \circ T_A \circ T_{Pr} \circ T_{\text{ТЭТ}}. \end{aligned} \tag{26.6}$$

При этом для каждого этапа *i* и аспекта формируется принцип эффективности (Ef_i) (26.4), отражающий представление ЛПР о качестве проектируемой структуры данного этапа. Эти принципы управляют процессом синтеза и постепенно выделяют из совокупности всех возможных структур (из множества универсум *Un*) подмножество все меньшей мощности.

После построения функции выбора для первого этапа реализуется отображение S_{pr}, которое соответствует синтезу или выбору одного из известных принципов построения проектируемой структуры. В настоящее время широко используются следующие принципы: последовательный и параллельный, с обратной связью, распределённый, иерархический и т.д. Для больших систем это иерархия уровней главного, функционального, элементного с повторением этих же уровней

иерархии при дальнейшей декомпозиции второго и третьего уровня. Универсальность модели состоит в том, что для каждого уровня проектирования системы справедлива композиция (26.6). Поэтому проектирование начинают с синтеза структур, использующих простейшие элементы и создающих набор компонентов и (или) подсистем для выполнения необходимых операций более высокого уровня. Очевидно, что компоненты и подсистемы выполняют роль простейших элементов для вышестоящего уровня.

Таким образом, образуется иерархия уровней проектирования и **иерархия элементов базы знаний** для разных уровней проектирования.

Отображение имеет область определения на множестве ТЭТ и универсальном множестве структур K_{Un}, а значение во множестве версий структур $K_{Pr} \subseteq K_{Un}$ способных реализовать синтезированный принцип. Синтез ведется под управлением функции выбора F_1, являющейся математическим выражением принципа эффективности Ef_1

$$S_{Pr}: F_{Ef1} \cap P_r \rightarrow K_{Pr};\ K_{Pr} = \{K_{Pri}\}, \qquad (26.7)$$

где F_{Ef1} – функция выбора в задаче синтеза множества принципов построения объекта.

Как показали исследования, сравнение синтезированных принципов, целесообразно осуществлять по их функции относительной чувствительности.В частности, при параллельном принципе построения и с контурами обратной связи чувствительность можно существенно уменьшить, при этом качество (надежность, стабильность, повторяемость характеристик и параметров и т. п.) системы улучшается, хотя возможно увеличение её стоимости.

Отображение S_A соответствует этапу формального описания вида объекта проектирования, некоторых его характеристик или параметров. В необходимых случаях можно прибегнуть к теории аппроксимации желаемого вида характеристик [1] и параметров объекта. Например, можно использовать полиномиальную аппроксимацию полиномами Чебышёва, Баттерворта, Бесселя и др.

Такими средствами будет создана математическая модель объекта проектирования. Для технических систем это достаточно частый путь создания моделей.

Отображение имеет область определения на множестве значений K_{Pr} и функции выбора F_{Ef2}, задающей критерии оптимальной аппроксимации и физической реализуемости на заданных в ТЭТ ограничениях и элементном базисе. В результате решения задачи F_{Ef2}, (K_{Pr}, Ef_2),выделяют из множества K_{Pr}подмножество версий моделей структур K_A, а область значений во множестве функций заданного класса $D(Z, p)$ (формальных описаний вида всего объекта, каких-то его частей, сторон или характеристик).

$$S_A: F_{Ef2} \cap K_{Pr} \rightarrow D(Z, p),\quad D(Z,p) = \{D_k(Z,p)\},\ k = \overline{(1,\psi)}, \qquad (26.8)$$

где p– комплексная переменная; Z– вектор коэффициентов.

Оператор синтеза способов построения структур S_{Met}выделяет из множества K_Aподмножество K_{pMet}структур. Они реализуют не только синтезированный прин-

[1] Напр.,***Рыжиков Ю.И.*** Решение научно-технических задач на персональном компьютере / Ю. И. Рыжиков. – СПб.: КОРОНА принт, 2000. – 272 с.

цип построения, но и удовлетворяют заданным ТЭТ $-E_{f3}$ и функции $D(Z, p)$, т. е.

$$S_{Met} : K_A \cap E_{f3} \cap D(Z, p) \rightarrow K_{pMet},$$

$$K_{pMet} = \{K_{pMetj}\},\ j=(1, 2, \ldots, \mu), \qquad (26.9)$$

где область значений K_{pMet} является множеством способов построения структур.

Способ построения K_{pMetj}–этото, что в патентной литературе называют способом, но в отличие от патента, здесь он должен быть изложен не столько вербально, сколько с помощью алфавита описания структур K_A, некоторых параметров функции $D(Z, p)$ и ТЭТ, задающих функции выбора F_{Ef3}. Фактически это означает, что коэффициенты $z_i \subset Z$ представляются в виде некоторых структур, анализ которых с помощью функции выбора F_{Ef3} позволяет выбрать эффективные.

Дальнейшее уменьшение мощности множества K_{pMet} достигается с помощью структурного анализа и выделения из множества эффективного, на взгляд ЛПР, способа j, предназначенного для последующей реализации в процедуре синтеза S_{App} множества возможных структур

$$S_{App} : F_{Ef4} \cap D(Z, p) \cap K_{pMetj} \rightarrow K_{pApp}. \qquad (26.10)$$

Выполнение этого отображения порождает множество эквивалентных, с точки зрения области значений S_{App} структур $K_{pApp} = \{K_{p1}, K_{p2}, \ldots, K_{p\gamma}\}$.

Каждая из этих r структур описывается функцией

$$K_j(p) = U_2(p)/U_1(p) = B(p)/A(p) =$$

$$= (b_0 + b_1 p + \ldots + b_m p^m)/(a_0 + a_1 p + \ldots + a_n p^n), \qquad (26.11)$$

где $U_1(p)$ и $U_2(p)$ – входные и выходные материальные потоки. Вид и порядок полиномов числителя и знаменателя функции (9) совпадают с соответствующими коэффициентами полиномов функции (6.46).

Последнее множество K_{pApp} совместно с исходными ТЭТ является областью определения отображения F_K, имеющее область значений во множестве эффективных структур с оптимальными параметрами X^*

$$F_K : K_{pApp} \cap F_{Ef5} \rightarrow K_p^*. \qquad (26.12)$$

Схемотехническое проектирование завершает этап определения допусков на параметры элементов. Этап описывается отображением S_{tol}, имеющим область определения на множестве X^* эффективной структуры, а область значений во множестве d^*_{tol}, или

$$S_{tol} : K_p^* \cap F_{Ef6} \rightarrow d^*_{tol}. \qquad (26.13)$$

Полная реализация системного подхода осуществляется, если на каждом шаге процедуры проектирования порождается множество эффективных решений, предоставляя тем самым возможность проводить поиск эффективных решений на последующих шагах синтеза.

Ниже приведены примеры применения теории структурного синтеза в различных областях, иллюстрирующие единство методологии.

Синтез структур электронных цепей, описываемых линейными, кусочно-линейными, динамическими и статическими уравнениями с постоянными и переменными во времени параметрами. Рассмотрим особенности процедуры (26.4) для этого класса систем. Этапы (26.5) и (26.6) решаются так же, как и выше. В результате выполнения отображения (26.6) получают функцию (чаще всего передаточную) $D(Z, p)$, описывающую желаемую временную, частотную или фазовую харак-

теристики синтезируемого устройства. Здесь удобно использовать преобразование Лапласа, которое позволяет относительно просто перейти во временную или частотную область в зависимости от решаемой задачи. В общем случае вид этой функции следующий

$$D(Z,p)=\frac{\sum_{i=0}^{m} z_i\, p^i}{\sum_{i=m+1}^{m+1+n} z_i\, p^{\,i-m-1}}\,, \qquad (26.14)$$

где p– оператор Лапласа; z_i– числа.

Заметим, функцию (26.12) нельзя было построить раньше определения принципа построения. Теперь известен вид математической модели проектируемого устройства и класс, к которому оно относится. Математической моделью всего класса является функция (26.9), совпадающая по виду с функцией (26.12), но имеющая символьные коэффициенты b_j и a_i.

С помощью отображения (26.7) порождаются способы построения структур систем и устройств K_{pm}, вся совокупность которых и образует класс. Число независимых способов чаще всего невелико, но все возможные их сочетания могут составить внушительное число. Поэтому уже здесь необходим структурный анализ с целью сравнения их между собой и выделения эффективных с точки зрения ТЭТ способов.

Конструирование способов, их исчисление, основано на представлении коэффициентов b_j и a_i в виде некоторых структур – слагаемых, создаваемых элементами будущей системы. Для электронных схем доказано [1], что способы описывают структуру формирования коэффициентов функции (26.9) с помощью алгебраических операций сложения, вычитания и умножения. Эти операции (правила объединения) выполняются с символами элементов схем (размерными и безразмерными величинами, отражающими, например, проводимость или безразмерный коэффициент передачи). Поэтому порождаемая ТЭТ функция выбора F_{Ef3} (правила вывода) выражается через взаимосвязи и структуру этих коэффициентов. Конечно, функция F_{Ef3} не могла быть составлена до решения задачи (26.12). Если же пропустить этап (26.7) и сразу же перейти к (26.8), то не будет решаться задача F_{Ef3}. Совокупность правил объединения и вывода образует исчисление способов построения. Их сравнение осуществляют по степени выполнения ТЭТ и по значениям функций относительной чувствительности. Последние характеризуют параметрическую надёжность, стабильность характеристик, их повторяемость при изготовлении, динамический диапазон сигналов и уровень шумов. Иными словами, минимизация функций относительной чувствительности к изменению параметров всех элементов схем приводит к улучшению большинства показателей качества устройств. Выделенные с помощью функции выбора F_{Ef3} эффективные способы будут использованы в процедурах (26.8). Каждый способ для данного уровня образует подкласс эквивалентных структур, среди которых в дальнейшем только часть

[1] ***Лыпарь Ю. И.*** Структурный синтез электронных цепей / Ю. И. Лыпарь. – Л.: Изд-во ЛПИ, 1982. – 84 с.; ***Лыпарь Ю. И.***Автоматизация проектирования избирательных усилителей и генераторов / Ю. И. Лыпарь.. – Л.: Изд-во Леннингр. ун-та, 1983. – 144 с.

окажется эффективной из-за того, что на данном уровне иерархии часто ещё невозможно учесть детальные конструкторские, технологические и т.п. требования.

Областью значений отображения (26.8) является множество схем, описываемых функцией, совпадающей по виду с функцией (26.12). Неявное задание множества схем K_pявляется исчислением, которое описано ниже. Комбинаторные алгоритмы порождения неизоморфных схем K_{pi} (p) основаны на теории графов и представляют собой композицию, вызывающую целый ряд последовательных процедур

$$S_{St}=S_{W_{\text{у}}} \circ S_{\Gamma W} \circ S_{\Gamma B} \circ S_{\Gamma A} \circ S_{B_{ак}} \circ S_{A_{ак}} \circ S_{A_{\text{п}}} \circ S_{\Gamma}, \qquad (26.13)$$

синтеза ненаправленных S_{Γ} и направленных $S_{A\,ак}$, $S_{B\,ак}$ графов, отражающих соответственно структуру пассивной части схемы, контуры обратной связи и пути прохождения сигналов. Именно последние два графа в основном отражают способы построения системы. Раскраска $S_{A\text{п}}$ графа Γ в цвета элементов допустимых по ТЭТ и совмещение $\left(S_{\Gamma A}, S_{\Gamma B}\right)$ его с графами, описывающих собственное поведение системы (Γ^A) и пути прохождения сигнала от входа к выходу (Γ^B), ведется под управлением функции выбора F_{Ef4}. Преобразование ($S_{\Gamma W}$) полученного графа в смешанный, позволяет по нему легко построить схему.

Для выполнения процедур не требуется иметь какие-либо сведения о структуре будущегоустройства, но и не отвергается использование уже накопленного опыта и знаний.

Последним этапом синтеза структуры является построение принципиальной схемы устройства S_{Wy} по смешанному графу, которое необходимо сделать для выполнения синтеза конструкции. Операция построения схемы по графу тривиальна.

При выполнении каждого оператора процедуры (26.13) необходимо контролировать соблюдение F_{Ef4}с помощью структурного[1]и символьного анализов, а также анализа чувствительности. В результате становятся известными предельно достижимые характеристики и вторичные параметры устройства, соответствие их, а также структуры критериям эффективности. В случае невыполнения критериев возвращаются на один шаг назад и повторяют процедуру с учетом полученного опыта. Иногда ТЭТ таковы, что невозможность их выполнения выясняется только на уровне K_p, тогда необходимо подняться на более высокий уровень иерархии и выбрать другой способ построения или изменить ТЭТ.

Отметим, что была разработана синтезированная универсальная (в смысле возможного элементного базиса и размерности реализуемых коэффициентов) каноническая по числу узлов и элементов исходная схема[1]. Анализ более 2000 электронных схем различного назначения показал, что более 87% из них были порождены из исходной схемы. Авторам этих схем она не была известна, а поэтому потребовала от разработчиков не малого времени и усилий, чтобы решить свою задачу. Важно также то, что среди проанализированных схем были аналоговые и цифровые, линейные и нелинейные, с постоянными и переменными во времени параметрами устройства. Это обстоятельство, конечно, имеет большое методическое

[1] ***Захаров В. К.*** Электронные устройства автоматики и телемеханики. Учебник для вузов / В. К. Захаров, Ю. И. Лыпарь. – 3-е изд. – Л.: Энергоатомиздат, 1984. – 432 с.

значение, так как формирует системное видение и взаимосвязь различных устройств, а кроме того позволяет чаще всего пропустить первые две процедуры (S_{Γ}, S_A), что ускоряет синтез.

Таким образом, для получения множества K_p разработаны процедуры синтеза, для которых достаточно сведений, имеющихся в ТЭТ.

Следовательно, чем более подробно будут сформулированы ТЭТ и чем более полно будут указаны недостатки известной схемы-прототипа, тем более качественное решение и за меньшее время оно будет получено. Заметим, пропуская этап (8), как это делается в методах прямой реализации, не реализуют F_{Ef3} и F_{Ef4}. На этом системный синтез структур (*SSS*) заканчивается и кратко его можно описать в виде композиции:

$$SSS = S_{App} \circ S_{Met} \circ S_A \circ S_{Pr} \circ F_{\text{ТЭТ}}.$$

Когда говорят о процедуре проектирования снизу, то неявно предполагают, что известен принцип построения и аппроксимирующая функция, задающая идеальный «облик» системы или устройства. Затем необходимо решать задачи (26.8) и (26.7), но в этом случае без критериев Φ_{Op4} и Φ_{Op3} они становятся совершенно неподъемными из-за очень большого числа вариантов.

Реализация отображения (26.9), т.е. расчет параметров элементов состоит в том, чтобы решить систему нелинейных алгебраических уравнений и неравенств – ограничений, получаемых путем уравнивания соответствующих коэффициентов в передаточных функциях $D(Z, p)$ и $K_{pi}(p)$. Обычно задача решается с использованием методов оптимизации. Критерию F_{Ef5} удовлетворяют далеко не все схемы из K_p, поэтому $K^*_p \subseteq K_p$.

Схемотехническое проектирование завершает этап определения допусков на параметры элементов, осуществляемый отображением (26.11), в котором F_{Ef6} – отражает компромисс между стоимостью и допуском на изготовление элементов, а также технологические ограничения на допуски.

Полученное множество эффективных схем оставляет конструктору печатной платы, чипа, технологу интегральной схемы возможность находить эффективные решения на своем уровне. Это приведет к дальнейшему уменьшению мощности множества схем K^*_p.

Итак, декомпозиция (26.13) построена так, что ее отображения структурируют цели синтеза, и выделяют пути их решения, не требуя априорных знаний о будущей схеме. Одновременно видна большая важность формализации ТЭТ. Системность подхода обеспечивает открытость системы проектирования, применимость к различным областям деятельности, накопление знаний и опыта, предполагает, но не обязывает проводить оптимизацию на всех этапах.

Накопление знаний и опыта в такой системе проектирования достаточно эффективно, так как запоминаются из ТЭТ предельные значения характеристик и параметров, реализованных схем, уровень выполнения ограничений; принципы, эффективные способы и схемы; конструкторские и технологические решения.

Опишем метод формализации технических требований (ТТ), которые играют важнейшую роль в процессе проектирования. Охарактеризуем типовые ТТ, которые определяют:

а) вид и параметры выходных сигналов и характеристик (время задержки, частоту генерации, добротность нулей и полюсов передаточной функции, амплитудно-частотную, фазовую, временную и т. д. характеристики);

б) порядок системы дифференциальных уравнений проектируемого устройства (порядок аппроксимирующей передаточной функции $D(Z, p)$ и наличие в определителях системы уравнений действительных, комплексно-сопряженных и мнимых корней;

в) условия перестройки параметров (в том числе электронной);

г) стабильность характеристик при изменении условий окружающей среды;

д) слабую зависимость расстояния между нулями и полюсами от параметров элементов;

е) набор элементов, из которых производится синтез схемы;

ж) удобство тестирования работы устройства.

Системное проектирование и управление портфелем ценных бумаг (ЦБ). Исследования фондового рынка США показывают, что при составлении портфеля ценных бумаг (ЦБ) большое значение для достижения его доходности выше среднерыночной имеют моменты выбора времени включения и вывода ЦБ из портфеля, а также перераспределения объёма инвестиций между его отдельными инструментами. Например, индекс S&P 500, являющийся главным стандартом сравнения для большинства взаимных и пенсионных фондов США, за период с 1975 по 2002 годы только в 10,7 % времени превосходил по доходности казначейские векселя. Поэтому, управляя составом портфеля, можно повысить его доходность, если вовремя спрогнозировать моменты времени для переформирования состава и объема отдельных инструментов в нём.

Изложим кратко основную процедуру проектирования портфеля ЦБ [1], показав, что помимо традиционных этапов она содержит и новые. Представим процесс проектирования как отображение П, имеющее область определения на следующих требованиях:

1.**Технические** – множество типов рынков, значения объёмов инвестиционного капитала (IC), сроки инвестиций (времени жизни инвестиций) $T_{ж}$, желаемой (норме) доходности $Y_{пок}$ к покупке, уровне допустимого риска снижения доходности σ.

2.**Технологические** – работа на бирже в режимах on-line, дневных торгов, объёмов лотов, и т.д.

3.**Экономические** – на внешне- и внутри-экономических, политических, отраслевых и региональных условиях, а также финансовых условиях инвестора.

4.**Эксплуатационные** – информационная система (ИС) о текущих экономических, политических, экологических событиях, могущих повлиять на тренды ЦБ; база данных (БД); внутренняя система, отражающая информацию о состоянии портфеля и его отдельных бумаг; подсистема прогнозирования и диверсификации рисков.

[1] Лыпарь Ю.И., Косенков А.Н., Ельцов А.А. Системное проектирование и управление портфелем ценных бумаг // М.: Физматлит, Труды междунар. конф. «Интеллектуальные САПР, интеллектуальные системы», т. II, 2005. – С. 491-498

Отображение Π_S имеет значение во множестве эффективных структур портфеля K^*_p, во множестве значений параметров X^* их инструментов, допустимых по ТЭТ (цены покупки $P^*_{\text{пок}}$, продажи $P^*_{\text{пр}}$; денежные объёмы V^* ЦБ и их число в портфеле N^*) и во множестве допусков d_{Tol} на технологическое отклонение параметров X^* ЦБ. Отображение Π_Sпредставимкомпозицией семи промежуточных отображений

$$\Pi_S = S_{Tol} \circ S_{Par} \circ S_{App} \circ S_{Met} \circ S_A \circ S_{Pr} \circ F_{\text{ТЭТ}},$$

где $F_{\text{ТЭТ}}$ – область определения, т.е. сформулированные ЛПР часть критериев Ef_i, $i=(\overline{1,\ 6})$ для построения функций выбораF_{Efi}, структур портфелей и параметров инструментов; S_{Pr}- соответствует синтезу или выбору из известных одного или нескольких принципов Pr (стратегий) построения портфеля ЦБ. Для диверсификации рисков обычно портфель инвестиций содержит одновременно, т.е. параллельно, несколько портфелей из различных типов ЦБ (акций, облигаций, опционов, фьючерсов, валюты) из взаимно дополняющих отраслей и взаимно заменяющих в одной отрасли; S_A - синтез математической модели желаемого идеального образа кривой доходности, изменяющейся во времени за время инвестиций $T_{\text{ж}}$; S_{Met} – синтез способов построения структур портфелей с помощью всех доступных инструментов, изменяющихся во времени в портфеле по составу и по параметрам: цен покупки и продажи $P^*_{\text{пок}}$, $P^*_{\text{пр}}$ и объему V^*;S_{App} – синтез в соответствии с S_Aи S_{Met}портфеля ЦБ в первоначальный и последующие моменты времени на основе анализа конъюнктуры рынков первичных и производных ЦБ; S_{Par} – синтез параметров инструментов с учётом необходимости диверсификации рисков; F_{tol}- синтез допусков σ на параметры каждого инструмента и выработке сигналов stop-lossorder, для ограничения убытков по каждой ЦБ.

Функции выбораF_{Efi}управляют процессом синтеза и постепенно выделяют из совокупности всех возможных структур (из множества универсум Un) подмножество всё меньшей мощности. Первоначально они формируются из ТЭТ, а потом в ходе реализации i-го отображения F_{Efi+1} дополняются критериями, полученными на i-ом и предыдущих этапах и описанных на языке соответствующего этапа или вербально.

В настоящее время при формировании портфеля чаще всего используются следующие принципыS_{Pr}: последовательный, параллельный, иерархический, с обратной связью для отдельных ЦБ и для всего портфеля, распределённый, и т. д. В случае, если известные принципы не позволяют достичь нужного качества портфеля, что выясняют на функциональном аспекте проектирования и управления портфелем, то на структурном аспекте синтезируется новый принцип предварительно выбрав из известных принципов наиболее близкий к необходимому и диагностировав причину, из-за которой не удаётся достичь целей. Для сравнения разных принципов построения используются функции относительной чувствительности и F_{Ef1},. В результате из всех возможных принципов остаются для дальнейшей реали-

зации несколько равносильных, с помощью которых и формируется функция выбора F_{Ef2}.

Аппроксимация ***A*** доходности – реализуется с помощью достаточно хорошо разработанного программного обеспечения, например MatLab. Однако необходимо учитывать, что в связи с изменяющимися экономическими и политическими условиями работы рынков за время $T_ж$ бывает необходимо неоднократно её менять (например, при изменении ставки рефинансирования, изменяющей условия налогообложения, изменении уровня инфляции, цен на энергоносителей). Аппроксимирующая функция времени D (***Z***, t) чаще всего является кусочно-линейной функцией, в которой ***Z*** является вектором численных коэффициентов. Кроме того, формируется функция выбора F_{Ef3}..

В результате синтеза **способов построения** S_{Met} структуры портфеля получают функцию времени доходности K (a,b,t) по виду, совпадающую с D (***Z***, t). Эта функция отличается от последней тем, что содержит вместо численных коэффициентов z_k их символьные выражения a_i, b_j, которые в общем виде описывают различные возможности достижения планируемой доходности при снижении рисков или при заданном уровне риска повышения доходности портфеля. Фактически коэффициенты a_i и b_j отражают взаимодействие базовых и их производных инструментов, создающих контуры обратной связи. Сравнение способов между собой осуществляют с помощью F_{Ef3} и функций относительной чувствительности, так как они отражают совокупное качество портфеля. Завершается этап формированием функции выбора F_{Ef4} и выделением подмножества эффективных способов.

Для этапа **синтеза структуры** S_{st} портфеля ЛПР выбирает из указанного подмножества один из способов построения и на основе анализа рынков и прогноза трендов цен на ближайшее время выбираются инструменты потенциально могущие принести заданную в $D(Z, t)$ доходность. С помощью теории графов синтезируется структура портфеля. Для уменьшения рисков потерь и в условиях ограниченности инвестиционного капитала IC синтез целесообразно начинать с исходных графов (рис. 26.3 а). Наиболее простым будет портфель, образованный по типу, отображенному на рисунке 26.3, в. В нём ветвь ак2 отражает опцион, по базовому элементу ак1. Если к графу на рисунке 26.3, в добавить ещё две параллельные ветви, отражающие индексы акций и облигаций на некоторых рынках, то получится граф – рисунок 26.3, г. При использовании нескольких независимых бирж и торговли только с акциями и облигациями получаются структуры, показанные на рисунке 26.3, б.

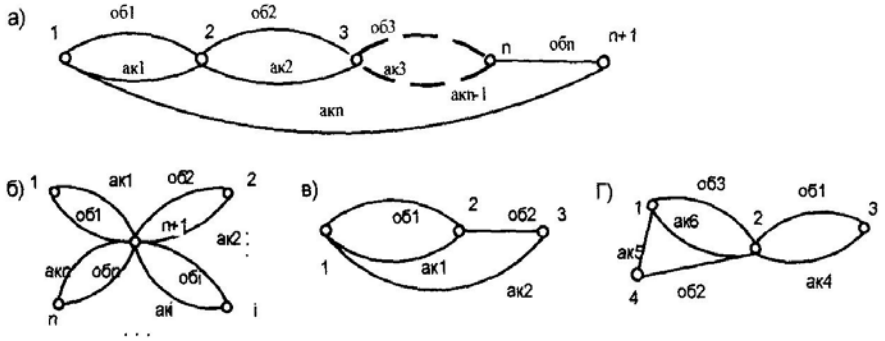


Рисунок 26.3 - Графы, отображающие структуры исходных портфелей

Синтез параметров *S* инструментов и портфеля осуществляется с помощью информационной системы (ИС) сбора и обработки биржевых данных.

В экспериментах использованы сводки ежедневных торгов с нью-йоркской фондовой биржи (*NYSE*). Произведен автоматический сбор данных о торгах, дивидендах и сплитах из Интернета, осуществлено приведение данных по ценам и объемам к сплитам.

Во-первых, построена процедура ранжирования акций компаний по степени доминирования в секторе рынка и уровню влияния на остальные компании, составлены наборы «доминанта – ведомые акции».

Во-вторых, проверена гипотеза о возможности прогнозирования направления ценового тренда акций компании на основе данных о текущем тренде доминирующей над ней акции.

Подтверждено предположение о том, что вероятность соответствия направлений предсказанного экстраполированного и базового трендов зависит от коэффициента взаимной корреляции между ценами ведущей и ведомой акции.

Также подтверждено предположение, что эта вероятность может быть разной в зависимости от длины тренда. С использованием имеющихся исторических данных был построен масштабный расчетный эксперимент.

Для этого выделен набор из 10 отрезков трендов (на основе ряда Фибоначчи) и произведена проверка гипотезы над данными о торгах 500 ведущих компаний США, котирующихся на NYSE, за последние 25 лет (с 1980г).

Всего произведено более $92 \cdot 10^6$ экспериментов, собрана статистика, и построена усредненная двумерная зависимость вероятности совпадения направления базового и предсказанного наведенного трендов от длины трендов и коэффициента взаимной корреляции.

Анализ полученной зависимости дал основания считать, что обе зависимости (от длины и коэффициента) можно аппроксимировать как линейные.

В-третьих, построена программная реализация нейронной сети слоистой архитектуры, использующей алгоритмы градиентного спуска и обратного распространения ошибки.

Правильный ответ о направлении тренда нейронная сеть давала в среднем на уровне 70%, что говорит о том, что эта информация может быть использована в качестве составной компоненты для принятия решения о покупке или продаже акций. Что касается государственных облигаций, то доминирующие облигации могут прогнозировать тренд облигации, которая только будет выпущена в обращение.

Синтез допусков S_{Tol} задает допустимые отклонения доходности каждой ЦБ и всего портфеля и обеспечивается постановкой приказов stoploss.

Резюмируя вышеизложенное, покажем результат проектирования информационно-управляющей системы для фондового рынка (рис. 26.4). Ее проектирование осуществлялось с помощью того же метода, что и синтез системы управления портфелем.

Эта система позволяет не только получать информацию о торгах, графически отображать ее и использовать методы технического анализа, но и создавать свои

методы прогноза и анализа рынка, отдавать приказы на куплю-продажу, оценивать состояние, а также качество управления портфелем.

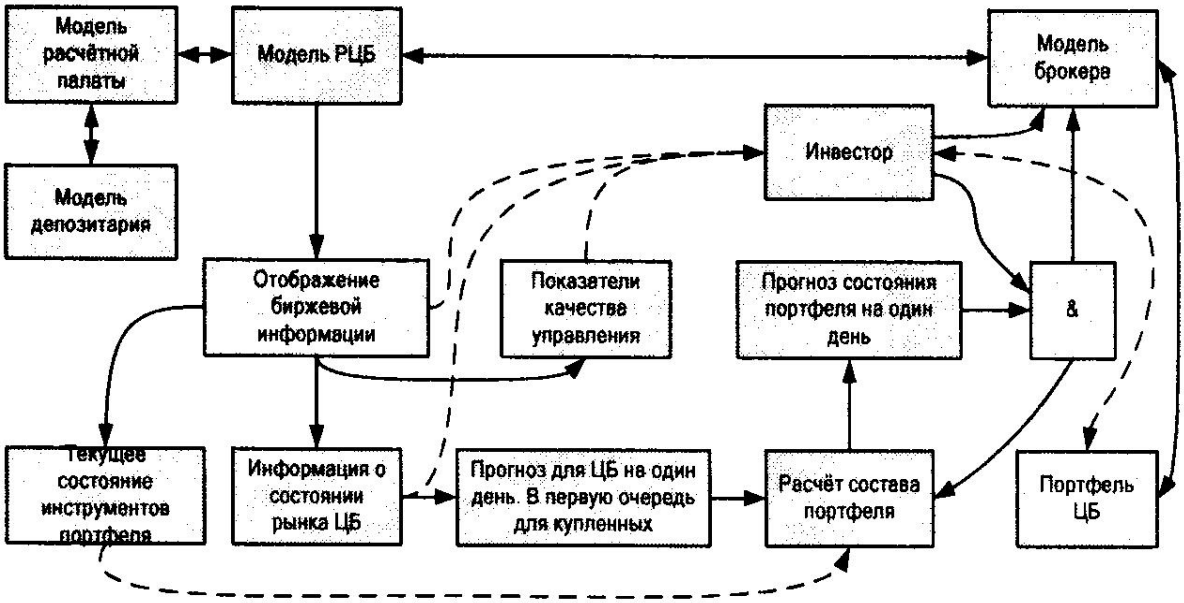


Рисунок 26.4 – Результат проектирования информационно-управляющей системы для фондового рынка

Вопросы для самоконтроля

- Понятие о системно-структурном синтезе.
- История и принципиальные особенности системно-структурного синтеза.
- Основные понятийные категории системно-структурного синтеза
- Сферы применения системно-структурного синтеза.

ЗАКЛЮЧЕНИЕ

В 1999-2000 годах, когда интернет-бум обернулся кризисом, родилась шутка: «Интернет – это телеграф, возомнивший себя новой экономикой». Н.Карр развивает этот тезис, низводя ИТ до уровня базовых технологий и ставя их в один ряд с электричеством или кондиционированием воздуха. Здоровый цинизм автора вызывает жгучее желание подискутировать, однако нельзя не признать: книга заставляет по-новому взглянуть на эффективность вложений в информационные технологии и их роль в развитии бизнеса.

Президент консорциума «Инфорус» А. Масалович о книге Н.Карра «Блеск и нищета информационных технологий».

Успешная практическая деятельность человека все в большей мере зависит от организации сбора, хранения, обработки и передачи информации - основных функций информационных технологий (ИТ). Совершенствование технологий записи и хранения данных, создание баз данных и знаний во всех сферах деятельности человека предъявляют новые требования к уровню подготовки специалистов. Большой объем информации, которой сопровождается деятельность практически любого предприятия и учреждения, обычно содержит полезные сведения, благодаря которым можно значительно повысить эффективность работы, совершенствуя технологию, управление и т.д. Современные коммерческие организации интенсивно внедряют информационные хранилища (банки) данных и знаний, многие из которых содержат средства интеллектуального анализа данных (ИАД) или предполагают возможность их применения.

Следует отметить, что в настоящее время существует как минимум две точки зрения на коммерческую привлекательность методов интеллектуального анализа данных и ИТ-технологий:

1) применение методов *DATA MINING* (ИТ-технологий) может дать ощутимое преимущество в конкурентной борьбе – мантра (магическое заклинание) романтиков ИТ и ИАД, а так - же бизнес-гуру, барыг и спекулянтов;

2) информационные технологии (в том числе и методы интеллектуального анализа данных) перешли (переходят или перейдут) из категории потенциального стратегического ресурса, обеспечивающего конкурентное преимущество, в категорию товарного ресурса – одну изстатей затрат на ведение бизнеса, которые несут практически все предприятия не, получая никаких конкурентных преимуществ.

Последний подход провозгласил в 2003 году Николас Дж. Карр в своей статье «ИТ ничего не значат» [25].Идея Карра не нова - рассмотреть ИТ в ряду инновационных технологий, изменивших мировую экономику за последние несколько столетий. В этом плане обычно рассматривается три глобальных революции: сельскохозяйственная, произошедшая около 10 тыс. лет назад в связи с изобретением землепашества; промышленная, произошедшая около 300 лет назад; информационная, свидетелями которой мы являемся уже более 50 лет. Прослеживая параллели, можно

сказать, что все инновации (землепашество, паровые двигатели, железные дороги, телеграф, телефон, линии электропередач, скоростные автомагистрали, ИТ (и ИАД)) способствовали некоторое время приоритету в конкурентной борьбе. Затем они становились более дешёвыми и стандартизированными – их возможности и потенциал стали значительно превышать потребности.

Последние примерно десять лет в крупных компаниях появилась должность директора по ИТ-технологиям (в вузах – проректора по информатизации). Одним из выводов Н. Дж. Карра является мысль, что их главная и парадоксальная профессиональная задача – самоликвидация, то есть обеспечение такого уровня стабильности и надёжности ИТ-технологий, когда вмешательство уже незаметно и ненужно. Современным компаниям, внедряющим (или думающим о внедрении) ИТ-технологий (и ИАД) рекомендуется следовать принципам: расходовать меньше; следовать за лидером, а не рваться вперёд; вводить инновации, если риски не значительны; думать о недостатках, а не о возможностях; критически относиться к любым рецептам в бизнесе (в том числе и выше перечисленным).

Следует отметить, что у нас в стране активное внедрение ИТ и средств ИАД наблюдается в крупных городах (Москва, Петербург). Регионы, может быть, полуинтуитивно, с опаской относятся ко всем подобным нововведениям. Интересно,что отечественные компании – производители средств ИАД (Megaputer, BaseGroupLabs) ориентируются на запад и столицы. Региональные компании не имеют тех объёмов информации (тех задач), которые под силу системам класса DataMining – возникает эффект,когда предложение превышает спрос. Современные возможности систем анализа данных превышают необходимые задачи в регионах. Будут ли внедряться средства ИАД в регионах? – Да, будут! Этот момент, скорее всего, наступит при дальнейшем удешевлении ИТ и превращении ИАД в столь же обязательное обеспечение работы организации, как телефон, факс, компьютер, ПО Microsoft, базы данных и, часто сейчас критикуемая, 1С бухгалтерия. То есть можно предположить, что по крайне мере для бизнес-процессов средства ИАД станут обязательными.

На современном этапе управление производством, фирмой, районом, регионом практически невозможно без системного подхода, разрабатывающего методики анализа целей, методы и модели совершенствования организационной структуры, управления функционированием социально-экономических объектов [14].

В зависимости от априорной информации об изучаемом объекте применяют следующие методы: мозговой атаки, построения сценариев, экспертной оценки, построения дерева целей, математической логики, теории множеств, теории игр, прикладной статистики, математического программирования, интеллектуального анализа данных и т.д. Разумеется, большинство методов пересекается. При этом рассматриваемые в практикуме методы в рамках системных исследований являются одними из возможных подходов перевода вербального (словесного) описания модели изучаемого объекта в формальное, для решения задач управления и принятия решений.

Многие ученые считают, что человеку изначально присуще образное мышление, поэтому визуализация и формализация представления данных и знаний, которым в основном и посвящен настоящий практикум, и представление их в виде графиков, формул, таблиц, правил позволяет человеку перейти от оперирования

нагромождением цифр, фактов, прецедентов и знаний экспертов к работе с объектами, соответствующими наиболее эффективному способу принятия решений.

Безусловно, не существует абсолютных рецептов построения моделей сложных обектов и процессов. Моделирование в большей степени искусство, овладеть которым можно только решая практические задачи. Закончим повествование несколько перефразированными словами отечественного классика анализа данных, Юрия Павловича Адлера [1]:

«… Единственный способ вырваться из «плена» какого-нибудь метода – это овладеть им. Поэтому стоит учиться *строить модели данных и знаний – это необходимые знания и умения для специалистов всех направлений (экономистов, инженеров, аналитиков, управленцев и др.) уже сегодня* (курсив наш)».

ЛИТЕРАТУРА

Раздел 1

1.Адлер Ю.П. Предисловие к русскому изданию книги Ф. Мостеллера, Дж. Тьюки. Анализ данных и регрессия: Вып.2/ Пер. с англ. Б.Л.Розовского; Под ред. и с предисл. Ю.П.Адлера.-М.: Финансы и статистика, 1982, - 239с.

2.Арнольд В.И. «Жесткие» и «мягкие» математические модели М.:МЦНМО, 2000.- 32с.

3.Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1985. – 472 с.

4.Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985. – 488 с.

5.Айвазян С.А., Бухштабер В.М., Енюков И.С. и др. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.: ил.

6.АйвазянС.А., Мхитарян В.С. Теория вероятностей и прикладная статистика.-М.:ЮНИТИ –ДАНА, 2001.-656с.

7. Айвазян С.А. Основы эконометрики- М.:ЮНИТИ –ДАНА, 2001.-432с.

8.Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. Пер. с англ. – М.: Мир, 1982. – 488 с., ил.

9.Берсегян А.А. и др. Методы и модели анализа данных: OLAP и DataMining. БХВ-Петербург, 2004. – 336 с.: ил.

10.Бокс Дж., Дженкинс Г. Анализ временных рядов.Прогноз и управление. – М.: Мир, 1974 – Вып.1,2.

11.Борисов В.В., Бычков И.А.и др. Компьютерная поддержка сложных организационно-технических систем.- М.: Горячая линия − Телеком, 2002. − 154 с.:ил.

12.Боровиков В.П., Боровиков И.П.. STATISTICA – Статистический анализ и обработка данных в среде Windows. Издание 2-е,стереотипное – М.: Информационно-издательский дом «Филинъ», 1998. – 608 с.

13.Боровиков В.П. Программа Statistica для студентов и инженеров. – 2-е изд. – М.: КомпьютерПресс, 2001. – 301 с. – илл.

14.Боровиков В.П. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688с.:ил.

15.Боровиков В.П., Ивченко Г.И. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере: Учеб пособие. 2-е изд., перераб. и доп. – М.: Финансы и статистика, 2006. – 368 с.: ил.

16.Волкова В.Н., Денисов А.А. Основы теории систем и системного анализа. - СПб.:Изд. СПбГТУ,1997. - 510с.:ил.

17.Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением EXCEL. Учеб. пособие для вузов. Издание 4-е испр. и доп. – Ростов н/Д: Феникс, 2006. – 476с.:ил.

18.Демиденко Е. З. Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981. – 302 с., ил.

19. Дрейпер И., Смит Г. Прикладной регрессионный анализ: В 2-х кн. Пер. с англ. – 2-е изд. перераб. и доп. – М.: Финансы и статистика, 1986 Кн. 1. – 366 с., ил; 1987. Кн.2. – 351 с., ил.
20. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 1998. – 352 с.: ил.
21. Дюк В. Обработка данных на ПК в примерах – СПб.: Питер, 1997. – 240 с., ил.
22. Дюк В., Самойленко А. В. DataMining: учебный курс. – СПб, 2001. – 368 с.:ил.
23. Заде Л.А.Роль мягких вычислений и нечеткой логики в понимании, конструировании и развитии информационных/ интеллектуальных систем. – // Новости Искусственного Интеллекта, №2–3, 2001.
24. Закс Л. Статистическое оценивание. Пер. с нем. В.Н. Варыгина. Под ред. Ю.П. Адлера, В. Г. Горского. М., «Статистика», 1976. – 598 с., ил.
25. Карр Н.Дж. Блеск и нищета информационных технологий: Почему ИТ не являются конкурентным преимуществом./ Пер.С англ.- М.: Изд.дом «Секрет фирмы», 2005.-176с.
26. Кацко И.А. К вопросу о новой парадигме в анализе данных (статья). – Таганрог, ТИУЭ, 2003.
27. Кацко И.А. Ковариационный анализ многолетнего, многфакторного опыта в северокавказском филиале КНИИСХ. Труды КубГАУ №1 – 2006.
28. Киселев М., Соломатин Е. Средства добычи знаний в бизнесе и финансах. - Открытые системы, № 4, 1997, с.41-44.
29. Кречетов Н. Продукты для интеллектуального анализа данных. - Рынок программных средств, N14-15_97, с.32-39.
30. Кендэл М. Временные ряды. / Пер. с англ. Ю.П. Лукашина. – М.: Финансы и статистика, 1981. – 199 с., ил.
31. Орлов А.И. Современная прикладная статистика (обобщающая статья). Заводская лаборатория. 1998. т.64. №3. – 52-60 с.
32. Орлов А.И. Прикладная статистика. – М.: Экзамен, 2006. – 611с.
33. Системный анализ в проектировании и управлении: Труды X Международной науч.- практ. конф. Ч.I. СПб.: Изд-во Политехн. ун-та, 2006. – 236с.
34. Сошникова Л.А., Тамашевич В.Н., Уебе Г., Шеффер М. Многомерный статистичекий анализ в экономике: Учеб. пособие для вузов/Под ред.проф. В.Н. Тамашевича. – М.:ЮНИТИ-ДАНА, 1999. – 598 с.
35. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. – М.: Мир,1981. – 693 с.:ил.
36. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере / Под. ред. В.Э. Фигурнова. – 3-е изд., перераб. и доп. – М.: ИНФРА – М, 2003. – 544 с., ил.
37. Халафян А.А. Статистический анализ данных. Statistica 6.0. 2-е изд. Испр. И доп.: Учеб.пособие.- Краснодар: КубГу, 2005. – 307с.
38. Харман Г. Современный факторный анализ. – М.: Статистика, 1972.- 486с.
39. Хьюстон А. Дисперсионный анализ. Пер с англ. А.Г. Кругликова. – М.: «Статистика», 1971. – 88 с.

40. StatSoft, Inc. (2001). Электронный учебник по промышленной статистике. Москва,StatSoft.WEB: http://www.statsoft.ru/home/portal/textbook_ind/default.htm.
41. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем./Ахим Бююль, Петер Цёфель – СПб.: ООО «ДиаСофтЮП», 2002. – 608 с.
42. Knowledge Discovery Through Data Mining: What Is Knowledge Discovery? - Tandem Computers Inc., 1996.
43. Boulding K. E. General Systems Theory - The Skeleton of Science//Management Science, 2, 1956.
44. http://www.ibusiness.ru/development/soft/19369/Гордиеко.И. Вскрытие данных.
45. http://www.russianenterprisesolutions.com/techno/dm.html В.Дюк. Что такое DataMining?
46. http://www.tora-centre.ru/library/dm/market.htmSergei M. Ananyan How database marketing could marry data mining?
47. http://www.biont.ru/docs/biosecrm.htm Иванов В. Аналитический CRM и Data Mining. Добыча знаний и денег из данных о клиентах.
48. http://www.permonline.ru/~enter/june/olap.htmКолчанов А.Чудесное превращение данных в информацию.
49. http://fx-trader.narod.ru/Kognit.htm Масалович А. Нечеткие когнитивные схемы − новый инструмент для моделирования экономических, политических, социальных ситуаций.
50. http://ipu51.chat.ru/projects.htm Сайт ИПУ РАН 51-сектор «когнитивное моделирование и анализ».
51. http://www.olap.ru/desc/oracle/oracle_dm.aspЧарльзБергер (Charles R.Berger), Oracle Magazine Online/RE. (В статье описывается набор продуктов и сервисов класса DataMining)
52. http://tora-centre.ru/- Сайт Тора центра. Приводится описание большого числа систем интеллектуального анализа данных.
53. http://wwwrus.sas.com/) - русскоязычный сайт SAS.
54. http://www.spss.ru/- русскоязычный сайт SPSS.
55. http://www.statsoft.ru- русскоязычный сайт Statistica.
56. http://www.megaputer.ru - сайт создателей DataMining системы PolyAnalyst.
57. http://www.basegroup.ru– сайт создателей аналитической платформы Deductor.
58. http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/- электронная библиотека временных рядов.

Раздел 2

1. Орлов А.И. Устойчивость в социально-экономических моделях. - М.: Наука, 1979. - 296 с.
2. Орлов А.И. Статистика объектов нечисловой природы и экспертные оценки. – В сб.: Экспертные оценки / Вопросы кибернетики. Вып.58. - М.: Научный Совет АН СССР по комплексной проблеме "Кибернетика", 1979. - С.17-33.

3. Кривцов В.С., Орлов А.И., Фомин В.Н. Современные статистические методы в стандартизации и управлении качеством продукции. – Журнал «Стандарты и качество». 1988. No.3. С.32-36.
4. Беляев Ю.К. Вероятностные методы выборочного контроля. - М.: Наука, 1975. - 408 с.
5. Лумельский Я.П. Статистические оценки результатов контроля качества. - М.: Изд-во стандартов, 1979. - 200 с.
6. Орлов А.И. Статистика объектов нечисловой природы (Обзор). – Журнал «Заводская лаборатория». 1990. Т.56. No.3. С.76-83.
7. Вероятность и математическая статистика: Энциклопедия / Гл. ред. Ю.В. Прохоров. - М.: Большая Российская энциклопедия, 1999. - 910 с.
8. Орлов А.И. Эконометрика. Учебник для вузов. Изд. 2-е, исправленное и дополненное. - М.: Изд-во "Экзамен", 2003. – 576 с.
9. Толстова Ю.Н. Анализ социологических данных. – М.: Научный мир, 2000. – 352 с.
10. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. - М.: Советское радио, 1972. - 192 с.
11. Орлов А.И. Асимптотика решений экстремальных статистических задач. – В сб.: Анализ нечисловых данных в системных исследованиях. Сборник трудов. Вып.10. - М.: Всесоюзный научно-исследовательский институт системных исследований, 1982. - С. 4-12.
12. Орлов А.И. Асимптотическое поведение статистик интегрального типа. – В сб.: Вероятностные процессы и их приложения. Межвузовский сборник. - М.: МИЭМ, 1989. С.118-123.
13. Кендэл М. Ранговые корреляции. - М.:Статистика,1975. - 216 с.
14. Раушенбах Г.В. Меры близости и сходства. - В сб.: Анализ нечисловой информации в социологических исследованиях. - М.: Наука, 1985. - С.169-203.
15. Маамяги А.В. Некоторые задачи статистического анализа классификаций. – Таллинн: АН ЭССР, 1982. – 24 с.
16. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983 (3-е изд.). - 474 с.
17. Крамер Г. Математические методы статистики. - М.: Мир, 1975. - 648 с.
18. Орлов А.И. Парные сравнения в асимптотике Колмогорова. – В сб.: Экспертные оценки в задачах управления. - М.: Изд-во Института проблем управления АН СССР, 1982. - С. 58-66.
19. Орлов А.И. Случайные множества с независимыми элементами (люсианы) и их применения. – В сб.: Алгоритмическое и программное обеспечение прикладного статистического анализа. Ученые записки по статистике, т.36. - М.: Наука, 1980. - С. 287-308.
20. Рыданова Г.В. Некоторые вопросы статистического анализа случайных бинарных векторов. Дисс. ... канд. физ.-мат. наук. - М.: МГУ, ф-т вычислит. математ. и кибернет., 1987. - 139 с.
21. Аксенова Г.А., Кузьмина Е.С., Орлов А.И., Розова Н.К. Кинетотопография в диагностике инфаркта миокарда. – В сб.: Актуальные вопросы клинической

и экспериментальной медицины. – М.: 4 Главное Управление при Минздраве СССР, 1979. С.24-26.
22. Попов В.Г., Аксенова Г.А., Орлов А.И., Розова Н.К., Кузьмина Е.С. Кинетокардиография в определении зон асинергии у больных инфарктом миокарда. - Журнал «Клиническая медицина». 1982. T.LX. No.3. С.25-30.
23. Методические рекомендации по проведению экспертной оценки планируемых и законченных научных работ в области медицины (по проблемам союзного значения) / Составители: Г.В. Раушенбах, О.В. Филиппов. - М.: АМН СССР - Ученый медицинский совет Минздрава СССР, 1982. - 36 с.
24. Леман Э. Проверка статистических гипотез. - М.: Наука, 1979. - 408 с.
25. Боровков А.А. Математическая статистика/Учебное пособие для вузов. - М.: Наука, 1984. - 472 с.
26. Любищев А.А. Дисперсионный анализ в биологии. - М.: Изд-во МГУ, 1986. - 200 с.
27. Дылько Т.Н. Проверка гипотез в экспертном оценивании / Вестник Белорусского государственного университета / Сер. 1: Физика, математика и механика. 1988, N 2. С. 36-40.
28. Орлов А.И., Раушенбах Г.В. Метрика подобия: аксиоматическое введение, асимптотическая нормальность. - В сб.: Статистические методы оценивания и проверки гипотез. Межвузовский сборник научных трудов. - Пермь: Изд-во Пермского государственного университета, 1986, с.148-157.
29. Орлов А.И., Миронова Н.Г., Фомин В.Н., Черчинцев А.Н. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики. - М.: ВНИИСтандартизации, 1987. - 62 с.
30. Тюрин Ю.Н., Василевич А.П. К проблеме обработки рядов ранжировок. - В сб.: Статистические методы анализа экспертных оценок. Ученые записки по статистике, т.29. - М.: Наука, 1977. - С. 96-111.
31. Орлов А.И. Некоторые вероятностные вопросы теории классификации. – В сб.: Прикладная статистика. Ученые записки по статистике, т.45. - М.: Наука, 1983. - С. 166-179.
32. Дэвид Г. Метод парных сравнений. - М.: Статистика, 1978.- 144 с.
33. Орлов А.И. Сходимость эталонных алгоритмов. – В сб.: Прикладной многомерный статистический анализ. Ученые записки по статистике, т.33. - М.: Наука, 1978. С. 361-364.
34. Орлов А.И. Задачи оптимизации и нечеткие переменные. – М.: Знание, 1980. - 64 с.
35. Горский В.Г., Гриценко А.А., Орлов А.И., Метод согласования кластеризованных ранжировок // Автоматика и телемеханика. 2000. №3. С.159-167.
36. Шрейдер Ю.А. Равенство, сходство, порядок. - М.: Наука, 1971. – 256 с.
37. Менеджмент. / Под ред. Ж.В. Прокофьевой. - М.: Знание, 2000. - 288 с.
38. Орлов А.И. Прикладная статистика. Учебник. / А.И.Орлов. – М.: Издательство «Экзамен», 2004. - 656 с.

39. Болотова Л.С**.** Системы искусственного интеллекта: модели и технологии, основанные на знаниях: учебник**. –** М.: Финансы и статистика, 2012. – 664 с.: ил.
40. Горелова Г.В. Захарова Е.Н., Радченко С.А.Исследование слабоструктурированных проблем социально-экономических систем: когнитивный подход. - Ростов н/Д: Изд-во РГУ, 2006. - 332с.
41.Волкова В.Н. Системный анализ и принятие решений: Словарь-справочник. Учеб. пособие для вузов / Под ред. В. Н. Волковой, В. Н. Козлова. – М.: Высш. шк. , 2004 – 616 с: ил.
42. Луценко Е.В. Интеллектуальные информационные системы: Учебное пособие для студентов специальности "Прикладная информатика (по областям)" и другим экономическим специальностям. 2-е изд., перераб. и доп.– Краснодар: КубГАУ, 2006. – 615 с.

Учебное издание

Болотова Людмила Сергеевна
Горелова Галина Викторовна
Гудков Георгий Владимирович
Жминько Альбина Евгеньевна
Захарова Юлия Николаевна
Кацко Игорь Александрович
Кацко Светлана Александровна
Луценко Евгений Вениаминович
Лыпарь Юрий Иванович
Орлов Александр Иванович
Паклин Николай Борисович
Сенникова Алина Евгеньевна
Чефранов Сергей Георгиевич

МОДЕЛИ И МЕТОДЫ ПРИКЛАДНЫХ СИСТЕМНЫХ ИССЛЕДОВАНИЙ (ПРАКТИКУМ)

Учебное пособие

В авторской редакции

Подписано в печать 17.02.14. Формат $60 \times 84\ {}^{1}/_{16}$.
Усл. печ. л. – 26,1. Уч.-изд. л. – 20,4
Тираж 500 экз. Заказ № 265

Типография Кубанского государственного аграрного университета.
350044, Краснодар, ул. Калинина, 13