

**Министерство сельского хозяйства РФ
федеральное государственное бюджетное
образовательное учреждение высшего образования
«Вологодская государственная молочнохозяйственная академия
имени Н.В. Верещагина»**

**УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО
ДИСЦИПЛИНЕ**

Анализ данных в системах искусственного интеллекта

название дисциплины

Направление подготовки:
35.04.06 Агроинженерия

Образовательная программа:
Искусственный интеллект

Форма обучения:
очная

1. МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

1.1. ЧТО ТАКОЕ «МОДЕЛЬ»?

Математическая модель — это абстракция реального мира, в которой интересующие исследователя отношения между реальными элементами заменены отношениями между подходящими математическими категориями. Эти отношения, как правило, представляются в форме уравнений и (или) неравенств между показателями (переменными), характеризующими функционирование моделируемой реальной системы.

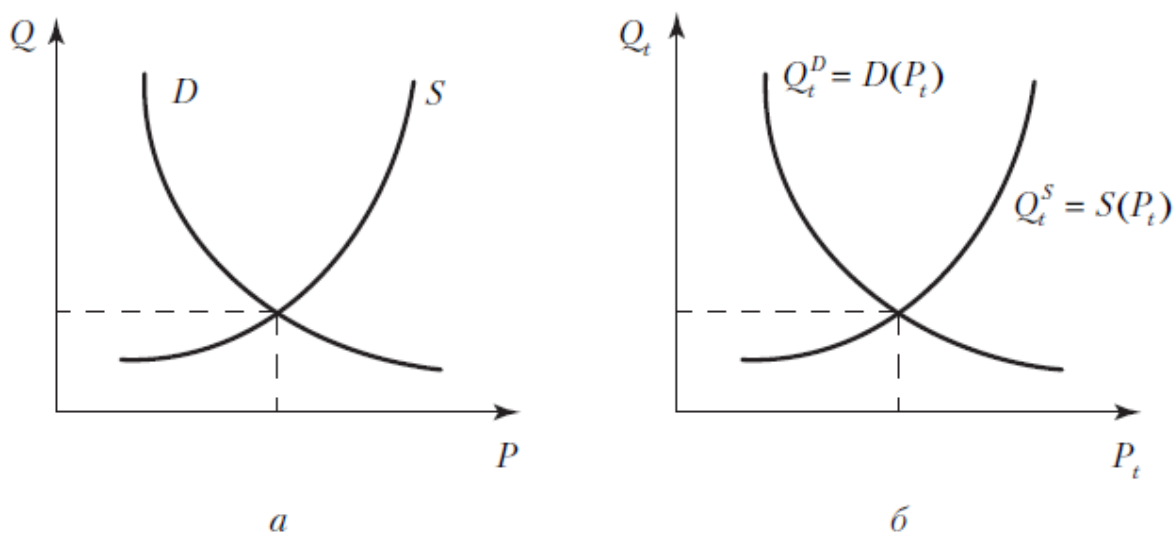
Искусство построения математической модели состоит в том, чтобы совместить как можно большую лаконичность в ее математическом описании с достаточной точностью модельного воспроизведения именно тех сторон анализируемой реальности, которые в наибольшей степени интересуют исследователя.

Вероятностная модель — это математическая модель, имитирующая механизм функционирования гипотетического (не конкретного) реального явления (или системы) стохастической, или вероятностной, природы.

Вероятностно-статистическая модель — это вероятностная модель, значения отдельных характеристик (параметров) которой оцениваются по результатам наблюдений, т.е. по имеющимся статистическим данным.

Рассмотрим пример. Пусть изучается традиционная модель спроса и предложения, объясняющая соотношение между ценой (P), объемами выпуска (S) и спроса (D).

Из экономической теории известно, что кривые спроса и предложения имеют вид, представленный на рис. 1.1, а. Это — экономическая модель. Если ввести конкретные функции, описывающие изменения спроса и предложения, то модель перейдет в класс экономико-математических моделей.



Для того чтобы эта модель стала эконометрической (рис. 1.1, б), следует говорить не о законе спроса и предложения вообще, а о конкретном его действии в четко определенный момент времени t и применительно к конкретному товару (услуге). Конкретизация вида функций спроса $D(P_t)$ и предложения $S(P_t)$ должна происходить с использованием реально

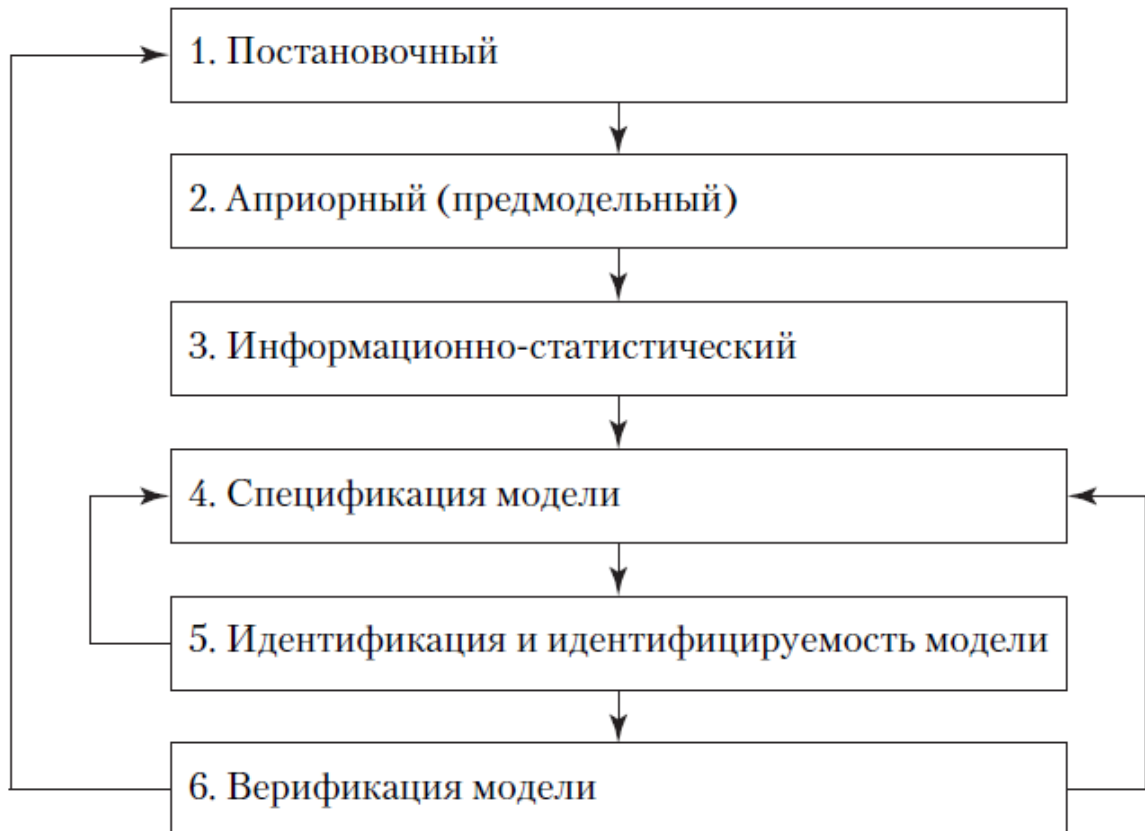
существующих статистических данных. Это позволит с помощью специальных статистических процедур четко верифицировать¹ получаемые выводы.

¹ Верификация — проверка, проверяемость, способ подтверждения, проверка с помощью доказательств, каких-либо теоретических положений, алгоритмов, программ и процедур путем их сопоставления с опытными (эталонными или эмпирическими) данными, алгоритмами и программами.

1.2. ЭТАПЫ АНАЛИЗА МОДЕЛЕЙ

Как же происходит построение модели? Анализ литературных источников и личный опыт позволил авторам выделить ряд этапов в процедуре построения практически любой модели.

Схематично процесс построения модели, или процесс анализа изучаемой системы (объекта), представлен на рис. 1.2.



Выполнение этих этапов на практике может быть обозначено явно, а может происходить и неявным образом, но в любом случае каждый из названных этапов выполняется в полном объеме.

Рассмотрим подробнее, в чем заключается каждый из этих этапов и решению каких задач он посвящен.

Этап 1. Постановочный

На этом этапе происходит определение конечной цели анализа и перечня решаемых задач. Формируется набор показателей, анализ которых позволит решать поставленные задачи. При этом за каждым показателем закрепляется его роль в предполагаемой модели.

Различают входные и выходные показатели. Входными показателями (факторами) считаются те, значения которых являются заданными, т.е. определяются вне предполагаемой модели. Для некоторых из них может присутствовать возможность установки желаемых значений – это так называемые регулируемые факторы.

Выходными показателями (факторами) принято считать показатели, значения которых формируются в процессе функционирования анализируемой системы (объекта) и могут зависеть от значений входных факторов.

Этап 2. Априорный (предмодельный)

Этот этап заключается в предварительном анализе содержательной сущности исследуемого явления (системы). На нем происходят формирование имеющейся априорной информации о данном явлении (системе) и ее дальнейшая формализация в виде ряда гипотез и исходных допущений.

Причем это должно быть обязательно подтверждено теоретическими рассуждениями о механизме функционирования изучаемого явления, о тех экономических, физических и других законах, которые должны быть учтены в эконометрической модели.

При возможности должна быть проведена экспериментальная проверка справедливости выдвинутых гипотез и допущений.

Этап 3. Информационно-статистический

Этот этап посвящен сбору требуемых статистических данных о тех показателях (факторах), которые были отобраны на постановочном этапе. На нем возможно использование различных статистических сборников, а при необходимости и проведение дополнительных исследований.

Следует отметить, что при сборе статистической информации необходимо обращать особое внимание на ее достоверность и полноту.

Этап 4. Спецификация модели

Этот этап включает в себя непосредственный вывод общего вида модельных соотношений. При этом нужно опираться на принятые на априорном этапе гипотезы и допущения. Полученные модельные соотношения должны связывать между собой интересующие исследователя входные и выходные факторы.

На данном этапе должна быть определена только структура эконометрической модели, т.е. ее символическая аналитическая запись, где наряду с показателями, для которых есть статистические данные, будут присутствовать величины, содержательный смысл которых четко определен, а конкретные числовые значения нет. Такие величины называют неизвестными параметрами модели, подлежащими статистическому оцениванию.

Этап 5. Идентификация и идентифицируемость модели

На этом этапе проводится статистический анализ модели с целью «настройки» или «подгонки» значений ее параметров, чтобы обеспечить наилучшее соответствие модели имеющимся исходным данным.

Однако перед решением этой задачи полезно ответить на следующий вопрос: возможно ли в принципе по имеющимся статистическим данным однозначно определить значения всех неизвестных параметров для принятой на этапе 4 общей структуры модели? Если

однозначное определение неизвестных параметров возможно, то модель называют идентифицируемой, в противном случае – неидентифицируемой.

Только в случае положительного ответа на поставленный вопрос следует приступать к процедуре идентификации модели, т.е. предлагать и использовать конкретные математически корректные процедуры нахождения значений оценок для всех неизвестных параметров эконометрической модели.

Если же модель является неидентифицируемой, то следует вновь вернуться к этапу 4 для корректировки общего вида модели.

Этап 6. Верификация модели

Этот этап состоит в использовании специальных процедур сопоставления модельных заключений, оценок, следствий и выводов с реально наблюдаемой действительностью. Этот этап часто называют этапом статистического анализа точности и адекватности модели.

При неудовлетворительных результатах можно порекомендовать перейти к этапу 4 с целью видоизменения структуры модели или к этапу 1 для привлечения к анализу дополнительных показателей, а возможно, и для пересмотра конечной цели и задач анализа.

Следует помнить, что при построении эконометрических моделей нельзя опираться только на методы прикладного статистического анализа, необходимо использовать экономические знания. Иначе могут быть построены модели, которые ни при каких условиях нельзя будет считать адекватными, даже если их верификация формальными, статистическими методами дает положительный результат.

Действуя таким образом, можно не учесть в модели (или, попросту говоря, «забыть») один или несколько показателей (факторов), относительно которых заведомо известно, что они оказывают существенное влияние на изучаемую ситуацию. Кроме того, для таких моделей может быть затруднена интерпретация отдельных выводов и результатов.

Верно и обратное: хорошо зная только специфику функционирования изучаемой экономической или социально-экономической системы (действующие экономические законы и закономерности) и не используя статистические методы, также нельзя построить хорошую эконометрическую модель.

В этом случае просто невозможно проверить адекватность выявленных закономерностей и сделанных выводов. Полученные выводы и результаты будут носить только форму гипотез, никак не сопоставленных (соотнесенных) с наблюдаемой ситуацией, т.е. с реально существующими статистическими данными. Могут возникнуть проблемы с оценкой точности прогнозов отдельных показателей. Проверка предположений на соответствие со статистическими данными проводится посредством использования статистических методов.

1.3. ИЗМЕРИТЕЛЬНЫЕ ШКАЛЫ

Как уже отмечалось ранее, эконометрика буквально означает «измерения в экономике», поэтому в основе любого эконометрического исследования лежат измерения тех или иных показателей. Но каждое такое измерение производится в определенной шкале, и необходимо четко понимать, в какой шкале измеряется тот или иной показатель. В зависимости от выбранной шкалы определяются операции, которые допустимо проводить над имеющимися данными, а также методы, которые могут быть использованы для получения корректных и обоснованных результатов.

Что же понимают под термином «измерение»? **Измерение** — это алгоритмическая операция, которая данному наблюдаемому состоянию объекта, процесса, явления ставит в соответствие определенное обозначение: число, символ и др.

Такое соответствие обеспечивает то, что результаты измерений содержат информацию о наблюдавшемся объекте.

1. Номинальная шкала

Существует и другое название этой шкалы – шкала наименований. Пусть число различных состояний наблюдаемой системы (классов эквивалентности) конечно. Каждому такому классу эквивалентности поставим в соответствие уникальное обозначение. При этом могут быть использованы: слова естественного языка, произвольные символы, номера или их комбинации. Тогда измерение будет состоять в том, чтобы, проведя эксперимент над объектом, четко определить принадлежность полученного результата к тому или иному состоянию и записать это соответствие с помощью уже выбранного символа, обозначающего данный класс эквивалентности. Указанное множество символов образует шкалу наименований.

Однако нужно помнить, что почти всегда выбранные обозначения – это только символы, а не числа! Для этой шкалы установлены только аксиомы тождества, поэтому допустимой операцией является лишь сравнение объектов на «равно» или «неравно». Даже когда в качестве обозначений выбраны номера, с ними нельзя работать как с числами. Статистическая обработка результатов возможна только после перехода к частотам появления различных значений, зная которые, можно выполнять более сложные преобразования: определять количества совпадений, вычислять относительные частоты классов, сравнивать частоты между собой (находя, например, моду), осуществлять проверку различных гипотез по критерию χ^2 -Пирсона и др.

Примером показателей, измеряемых в номинальной шкале, могут служить профессия, пол, адрес, номера и марки машин и др.

2. Ранговая шкала

Кроме аксиом тождества для состояний должны быть справедливы *аксиомы упорядоченности*:

- Если $A > B$, то $B < A$.
- Если $A > B$, $B > C$ и $A < C$.

В этом случае говорят, что измерения проводятся в ранговой (порядковой) шкале. Запись $A > B$ означает, что состояние A больше, лучше или предпочтительнее состояния B . Шкала получила такое название от термина «ранг наблюдения». Под рангом наблюдения понимают номер данного наблюдения в упорядоченном ряду наблюдений.

Важной особенностью ранговых шкал является то, что отношение порядка ничего не говорит о расстоянии между отдельными классами.

Таким образом, мы не можем точно сказать, насколько одно состояние лучше или хуже другого. Поэтому порядковые экспериментальные данные нельзя рассматривать как числа, даже если они представлены в виде чисел.

Примером показателей, измеряемых в ранговой шкале, могут служить нумерация очередности, уровни образования, «должностная лестница», воинские звания, призовые места.

3. Интервальная шкала

Если упорядочение объектов можно выполнить настолько точно, что становятся известными расстояния между любыми двумя из них, то измерение будет более сильным, чем в ранговой шкале.

Расстояния могут измеряться в различных единицах – метрах, футах, аршинах и т.д. При этом будут получены разные числовые значения. Но одно останется неизменным: *если одно расстояние больше другого, то это будет верно для любых единиц измерения.*

Еще одним свойством интервальной шкалы является отсутствие в ней естественного начала отсчета. При всех измерениях исследователь должен сам задать точку отсчета и выбрать единицу измерения.

Иными словами, можно сказать так: шкала интервалов единственная с точностью до линейных преобразований.

Примером показателей, измеряемых в ранговой шкале, могут служить температура, время, высота местности.

4. Шкала отношений

Пусть наблюдения удовлетворяют не только аксиомам тождества и упорядоченности, но и аксиомам сложения (аддитивности)

- Если $A = P$ и $B > 0$, то $A + B > 0$.
- $A + B = B + A$ (аксиома коммутативности).
- Если $A = P$ и $B = Q$, то $A + B = P + Q$.
- $(A + B) + C = A + (B + C)$ (аксиома ассоциативности).

В такой шкале существенно «усиливаются» измерения, они являются полноправными числами, с ними можно выполнять любые арифметические операции. Но и эта шкала имеет одну особенность: отношение двух наблюдаемых значений измеряемого показателя не зависит от того, в какой конкретно из таких шкал произведены эти измерения:

$$\frac{x_1}{x_2} = \frac{y_1}{y_2}.$$

Следовательно, можно говорить о том, что в данной шкале существует абсолютный нуль, но остается свобода в выборе единиц измерения.

Примером показателей, измеряемых в шкале отношений, могут служить длина, вес, деньги.

5. Абсолютная шкала

Эта шкала имеет и абсолютное начало отсчета, и абсолютную единицу. Именно такими качествами обладает числовая ось. Важной особенностью такой шкалы являются отвлеченность (безразмерность) и абсолютность ее единицы.

Внутренние свойства числовой оси, при всей кажущейся ее простоте, оказываются очень разнообразными, и теория чисел до сих пор не изучена до конца. Некоторые безразмерные числовые отношения, обнаруживаемые в природе, вызывают восхищение и изумление (явление резонанса, гармонические отношения размеров). В абсолютной шкале измеряются различные индексы, удельные экономические показатели и др.

Подводя итог, можно сказать, что чем сильнее шкала, тем больше сведений об изучаемом объекте дают измерения. Но нужно всегда стремиться проводить измерения именно в той шкале, которая максимально согласована с наблюдаемой величиной.

Если измерения проводятся в более слабой шкале, чем это возможно, то будет происходить потеря информации.

Применять же более сильную шкалу просто опасно, поскольку возможно получение некорректных, а иногда и просто курьезных результатов, что в большинстве случаев приводит к грубым ошибкам в расчетах.

1.4. ОБЩИЕ ПОНЯТИЯ МОДЕЛЕЙ

В исследованиях используют следующие типы данных:

Пространственные – характеризуют ситуацию по конкретной переменной (или набору переменных), относящейся к пространственно разделенным сходным объектам в один и тот же момент времени.

Примеры:

1. *Данные по курсам покупки или продажи наличной валюты в конкретный день по разным обменным пунктам г. Москвы.*
2. *Набор сведений (объем производства, количество работников, доход и др.) по разным фирмам в один и тот же момент времени.*

Временные ряды – отражают изменения (динамику) какой-либо переменной на промежутке времени.

Примеры:

1. *Ежеквартальные данные по инфляции.*
2. *Данные по средней заработной плате, национальному доходу и денежной эмиссии за несколько лет.*

Панельные данные – это разновидность пространственно-временных данных. Панельные данные состоят из наблюдений одних и тех же экономических единиц или объектов (индивидуумов, домашних хозяйств, фирм, регионов, стран и т.п.) на протяжении нескольких периодов времени.

Панельные данные насчитывают три измерения: **признаки – объекты – время**.

Их использование дает ряд существенных преимуществ при оценке параметров регрессионных зависимостей, поскольку они позволяют проводить анализ как временных рядов, так и пространстве иных выборок.

В эконометрике решаются задачи:	<ul style="list-style-type: none">• описания данных, проверки гипотез;• восстановления зависимостей;• классификации объектов и признаков;• прогнозирования;• принятия статистических решений и др.
--	--

При выборе метода анализа конкретных экономических данных следует учитывать, что экономические процессы развиваются во времени, поэтому большое место в эконометрике занимают вопросы анализа и прогнозирования временных рядов, в том числе многомерных. При этом следует отметить, что временные ряды качественно отличаются от простых статистических выборок.

Особенности временных рядов:

- 1) последовательные по времени уровни временных рядов являются **взаимозависимыми**, особенно это относится к близко расположенным наблюдениям;
- 2) в зависимости от момента наблюдения уровни во временных рядах обладают **разной информативностью**: информационная ценность наблюдений убывает по мере их удаления от текущего момента времени;
- 3) с увеличением количества уровней временного ряда **точность статистических характеристик не увеличивается пропорционально числу наблюдений**, а при появлении новых закономерностей развития может даже уменьшаться.

Переменные, участвующие в любой модели:

Результирующая (зависимая, эндогенная) переменная Y – характеризует результат или эффективность функционирования экономической системы.

Значения ее формируются внутри системы (поэтому ее и называют эндогенной) под воздействием ряда других переменных и факторов, часть из которых поддается регистрации, управлению и планированию.

В регрессионном анализе результирующая переменная (еще ее называют результативным признаком) играет роль функции, значение которой определяется значениями объясняющих переменных. По своей природе результирующая переменная всегда случайна (стохастична).

Объясняющие (независимые, экзогенные) переменные X – поддаются регистрации и описывают условия функционирования реальной экономической системы.

Они в значительной мере определяют значения результирующих переменных. Обычно независимые переменные поддаются регулированию и управлению. Значения этих переменных могут задаваться вне анализируемой системы (поэтому их и называют экзогенными; другое название – факторные признаки).

В регрессионном анализе объясняющие переменные – это аргументы результирующей функции Y . По своей природе они могут быть как случайными, так и неслучайными.

Можно выделить **три основных класса моделей**, которые применяются для анализа и прогнозирования экономических систем:

- 1) модели временных рядов;
- 2) регрессионные модели с одним уравнением;
- 3) системы эконометрических уравнений.

Модели временных рядов представляют собой модели зависимости результирующего признака от времени. К ним относятся адаптивные модели, модели кривых роста (трендовые) и модели авторегрессии и скользящего среднего. С помощью таких моделей можно решать задачи прогнозирования объема продаж, спроса на продукцию, краткосрочного прогноза процентных ставок и др.

В **регрессионных моделях с одним уравнением** зависимая (объясняемая) переменная Y может быть представлена в виде функции $f(X_1, X_2, \dots, X_k)$, где X_1, X_2, \dots, X_k – независимые (объясняющие) переменные, или факторы; k — количество факторов.

В качестве зависимой переменной может выступать практически любой показатель, характеризующий, например, деятельность предприятия или курс ценной бумаги.

В зависимости от вида функции $f(X_1, X_2, \dots, X_k)$, модели делятся на **линейные** и **нелинейные**, а в зависимости от количества включенных в модель факторов X – на **однофакторные** (парная модель регрессии) и **многофакторные** (модель множественной регрессии).

Примеры задач, решаемых с помощью регрессионных моделей:

1. Исследование зависимости заработной платы Y от возраста X_1 уровня образования X_2 , пола X_3 стажа работы X_4 :

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4.$$

2. Прогноз и планирование выпускаемой продукции по факторам производства (производственная функция Кобба-Дугласа)

$$Y = a_0K^{a_1}L^{a_2}$$

означает, что объем выпуска продукции Y является функцией количества капитала K и количества труда L).

3. Прогноз объемов потребления продукции или услуг определенного вида (кривая Энгеля)

$$y = \frac{a_0}{1+a_1e^{-a_2x}}$$

где Y – удельная величина спроса; X – среднедушевой доход).

Системы уравнений применяются в том случае, когда явления настолько сложны, что невозможно адекватно описать их с помощью только одного соотношения (уравнения). Модели с одним уравнением не отражают взаимосвязей между объясняющими переменными или их связей с другими переменными. Кроме того, некоторые переменные могут оказывать взаимные воздействия, и трудно однозначно определить, какая из них является зависимой, а какая — независимой переменной. Поэтому при построении эконометрической модели прибегают к системам уравнений.

Выделяют следующие три вида систем:

- 1) системы независимых уравнений;
- 2) системы рекурсивных уравнений;
- 3) системы взаимосвязанных уравнений.

В системах независимых уравнений каждая зависимая переменная $y_i (i = \overline{1, n})$ представлена как функция одного и того же набора независимых переменных $x_j (j = \overline{1, m})$:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m + \varepsilon_2, \\ \dots \dots \dots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Заметим, что отдельные коэффициенты при переменных могут быть равны нулю. Каждое уравнение этой системы можно рассматривать самостоятельно как уравнение регрессии. В него может быть введен свободный член, и коэффициенты регрессии могут быть найдены – методом наименьших квадратов (МНК).

В **системах рекурсивных уравнений** зависимые переменные $y_i (i = \overline{1, n})$ представлены как функции независимых переменных $x_j (j = \overline{1, m})$ и определенных ранее зависимых переменных $y_1, y_2, \dots, y_{i-1}, \dots, y_{n-1}$:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \dots + a_{2m}x_m + \varepsilon_2, \\ \dots \dots \dots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \dots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Пример:

$$\begin{cases} P = a_0 + a_3W + u_1, \\ P' = b_0 + b_1P + b_4T + u_2, \\ Q = c_0 + c_1P + c_2P' + c_3W + u_3 \end{cases}$$

где P – цена на хлопок; P' – цена на хлопковые продукты; Q – количество проданных хлопковых товаров; W – индекс погодных условий; T – налоговые тарифы на хлопковые товары.

Цена на хлопок определяется погодой, а цена на хлопковые товары – ценой на хлопок и налогами и т.д.

2. КОРРЕЛЯЦИЯ, ВЫЧИСЛЕНИЕ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

2.1. ОЦЕНКА ТЕСНОТЫ ЛИНЕЙНОЙ СВЯЗИ.

Функциональные – характеризуются полным соответствием между изменением факторного признака и изменением результативной величины.

Пример Величина заработной платы при повременной оплате труда зависит от количества отработанных часов

Корреляционные – между изменением двух признаков нет полного соответствия, воздействие отдельных факторов проявляется лишь в среднем, при массовом наблюдении фактических данных.

Изучая взаимосвязи между признаками, их классифицируют по направлению, форме, числу факторов:

По **направлению** связи делятся на **прямые и обратные**. При **прямой связи** направление изменения результативного признака совпадает с направлением изменения признака-фактора. При **обратной связи** направление изменения результативного признака противоположно направлению изменения признака-фактора.

Пример Чем выше квалификация рабочего, тем выше уровень производительности его труда (прямая связь). Чем выше производительность труда, тем ниже себестоимость единицы продукции (обратная связь).

По **форме (виду функции)** связи делят на линейные (прямолинейные) и нелинейные (криволинейные). Линейная связь отображается прямой линией, нелинейная – кривой (параболой, гиперболой и т.п.). При линейной связи с возрастанием значения факторного признака происходит равномерное возрастание (убывание) значения результативного признака.

По **количеству факторов**, действующих на результативный признак, связи подразделяют на однофакторные (парные) и многофакторные.

Изучение зависимости вариации признака от окружающих условий и составляет содержание теории корреляции.

При проведении корреляционного анализа вся совокупность данных рассматривается как множество переменных (факторов), каждая из которых содержит n наблюдений.

При изучении взаимосвязи между двумя факторами их, как правило, обозначают $X = (x_1, x_2, \dots, x_n)$ и $Y = (y_1, y_2, \dots, y_n)$

Ковариация – это статистическая мера взаимодействия двух переменных.

Пример Положительное значение ковариации доходности двух ценных бумаг показывает, что доходности этих ценных бумаг имеют тенденцию изменяться в одну сторону.

Ковариация между двумя переменными X и Y рассчитывается следующим образом:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

где $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ – фактические значения переменных X и Y ;

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Если случайные величины X и Y независимы, теоретическая ковариация равна нулю. Ковариация зависит от единиц, в которых измеряются переменные X и Y она является ненормированной величиной. Поэтому для измерения силы связи между двумя переменными используется другая статистическая характеристика, называемая коэффициентом корреляции.

Коэффициент парной корреляции

$$\begin{aligned} r_{y,x} &= \frac{\text{Cov}(X,Y)}{S_x S_y} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

Дисперсия (оценка дисперсии) определяется по формуле

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Среднеквадратическое отклонение (стандартное отклонение) или **стандартная ошибка**

$$S_x = \sqrt{S_x^2}$$

Для **качественной оценки коэффициента корреляции** применяются различные шкалы, например, **шкала Чеддока**

0,1–0,3 — слабая;
0,3–0,5 — заметная;
0,5–0,7 — умеренная;
0,7–0,9 — высокая;
0,9–1,0 — весьма высокая.

Оценка существенности линейного коэффициента корреляции

***t*-критерий Стьюдента**

$$t_{набл} = \sqrt{\frac{r_{y,x}^2}{1-r_{y,x}^2}} (n - 2).$$

Матрица коэффициентов парной корреляции

<div style="display: inline-block; transform: rotate(-45deg); font-size: 0.8em;"> Переменная Номер наблюдения </div>	Y	X_1	X_2	...	X_m
1	y_1	x_{11}	x_{21}	...	x_{m1}
2	y_2	x_{12}	x_{22}	...	x_{m2}
3	y_3	x_{13}	x_{23}	...	x_{m3}
...
n	y_n	x_{1n}	x_{2n}	...	x_{mn}

$$R = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_1x_2} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_1x_m} & r_{x_2x_m} & \dots & 1 \end{pmatrix}.$$

Множественный коэффициент корреляции

$$R_{j,1,2,\dots,j-1,j+1,\dots,m} = \sqrt{1 - \frac{|R|}{R_{jj}}},$$

Проверка значимости коэффициента детерминации

Проверка значимости коэффициента детерминации осуществляется путем сравнения расчетного значения **F -критерия Фишера**

$$F_{\text{расч}} = \frac{R^2/m}{(1 - R^2)/(n - m - 1)}$$

с табличным $F_{\text{табл}}$.

Табличное значение критерия определяется заданным уровнем значимости α и степенями свободы $\nu_1 = m$ и $\nu_2 = n - m - 1$.

Коэффициент R^2 значимо отличается от нуля, если выполняется неравенство

$$F_{\text{расч}} > F_{\text{табл}}$$

Частный коэффициент корреляции

Выборочный частный коэффициент корреляции

$$r_{jk(1,2,\dots,m)} = \frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}},$$

Выражение (2.9) при условии $m = 3$ будет иметь вид

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}.$$

Пример 2.1. Вычисление коэффициентов парной, множественной и частной корреляции.

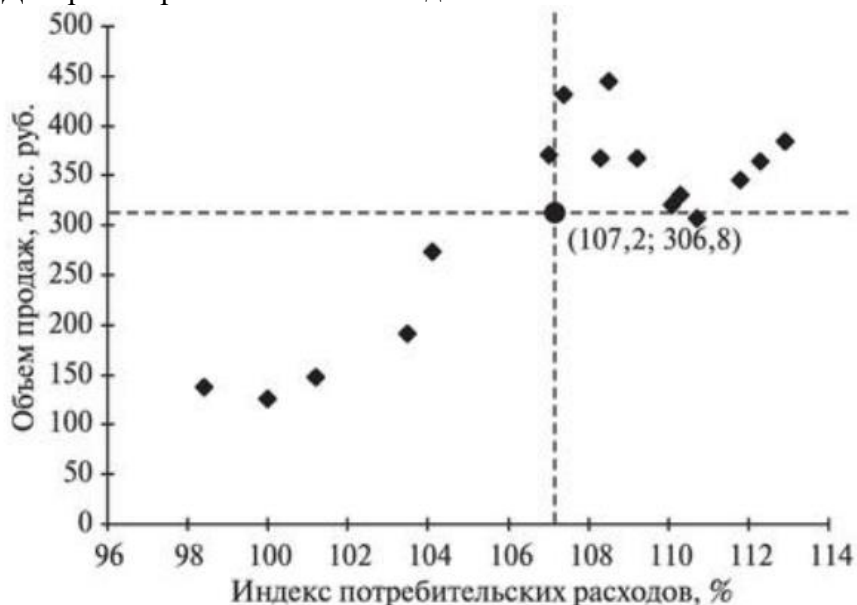
В табл. 2.2 представлена информация об объемах продаж и затратах на рекламу одной фирмы, а также индекс потребительских расходов за ряд текущих лет.

Объем продаж Y , тыс. руб.	Затраты на рекламу X_1 , тыс. руб.	Индекс потребительских расходов X_2 , %	Объем продаж Y , тыс. руб.	Затраты на рекламу X_1 , тыс. руб.	Индекс потребительских расходов X_2 , %
126	4	100,0	367	19,8	108,3
137	4,8	98,4	367	10,6	109,2
148	3,8	101,2	321	8,6	110,1
191	8,7	103,5	307	6,5	110,7
274	8,2	104,1	331	12,6	110,3
370	9,7	107,0	345	6,5	111,8
432	14,7	107,4	364	5,8	112,3
445	18,7	108,5	384	5,7	112,9

1. Построить диаграмму рассеяния (корреляционное поле) для переменных «объем продаж» и «индекс потребительских расходов».
2. Определить степень влияния индекса потребительских расходов на объем продаж (вычислить коэффициент парной корреляции).
3. Оценить значимость вычисленного коэффициента парной корреляции.
4. Построить матрицу коэффициентов парной корреляции по трем переменным.
5. Найти оценку множественного коэффициента корреляции.
6. Найти оценки коэффициентов частной корреляции.

Решение

1. Диаграмма рассеяния имеет вид



2. Промежуточные расчеты приведены в табл. 2.3

№	Y	X ₂	y _i - \bar{y}	x _i - \bar{x}	(y _i - \bar{y})(x _i - \bar{x})	(x _i - \bar{x}) ²	(y _i - \bar{y}) ²
1	126	100,0	-180,813	-7,231	1307,500	52,291	32 693,160
2	137	98,4	-169,813	-8,831	1499,657	77,991	28 836,285
3	148	101,2	-158,813	-6,031	957,838	36,376	25 221,410
4	191	103,5	-115,813	-3,731	432,125	13,922	13 412,535
5	274	104,1	-32,813	-3,131	102,744	9,805	1076,660
6	370	107,0	63,188	-0,231	-14,612	0,053	3992,660
7	432	107,4	125,188	0,169	21,125	0,028	15 671,910
8	445	108,5	138,188	1,269	175,325	1,610	19 095,785
9	367	108,3	60,188	1,069	64,325	1,142	3622,535
10	367	109,2	60,188	1,969	118,494	3,876	3622,535
11	321	110,1	14,188	2,869	40,700	8,230	201,285
12	307	110,7	0,188	3,469	0,650	12,032	0,035
13	331	110,3	24,188	3,069	74,225	9,417	585,035
14	345	111,8	38,188	4,569	174,469	20,873	1458,285
15	364	112,3	57,188	5,069	289,869	25,692	3270,410
16	384	112,9	77,188	5,669	437,557	32,135	5957,910
Сумма	4909	1715,7	0	0	5681,994	305,474	158 718,438
Среднее	306,81	107,23					

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 107,23; \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 306,81.$$

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{305,474}{15} = 20,36;$$

$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{158 718,438}{15} = 10 581,23.$$

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 4,51;$$

$$S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = 102,87.$$

$$r_{y,x} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y} = \frac{\frac{1}{15} \cdot 5681,99}{4,51 \cdot 102,87} = 0,816.$$

3. Оценим значимость коэффициента корреляции

$$t_{\text{расч}} = \frac{r_{y,x} \sqrt{n-2}}{\sqrt{1-r_{y,x}^2}} = \frac{0,816 \cdot \sqrt{14}}{\sqrt{1-0,666}} = 5,282.$$

4. Матрица R коэффициентов парной корреляции [см. формулу (2.2)]

$$R = \begin{pmatrix} 1 & 0,646 & 0,816 \\ 0,646 & 1 & 0,273 \\ 0,816 & 0,273 & 1 \end{pmatrix}.$$

5. Вычислим множественный коэффициент корреляции Y с X₁ и X₂:

$$R_{1,2,3} = \sqrt{1 - \frac{|R|}{R_{11}}} = \sqrt{1 - \frac{0,1304}{0,9253}} = 0,9269,$$

6. Вычислим коэффициенты частной корреляции по формуле (2.9)

$$r_{12(3)} = -\frac{R_{12}}{\sqrt{R_{11}R_{22}}} = -\frac{-0,423}{\sqrt{0,925 \cdot 0,334}} = 0,706;$$

$$r_{13(2)} = -\frac{R_{13}}{\sqrt{R_{11}R_{22}}} = -\frac{-0,639}{\sqrt{0,925 \cdot 0,334}} = 0,871,$$

2.2. ОЦЕНКА ТЕСНОТЫ НЕЛИНЕЙНОЙ СВЯЗИ

1. Вычислим среднее значение Y в j -й группе

$$\bar{y}_j = \sum_{i=1}^{m_j} y_{ji} / m_j.$$

2. Вычислим \bar{Y} , используя средние значения в каждой группе:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^k m_j \bar{y}_j.$$

3. Найдем межгрупповую дисперсию и общую дисперсию:

$$S_{y_j}^2 = \frac{1}{n} \sum_{j=1}^k m_j (\bar{y}_j - \bar{y})^2; \quad S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Корреляционное отношение η зависимой переменной Y по независимой переменной X

$$\eta = \sqrt{\frac{S_{y_j}^2}{S_y^2}} = \frac{S_{y_j}}{S_y}.$$

Пример 2.2. Вычисление объема выпускаемой продукции и температуры

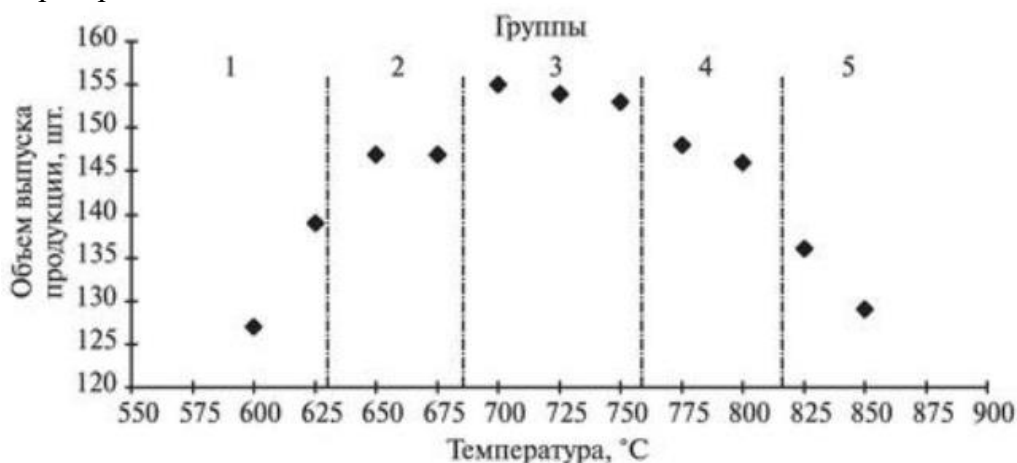
1. Построить диаграмму рассеяния (корреляционное поле) для совокупности данных (табл. 2.4)

Температура X , °С	600	625	650	675	700	725	750	775	800	825	850
Объем выпуска продукции Y , шт.	127	139	147	147	155	154	153	148	146	136	129

2. Оценить тесноту связи между объемом выпуска продукции и температурой

Решение

1. Корреляционное поле, показанное на рис. 2.6, иллюстрирует сильную нелинейную взаимосвязь, характеризующуюся незначительным случайным разбросом.



2. Оценим тесноту связи между объемом выпуска продукции и температурой с помощью корреляционного отношения.

Номер группы	Количество элементов в j -й группе m_j	Значения y_j , попавшие в j -ю группу	Среднее значение Y в j -й группе \bar{y}_j	$(\bar{y}_j - \bar{y})^2$
1	2	127; 139	133	$(133 - 143,727)^2$
2	2	147; 147	147	$(147 - 143,727)^2$
3	3	155; 154; 153	154	$(154 - 143,727)^2$
4	2	148; 146	147	$(147 - 143,727)^2$
5	2	136; 139	132,5	$(132,5 - 143,727)^2$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^k m_j \bar{y}_j = \frac{1}{11} (2 \cdot 133 + 2 \cdot 147 + 3 \cdot 154 + 2 \cdot 147 + 2 \cdot 132,5) = 143,727.$$

$$S_{y_j}^2 = \frac{1}{n} \sum_{j=1}^k m_j (\bar{y}_j - \bar{y})^2 = \frac{841,6816}{11} = 76,5165.$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{942,18}{11} = 85,6529.$$

$$\eta = \sqrt{\frac{S_{y_j}^2}{S_y^2}} = \sqrt{\frac{76,5165}{85,6529}} = \sqrt{0,893} = 0,945.$$

3. ЛИНЕЙНАЯ МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

Сформулируем регрессионную задачу для случая одного факторного признака.

Пусть имеется набор значений двух переменных: Y и X , где $Y = (y_1, y_2, \dots, y_n)$ – объясняемая переменная и $X = (x_1, x_2, \dots, x_n)$ объясняющая переменная, каждая из которых содержит n наблюдений. Пусть между переменными Y и X теоретически существует некоторая линейная зависимость

$$Y = f(X) = f(x_1, x_2, \dots, x_k) = \alpha + \beta x.$$

Это уравнение будем называть «*истинным*» *уравнением регрессии*. Однако в действительности между X и Y наблюдается не столь жесткая связь. Отдельные наблюдения y_i будут отклоняться от линейной зависимости в силу воздействия различных причин.

Обычно зависимая переменная находится под влиянием целого ряда факторов, в том числе и неизвестных исследователю, а также случайных причин (возмущения и помехи). Существенным источником отклонений в ряде случаев являются ошибки измерения. Отклонения от предполагаемой формы связи, естественно, могут возникнуть и в силу неправильного выбора вида уравнения, описывающего эту зависимость. Учитывая возможные отклонения, линейное уравнение связи двух переменных (парную регрессию) представим в виде

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

где α — постоянная величина (или свободный член уравнения); β — коэффициент регрессии, определяющий наклон линии, вдоль которой рассеяны данные наблюдений; ε_i — случайная переменная (случайная составляющая, остаток, или возмущение).

Коэффициент регрессии β характеризует изменение переменной y_i при изменении значения x_i на единицу. Если $\beta > 0$, переменные x_i и y_i *положительно* коррелированы, если $\beta < 0$ — *отрицательно* коррелированы.

Случайная составляющая ε_i отражает тот факт, что изменение, будет неточно описываться изменением x_i , так как присутствуют другие факторы, не учтенные в данной модели.

Таким образом, в уравнении (1) значение каждого наблюдения y_i представлено как сумма двух частей – *систематической* $\alpha + \beta x_i$ и *случайной* ε_i .

Можно сказать, что общим моментом для любой эконометрической модели является разбиение зависимой переменной на две части – *объясненную* и *случайную*.

3.1. ОСНОВНЫЕ ПРЕДПОСЫЛКИ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Свойства коэффициентов регрессии существенным образом зависят от свойств случайной составляющей. Для того чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов (МНК), давал наилучшие из всех возможных результаты, должны выполняться следующие условия, известные как **условия Гаусса-Маркова**:

1. Математическое ожидание случайной составляющей в любом наблюдении должно быть равно нулю. Иногда случайная составляющая e_i будет положительной, иногда отрицательной, но она не должна иметь систематического смещения ни в одном из двух возможных направлений. Условие записывается следующим образом:

$$M(\varepsilon_i) = 0.$$

Фактически, если уравнение регрессии включает постоянный член, то обычно условие (2) выполняется автоматически, так как роль константы состоит в определении любой систематической тенденции Y , которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

2. В модели (3.1) возмущение ε_i (или зависимая переменная y_i) есть величина случайная, а объясняющая переменная x_i – величина не случайная. Если это условие выполнено, то теоретическая ковариация между независимой переменной и случайным членом равна нулю.

3. В любых двух наблюдениях отсутствует систематическая связь между значениями случайной составляющей. Например, если случайная составляющая велика и положительна в одном наблюдении, это не должно обуславливать систематическую тенденцию к тому, что она будет большой и положительной в следующем наблюдении. Случайные составляющие должны быть независимы друг от друга.

В силу того что $M(\varepsilon_i) = M(\varepsilon_j) = 0$, данное условие можно записать следующим образом:

$$M(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j).$$

Возмущения ε_i и ε_j некоррелированы (**условие независимости случайных составляющих** в различных наблюдениях). Это условие означает, что отклонения регрессии (а значит, и сама зависимая переменная) не коррелируют. В случае временного ряда y_i (условие (3.3) означает отсутствие автокорреляции ряда ε_i

4. Дисперсия случайной составляющей должна быть постоянна для всех наблюдений. Иногда случайная составляющая будет больше, иногда меньше, однако не должно быть априорной причины для того, чтобы она порождала большую ошибку в одних наблюдениях, чем в других.

Постоянная дисперсия обычно обозначается $\sigma^2(\varepsilon)$ или σ_ε^2 условие записывается следующим образом:

$$D(\varepsilon_i) = \sigma_\varepsilon^2$$

Это условие **гомоскедастичности**, или **равноизменчивости**, случайной составляющей (возмущения).

Величина σ_ε^2 , конечно, неизвестна. Одна из задач регрессионного анализа состоит в оценке стандартного отклонения случайной составляющей.

Наряду с условиями Гаусса-Маркова, обычно также предполагается нормальность распределения случайного члена. Дело в том, что если случайный член нормально распределен, то так же будут распределены и коэффициенты регрессии.

В тех случаях, когда предпосылки выполняются, оценки, полученные по МНК, будут обладать **свойствами несмещенности, эффективности и состоятельности**.

Несмещенность оценки означает, что математическое ожидание остатков равно нулю. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным исследованиям.

Оценки считаются **эффективными**, если они характеризуются наименьшей дисперсией. Поэтому несмещенность оценки должна дополняться минимальной дисперсией.

Достоверность доверительных интервалов параметров регрессии обеспечивается, если оценки будут не только несмещенными и эффективными, но и состоятельными.

Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки.

3.2. ОЦЕНКА ПАРАМЕТРОВ РЕГРЕССИОННОГО УРАВНЕНИЯ

Для оценки параметров регрессионного уравнения наиболее часто используют метод наименьших квадратов.

Метод наименьших квадратов (МНК) дает оценки, имеющие наименьшую дисперсию в классе всех линейных оценок, если выполняются предпосылки нормальной линейной регрессионной модели. МНК минимизирует сумму квадратов отклонения наблюдаемых значений y_i от модельных значений \hat{y}_i .

Оценки $\hat{\alpha}, \hat{\beta}$ находят путем минимизации суммы квадратов

$$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

по всем возможным значениям α и β при заданных (наблюдаемых) значениях $x_1, \dots, x_n, y_1, \dots, y_n$. Задача сводится к математической задаче поиска точки минимума функции двух переменных. Точка минимума находится путем приравнивания к нулю частных производных функции $z = Q(\alpha, \beta)$ по переменным α и β . Это приводит к системе нормальных уравнений

$$\begin{cases} \frac{\partial Q(\alpha, \beta)}{\partial \alpha} = 0, \\ \frac{\partial Q(\alpha, \beta)}{\partial \beta} = 0, \end{cases}$$

решением которой и является пара $\hat{\alpha}, \hat{\beta}$.

Согласно правилам вычисления производных имеем

$$\begin{cases} \frac{\partial Q(\alpha, \beta)}{\partial \alpha} = 2 \sum_{i=1}^n (y_i - \alpha - \beta x_i)(-1), \\ \frac{\partial Q(\alpha, \beta)}{\partial \beta} = 2 \sum_{i=1}^n (y_i - \alpha - \beta x_i)(-x_i), \end{cases}$$

так что искомые значения $\hat{\alpha}, \hat{\beta}$ удовлетворяют соотношениям

$$\begin{cases} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0, \\ \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0. \end{cases}$$

Эту систему двух уравнений можно записать также в виде

$$\begin{cases} n\hat{\alpha} + \left(\sum_{i=1}^n x_i \right) \hat{\beta} = \sum_{i=1}^n y_i, \\ \left(\sum_{i=1}^n x_i \right) \hat{\alpha} + \left(\sum_{i=1}^n x_i^2 \right) \hat{\beta} = \sum_{i=1}^n x_i y_i. \end{cases}$$

Она является системой двух линейных уравнений с двумя неизвестными и может быть легко решена, например, методом подстановки. В результате получаем так называемые оценки наименьших квадратов:

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}.$$

Такое решение может существовать только при выполнении условия

$$\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0,$$

что равносильно отличию от нуля определителя системы нормальных уравнений. Действительно, этот определитель равен

$$n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = n \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] = n \sum_{i=1}^n (x_i - \bar{x})^2.$$

Условие (3.6) называется условием идентифицируемости модели наблюдений $y_i = (\alpha + \beta_i) + \varepsilon_i, i = \overline{1, n}$, и означает, что не все значения x_1, \dots, x_n совпадают между собой. При нарушении этого условия все точки $(x_i, y_i), i = \overline{1, n}$ лежат на одной вертикальной прямой $x = \bar{x}$.

Обратим внимание на полученное выражение для параметра [см. формулу (3.5)]. Сюда входят выражения, участвовавшие ранее в определении выборочной дисперсии

$$S_x^2 = \text{Var}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

и выборочной ковариации

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Так что в этих терминах параметр $\hat{\beta}$ можно получить следующим образом:

$$\hat{\beta} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x^2} =$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r_{y,x} \frac{S_y}{S_x} = \frac{\overline{y\bar{x}} - \bar{y}\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\sum_{i=1}^n y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}.$$

В матричной форме модель парной регрессии имеет вид

$$Y = X \cdot a + \varepsilon,$$

где Y – вектор-столбец размерности $n \times 1$ наблюдаемых значений зависимой переменной; X – матрица размерности $n \times 2$ наблюдаемых значений факторных признаков (дополнительный фактор x_0 , состоящий из одних единиц, вводится для вычисления свободного члена); a – вектор-столбец размерности 2×1 неизвестных, подлежащих оценке коэффициентов регрессии; ε – вектор-столбец размерности $n \times 1$ ошибок наблюдений.

Таким образом.

$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_i \\ \dots & \dots \\ 1 & x_n \end{pmatrix}, \quad a = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_i \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Решение системы нормальных уравнений в матричной форме имеет вид

$$A = (X'X)^{-1}X'Y = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}.$$

3. ОЦЕНКА КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Качество модели регрессии связывают с ее адекватностью наблюдаемым (эмпирическим) данным. Проверка адекватности (или соответствия) модели регрессии наблюдаемым данным проводится на основе анализа остатков e_i .

После построения уравнения регрессии мы можем разбить значение Y в каждом наблюдении на две составляющие – \hat{y}_i и e_i :

$$y_i = \hat{y}_i + e_i.$$

Остаток e_i представляет собой отклонение фактического значения зависимой переменной от ее значения, полученного расчетным путем:

$$e_i = y_i - \hat{y}_i.$$

Если $e_i = 0$ ($i = \overline{1, n}$), то для всех наблюдений фактические значения зависимой переменной совпадают с расчетными (теоретическими) значениями. Графически это означает, что теоретическая линия регрессии (линия, построенная по функции $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ проходит через все точки корреляционного поля, что возможно только при строго функциональной связи. Следовательно, результативный признак Y полностью обусловлен влиянием фактора X .

На практике, как правило, имеет место некоторое рассеивание точек корреляционного поля относительно теоретической линии регрессии, т.е. отклонения эмпирических данных от теоретических ($e_i \neq 0$). Величина этих отклонений и лежит в основе расчета показателей качества (адекватности) уравнения регрессии.

При анализе качества модели регрессии используется основное положение дисперсионного анализа, согласно которому общая сумма квадратов отклонений зависимой переменной от среднего значения y

может быть разложена на две составляющие – объясненную и не объясненную уравнением регрессии:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

где \hat{y}_i – значения, y вычисленные по модели $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$.

Часто уравнение (3.11) записывают в следующих обозначениях:

$$TSS = RSS + ESS,$$

<p>где</p> $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ $RSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ $ESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$	<p>общая сумма квадратов отклонений зависимой переменной от ее среднего выборочного значения;</p> <p>объясненная регрессией сумма квадратов отклонений;</p> <p>не объясненная регрессией (остаточная) сумма квадратов отклонений.</p>
--	---

Коэффициент детерминации определяют следующим образом:

$$R^2 = \frac{\text{Объясненная сумма квадратов}}{\text{Общая сумма квадратов}} =$$

$$= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}.$$

Данный коэффициент показывает *долю вариации результативного признака, находящегося под воздействием изучаемых факторов*, т.е. определяет, какая доля вариации признака Y учтена в модели и обусловлена влиянием на него факторов.

Чем ближе R^2 к единице, тем выше качество модели.

Для оценки качества регрессионных моделей целесообразно также использовать коэффициент **множественной корреляции** (индекс корреляции):

$$R = \sqrt{1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Данный коэффициент универсален, так как он отражает тесноту связи и точность модели, а также может использоваться при любой форме связи переменных.

Для парной модели регрессии индекс корреляции равен *коэффициенту парной корреляции*.

$$R = |r_{y,x}|.$$

Очевидно, что чем меньше влияние неучтенных факторов, тем лучше модель соответствует фактическим данным.

Для оценки качества регрессионных моделей может использоваться **средняя относительная ошибка аппроксимации**:

$$E_{\text{отн}} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\% = \frac{1}{n} \sum_{i=1}^n \frac{|e_i|}{y_i} \cdot 100\%.$$

Чем меньше рассеяние эмпирических точек вокруг теоретической линии регрессии, тем меньше средняя ошибка аппроксимации; $E_{\text{отн}} < 7\%$ свидетельствует о хорошем качестве модели.

После того как уравнение регрессии построено, выполняется **проверка значимости построенного уравнения в целом и отдельных параметров**.

Во-первых, необходимо оценить значимость уравнения регрессии, т.е. установить, соответствует ли математическая модель, выражающая зависимость между Y и X , фактическим данным и достаточно ли

включенных в уравнение объясняющих переменных X для описания зависимой переменной Y .

Оценка значимости уравнения регрессии позволяет узнать, пригодно уравнение регрессии для практического использования (например, для прогноза) или нет. При этом выдвигают основную гипотезу о *незначимости* уравнения в целом, которая формально сводится к гипотезе о равенстве нулю параметров регрессии, или, что то же самое, о равенстве нулю коэффициента детерминации: $R^2 = 0$. Альтернативная гипотеза о значимости уравнения – гипотеза о неравенстве нулю параметров регрессии.

Для проверки значимости модели регрессии используется F -критерий Фишера, вычисляемый как отношение дисперсии исходного ряда и несмещенной дисперсии остаточной компоненты. Если расчетное значение с $v_1 = k$ и $v_2 = n - k - 1$ степенями свободы, где k – количество факторов, включенных в модель, больше табличного при заданном уровне значимости, то модель считается *значимой*.

Для модели парной регрессии

$$F = \frac{R^2}{1 - R^2} (n - 2) = \frac{r_{y,x}^2}{1 - r_{y,x}^2} (n - 2).$$

В качестве меры точности применяют несмещенную оценку дисперсии остаточной компоненты σ_e^2 , которая представляет собой отношение суммы квадратов уровней остаточной компоненты к величине $(n - k - 1)$, где k – количество факторов, включенных в модель. Квадратный корень из этой величины называется **стандартной ошибкой**:

$$\sigma_e = \sqrt{\frac{1}{n - k - 1} \sum_{i=1}^n e_i^2}.$$

Для модели парной регрессии

$$\sigma_e = \sqrt{\frac{1}{n - 2} \sum_{i=1}^n e_i^2}.$$

Во-вторых, необходим анализ статистической значимости параметров модели парной регрессии $y_i = \alpha + \beta x_i + \varepsilon_i$.

Значения y_i соответствующие данным x_i при теоретических значениях α и β . являются случайными. Случайными являются и рассчитанные по ним значения коэффициентов $\hat{\alpha}$ и $\hat{\beta}$.

Надежность получаемых оценок α и β зависит от дисперсии случайных отклонений (ошибок).

По данным выборки эти отклонения и, соответственно, их дисперсия не оцениваются – в расчетах используются отклонения зависимой переменной y_i от ее расчетных значений \hat{y}_i .

$$e_i = y_i - \hat{\alpha} - \hat{\beta}x_i.$$

Так как ошибки (остатки) e_i нормально распределены, то среднеквадратическое отклонение ошибок используется для измерения этой вариации.

Среднеквадратические отклонения коэффициентов известны как стандартные ошибки коэффициентов:

$$\sigma_{\hat{\alpha}} = \frac{\sigma_e \sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}} = \sqrt{\frac{\sigma_e^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}},$$

$$\sigma_{\hat{\beta}} = \frac{\sigma_e \sqrt{n}}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}} = \sqrt{\frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

где \bar{x} – среднее значение независимой переменной x ; σ_e – стандартная ошибка, вычисляемая по формуле (3.16).

Проверка значимости отдельных коэффициентов регрессии связана с определением расчетных значений t -критерия (t -статистики) для соответствующих коэффициентов регрессии:

$$t_{\alpha \text{ расч}} = \frac{|\hat{\alpha}|}{\sigma_{\hat{\alpha}}}; \quad t_{\beta \text{ расч}} = \frac{|\hat{\beta}|}{\sigma_{\hat{\beta}}}.$$

Затем расчетные значения $t_{\text{расч}}$ сравниваются с табличными $t_{\text{табл}}$.

Табличное значение критерия определяется при $(n - 2)$ степенях свободы (n – число наблюдений) и соответствующем уровне значимости α .

Если расчетное значение t -критерия с $(n - k - 1)$ степенями свободы больше его табличного значения при заданном уровне значимости, коэффициент регрессии считается значимым. В противном случае фактор, соответствующий этому коэффициенту, должен быть исключен из модели, а оставшиеся в модели параметры пересчитаны.

Интервальная оценка параметров модели выполняется для значимого уравнения по формулам

$$\hat{\alpha} \pm t_{\text{кр}} \sigma_{\hat{\alpha}}, \quad \hat{\beta} \pm t_{\text{кр}} \sigma_{\hat{\beta}},$$

где $\sigma_{\hat{\alpha}}, \sigma_{\hat{\beta}}$ – стандартные ошибки параметров модели. Полученные таким образом доверительные интервалы с вероятностью $(1 - \alpha)$ накрывают истинные значения параметров α и β .

3.4. ПРОГНОЗИРОВАНИЕ С ПРИМЕНЕНИЕМ УРАВНЕНИЯ РЕГРЕССИИ

$$\hat{y}_{\text{прогн}} = \hat{\alpha} + \hat{\beta} x_{\text{прогн}}.$$

$$y_{\text{прогн}} \in \left[\hat{y}_{\text{прогн}} \pm \sigma_e t_{\alpha} \sqrt{1 + \frac{1}{n} + \frac{(x_{\text{прогн}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$

Пример 1.

В табл. 1 приведена информация о среднедушевых месячных доходах и расходах по Центральному федеральному округу в 2002 г.

Требуется:

- **построить** однофакторную модель регрессии зависимости расходов от доходов;

- **проверить** качество уравнения регрессии и оценить его значимость. Оценить точность модели;
- **проверить** значимость коэффициента модели регрессии, вычислить доверительные интервалы с вероятностью 95% коэффициента модели регрессии;
- **построить** доверительный интервал для полученной модели регрессии ($\alpha = 0,1$);
- **оценить** расходы, если доход составит 3600 руб. ($\alpha = 0,1$);
- **отобразить** на графике исходные данные, результаты моделирования и прогнозирования.

Таблица 1

<i>Область</i>	<i>№ п/п</i>	<i>Доходы, руб.</i>	<i>Расходы, руб.</i>
Белгородская	1	2784	2478
Брянская	2	2255	2034
Владимирская	3	2062	2019
Воронежская	4	2553	2501
Ивановская	5	1595	1668
Калужская	6	2254	2188
Костромская	7	2371	2217
Курская	8	2518	2202
Липецкая	9	2742	2392
Московская (без г. Москва)	10	3416	3354
Орловская	11	2540	2347
Рязанская	12	2510	2309
Смоленская	13	2843	2671
Тамбовская	14	2648	2201
Тверская	15	2204	1932
Тульская	16	2561	2160
Ярославская	17	3311	2921

Источник: Россия в цифрах. 2004 / Росстат. М., 2004.

РЕШЕНИЕ

1. Для вычисления параметров модели воспользуемся формулами (3.5).

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Промежуточные расчеты приведены в табл. 3.2.

Таблица 2

№ п/п	Доходы x_i	Расходы y_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})(x_i - \bar{x})$	\hat{y}_i	Остатки e_i	e_i^2
1	2784	2478	148,94	244,76	59 909,76	36 455,54	2537,13	-59,13	3496,59
2	2255	2034	-295,06	-284,24	80 789,70	83 866,13	2087,43	-53,43	2854,98
3	2062	2019	-310,06	-477,24	227 753,53	147 971,01	1923,36	95,64	9146,30
4	2553	2501	171,94	13,76	189,47	2366,72	2340,76	160,24	25 676,82
5	1595	1668	-661,06	-944,24	891 580,29	624 195,07	1526,37	141,63	20 059,17
6	2254	2188	-141,06	-285,24	81 359,17	40 234,96	2086,58	101,42	10 285,64
7	2371	2217	-112,06	-168,24	28 303,11	18 852,25	2186,04	30,96	958,34
8	2518	2202	-127,06	-21,24	450,94	2698,13	2311,01	-109,01	11 882,49
9	2742	2392	62,94	202,76	41 113,53	12 762,25	2501,43	-109,43	11 974,48
10	3416	3354	1024,94	876,76	768 716,35	898 632,25	3074,39	279,61	78 180,82
11	2540	2347	17,94	0,76	0,58	13,72	2329,71	17,29	298,98
12	2510	2309	-20,06	-29,24	854,70	586,43	2304,21	4,79	22,98
13	2843	2671	341,94	303,76	92 273,00	103 869,66	2587,29	83,71	7007,78
14	2648	2201	-128,06	108,76	11 829,76	-13 928,28	2421,52	-220,52	48 628,67
15	2204	1932	-397,06	-335,24	112 382,70	133 108,13	2044,08	-112,08	12 561,29
16	2561	2160	-169,06	21,76	473,70	-3679,52	2347,56	-187,56	35 179,08
17	3311	2921	591,94	771,76	595 620,76	456 839,31	2985,13	-64,13	4112,88
Сумма	43 167,00	39 594,00	0	0	2 993 601,06	2 544 843,76	39 594,00	0	282 327,28
Среднее	2539,24	2329,06	0	0	—	149 696,69	—	0	—

Получаем

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{2\,544\,843,76}{2\,993\,601,06} = \mathbf{0,85},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2329,06 + 0,85 \cdot 2539,24 = \mathbf{170,47}.$$

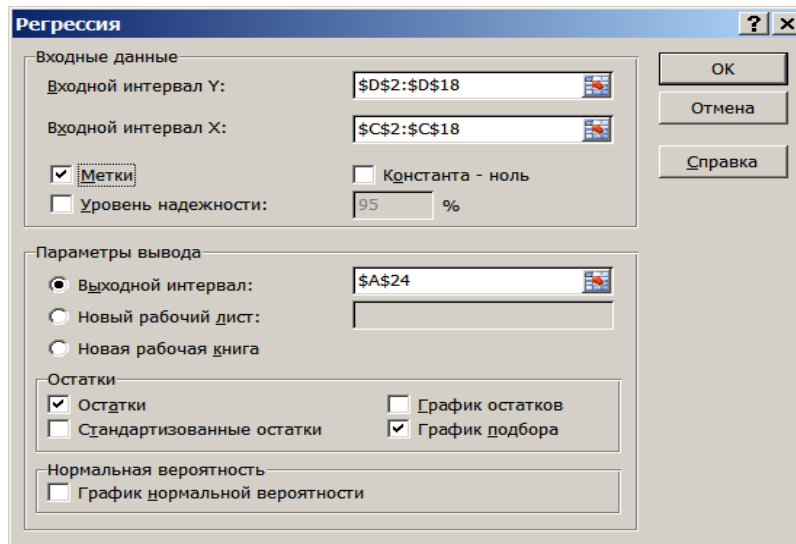
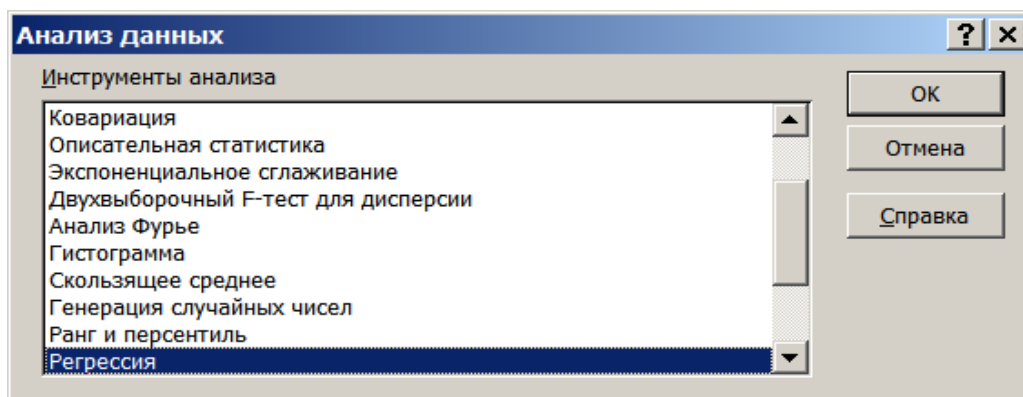
Представим расчет параметров модели в матричной форме:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{2\,544\,843,76}{2\,993\,601,06} = \mathbf{0,85},$$

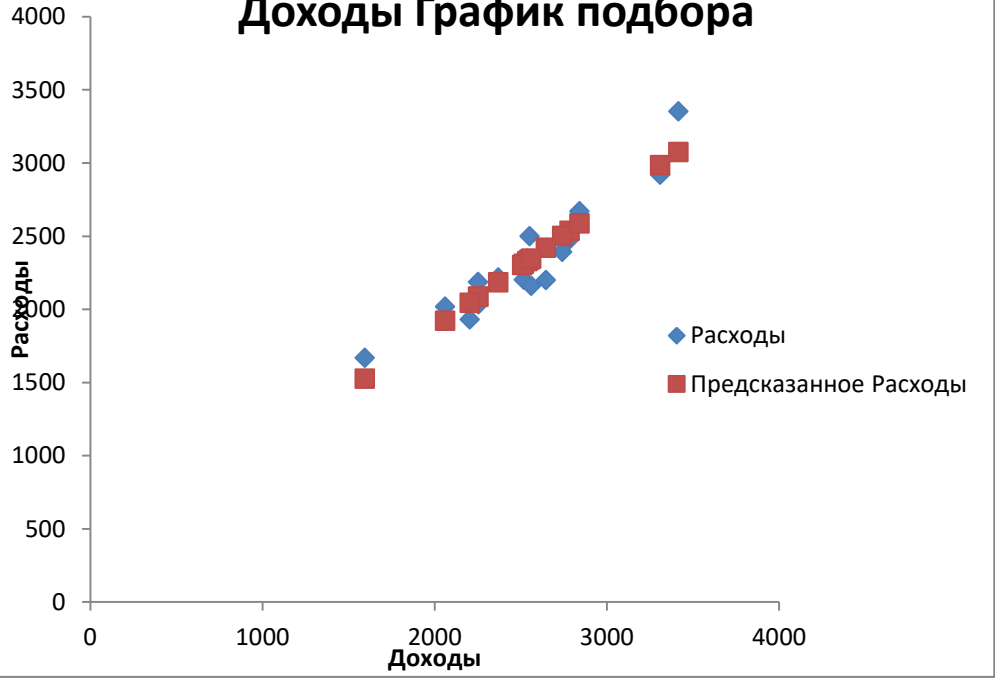
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 2329,06 + 0,85 \cdot 2539,24 = \mathbf{170,47}.$$

Построена модель зависимости расходов от дохода:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = \mathbf{170,47 + 0,85x_i}.$$



Доходы График подбора



ВЫВОД ИТОГОВ								
<i>Регрессионная статистика</i>								
Множественный	0,940511063							
R-квадрат	0,884561059							
Нормированный	0,87686513							
Стандартная ош	137,1926333							
Наблюдения	17							
<i>Дисперсионный анализ</i>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>			
Регрессия	1	2163357,662	2163357,662	114,938822	1,98454E-08			
Остаток	15	282327,2794	18821,81863					
Итого	16	2445684,941						
	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-Значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95,0%</i>	<i>Верхние 95,0%</i>
Y-пересечение	170,4688917	204,0740415	0,835328641	0,416640012	-264,5046293	605,442413	-264,5046293	605,4424127
Доходы	0,85009449	0,079292814	10,72095248	1,98454E-08	0,681085858	1,01910312	0,681085858	1,019103121

Регрессионная статистика		
Наименование показателя в отчете Excel	Принятые наименования	Формула
Множественный R	Коэффициент множественной корреляции, индекс корреляции	$R = \sqrt{R^2}$ 0,941
R-квадрат	Коэффициент детерминации R^2	$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{ESS}{TSS}$ $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{RSS}{TSS}$ 0,885
Нормированный R-квадрат	Скорректированный R^2	$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k-1}$ 0,878
Стандартная ошибка	Среднеквадратическое отклонение от модели	$\sigma_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-k-1}} = \sqrt{\frac{ESS}{n-k-1}}$ 137,193
Наблюдения	Количество наблюдений	$n = 17$

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{282\,327,28}{2\,445\,684,94} = \mathbf{0,885}.$$

	<i>df</i> – число степеней свободы	<i>SS</i> – сумма квадратов отклонений	$MS = SS/df$	F -критерий Фишера = $\frac{MS \text{ (регрессия)}}{MS \text{ (остаток)}}$	Значимость <i>F</i>
Регрессия	$k = 1$	$RSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 2 163 357,66	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{k} = \frac{RSS}{k}$ 2 163 357,66	$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$ 114,94	Уровень значимости, соответствующий вычисленной <i>F</i> -статистике 1,98454E-08
Остаток	$n - k - 1 = 15$	$ESS = \sum_{i=1}^n e_i^2$ 282 327,28	$\frac{\sum_{i=1}^n e_i^2}{n - k - 1} = \frac{ESS}{n - k - 1}$ 18 821,82		
Итого	$n - 1 = 16$	$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ 2 445 684,94			

	Коэффициенты	Стандартная ошибка	<i>t</i> -статистика	<i>P</i> -значение	Нижние 95%	Верхние 95%
У-пересечение	170,47	204,07	0,84	0,42	-264,50	605,44
Доходы	0,85	0,08	10,72	0,00	0,68	1,02

ВЫВОД ОСТАТКА		
<i>Наблюдение</i>	<i>Предсказанное Расходы</i>	<i>Остатки</i>
1	2537,131951	-59,13195129
2	2087,431966	-53,4319662
3	1923,36373	95,63627033
4	2340,760124	160,2398758
5	1526,369603	141,6303971
6	2086,581872	101,4181283
7	2186,042927	30,95707299
8	2311,006817	-109,006817
9	2501,427983	-109,4279827
10	3074,391669	279,6083312
11	2329,708896	17,29110421
12	2304,206061	4,793938907
13	2587,287526	83,71247381
14	2421,519101	-220,5191007
15	2044,077147	-112,0771472
16	2347,56088	-187,5608801
17	2985,131747	-64,13174741

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0,885/1}{(1-0,885)/(17-2)} = \mathbf{114,939}.$$

$$t_{\beta_{расч}} = \frac{|\hat{\beta}|}{\sigma_{\hat{\beta}}} = \frac{0,85}{0,079} = \mathbf{10,72},$$

где $\sigma_{\hat{\beta}} = \sqrt{\frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \sqrt{\frac{137,19^2}{2993601,06}} = 0,079,$

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}} = \sqrt{\frac{282327,28}{15}} = 137,19.$$

$$y_i \in [\hat{y}_i \pm U_i] = \left[\hat{y}_i \pm 137,19 \cdot 1,75 \cdot \sqrt{1 + \frac{1}{17} + \frac{(x_i - 2539,24)^2}{2993601,06}} \right],$$

$$\hat{y}_{\text{прогн}} = 170,47 + 0,85 \cdot 3600 \approx \mathbf{3230,81}.$$

Тогда

$$U_{(x=3600, n=17, \alpha=0,1)} = 137,19 \cdot 1,75 \cdot \sqrt{1 + \frac{1}{17} + \frac{(3600 - 2539,24)^2}{2993601,06}} = \mathbf{288,08}.$$

i	U_i	Верхняя граница	Нижняя граница	i	U_i	Верхняя граница	Нижняя граница
1	249,81	2786,94	2287,33	10	275,86	3350,25	2798,53
2	250,61	2338,04	1836,82	11	247,48	2577,19	2082,23
3	256,22	2179,58	1667,15	12	247,51	2551,72	2056,69
4	247,49	2588,25	2093,27	13	251,05	2838,34	2336,23
5	280,13	1806,50	1246,24	14	247,94	2669,46	2173,58
6	250,63	2337,22	1835,95	15	251,83	2295,90	1792,25
7	248,58	2434,62	1937,46	16	247,50	2595,06	2100,06
8	247,50	2558,50	2063,51	17	269,73	3254,86	2715,40
9	249,08	2750,51	2252,35				

Таким образом, прогнозируемое значение $\hat{y}_{\text{прогн}} = 3230,81$ с вероятностью 90% будет находиться между верхней границей, равной $3230,81 + 288,08 = 3518,88$, и нижней границей, равной $3230,81 - 288,08 = 2942,73$ (рис. 3.3.3).

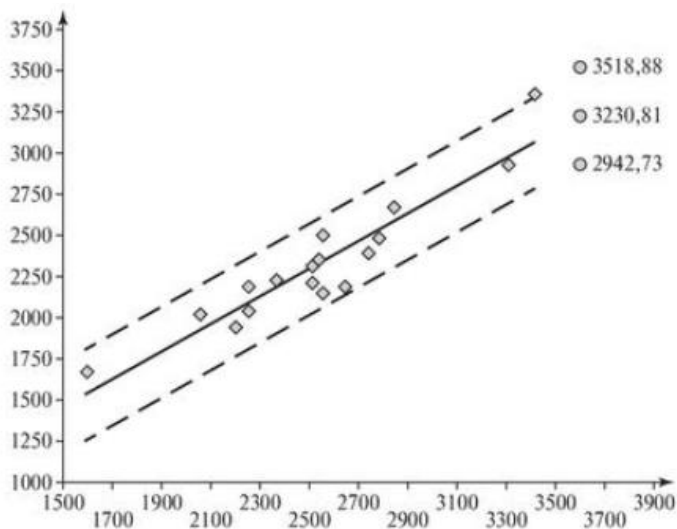


Рис. 3.3.3. График исходных данных (◊), результатов моделирования (—), прогнозирования (◊) и доверительные интервалы (— — —)

4. МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

4.1. ОТБОР ФАКТОРОВ В МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ. ОЦЕНКА ПАРАМЕТРОВ МОДЕЛИ

При построении модели множественной регрессии для отображения зависимости между объясняемой переменной Y и независимыми (объясняющими) переменными X_1, X_2, \dots, X_k могут использоваться показательная, параболическая и многие другие функции. Однако наибольшее распространение получили модели линейной взаимосвязи, когда факторы входят в модель линейно.

Линейная модель множественной регрессии имеет вид

$$y_i = a_0 + a_1x_{i1} + a_2x_{i2} + \dots + a_kx_{ik} + \varepsilon_i, \quad i = \overline{1, n}, \quad (4.1)$$

где k – количество включенных в модель факторов.

Коэффициент регрессии a_j показывает, на какую величину в среднем изменится результирующий признак Y , если переменную X_j увеличить на единицу измерения, т.е. является нормативным коэффициентом.

Анализ уравнения (1) и методика определения параметров становятся более наглядными, а расчетные процедуры существенно упрощаются, если воспользоваться матричной формой записи уравнения:

$$Y = X \cdot a + \varepsilon$$

где Y – это вектор зависимой переменной размерности $n \times 1$, представляющий собой n наблюдений значений y_i ; X – матрица n наблюдений независимых переменных X_1, X_2, \dots, X_k , размерность матрицы X равна $n \times (k + 1)$; a – подлежащий оцениванию вектор неизвестных параметров размерности $(k + 1) \times 1$; ε – вектор случайных отклонений (возмущений) размерности $n \times 1$.

Таким образом,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Уравнение (4.1) содержит значения неизвестных параметров $a_0, a_1, a_2, \dots, a_k$. Эти величины оцениваются на основе выборочных наблюдений, поэтому полученные расчетные показатели не являются истинными, а представляют собой лишь их статистические оценки.

Модель линейной регрессии, в которой вместо истинных значений параметров подставлены их оценки (а именно такие регрессии и применяются на практике), имеет вид

$$Y = XA + e = \hat{Y} + e, \quad (4.2)$$

где A – вектор оценок параметров; e – вектор «оцененных» отклонений регрессии, остатки регрессии $e = Y - XA$; \hat{Y} – оценка значений Y , равная XA .

Оценка параметров модели множественной регрессии проводится с помощью метода наименьших квадратов. Формулу для вычисления параметров регрессионного уравнения приведем без вывода:

$$A = (X'X)^{-1}X'Y \quad (4.3)$$

Отбор факторов, включаемых в регрессию – один из важнейших этапов построения модели регрессии. Подходы к отбору факторов могут быть разные: один из них основан на анализе матрицы коэффициентов парной корреляции, другой – на процедурах пошагового отбора факторов.

Перед построением модели множественной регрессии вычисляются парные коэффициенты линейной корреляции между всеми исследуемыми переменными Y, X_1, X_2, \dots, X_m , и из них формируется матрица

$$R = \begin{pmatrix} 1 & r_{x_1,y} & r_{x_2,y} & \dots & r_{x_m,y} \\ r_{y,x_1} & 1 & r_{x_2,x_1} & \dots & r_{x_m,x_1} \\ r_{y,x_2} & r_{x_1,x_2} & 1 & \dots & r_{x_m,x_2} \\ \dots & \dots & \dots & 1 & \dots \\ r_{y,x_m} & r_{x_1,x_m} & r_{x_2,x_m} & \dots & 1 \end{pmatrix}.$$

Вначале анализируют коэффициенты корреляции, отражающие тесноту связи зависимой переменной со всеми включенными в анализ факторами, с целью отсева незначимых переменных.

Затем переходят к анализу остальных столбцов матрицы с целью выявления мультиколлинеарности.

Ситуация, когда два фактора связаны между собой тесной линейной связью (парный коэффициент корреляции между ними превышает по абсолютной величине 0,8), называется **коллинеарностью факторов**. Коллинеарные факторы фактически дублируют друг друга в модели, существенно ухудшая ее качество.

Наибольшие трудности возникают при наличии мультиколлинеарности факторов, когда тесной связью одновременно связаны несколько факторов, т.е. когда нарушается одна из предпосылок регрессионного анализа, состоящая в том, что объясняющие переменные должны быть независимы.

Под **мультиколлинеарностью** понимается высокая взаимная коррелированность объясняющих переменных, которая приводит к линейной зависимости нормальных уравнений. Мультиколлинеарность может проявляться в двух формах:

- **функциональной** – определитель матрицы $X'X$ равен нулю. Это приводит к невозможности решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели;
- **стохастической**, когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. В этом случае определитель матрицы $X'X$ не равен нулю, но очень мал. Экономическая интерпретация параметров уравнения регрессии при этом затруднена, так как некоторые из его коэффициентов могут иметь неправильные с точки зрения экономической теории знаки и неоправданно большие значения. Оценки

параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Мультиколлинеарность может возникать в силу разных причин. Например, несколько независимых переменных могут иметь общий временной тренд, относительно которого они совершают малые колебания.

Существует несколько **способов для определения наличия или отсутствия мультиколлинеарности:**

- анализ матрицы коэффициентов парной корреляции. Явление мультиколлинеарности в исходных данных считают установленным, если коэффициент парной корреляции между двумя переменными больше 0,8:

$$r_{x_i x_k} > 0,8;$$

- исследование матрицы $X'X$. Если определитель матрицы $X'X$ близок к нулю, это свидетельствует о наличии мультиколлинеарности.

Для выявления второй ситуации служит тест на мультиколлинеарность Фаррара-Глоубера. С помощью этого теста проверяют, насколько значимо определитель матрицы парных коэффициентов корреляции отличается от единицы. Если он равен нулю, то столбцы матрицы X линейно зависимы и вычислить оценку коэффициентов множественной регрессии по методу наименьших квадратов становится невозможно.

Этот алгоритм содержит **три вида статистических критериев проверки наличия мультиколлинеарности:**

- 1) всего массива переменных (критерий «хи-квадрат»);
- 2) каждой переменной с другими переменными (F -критерий);
- 3) каждой пары переменных (t -тест).

Опишем алгоритм для каждого вида критериев.

1. *Проверка наличия мультиколлинеарности всего массива переменных (критерий «хи-квадрат»):*

- 1) Построить корреляционную матрицу R и найти ее определитель $\det[R]$.
- 2) Вычислить наблюдаемое значение статистики Фаррара-Глоубера по формуле

$$FG_{\text{набл}} = - \left[n - 1 - \frac{1}{6}(2k + 5) \right] \ln(\det[R]).$$

Эта статистика имеет распределение χ^2 (хи-квадрат).

- 3) Фактическое значение χ^2 -критерия сравнить с табличным значением χ^2 при $0,5k(k - 1)$ степенях свободы и уровне значимости α . Если $FG_{\text{набл}}$ больше табличного, то в массиве объясняющих переменных существует мультиколлинеарность.

2. *Проверка наличия мультиколлинеарности каждой переменной другими переменными (F-критерий):*

- 1) Вычислить обратную матрицу $C = R^{-1}$.
- 2) Вычислить F -критерии

$$F_j = (c_{jj} - 1) \frac{n - k - 1}{k},$$

где c_{ij} – диагональные элементы матрицы C .

- 3) Фактические значения F -критериев сравнить с табличным значением при $v_1 = k$, $v_2 = n - k - 1$ степенях свободы и уровне значимости α , где k – количество факторов. Если $F_j > F_{\text{табл}}$, то соответствующая j -я независимая переменная мультиколлинеарна с другими.

3. Проверка наличия мультиколлинеарности каждой пары переменных (t -тест).

- 1) Вычислить коэффициент детерминации для каждой переменной:

$$[R(x_{jj})]^2 = 1 - \frac{1}{c_{jj}}.$$

- 2) Найти частные коэффициенты корреляции:

$$r_{ij(i)} = \frac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}},$$

где c_{ij} — элемент матрицы C , содержащийся в i -й строке и j -м столбце; c_{ii} и c_{jj} – диагональные элементы матрицы C .

- 3) Вычислить t -критерии:

$$t_{ij} = \frac{r_{ij(i)} \sqrt{n - k - 1}}{\sqrt{1 - r_{ij(i)}^2}}.$$

- 4) Фактические значения критериев t_{ij} сравнить с табличным $t_{\text{табл}}$ при $(n - k - 1)$ степенях свободы и уровне значимости α . Если $|t_{ij}| > t_{\text{табл}}$, то между независимыми переменными i и j существует мультиколлинеарность.

Разработаны различные методы устранения или уменьшения мультиколлинеарности. Самый простой из них, но не всегда самый эффективный, состоит в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0,8), одну переменную исключают из рассмотрения. При этом какую переменную оставить, а какую удалить из анализа, решают исходя из экономических соображений.

Для устранения мультиколлинеарности можно также:

- добавить в модель важный фактор для уменьшения дисперсии случайного члена;

- изменить или увеличить выборку;

- преобразовать мульти коллинеарные переменные и др.

Другой метод устранения или уменьшения мультиколлинеарности – использование стратегии шагового отбора, реализованной в ряде алгоритмов пошаговой регрессии.

Наиболее широкое применение получили следующие схемы построения уравнения множественной регрессии:

- *метод включения* – дополнительное введение фактора;

- *метод исключения* – отсеивание факторов из полного его набора.

В соответствии с *первой схемой* признак включается в уравнение в том случае, если его включение существенно увеличивает значение множественного коэффициента корреляции. Это позволяет последовательно отбирать факторы, оказывающие существенное влияние на результативный признак даже в условиях мультиколлинеарности системы признаков, отобранных в качестве аргументов. При этом первым в уравнение включается фактор, наиболее тесно коррелирующий с Y вторым – тот фактор, который в паре с первым из отобранных дает максимальное значение множественного коэффициента корреляции, и т.д. Существенно, что на каждом шаге получают новое значение множественного коэффициента (большее, чем на предыдущем шаге); тем самым определяется вклад каждого отобранного фактора в объясненную дисперсию Y .

Вторая схема пошаговой регрессии основана на последовательном исключении факторов с помощью t -критерия. Она заключается в том, что после построения уравнения регрессии и оценки значимости всех коэффициентов регрессии из модели исключают тот фактор, коэффициент при котором незначим и имеет наименьшее по модулю значение t -критерия. После этого получают новое уравнение множественной регрессии и снова производят оценку значимости всех оставшихся коэффициентов регрессии. Если и среди них окажутся незначимые, то опять исключают фактор с наименьшим значением t -критерия. Процесс исключения факторов останавливается на том шаге, при котором все регрессионные коэффициенты значимы.

Ни одна из этих процедур не гарантирует получения оптимального набора переменных. Однако при практическом применении они позволяют получить достаточно хорошие наборы существенно влияющих факторов.

При отборе факторов также рекомендуется пользоваться следующим правилом: число включаемых факторов обычно в 6-7 раз меньше объема совокупности, по которой строится регрессия.

Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F -критерий меньше табличного значения.

4.2. ОЦЕНКА КАЧЕСТВА МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Качество модели регрессии проверяется на основе анализа остатков регрессии ε . Анализ остатков позволяет получить представление, насколько хорошо подобрана сама модель и насколько правильно выбран метод оценки коэффициентов. Согласно общим предположениям регрессионного анализа остатки должны вести себя как независимые (в действительности – почти независимые) одинаково распределенные случайные величины.

Исследование полезно начинать с изучения графика остатков. Он может показать наличие какой-то зависимости, не учтенной в модели. Скажем, при подборе простой линейной зависимости между Y и X график остатков может показать необходимость перехода к нелинейной модели (квадратичной, полиномиальной, экспоненциальной) или включения в модель периодических компонент.

График остатков хорошо показывает и резко отклоняющиеся от модели наблюдения – **выбросы**. Подобным аномальным наблюдениям надо уделять особо пристальное внимание, так как они могут грубо исказить значения оценок. Чтобы устранить эффект выбросов, надо либо удалить эти точки из анализируемых данных (эта процедура называется цензурированием), либо применять методы оценивания параметров, устойчивые к подобным грубым отклонениям.

Качество модели регрессии оценивается по следующим направлениям:

- проверка качества уравнения регрессии;
- проверка значимости уравнения регрессии;
- анализ статистической значимости параметров модели;
- проверка выполнения предпосылок МНК.

Проверка качества уравнения регрессии

Для проверки качества уравнения регрессии вычисляют коэффициент множественной корреляции (индекс корреляции) R и коэффициент детерминации R^2 . Чем ближе к единице значения этих характеристик, тем выше качество модели.

В многофакторной регрессии добавление дополнительных объясняющих переменных увеличивает коэффициент детерминации. Следовательно, коэффициент детерминации должен быть скорректирован с учетом числа независимых переменных. Скорректированный R^2 , или \bar{R}^2 рассчитывается так:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}, \quad (4.4)$$

где n – число наблюдений; k – число независимых переменных.

Проверка значимости уравнения регрессии

Для проверки значимости уравнения регрессии используется **F-критерий Фишера**, вычисляемый по формуле

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (4.5)$$

Если расчетное значение с $\nu_1 = k$ и $\nu_2 = n - k - 1$ степенями свободы, где k – количество факторов, включенных в модель, больше табличного при заданном уровне значимости α , то модель считается значимой.

Анализ статистической значимости параметров модели

Анализ статистической значимости параметров модели (коэффициентов регрессии) проводится с использованием t -статистики путем проверки гипотезы о равенстве нулю j -го параметра уравнения (кроме свободного члена):

$$t_{a_j} = \hat{a}_j / \sigma_{a_j}, \quad (4.6)$$

где σ_{a_j} – это стандартное (среднеквадратическое) отклонение коэффициента уравнения регрессии a_j

Величина σ_{a_j} представляет собой квадратный корень из произведения несмещенной оценки дисперсии σ_e^2 и j -го диагонального элемента матрицы, обратной матрице системы нормальных уравнений:

$$\sigma_{a_j} = \sigma_e \sqrt{b_{jj}}, \quad (4.7)$$

где b_{ij} – диагональный элемент матрицы $(X'X)^{-1}$.

Если расчетное значение t -критерия с $(n - k - 1)$ степенями свободы больше его табличного значения при заданном уровне значимости α , коэффициент регрессии считается значимым. В противном случае фактор, соответствующий этому коэффициенту, следует исключить из модели (при этом ее качество не ухудшится).

Проверка выполнения предпосылок МНК

Условия, необходимые для получения несмещенных, состоятельных и эффективных оценок, представляют собой предпосылки МНК (см. вопрос 3.1). Выполнение этих предпосылок проверяется на основе анализа остатков e_i . Выполнение условия равенства нулю математического ожидания остатков [см. формулу (3.2)] обеспечивается всегда при использовании МНК для линейных моделей. Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии с помощью критериев t и F . Вместе с тем оценки регрессии, полученные методом наименьших квадратов, обладают хорошими свойствами даже при отсутствии нормального распределения остатков. Таким образом, самым важным для получения по МНК состоятельных параметров регрессии является

соблюдение третьей и четвертой предпосылок (условие независимости и условие гомоскедастичности).

1. Проверка условия независимости случайных составляющих в различных наблюдениях. Зависимость текущих значений случайного члена от их непосредственно предшествующих значений называется автокорреляцией. Автокорреляция случайной составляющей нарушает третью предпосылку нормальной линейной модели регрессии (см. вопрос 3.1).

В эконометрических исследованиях часто возникают и такие ситуации, когда дисперсия остатков постоянная, но наблюдается их ковариация. Это явление называют автокорреляцией остатков. Чаше всего она наблюдается тогда, когда эконометрическая модель строится на основе временных рядов. Если существует корреляция между последовательными значениями некоторой независимой переменной, то будет наблюдаться и корреляция последовательных значений остатков.

Автокорреляция может быть также следствием ошибочной спецификации эконометрической модели. Кроме того, наличие автокорреляции остатков может означать, что необходимо ввести в модель новую независимую переменную.

Существуют различные способы устранения автокорреляции, например:

- введение в модель фактора времени;
- переход к темповым или относительным показателям;
- включение в модель неучтенных факторов;
- построение авторегрессионных уравнений.

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

1. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Динамический ряд – совокупность наблюдений некоторого явления (показателя), упорядоченная в зависимости от последовательности значений другого явления (признака).

Временной ряд – динамические ряды, признаком упорядочения которых является время.

В экономике и бизнесе временные ряды – это очень распространенный тип данных. Во временном ряде содержится информация об особенностях и закономерностях протекания процесса, а статистический анализ позволяет выявить закономерности и использовать их для оценки характеристик процесса в будущем, т.е. для прогнозирования.

Временной ряд – набор чисел, привязанный к последовательным, обычно равноотстоящим моментам времени.

Уровни (элементы) временного ряда – числа, составляющие временной ряд и полученные в результате наблюдения за ходом некоторого процесса.

Длина временного ряда – количество уровней n , входящих во временной ряд.

Временной ряд обычно обозначают $Y(t)$ или y_t , где $t = \overline{1, n}$. В общем случае каждый уровень временного ряда можно представить как функцию четырех компонент $f(t), S(t), U(t), \varepsilon(t)$, отражающих закономерность и случайность развития, где $f(t)$ – тренд (долговременная тенденция) развития; $S(t)$ – сезонная компонента; $U(t)$ – циклическая компонента; $\varepsilon(t)$ – остаточная компонента.

В модели временного ряда принято выделять две основные составляющие: *детерминированную* (систематическую) и *случайную*.

Детерминированная составляющая временного ряда y_1, y_2, \dots, y_n – числовая последовательность, элементы которой вычисляются по определенному правилу как функция времени t .

Детерминированная составляющая может содержать следующие структурные компоненты (рис. 1):

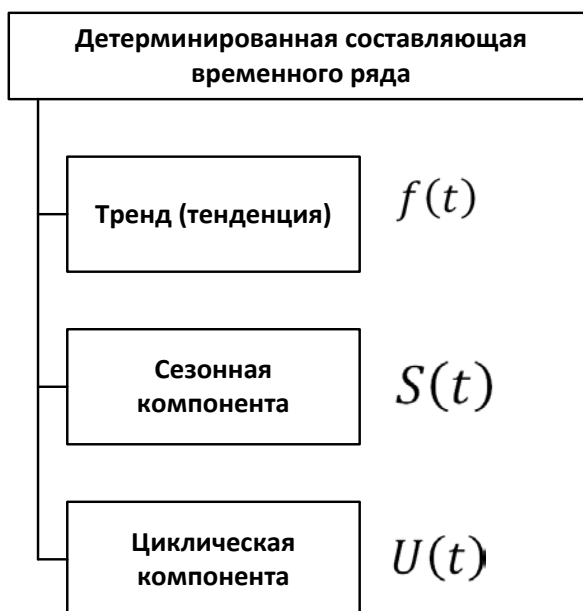


Рис. 1. Детерминированная составляющая временного ряда

1. **Тренд** или **тенденция** $f(t)$, представляет собой устойчивую закономерность, наблюдаемую в течение длительного периода времени.

Обычно тренд (тенденция) описывается с помощью той или иной неслучайной функции $f(t)$ (аргументом которой является время), как правило, монотонной. Эту функцию называют функцией тренда или просто – трендом.

2. **Сезонная компонента** $S(t)$ связана с наличием факторов, действующих с заранее известной периодичностью. Это регулярные колебания, которые носят периодический или близкий к нему характер и заканчиваются в течение года.

Типичные примеры сезонного эффекта: изменение загруженности автотрассы по временам года, пик продаж товаров для школьников в конце августа – начале сентября.

Сезонная компонента со временем может меняться либо иметь плавающий характер.

3. Циклическая компонента $U(t)$ – неслучайная функция, описывающая длительные периоды (более одного года) относительного подъема и спада и состоящая из циклов переменной длительности и амплитуды.

Примером циклической (конъюнктурной) компоненты являются волны Кондратьева, демографические «ямы» и т.п. Подобная компонента весьма характерна для рядов макроэкономических показателей. Здесь циклические изменения обусловлены взаимодействием спроса и предложения, а также наложением таких факторов, как истощение ресурсов, погодные условия, изменения в налоговой политике и т.п. Отметим, что циклическую компоненту крайне трудно идентифицировать формальными методами исходя только из данных изучаемого ряда.

Случайная компонента $\varepsilon(t)$ – это составная часть временного ряда, оставшаяся после выделения систематических компонент.

Она отражает воздействие многочисленных факторов случайного характера. Случайная компонента является обязательной составной частью любого временного ряда в экономике, так как случайные отклонения неизбежно сопутствуют любому экономическому явлению.

Если систематические компоненты временного ряда определены правильно, то остающаяся после выделения из временного ряда этих компонент так называемая остаточная последовательность (ряд остатков) будет случайной компонентой ряда.

В анализе случайной компоненты экономических временных рядов важную роль играет сравнение случайной величины ε_t , с хорошо изученной формой случайных процессов – стационарными случайными процессами.

Стационарным процессом в узком смысле называется такой случайный процесс, вероятностные свойства которого с течением времени не изменяются.

Он протекает в приблизительно однородных условиях и имеет вид непрерывных случайных колебаний вокруг некоторого среднего значения. Причем ни средняя амплитуда, ни его частота не обнаруживают с течением времени существенных изменений.

Однако на практике чаще встречаются процессы, вероятностные характеристики которых подчиняются определенным закономерностям и не являются постоянными величинами. Поэтому в прикладном эконометрическом анализе используется понятие слабой стационарности (или стационарности в широком смысле), которое предполагает неизменность во времени среднего значения, дисперсии и ковариации временного ряда.

Случайный процесс называется **стационарным в широком смысле**, если его математическое ожидание постоянно и автокорреляционная функция $r(\tau)$ зависит только от длины временного интервала τ .

В зависимости от вида связи (рис. 2) между компонентами может быть построена либо *аддитивная модель* временного ряда

$$Y(t) = f(t) + S(t) + U(t) + \varepsilon(t), \quad (1)$$

либо *мультипликативная модель*

$$Y(t) = f(t)S(t)U(t) + \varepsilon(t). \quad (2)$$

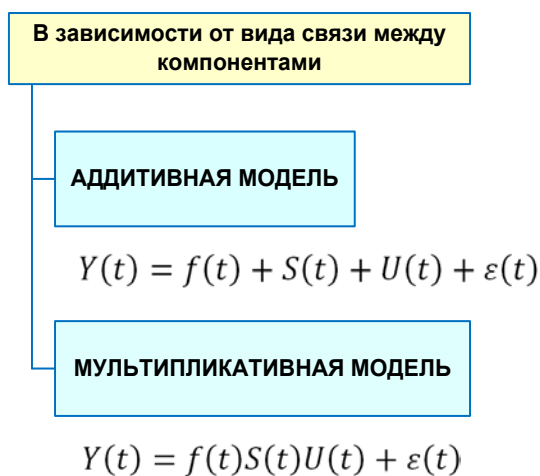


Рис. 2. Классификация моделей в зависимости от вида связи между компонентами

В процессе формирования значений временных рядов не всегда участвуют все четыре компонента. Однако во всех случаях предполагается наличие случайной составляющей.

Основная цель статистического анализа временных рядов – изучить соотношение между закономерностью и случайностью в формировании значений уровней ряда и оценить количественную меру их влияния.

Закономерности, объясняющие динамику показателя в прошлом, используются для прогнозирования его значений в будущем, а учет случайности позволяет определить вероятность отклонения от закономерного развития и его возможную величину.

Применяемые при обработке временных рядов методы во многом опираются на методы математической статистики, которые базируются на достаточно жестких **требованиях к исходным данным**:

1) **сопоставимость данных** – достигается в результате одинакового подхода к наблюдениям на разных этапах формирования динамического ряда. Уровни во временных рядах должны иметь одинаковые единицы измерения, шаг наблюдений, интервал времени, методику расчета и элементы, относящиеся к неизменной совокупности;

2) **однородность данных** – означает отсутствие сильных изломов тенденций, а также аномальных (т.е. резко выделяющихся, нетипичных для данного ряда) наблюдений. Аномальные наблюдения проявляются в виде сильного изменения уровня – скачка или спада – с последующим приблизительным восстановлением предыдущего уровня. Наличие аномалии резко искажает результаты моделирования, поэтому аномальные наблюдения необходимо исключить из временного ряда, заменив их расчетными значениями;

3) **устойчивость тенденции** – характеризуется преобладанием закономерности над случайностью в изменении уровней ряда. На графиках устойчивых временных рядов закономерность прослеживается визуально, на графиках неустойчивых рядов изменения последовательных уровней

представляются хаотичными, и поэтому поиск закономерностей в формировании значений уровней таких рядов лишен смысла:

4) **полнота данных** – требование обусловлено тем, что закономерность может обнаружиться лишь при наличии минимально допустимого объема наблюдений.

Следует иметь в виду, что при исследовании временных рядов экономических данных зачастую невозможно в должной мере проверить выполнимость перечисленных требований. Поэтому выводы, полученные на базе формально-статистического инструментария, должны восприниматься с осторожностью и дополняться содержательным анализом.

2. ЭТАПЫ ПОСТРОЕНИЯ ПРОГНОЗА ПО ВРЕМЕННЫМ РЯДАМ

Экстраполяционное¹ прогнозирование экономических процессов, представленных одномерными временными рядами, сводится к выполнению следующих **основных этапов**:

- 1) предварительный анализ данных;
- 2) построение моделей временных рядов: формирование набора аппроксимирующих функций (кривых роста) и численное оценивание параметров моделей;
- 3) оценка качества моделей (проверка их адекватности и оценка точности);
- 4) построение точечного и интервального прогнозов.

Предварительный анализ данных

В ходе предварительного анализа определяют, соответствуют ли имеющиеся данные требованиям, предъявляемым к ним математическими методами (сопоставимость данных, их полнота, однородность и устойчивость); строят график динамики и рассчитывают основные

¹ *Экстраполяция* – это распространение выявленных при анализе рядов динамики закономерностей развития изучаемого объекта на будущее (при предположении, что выявленная закономерность, выступающая в качестве базы прогнозирования, сохраняется и в дальнейшем).

динамические характеристики (приросты, темпы роста, темпы прироста, коэффициенты автокорреляции).

Для получения общего представления о динамике исследуемого показателя во времени целесообразно построить его график: по оси абсцисс откладываются значения переменной t , а по оси ординат – соответствующие значения показателя $Y(t)$.

К процедурам предварительного анализа относятся:

- выявление аномальных наблюдений;
- проверка наличия тренда;
- сглаживание временных рядов;
- расчет показателей динамики экономических процессов.

1. Выявление аномальных наблюдений – обязательная процедура этапа предварительного анализа данных. Так как наличие аномальных наблюдений приводит к искажению результатов моделирования, то необходимо убедиться в отсутствии аномалий данных.

Для диагностики аномальных наблюдений разработаны различные критерии, например метод Ирвина. Для всех или только для подозреваемых в аномальности наблюдений вычисляется величина λ_t :

$$\lambda_t = \frac{|y_t - y_{t-1}|}{S_y}, \quad (3)$$

где $S_y = \sqrt{\frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y})^2}$, $\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$.

Если рассчитанная величина λ_t превышает табличное значение (см. Приложение б), то уровень y_t считается *аномальным*. Аномальные наблюдения необходимо исключить из временного ряда и заменить их расчетными значениями (самый простой способ замены – в качестве нового значения принять среднее из двух соседних значений).

Пример 1. *Выявление аномальных наблюдений.*

На основании данных об изменении индекса потребительских цен $Y(t)$ (в % к предыдущему периоду), приведенных в табл. 1, проверить наличие аномальных наблюдений.

Таблица 1

Квар-тал	Год	t	$Y(t)$	Квар-тал	Год	t	$Y(t)$
IV	1994	1	100	I	1999	18	116
I	1995	2	142,77	II		19	107,3
II		3	124,92	III		20	105,6
III		4	115,21	IV		21	103,9
IV		5	113,02	I	22	103,94	
I	1996	6	110,01	II	2000	23	105,4
II		7	105,08	III		24	104,2
III		8	100,8	IV		25	105,4
IV		9	104,57	I		26	107,1
I	1997	10	105,29	II	2001	27	105,3
II		11	103,03	III		28	101,1
III		12	100,5	IV		29	104,1
IV		13	101,81	I		30	105,5
I	1998	14	103,03	II	2002	31	103,4
II		15	101	III		32	101,2
III		16	143,81	IV		33	104,26
IV		17	123,27	I		2003	34

Источник: Краткосрочные экономические показатели Российской Федерации. М.: Госкомстат (<http://www.gks.ru/>)

Решение. Результаты расчетов по методу Ирвина приведены в табл. 2.

Таблица 2

t	$Y(t)$	$\lambda(t)$	t	$Y(t)$	$\lambda(t)$
1	100	—	18	116	0,685
2	142,77	4,028	19	107,3	0,819
3	124,92	1,681	20	105,6	0,160
4	115,21	0,915	21	103,9	0,160
5	113,02	0,206	22	103,94	0,004
6	110,01	0,284	23	105,4	0,138
7	105,08	0,464	24	104,2	0,113
8	100,8	0,403	25	105,4	0,113
9	104,57	0,355	26	107,1	0,160
10	105,29	0,068	27	105,3	0,170
11	103,03	0,213	28	101,1	0,396
12	100,5	0,238	29	104,1	0,283
13	101,81	0,123	30	105,5	0,132
14	103,03	0,115	31	103,4	0,198
15	101	0,191	32	101,2	0,207
16	143,81	4,032	33	104,26	0,288
17	123,27	1,935	34	105,2	0,089

В нашем примере

$$S_y = \sqrt{\frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y})^2} = 10,62, \quad \bar{y} = \frac{1}{n} \sum_{t=1}^n y_t = 108,44.$$

Аномальными являются наблюдения 2, 3, 16 и 17. На рис. 1 наблюдениям 2 и 16 соответствуют резкие выбросы.

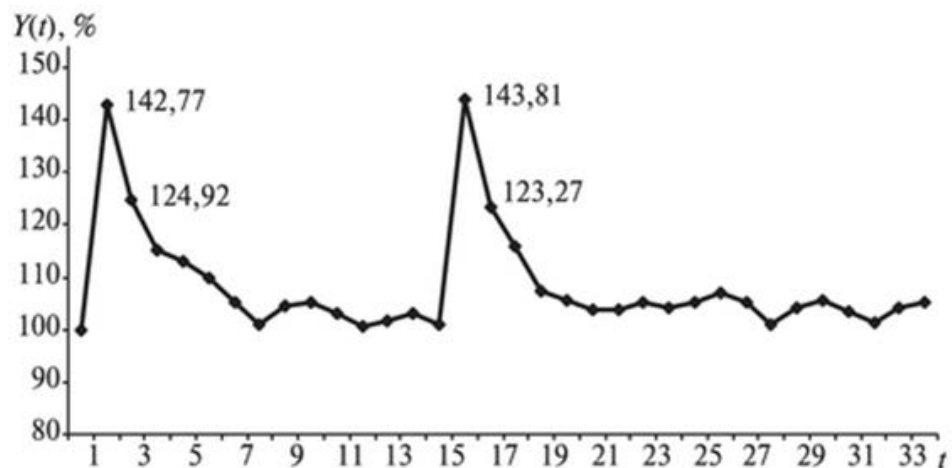


Рис. 1. График динамики временного ряда «Индекс потребительских цен»

2. Проверка наличия тренда – следующая процедура предварительного анализа данных. Отметим, что тенденция (тренд) в развитии исследуемого показателя прослеживается не только в увеличении

или уменьшении среднего текущего значения временного ряда, она присуща и другим его характеристикам: дисперсии, автокорреляции, корреляции с другими показателями и т.д.

Тенденцию среднего визуально можно определить из графика исходных данных.

Проверка наличия или отсутствия неслучайной (и зависящей от времени t) составляющей сводится к проверке гипотезы о неизменности среднего значения временного ряда. Процедура проверки может быть осуществлена с помощью различных критериев.

Критерий серий, основанный на медиане. Расположим члены анализируемого временного ряда в порядке возрастания, т.е. образуем ряд

$$y_{(1)}, y_{(2)}, \dots, y_{(n)}.$$

Определим выборочную медиану по формуле

$$y_{med} = \begin{cases} y_{\left(\frac{n+1}{2}\right)}, & \text{если } n \text{ нечетно,} \\ \frac{1}{2} \left(y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)} \right), & \text{если } n \text{ четно.} \end{cases} \quad (4)$$

После этого по исходному временному ряду образуем «серии» из плюсов и минусов, на статистическом анализе которых основана процедура проверки гипотезы о неизменности среднего значения временного ряда.

Схема построения последовательности из плюсов и минусов такова: вместо y_t , ставится знак «плюс», если $y_t > y_{med}$ и «минус», если $y_t < y_{med}$ (члены временного ряда, равные y_{med} , в полученной таким образом последовательности плюсов и минусов не учитываются).

Образованная последовательность плюсов и минусов характеризуется общим числом серии $v(n)$ и протяженностью самой длинной серии K_{max} . При этом под «серией» понимается последовательность подряд идущих плюсов и подряд идущих минусов. Если исследуемый ряд состоит из статистически независимых наблюдений, случайно варьирующих около некоторого постоянного уровня (т.е. справедлива гипотеза о неизменности среднего значения временного ряда), то чередование плюсов и минусов в

построенной последовательности должно быть случайным, т.е. эта последовательность не должна содержать слишком длинных серий подряд идущих плюсов и минусов, и, соответственно, общее число серий не должно быть слишком малым. Так что в данной критерии целесообразно рассматривать одновременно пару критических статистик $(v(n); K_{max})$.

Справедлив следующий приближенный статистический **критерий проверки гипотезы о неизменности среднего значения временного ряда:** если хотя бы одно из неравенств

$$\boxed{\begin{cases} v(n) > \left[\frac{1}{2}(n + 2 - 1,96\sqrt{n-1}) \right], \\ K_{max} < [3,3(\lg n + 1)] \end{cases}} \quad (5)$$

окажется нарушенным, то гипотеза о неизменности среднего значения временного ряда отвергается с вероятностью ошибки α , такой, что $0,05 < \alpha < 0,0975$, и тем самым подтверждается наличие зависящей от времени неслучайной составляющей в разложении $Y(t) = f(t) + S(t) + U(t) + \varepsilon(t)$.

Квадратные скобки в неравенствах (5) означают целую часть от числа.

Критерий «восходящих» и «нисходящих» серий. Этот критерий «улавливает» постепенное смещение среднего значения в исследуемом распределении не только монотонного, но и более общего, например периодического, характера.

Как и в предыдущем критерии, исследуется последовательность знаков – плюсов и минусов, однако правило образования этой последовательности в данном критерии иное. Здесь на t -м месте вспомогательной последовательности ставится знак «плюс», если $y_{t+1} - y_t > 0$, и «минус», если $y_{t+1} - y_t < 0$ (если два или несколько следующих друг за другом наблюдений равны между собой, то принимается во внимание только одно из них).

Последовательность подряд идущих плюсов («восходящая» серия) будет соответствовать возрастанию результатов наблюдения, а последовательность минусов («нисходящая» серия) – их убыванию.

Критерий основан на том же соображении, что и предыдущий: если выборка случайна, то в образованной последовательности знаков общее число серий не может быть слишком малым, а их протяженность – слишком большой.

При уровне значимости $0,05 < \alpha < 0,0975$ критерий имеет вид

$$\left\{ \begin{array}{l} v(n) > \left[\frac{2n-1}{3} - 1,96\sqrt{\frac{16n-29}{90}} \right], \\ K_{\max} < [K_0(n)], \end{array} \right. \quad (6)$$

где величина $K_0(n)$ определяется следующим образом:

$$K_0(n) = 5 \quad \text{при } n \leq 26;$$

$$K_0(n) = 6 \quad \text{при } 26 < n \leq 153;$$

$$K_0(n) = 7 \quad \text{при } 153 < n \leq 1170.$$

Сравнение средних уровней ряда. Для проверки обнаружения тренда временной ряд разбивают на две примерно равные по числу уровней части, каждая из которых рассматривается как самостоятельная выборочная совокупность, имеющая нормальное распределение. Если временной ряд имеет тенденцию к тренду, то средние, вычисленные для каждой совокупности, должны существенно (значимо) различаться между собой. Если же расхождение несущественно (случайно), то временной ряд не имеет тенденции.

Таким образом, проверка наличия тренда в исследуемом ряду сводится к **проверке гипотезы о равенстве средних двух нормально распределенных совокупностей.**

Рассмотрим применение этого метода на следующем примере.

Пример 2. Проверка наличия тренда.

Определить наличие основной тенденции (тренда) по данным табл. 3, где приведена урожайность ячменя в одной из областей Среднего Поволжья.

Таблица 3

Год	Урожайность ячменя, ц/га	Год	Урожайность ячменя, ц/га
1	14,1	9	14,7
2	9,3	10	16,6
3	19,4	11	5,6
4	19,7	12	16,2
5	5,4	13	25,3
6	24,2	14	11,9
7	13,8	15	18,5
8	24,5		

Решение. На рис. 2 приведен график динамики временного ряда «Урожайность ячменя».

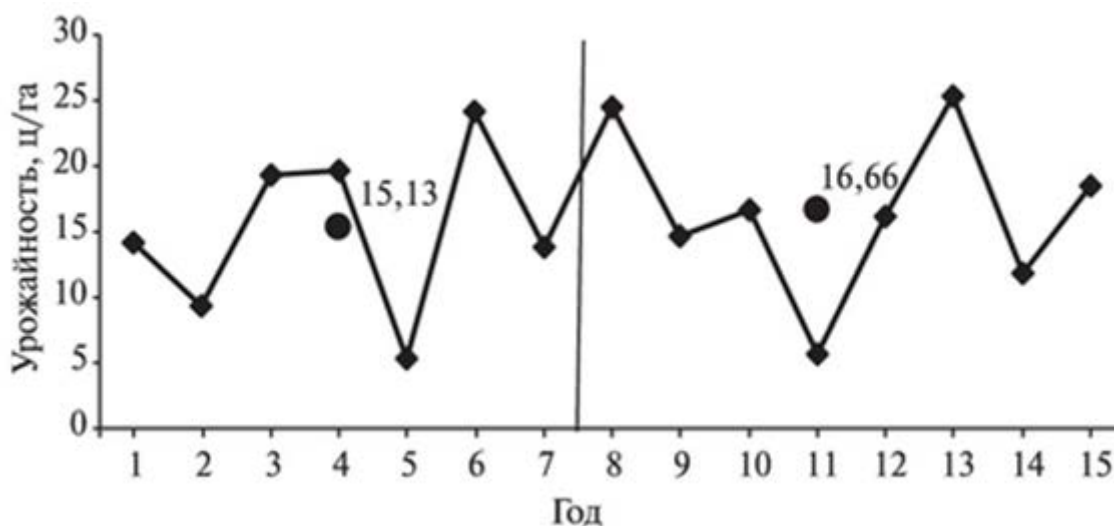


Рис. 2. График урожайности ячменя

1. Делим исходный временной ряд на две примерно равные по числу уровней части:

$$n_1 = 7, n_2 = 8 (n_1 + n_2 = n = 15).$$

2. Для каждой из этих частей вычисляем средние значения и дисперсии:

$$\bar{y}_1 = \frac{1}{n_1} \sum_{t=1}^{n_1} y_t = 15,13; \quad \bar{y}_2 = \frac{1}{n_2} \sum_{t=n_1+1}^n y_t = 16,66;$$

$$S_{y_1}^2 = \frac{1}{n_1 - 1} \sum_{t=1}^{n_1} (y_t - \bar{y}_1)^2 = 42,15; \quad S_{y_2}^2 = \frac{1}{n_2 - 1} \sum_{t=n_1+1}^n (y_t - \bar{y}_2)^2 = 41,22.$$

3. Проверяем гипотезу о равенстве (однородности) дисперсий обеих частей ряда с помощью F -критерия Фишера. Для вычисления F -критерия большую дисперсию делим на меньшую:

$$F_{\text{расч}} = S_{y_1}^2 / S_{y_2}^2 = 42,15/41,22 = 1,022,$$
$$F_{\text{табл}(0,05; 6, 7)} = 3,87.$$

Так как $F_{\text{расч}} < F_{\text{табл}(0,05;6,7)}$, то с вероятностью 95% нет оснований отвергать нулевую гипотезу, выборочные дисперсии ($S_{y_1}^2$ и $S_{y_2}^2$) различаются незначимо (расхождение между ними есть величина случайная).

4. Проверяем основную гипотезу о равенстве средних значений с использованием t -критерия Стьюдента:

$$t = \frac{|\bar{y}_1 - \bar{y}_2|}{\sqrt{(n_1 - 1)S_{y_1}^2 + (n_2 - 1)S_{y_2}^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \quad (7)$$

Подставляя числовые значения, получим

$$t = \frac{|15,13 - 16,66|}{\sqrt{6 \cdot 42,15 + 7 \cdot 41,22}} \sqrt{\frac{7 \cdot 8 \cdot 13}{15}} = 0,459,$$
$$t_{\text{табл}(0,05; 13)} = 2,16.$$

(Примечание. Табличное значение t -критерия можно получить с помощью функции Excel СТЬЮДРАСПОБР)

Так как $|t_{\text{расч}}| < t_{\text{табл}}$, то нет оснований отвергать нулевую гипотезу о равенстве средних, расхождение между вычисленными средними незначимо. Отсюда **вывод**: *тренд урожайности ячменя отсутствует.*

Рассмотрим решение примера 2 в Excel.

1. Гипотезу о равенстве дисперсий проверим с помощью F -теста, который можно найти среди инструментов Анализа данных (рис. 3).

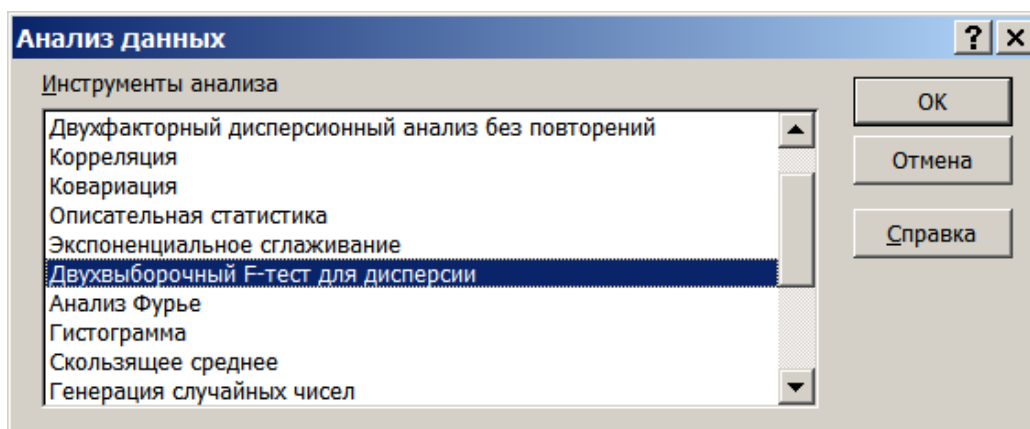


Рис. 3. Вызов надстройки Excel Анализ данных

2. Вводим данные для выполнения F -теста, указывая интервал для первой и второй переменных (рис. 4). Результат выполнения теста приведен на рис. 5. Анализируя результаты выполнения двухвыборочного F -теста для проверки гипотезы о равенстве дисперсий, приходим к выводу, что исправленные выборочные дисперсии ($S_{y_1}^2$ и $S_{y_2}^2$) и различаются незначимо.

Рис. 4. Введены данные для двухвыборочного t -теста

Рис. 5. Результат выполнения двухвыборочного F -теста для дисперсии

3. Выбираем инструмент анализа Двухвыборочный t -тест с одинаковыми дисперсиями (рис. 6). Вводим данные. Результат выполнения t -теста приведен на рис. 7, анализируя который убеждаемся, что тренда нет.

Рис. 6. Введены данные для двухвыборочного t -теста с одинаковыми дисперсиями

Рис. 7. Результат выполнения t -теста

Наличие тенденции среднего уровня на графике становится более заметным, когда на нем отражены сглаженные значения исходных данных.

3. Сглаживание временного ряда, т.е. замена фактических уровней расчетными значениями, имеющими меньшую колеблемость, чем исходные данные, является простым методом выявления тенденции развития. Соответствующее преобразование называется *фильтрацией*.

Сглаживание временных рядов проводится в следующих случаях:

- при графическом изображении временного ряда тренд прослеживается недостаточно отчетливо. Поэтому ряд сглаживают, на график наносят сглаженные значения, и как правило, тенденция проявляется более четко;

- применяются методы анализа и прогнозирования, требующие в качестве предварительного условия сглаживания временного ряда;

- при устранении аномальных наблюдений;

- при непосредственном прогнозировании экономических показателей и прогнозировании изменения тренда — «точек поворота».

Существующие **методы сглаживания** делят на две группы:

1) *аналитические методы*. Для сглаживания используется кривая, проведенная относительно фактических значений ряда так, чтобы она отображала тенденцию, присущую ряду, и одновременно освобождала его от мелких, незначительных колебаний. Такие кривые называют еще кривыми роста, применяются они главным образом для прогнозирования экономических показателей;

2) *методы механического сглаживания*. Сглаживается каждый отдельный уровень ряда с использованием фактических значений соседних с ним уровней. Для сглаживания временных рядов часто используются методы простой и взвешенной скользящей средней, экспоненциального сглаживания.

Метод простой скользящей средней включает в себя следующие этапы:

1. Определяется количество наблюдений, входящих в интервал сглаживания. При этом используют следующее правило: если необходимо сгладить мелкие, беспорядочные колебания, то интервал сглаживания берут по возможности большим, и наоборот, интервал сглаживания уменьшают, когда нужно сохранить более мелкие волны и освободиться от периодически повторяющихся колебаний, возникающих, например, из-за автокорреляций уровней.

2. Вычисляется среднее значение наблюдений, образующих интервал сглаживания, которое одновременно является сглаживающим значением уровня, находящегося в центре интервала сглаживания, при условии, что m – нечетное число, по формуле

$$\tilde{y}_t = \frac{1}{m} \sum_{i=t-p}^{t+p} y_i, \quad (8)$$

где m – количество наблюдений, входящих в интервал сглаживания; p – количество наблюдений, стоящих по разные стороны от сглаживаемого.

При нечетном m значение параметра p вычисляют следующим образом:

$$p = \frac{m - 1}{2}.$$

Первым сглаженным будет наблюдение \tilde{y}_t , где $t = p + 1$.

3. Интервал сглаживания сдвигается на один член вправо, и по формуле (8) находится сглаженное значение для $(t + 1)$ -го наблюдения. Затем снова производят сдвиг и т.д.

Процедура продолжается до тех пор, пока в интервал сглаживания не войдет последнее наблюдение временного ряда.

Недостаток метода: первые и последние p наблюдений ряда остаются несглаженными.

Метод простой скользящей средней можно использовать, если графическое изображение ряда напоминает прямую линию. В этом случае не искажается динамика развития исследуемого процесса. Однако, когда тренд выравниваемого ряда имеет изгибы и к тому же желательно сохранить мелкие волны, использовать для сглаживания ряда метод простой скользящей средней нецелесообразно, поскольку при этом:

- выравняются и выпуклые, и вогнутые линии;
- происходит сдвиг волны вдоль ряда;
- изменяется знак волны, т.е. на кривой, соединяющей сглаженные точки, вместо выпуклого участка образуется вогнутый и наоборот.

Последнее имеет место в случаях, когда интервал сглаживания в полтора раза превышает длину волны.

Метод взвешенной скользящей средней отличается от предыдущего тем, что сглаживание внутри интервала производится не по прямой, а по кривой более высокого порядка. Это обусловлено тем, что суммирование членов ряда, входящих в интервал сглаживания, производится с определенными весами, рассчитанными по методу наименьших квадратов.

Если сглаживание производится с помощью полинома (многочлена) второго или третьего порядка, то веса берутся следующие:

$$\frac{1}{35}(-3; 12; 17; 12; -3) \quad \text{для } m = 5;$$

$$\frac{1}{21}(-2; 3; 6; 7; 3; -2) \quad \text{для } m = 7.$$

Особенности весов:

- 1) симметричны относительно центрального члена;
- 2) сумма весов с учетом общего множителя равна единице.

Недостаток метода: первые и последние p наблюдений ряда остаются несглаженными.

Рассмотренные методы простой и взвешенной скользящей средней не дают возможности сгладить первые и последние p наблюдений временного ряда. Отсутствие сглаженных первых наблюдений не так важно по сравнению с последними наблюдениями, особенно если целью исследования является прогнозирование развития процесса. Есть методы, позволяющие получить сглаженные значения последних уровней так же, как и всех остальных. К их числу относится метод экспоненциального сглаживания.

Особенность метода экспоненциального сглаживания заключена в том, что в процедуре выравнивания каждого наблюдения используются только значения предыдущих уровней, взятых с определенным весом. Относительный вес каждого наблюдения уменьшается по экспоненте по мере его удаления от момента, для которого определяется сглаженное значение.

Сглаженное значение наблюдения ряда S_t на момент времени t определяется по формуле

$$S_t(y) = \alpha y_t + (1 - \alpha)S_{t-1}(y), \quad (9)$$

где α – сглаживающий параметр, характеризующий вес выравниваемого наблюдения, причем $0 < \alpha < 1$.

Величину S_{t-1} в формуле (9) можно представить в виде суммы фактического значения уровня y_{t-1} и сглаженного значения предшествующего ему наблюдения S_{t-2} , взятых с соответствующими весами.

Процесс такого разложения можно продолжить для членов S_{t-2} , S_{t-3} и т.д.

Используя рекуррентное соотношение (9) для всех уровней ряда, начиная с первого и кончая моментом времени t , можно получить, что экспоненциальная средняя, т.е. сглаженное данным методом значение уровня ряда, является взвешенной средней всех предшествующих уровней:

$$S_t(y) = \alpha \sum_{k=0}^{t-k} (1 - \alpha)^k y_{t-k} + (1 - \alpha)^k S_0(y), \quad (10)$$

где $0 < k < t - 1$ – число периодов отставания от момента t ; $S_0(y)$ – величина, характеризующая начальные условия.

При использовании метода экспоненциального сглаживания возникают следующие затруднения:

- выбор сглаживающего параметра α ;
- определение начальных условий $S_0(y)$.

От численного значения параметра α зависит, насколько быстро будет уменьшаться вес предшествующих наблюдений и в соответствии с этим степень их влияния на сглаживаемый уровень. Чем больше значение параметра α , тем меньше сказывается влияние предшествующих уровней и, соответственно, меньшим оказывается сглаживающее воздействие экспоненциальной средней.

Задачу выбора параметра $S_0(y)$, определяющего начальные условия, предлагается решать следующим образом: если есть данные о развитии процесса в прошлом, то их среднее значение можно принять в качестве $S_0(y)$; если таких сведений нет, то в качестве $S_0(y)$ используют исходное (первое) значение y_1 , наблюдения временного ряда.

4. Расчет показателей динамики экономических процессов – заключительный этап предварительного анализа данных. Традиционными показателями, характеризующими развитие экономических процессов, были и остаются показатели роста и прироста, формулы расчета которых приведены в табл. 4. Эти показатели используются для характеристики динамики изменения уровней временного ряда.

Таблица 4

Основные показатели динамики

	<i>Абсолютный прирост</i>	<i>Темп роста</i>	<i>Темп прироста</i>
Цепной	$\Delta y_t^u = y_t - y_{t-1}$	$T_t^u = \frac{y_t}{y_{t-1}} \cdot 100\%$	$T_{прt}^u = T_t^u - 100\%$
Базисный	$\Delta y_t^b = y_t - y_1^b$	$T_t^b = \frac{y_t}{y_1^b} \cdot 100\%$	$T_{прt}^b = T_t^b - 100\%$
Средний	$\overline{\Delta y} = \frac{y_n - y_1}{n - 1}$	$\overline{T} = \sqrt[n-1]{\frac{y_n}{y_1}} \cdot 100\%$	$\overline{T}_{пр} = \overline{T} - 100\%$

Показатель *среднего абсолютного прироста* $\overline{\Delta y}$ используется для построения простейших, так называемых наивных, прогнозов.

Прогноз на k шагов вперед на момент времени $t = n + 1, n + 2, \dots, n + k$ получается по формулам

$$\begin{aligned}
 y_{n+1} &= y_n + \overline{\Delta y}, \\
 y_{n+2} &= y_n + 2\overline{\Delta y}, \\
 &\dots\dots\dots \\
 y_{n+k} &= y_n + k\overline{\Delta y}.
 \end{aligned}$$

Этот способ очень привлекателен из-за своей простоты, однако он имеет несколько существенных недостатков:

1) все фактические наблюдения являются результатом закономерности и случайности. Следовательно, «отталкиваться» от последнего наблюдения неправомерно;

2) нет возможности оценить правомерность использования среднего прироста в каждом конкретном случае;

3) невозможно сформировать интервал, внутрь которого попадет прогнозируемая величина, и указать степень уверенности в этом.

В связи с этим данный подход используется лишь как первый ориентир будущего развития или же в условиях очень малого объема наблюдений при невозможности использования описываемых ниже статистических методов.

Кроме того, для характеристики динамики изменения экономических показателей часто используется понятие автокорреляции, которая характеризует не только взаимозависимость уровней одного и того же ряда, относящихся к разным моментам наблюдений, но и степень устойчивости развития процесса во времени, величину оптимального периода прогнозирования и т.п.

Степень тесноты статистической связи между уровнями временного ряда, сдвинутыми на τ единиц времени, определяется величиной *коэффициента корреляции* $r(\tau)$. Так как $r(\tau)$ измеряет тесноту связи между уровнями одного и того же временного ряда, его принято называть коэффициентом автокорреляции. При этом τ – длину временного смещения – называют обычно лагом.

Коэффициент автокорреляции вычисляют по формуле

$$r(\tau) = \frac{(n - \tau) \sum_{t=1}^{n-\tau} y_t y_{t+\tau} - \sum_{t=1}^{n-\tau} y_t \sum_{t=1}^{n-\tau} y_{t+\tau}}{\sqrt{\left[(n - \tau) \sum_{t=1}^{n-\tau} y_t^2 - \left(\sum_{t=1}^{n-\tau} y_t \right)^2 \right] \left[(n - \tau) \sum_{t=1}^{n-\tau} y_{t+\tau}^2 - \left(\sum_{t=1}^{n-\tau} y_{t+\tau} \right)^2 \right]}}. \quad (11)$$

При большой протяженности исследуемого ряда расчет коэффициентов автокорреляции можно упростить. Для этого находят отклонения не от средних коррелируемых рядов, а от общей средней всего ряда. В этом случае

$$r(\tau) \approx \frac{\sum_{t=1}^{n-\tau} (y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}. \quad (11)$$

Порядок коэффициентов автокорреляции определяется временным лагом: первого порядка (при $\tau = 1$), второго порядка (при $\tau = 2$) и т.д.

Последовательность коэффициентов автокорреляции уровней первого, второго и последующих порядков называют автокорреляционной функцией.

Значения автокорреляционной функции могут колебаться от -1 до $+1$, но из стационарности следует, что $r(\tau) = -r(\tau)$. График автокорреляционной функции называется **коррелограммой**.

Для расчета коэффициента автокорреляции по формуле (11) в Excel можно воспользоваться функцией КОРРЕЛ. Предположим, что базовая переменная включает диапазон A1:A34. Тогда коэффициент автокорреляции равен

$$= \text{КОРРЕЛ}(A1:A33;A2:A34).$$

На практике, как правило, при вычислении автокорреляции используется формула (12).

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, т.е. с помощью анализа автокорреляционной функции и коррелограммы можно выявить структуру 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 колебания с периодичностью в τ моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, то можно

сделать одно из двух предположений относительно структуры этого ряда: либо ряд не содержит тенденции и сезонных колебаний, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ. Поэтому коэффициент автокорреляции уровней и автокорреляционную функцию целесообразно использовать для выявления во временном ряде наличия или отсутствия трендовой компоненты $f(t)$ и сезонной компоненты $S(t)$.

Пример 3. Анализ временного ряда валового внутреннего продукта.

Пусть имеются следующие данные: номинальный объем валового внутреннего продукта, млрд руб. (с 1998 г. – млн руб.) – квартальные данные с 1994 по 2003 г. (табл. 5).

Таблица 5

Квар-тал	Год	t	ВВП	Квар-тал	Год	t	ВВП
IV	1994	1	225	I	1999	18	901
I	1995	2	235	II		19	1102
II		3	325	III		20	1373
III		4	421	IV		21	1447
IV		5	448	I	22	1527	
I	1996	6	425	II	2000	23	1697
II		7	469	III		24	2038
III		8	549	IV		25	2044
IV		9	565	I		26	1922
I	1997	10	513	II	2001	27	2120
II		11	555	III		28	2536
III		12	634	IV		29	2461
IV		13	641	I		30	2268
I	1998	14	551	II	2002	31	2523
II		15	602	III		32	3074
III		16	676	IV		33	2998
IV		17	801	I		2003	34

График этого ряда приведен на рис. 8.

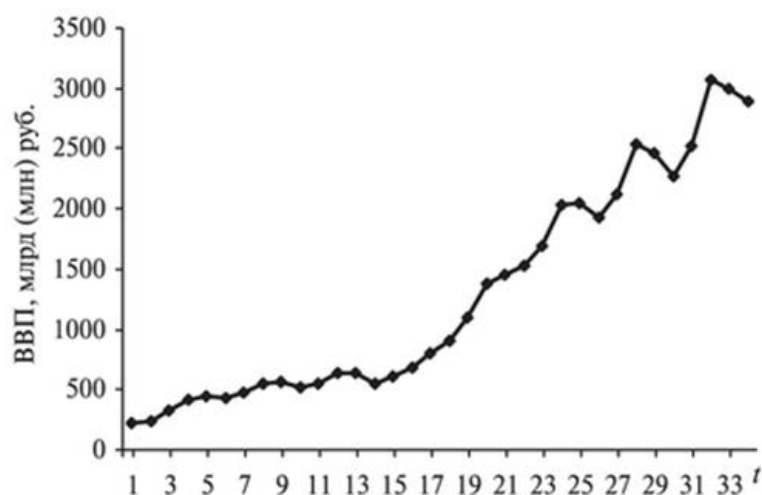


Рис. 8. График динамики ВВП

Из графика видно, что данные обладают повышающим трендом. Таким образом, уже визуальный анализ позволяет сделать вывод о нестационарности исходного временного ряда.

Проверим данное предположение, вычислим коэффициенты автокорреляции (табл. 6) и построим график автокорреляционной функции временного ряда ВВП (коррелограмму) (рис. 9).

Таблица 6

<i>Лаг</i>	<i>Коэффициент автокорреляции</i>	<i>Лаг</i>	<i>Коэффициент автокорреляции</i>
1	0,914	5	0,576
2	0,811	6	0,480
3	0,717	7	0,387
4	0,651	8	0,315

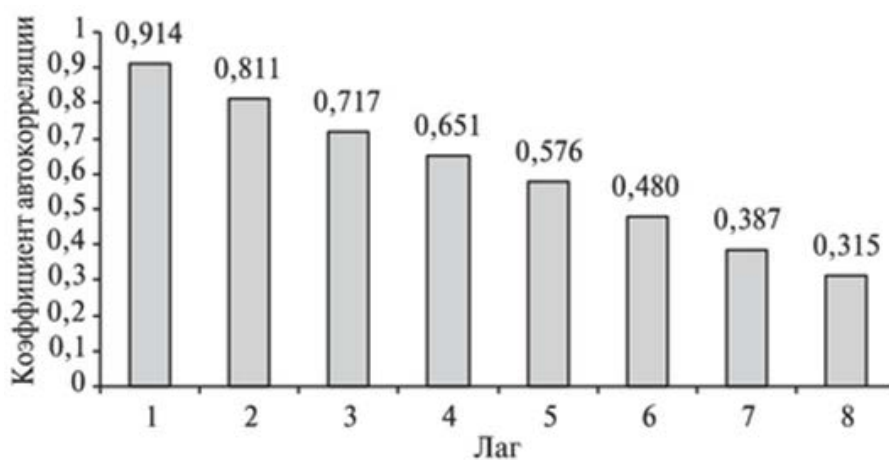


Рис. 9. Коррелограмма

Коррелограмма автокорреляционной функции в случае стационарного временного ряда должна быстро убывать с ростом t после нескольких первых значений. На рис. 9 видно, что исследуемый ряд не является стационарным. Временной ряд валового внутреннего продукта содержит трендовую компоненту.

Построение моделей временных рядов

Рассмотрим аналитические методы выделения неслучайной составляющей временного ряда.

Формирование уровней ряда определяется закономерностями трех основных типов:

- инерцией тенденции:
- инерцией взаимосвязи между последовательными уровнями ряда;
- инерцией взаимосвязи между исследуемым показателем и показателями-факторами, оказывающими на него причинное воздействие.

ВВЕДЕНИЕ

Методические указания к лабораторным работам (ЛР) по дисциплине "Модели и алгоритмы распознавания и обработки данных" представляют описание девяти четырехчасовых лабораторных работ, которые расположены последовательно, начиная от знакомства с аналитической платформой Deductor, выполнения исследований по распознаванию образов данных и их обработке с применением современных параллельных систем в виде нейронных сетей. В комплекс включены работы по предобработке данных, индуктивному обучению и поиску прецедентов в данных фильтрации на основе алгоритма Кальмана.

Лабораторные работы ориентированы на использование IBM PC, совместимых ПЭВМ, реализованных на микропроцессорах семейства 8086.

Целью проведения лабораторных работ является закрепление практических навыков распознавания и обработки данных путем программирования и визуального моделирования на аналитической платформе Deductor.

Организация и проведение лабораторных работ

Студенты объединяются в группы из 2–3 человек, работающих на закрепленном компьютере. Каждый студент получает индивидуальное задание в соответствии с номером в журнале и оформляет отчет по лабораторной работе.

Выполнение лабораторной работы предполагает предварительное изучение соответствующего раздела дисциплины и методических указаний к очередной работе.

Для допуска к выполнению лабораторной работы студент должен ознакомиться с темами для проработки и предварительно подготовить план работ и текст программы в соответствии с индивидуальным заданием.

Текст программы составляется на одном из языков программирования по указанию преподавателя или желанию студента с учетом уровня знаний конкретного языка.

В ходе выполнения лабораторной работы студент должен ответить на контрольные вопросы по предыдущей лабораторной работе. К лабораторной работе не допускаются студенты, не сдавшие более двух лабораторных работ.

Пропущенные лабораторные работы выполняются в конце семестра.

В процессе выполнения лабораторных работ следует ограничить перемещения студентов в лаборатории.

Лабораторная работа № 1. Знакомство с Аналитической Платформой «Deductor (АП DD)

Целью выполнения данной лабораторной работы является:

- получение первоначальных сведений о возможностях аналитической платформы;
- изучение основных модулей; работа с мастерами импорта, экспорта, обработки и визуализации данных.

Теоретическая часть

АП «Deductor» применима для решения задач распознавания и обработки данных, таких как парциальная обработка данных (подготовка к анализу) прогнозирование, поиск закономерностей и пр. Платформа применима в задачах, где требуется консолидация и отображение данных различными способами, построение моделей и последующее применение полученных моделей к новым данным.

Задачи, решаемые АП:

- Системы корпоративной отчетности. Готовое хранилище данных и гибкие механизмы предобработки, очистки, загрузки, визуализации позволяют быстро создавать законченные системы отчетности в сжатые сроки.
- Обработка нерегламентированных запросов. Конечный пользователь может получить ответ на вопросы типа "Сколько было продаж товара по группам за прошлый год с разбивкой по месяцам?" и просмотреть результаты наиболее удобным для него способом.
- Анализ тенденций и закономерностей, планирование, ранжирование. Простота использования и интуитивно понятная модель данных позволяет вам проводить анализ по принципу «Что, если...?», соотносить ваши гипотезы со сведениями, хранящимися в базе данных, находить аномальные значения, оценивать последствия принятия бизнес-решений.
- Прогнозирование. Построив модель на исторических примерах, можно использовать ее для прогнозирования ситуации в будущем. По мере

изменения ситуации нет необходимости перестраивать все, необходимо всего лишь дообучить модель.

- Управление рисками. Реализованные в системе алгоритмы дают возможность достаточно точно определиться с тем, какие характеристики объектов и как влияют на риски, благодаря чему можно прогнозировать наступление рискованного события и заблаговременно принимать необходимые меры к снижению размера возможных неблагоприятных последствий.

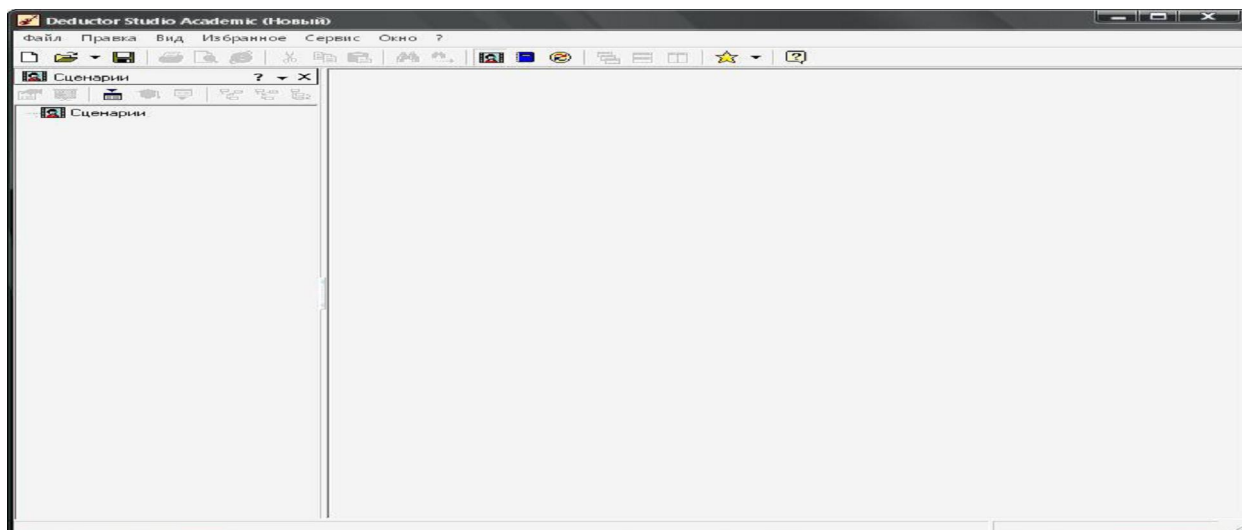
- Анализ данных маркетинговых и социологических исследований. Анализируя сведения о потребителях, можно определить, кто является вашим клиентом и почему. Как изменяются их пристрастия в зависимости от возраста, образования, социального положения, материального состояния и множества других показателей.

- Диагностика. Механизмы анализа, имеющиеся в системе Deductor, с успехом применяются в медицинской диагностике и диагностике сложного оборудования. Например, можно построить модель на основе сведений об отказах. При ее помощи быстро локализовать проблемы и находить причины сбоев.

- Обнаружение объектов на основе нечетких критериев. Часто встречается ситуация, когда необходимо обнаружить объект, основываясь не на таких четких критериях, как стоимость, технические характеристики продукта, а на размытых формулировках, например, найти продукты, похожие на ваши с точки зрения потребителя.

Ход работы

После запуска «Deductor Studio Academic» появится главное окно программы. Главное окно после запуска программы Deductor Studio__



Главное окно после запуска программы Deductor Studio

Выполнив вышеуказанные действия по импорту данных, на панели «Сценарии» формируется новый узел, с заданными именем, меткой и описанием.

Метка столбца		Мини...	Макс...	Сред...	Стан...	Сумма	Сумм...	s Кол
1	9.0 Код	1	150	75.5	3679924569	11325	1136275	
2	ab Проект по инвалидам	2	11	2,673	1,09	401	1249	
3	ab Проект по водным ре...	2	11	3,42	2,759	513	2889	
4	ab Проект по усыновлен...	2	11	2,553	1,303	383	1231	
5	ab Закон о врачах	2	11	2,82	1,443	423	1503	
6	ab Проект по Сальвадору	2	11	2,76	1,612	414	1530	
7	ab Закон о религиях	2	11	2,5	1,304	375	1191	
8	ab Антиспутниковый про...	2	11	2,553	1,102	383	1159	
9	ab Проект помощи Ника...	2	11	2,527	1,103	379	1139	
10	ab Проект по ракетам	2	11	2,82	1,601	423	1575	
11	ab Закон об иммигрантах	2	11	2,553	1,102	383	1159	
12	ab Проект по альтернат...	2	11	2,94	1,573	441	1665	
13	ab Закон об образовании	2	11	3,307	2,425	496	2516	
14	ab Проект по фондам	2	11	2,9	1,864	435	1779	
15	ab Проект по преступно...	2	11	2,753	1,757	413	1597	
16	ab Проект по таможенн...	2	11	2,933	1,856	440	1804	
17	ab Проект по экспорту	2	11	4,247	3,754	637	4805	
18	ab Класс	8	13	9,933	2,443	1490	15690	

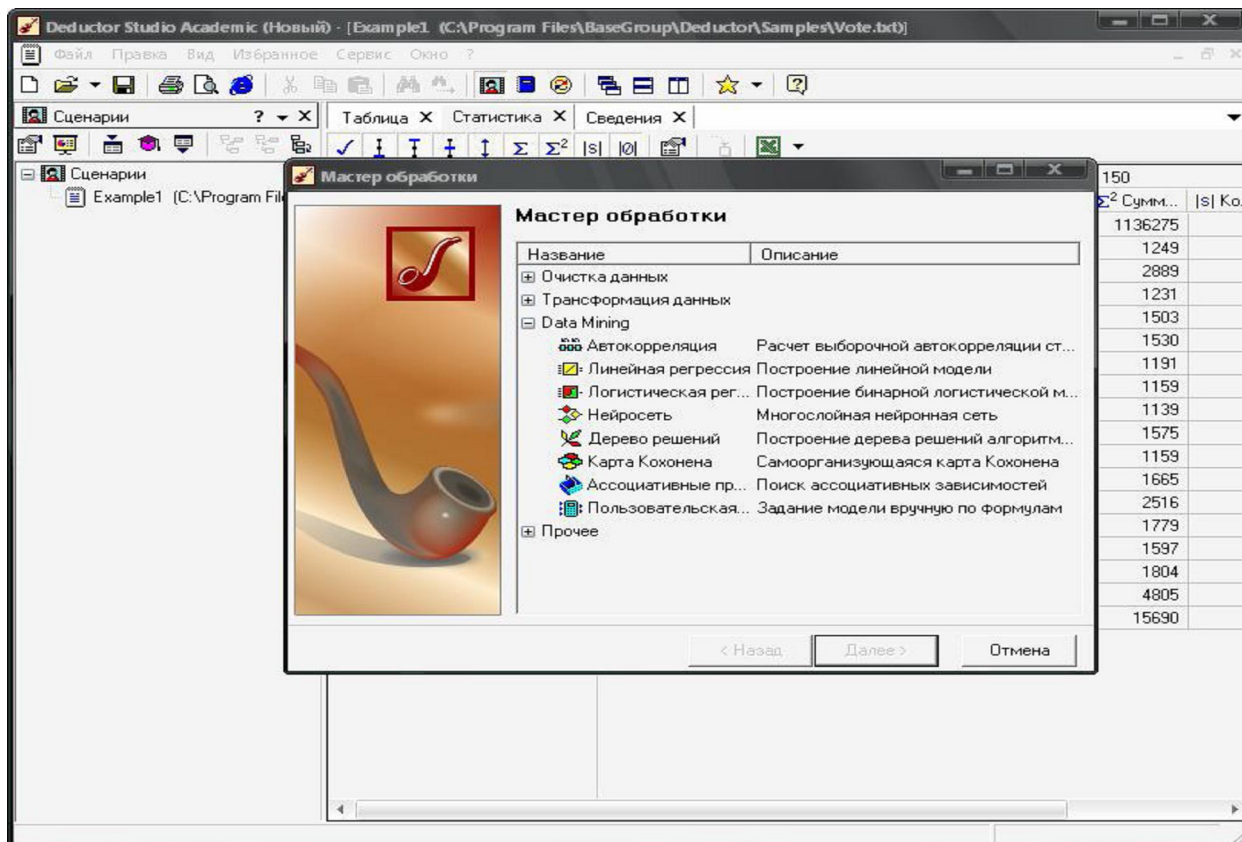
Пример создания сценария, вкладка «Статистика»

Для изучения возможности мастера обработки (кнопка в левой части главного окна либо клавиша F7). После запуска мастера обработки появится список возможных способов обработки данных.

Все способы разделены на четыре основные группы: очистка данных, трансформация данных, Data Mining, пр. Каждый способ обработки имеет название и краткое описание. Выбор способа зависит от целей обработки

данных (например, сортировка и фильтрация данных, построение дерева решений и пр.).

Мастер визуализации позволяет определить способ отображения данных, указать метки и добавить описание к проекту. Запустить его можно с помощью кнопки либо клавишей F5.



Список доступных способов обработки данных

Готовый проект можно экспортировать, воспользовавшись мастером экспорта (кнопка основного окна либо клавиша F8).

Указав параметры, проект можно перенести в один из доступных форматов.

Задание

1. Опишите назначение и возможности АП «Deductor».
2. Запустите программу «Deductor Studio Academic», ознакомьтесь с назначением кнопок и контекстным меню главного окна программы.
3. Воспользуйтесь мастером импорта данных (импортируйте файл с данными Вашей предметной области или из C:\Program Files\ BaseGroup\ Deductor\Samples\ *.txt), или из репозитория данных.
4. Ознакомьтесь с доступными способами обработки данных.

5. Изучите возможности мастера визуализации и экспорта.

Содержание отчета

1. Цель работы.
2. Краткое описание хода работы с описанием возможности Deductor для распознавания и обработки данных и приведением скриншотов.
3. Ответы на вопросы.

Вопросы:

1. Какие существуют другие платформы для распознавания и обработки данных?
2. Какие возможности имеет API Deductor для распознавания данных?
3. Какие возможности имеет API Deductor для обработки данных?
4. Какие параметры доступны для мастера экспорта данных?
4. В чем заключается процедура визуализации данных?

ЛАБОРАТОРНАЯ РАБОТА №2.

ХРАНИЛИЩЕ ДАННЫХ В АНАЛИТИЧЕСКОЙ ПЛАТФОРМЕ DEDUCTOR.

Цель работы: изучить программную среду хранилища данных в DeductorWarehouse, ознакомиться с архитектурой научиться создавать, и наполнять информацию из хранилища данных.

Ход работы:

1. Хранилище данных (ХД) **DeductorWarehouse** - это специально организованная база данных, ориентированная на решение задач анализа данных и поддержки принятия решений, обеспечивающая максимально быстрый и удобный доступ к информации. **DeductorWarehouse** 6 соответствует модели ROLAP (схема «снежинка»).

Хранилище данных **DeductorWarehouse** включает в себя потоки данных, поступающие из различных источников, и специальный семантический слой, содержащий так называемые метаданные (данные о данных). Семантический слой и сами данные хранятся в одной базе данных. Все данные в хранилище

DeductorWarehouse 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.4.9.).

Объекты хранилища данных DeductorWarehouse следующие.

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 - это свойство измерения (т.е. 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.

Факт - значение, соответствующее измерению. Факты - это данные, отражающие сущность события. Как правило, фактами являются численные значения, например, сумма и количество отгруженного товара, скидка.

Ссылка на измерение - это установленная связь между двумя и более измерениями. Бизнес-понятия (соответствующие измерениям в хранилище данных) могут образовывать иерархии,

например, Товары могут 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.

Процесс - совокупность измерений, фактов и атрибутов. По сути, процесс и есть «снежинка». Процесс описывает определенное действие, например, продажи товара, отгрузки, поступления денежных средств и прочее.

Атрибут процесса - свойство процесса. Атрибут процесса в отличие от измерения не определяет координату в многомерном пространстве. Это справочное значение, относящееся к процессу, например, № накладной,

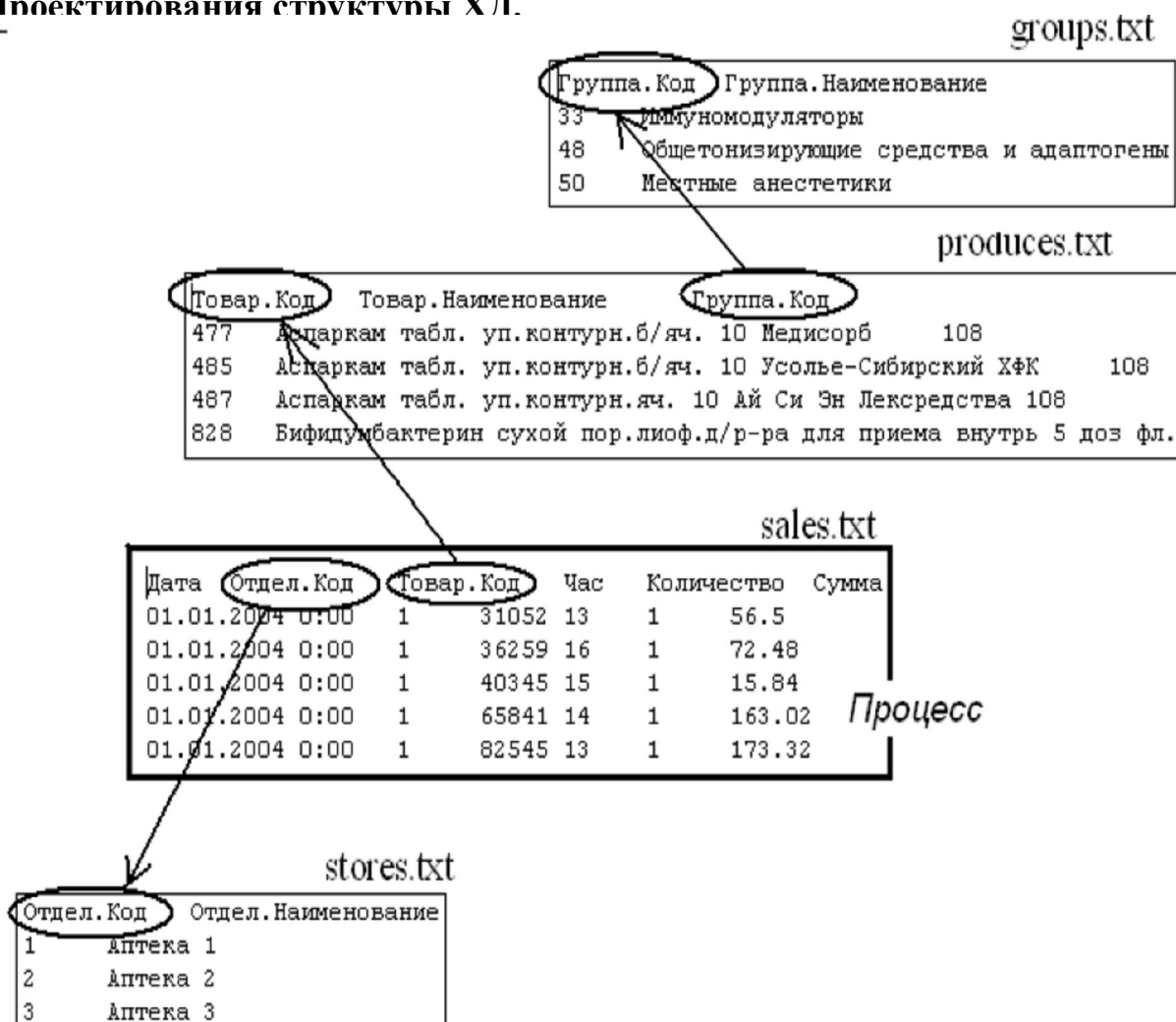
Валюта документа и так далее. Значение атрибута процесса в отличие от измерения может быть не всегда определено.

В DeductorWarehouse может одновременно храниться множество процессов, имеющих общие измерения, например, измерение *Товар*, фигурирующее в процессах *Поступления* и *Отгрузки*.

Все загружаемые в ХД данные обязательно должны быть определены как измерение, атрибут либо факт. 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).

Проектирования структуры ХД.




В таблице groups.txt *Код группы* является измерением, а *Наименование группы* - его атрибутом.

В таблице produces.txt *Код товара* является измерением, а *Наименование товара* - его атрибутом, а *Код группы* - ссылкой на одноименное измерение.

В таблице stores.txt *Код отдела* является измерением, а *Наименование отдела* - его атрибутом.

В таблице sales.txt *Дата* является измерением, *Отдел*, *Код товара* и *Код группы* как было сказано выше – измерения. *Час покупки* - измерение, *Количество* и *Сумма*- факты, т.е. таблица sales.txt является описанием процесса продаж в трех аптеках.

2. Создание хранилища данных в DeductorWarehouse.

Откройте программу DeductorStudio, используя ярлык на рабочем столе или через кнопку Пуск.  Deductor Studio

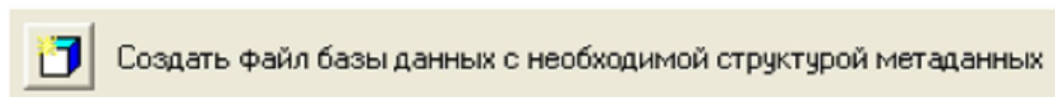
Для создания нового хранилища данных или подключения к существующему в DeductorStudio необходимо перейти на закладку **Подключения** и запустить **Мастер подключений**.

На экране появится первый шаг **Мастера**, в котором следует выбрать тип источника (приемника), к которому нужно подключиться. Выберите **DeductorWarehouse** и нажмите кнопку **Далее**.

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. 2), Нажмите **Далее**.

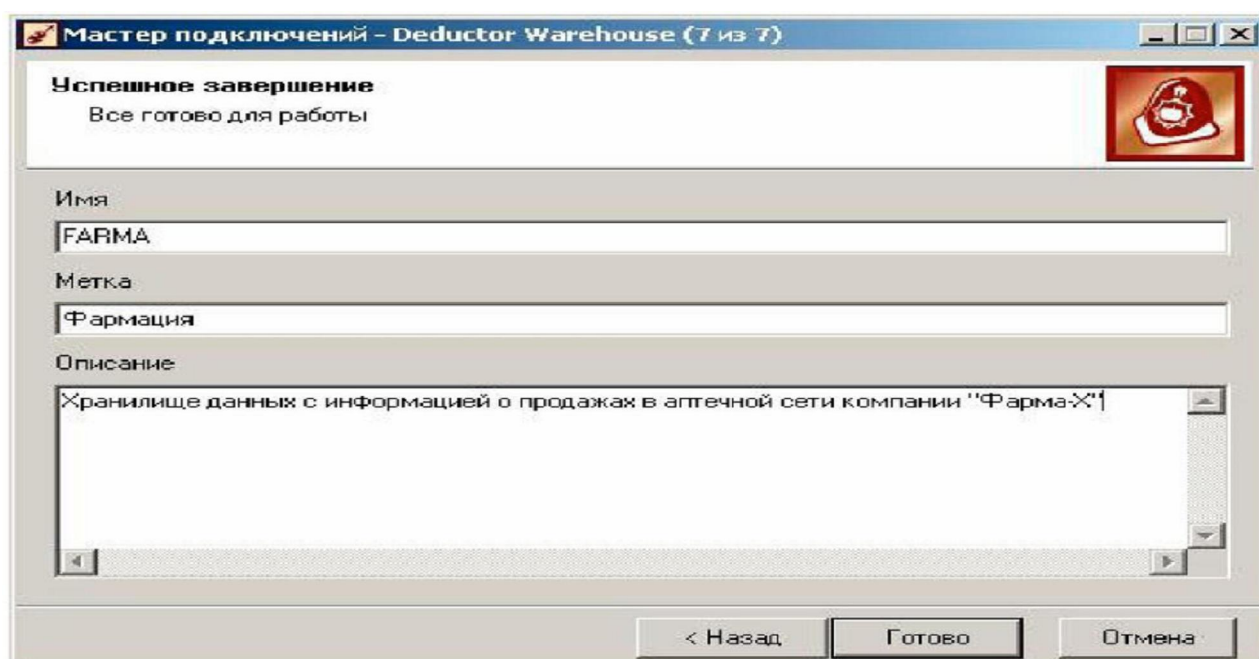
На следующей вкладке выберем последнюю версию для работы с ХД **DeductorWarehouse6** (предыдущие версии необходимы для совместимости с ранними версиями хранилищ).

На следующем шаге при нажатии на кнопку



По указанному ранее пути будет создан файл **farma.gdb** (появится сообщение об успешном создании). Это и есть пустое хранилище данных, готовое к работе.

На последних двух шагах осталось выбрать визуализатор для подключения (здесь это **Сведения** и **Метаданные**) и задать имя, метку и описание для нового хранилища.



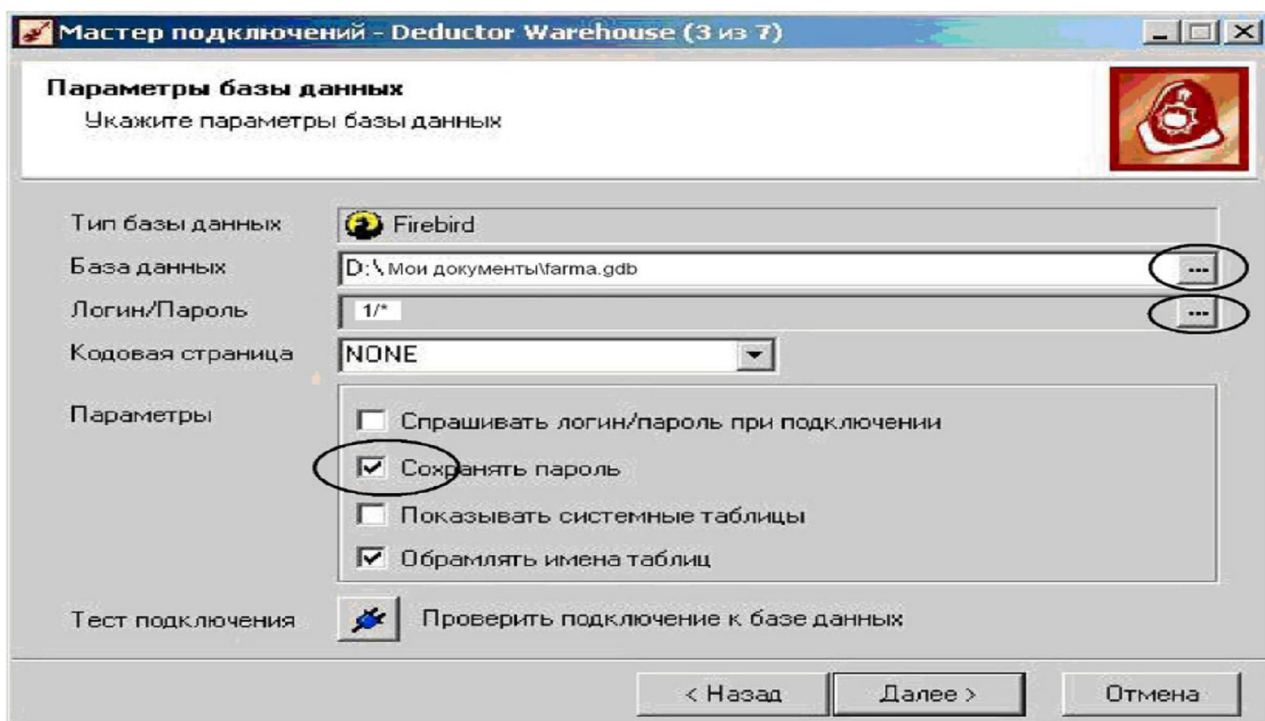
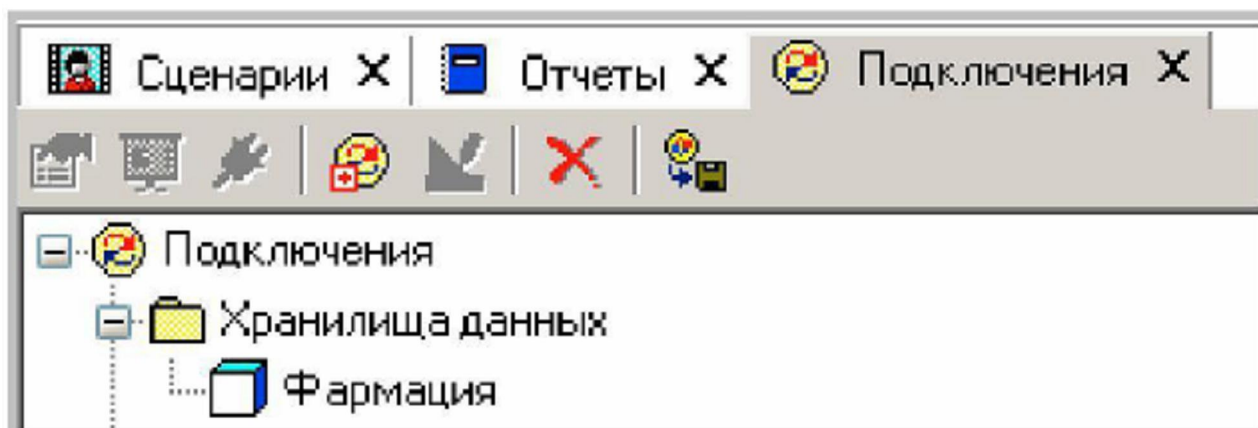


Рис.2. Установка параметров базы данных

После нажатия на кнопку **Готово** на дереве узлов подключений появится метка хранилища.



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 кнопку **Настроить подключение**).

Для проверки доступа к новому хранилищу данных воспользуйтесь кнопкой **Проверить подключение**. Если спустя некоторое время появится сообщение «Тестирование соединения прошло успешно», то хранилище готово к работе.

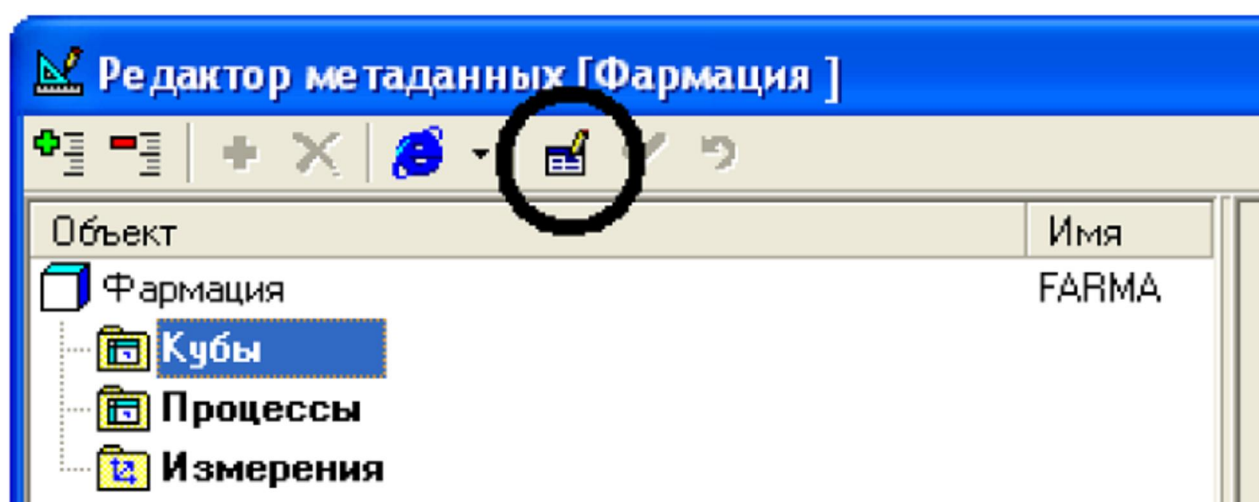
Сохраните настройки подключений, нажав на кнопку сохранения .

После создания хранилища необходимо спроектировать его структуру, т.к. в пустом хранилище нет ни одного объекта (процессов, измерений, фактов).

Для этого предназначен «Редактор метаданных», который вызывается кнопкой на вкладке **Подключения**. Нажмите ее.



Для перехода в режим внесения изменений в структуру хранилища нажмем кнопку **Разрешить редактировать**.



Появится диалоговое окно с предупреждением. Нажмем **Да** и в открывшемся окне редактора метаданных, встав на узле **Измерения**, при помощи кнопки **Добавить** добавим в метаданные первое измерение Код группы со следующими параметрами:

имя – GR_ID;

метка - Группа.Код;

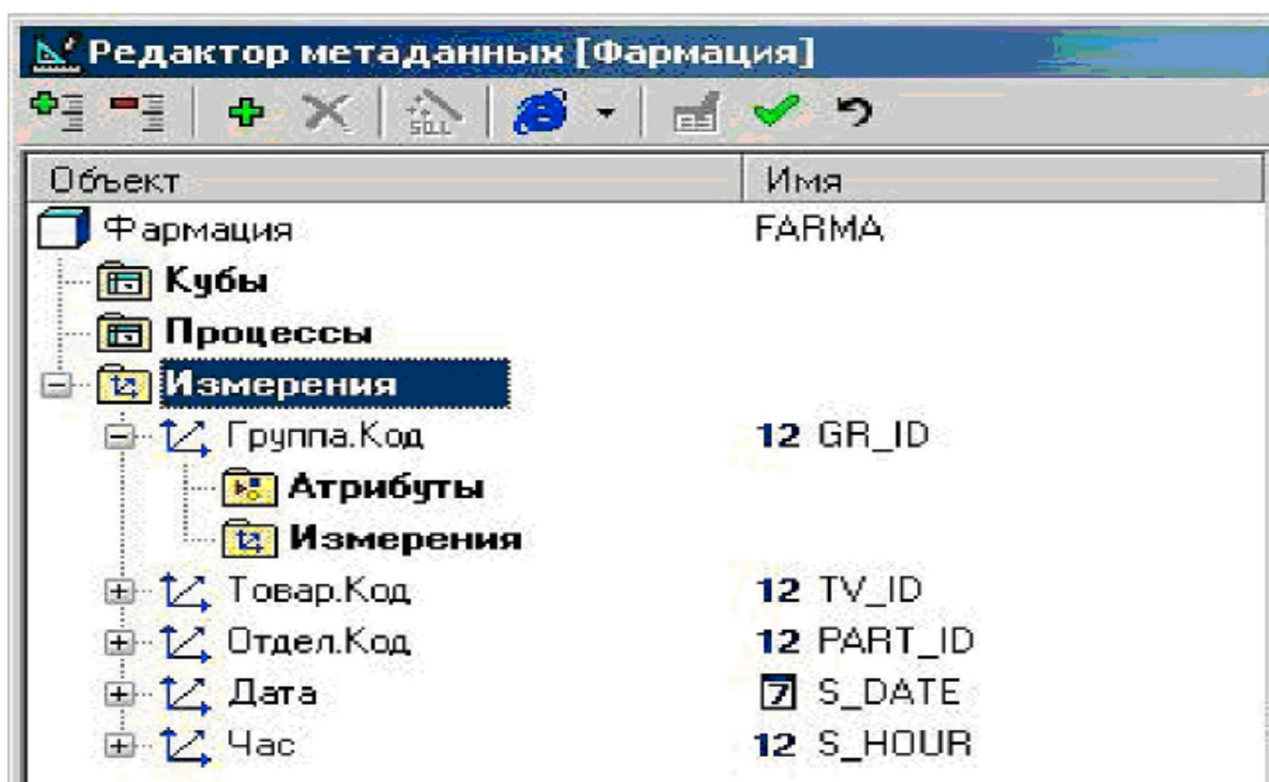
тип данных-целый.

Имя - это семантическое название объекта хранилища данных, которое увидит пользователь, работающий с ХД. (Эти параметры для таблицы «Товарные группы»).

Выполните аналогичные действия для создания всех остальных измерений, взяв параметры из таблицы 1.

Таблица	1.	Имя	Метка	Тип данных
Параметры измерений				
Измерение				
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.



К каждому измерению, кроме Дата и Час, теперь добавим по текстовому атрибуту. Для этого в измерении «Группа.Код» правой кнопкой мыши откроем **Атрибуты** и справа в поле «Метка» введем название атрибута - Группа.Наименование. Тип данных оставим строковым. Размер поля в строковых атрибутах предлагается равным 100, оставим это без изменений. Аналогично введите названия атрибутов :для измерения Товар.Код - Товар.Наименование, для измерения

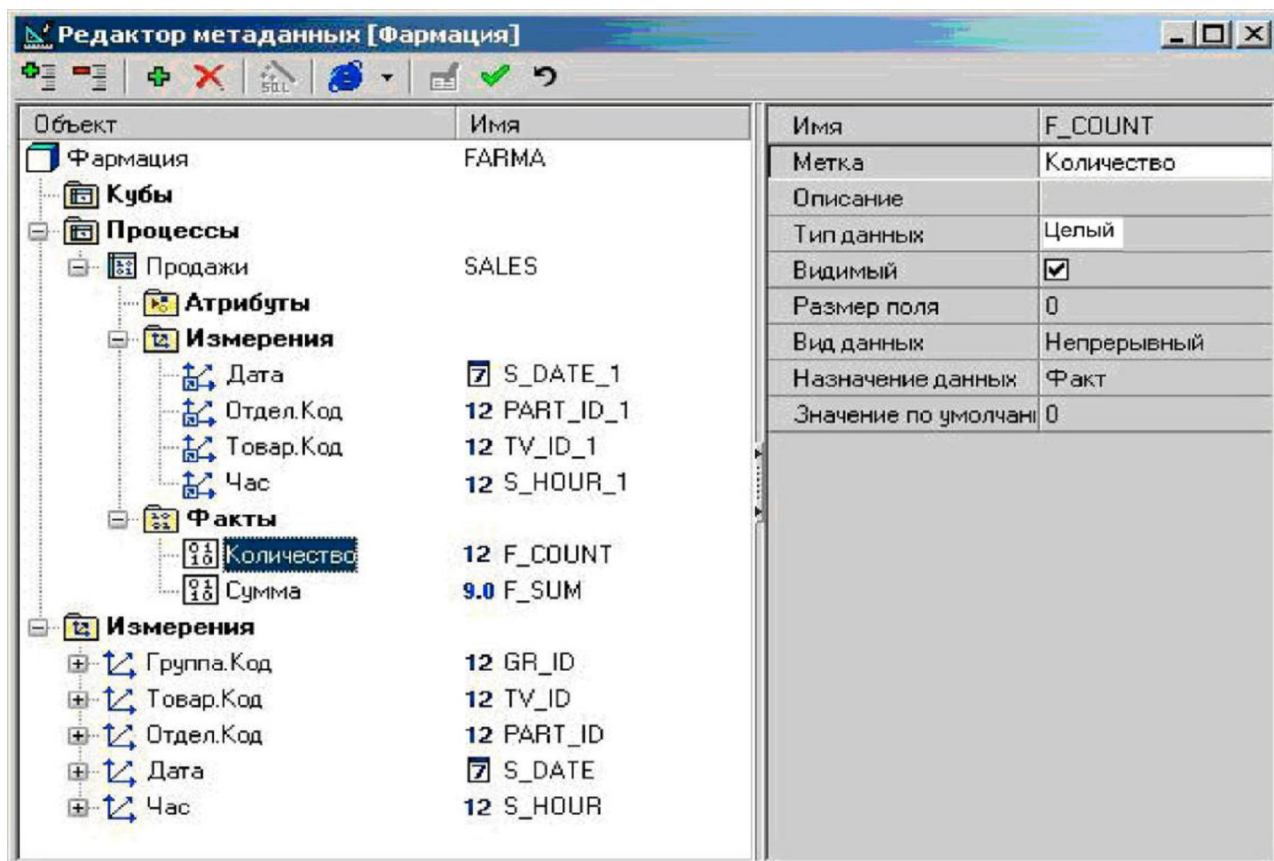
Отдел.Код - Отдел.Наименование..

Каждое измерение может ссылаться на другое измерение, реализуя тем самым иерархию измерений (схема «снежинка»). В нашем случае измерение Товар.Код ссылается на Группа.Код (см. табл. 1 и табл. 2). Эту ссылку и установим путем добавления объекта к измерению, для этого в измерении «Товар.Код» правой кнопкой мыши откроем Измерение и выберем пункт Добавить. Имя ссылки зададим GR_ID_1, а метку - Группа.Код. Ссылка на

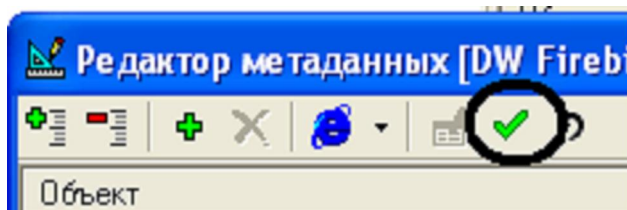
измерение отображается иконкой  .



После того как все измерения и ссылки на измерения созданы, приступают к формированию процесса. Назовем его *Продажи* и «соберем» его из 4 существующих измерений: Дата, Отдел.Код, Товар.Код, Час (кнопка). Кроме них в нашем процессе присутствуют два факта: Количество и Сумма, причем первый - целочисленный, второй – вещественный. Результат представлен на рисунке 3.



На этом проектирование структуры и метаданных ХД закончено. Для того чтобы принять все изменения, нужно нажать кнопку **Принять изменения**



После этого закройте окно редактора. Структура хранилища данных готова.

Задание. Для выбранной предметной области сформировать хранилище данных, заполнить его. Привести примеры вывода данных.

Содержание отчета

1. Цель работы.
2. Ход работы.
3. Ответы на вопросы.
4. Листинг программы.
5. Заключение.

Вопросы

1. Что такое «Редактор метаданных» в DeductorStudio?
2. Как создать новое пустое хранилище данных?

3. Как сделать иерархию измерений?
4. Какие типы данных могут быть у объектов хранилища?
5. Чем факт отличается от измерения?

ЛАБОРАТОРНАЯ РАБОТА №3. Поиск ассоциативных правил

Цель работы. Изучить возможность поиска ассоциативных правил используя аналитическую платформу **DEDUCTOR**

Теоретическая часть. В последнее время растет интерес к методам «обнаружения знаний в базах данных». Большие объемы современных баз данных вызывают спрос на новые алгоритмы распознавания и обработки данных. Одним из распространенных аналитических методов обработки данных является аффинитивный анализ (англ: affinityanalysis), название произошедшее от английского слова affinity – близость, сходство. Метод определяет взаимные связи между событиями, происходящие совместно. Одним из применения аффинитивного анализа является анализ рыночной корзины (англ: marketbasketanalysis), цель которого – обнаружить ассоциации между различными данными, т.е. найти правила для количественного описания взаимной связи между двумя или более данными. Такие правила называются **ассоциативными правилами** (англ.: associationrules) и применяются в data mining.

Примерами приложения ассоциативных правил могут быть следующие задачи:

1. Обнаружение наборов товаров, которые в супермаркетах часто покупаются вместе или никогда не покупаются вместе.
2. Определение доли клиентов, положительно относящихся к нововведениям в их обслуживании.
3. Определение профиля посетителя веб-ресурса.
4. Определение доли случаев, в которых новое лекарство показывает опасный побочный эффект.

Базовым понятием в теории ассоциативных правил является транзакция.

Транзакция – некоторое множество событий, происходящих совместно. Типичной транзакцией является приобретение клиентом некоторого товара в супермаркете. В табл. 1 представлен простой пример набора транзакций. В каждой строке содержится комбинация продуктов, приобретенных за одну покупку.

Таблица 1. Пример набора транзакций.

№ транзакции	Товары
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.помидоры
9	яблоки, сливы, морковь, помидоры, лук, конфеты
10	яблоки, картофель

На практике обрабатываются миллионы транзакций, в которых участвуют десятки и сотни различных продуктов. Данный пример ограничен 10 транзакциями, содержащими 13 видов продуктов, что достаточно для иллюстрации методики обнаружения ассоциативных правил. В большинстве случаев клиент приобретает не один товар, а некоторый набор товаров, называемых рыночной корзиной. Существует связь между спросом на товары, которую может обнаружить ассоциативное правило, утверждающее, что покупатель, купивший молоко, с вероятностью 75% купит и хлеб. Такие связи существуют и в других областях, например в медицинской или технической диагностике, выборе профессий и т.д.

Анализ рыночной корзины – это анализ наборов данных для определения комбинаций товаров, связанных между собой. Иными словами, производится поиск товаров, присутствие которых в транзакции влияет на вероятность наличия других товаров или комбинаций товаров [4].

Современные кассовые аппараты в супермаркетах позволяют собирать информацию о покупках, которая может храниться в базе данных. Накопленные данные затем могут использоваться для построения систем поиска ассоциативных правил.

Визуальный анализ примера (табл.1) показывает, что все четыре транзакции, в которых фигурирует салат, также включают и помидоры, и что четыре из семи транзакций, содержащих помидоры, также содержат и салат. Салат и помидоры в большинстве случаев покупаются вместе. Ассоциативные правила позволяют обнаруживать и количественно описывать такие совпадения.

Ассоциативное правило состоит из двух наборов предметов, называемых условие (англ: antecedent) и следствие (англ: consequent), записываемых в виде $X \rightarrow Y$, что читается «из X следует Y». Таким образом, ассоциативное правило формулируется в виде «Если условие, то следствие».

Условие часто ограничивают содержанием только одного предмета. Правила обычно отображаются с помощью стрелок, направленных от условия к следствию, например, (помидоры) \rightarrow (салат). Условие и следствие часто называются соответственно левосторонним (LHS – left-handside) и правосторонним (RHS – right-handside) компонентом ассоциативного правила.

Ассоциативные правила описывают связь между наборами предметов, соответствующим условию и следствию. Эта связь характеризуется двумя показателями – поддержкой и достоверностью.

Обозначим D как базу данных транзакций, а N как число транзакций в этой базе. Каждая транзакция T_i представляет собой некоторый набор предметов.

Зададим, что S (англ.: support) – поддержка, C (англ.: confidence) – достоверность.

Поддержка ассоциативного правила – это число транзакций, содержащих как условие, так и следствие.

Например, для ассоциации $A \rightarrow B$ можно записать:

$$S(A \rightarrow B) = P(A \cap B) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{общее количество транзакций}}$$

Достоверность ассоциативного правила – это мера точности правила, которая определяется как отношение количества транзакций, содержащих как условие, так и следствие, к количеству транзакций, содержащих только условие.

Например, для ассоциации $A \rightarrow B$ можно записать:

$$C(A \rightarrow B) = P(A|B) = P(A \cap B) / P(A) = \frac{\text{количество транзакций, содержащих } A \text{ и } B}{\text{количество транзакций, содержащих только } A}$$

Если поддержка и достоверность достаточно высоки, то это позволяет с большой вероятностью утверждать, что любая будущая транзакция, которая включает условие, будет также содержать и следствие.

Рассмотрим пример для вычисления поддержки и достоверности для ассоциаций из табл.1. Возьмем ассоциацию (салат) \rightarrow (помидоры). Поскольку количество транзакций, содержащее как (салат), так и (помидоры), равно 4, а общее число транзакций 10, то поддержка данной ассоциации будет:

$$S((\text{салат}) \rightarrow (\text{помидоры})) = 4/10 = 0,4 .$$

Поскольку количество транзакций, содержащее только (салат) как условие, равно 4, то достоверность данной ассоциации будет:

$$C((\text{салат}) \rightarrow (\text{помидоры})) = 4/4 = 1.$$

Все наблюдения, содержащие салат, также содержат и помидоры, что позволяет сделать вывод о том, что данная ассоциация может рассматриваться как правило. С точки зрения интуитивного поведения такое правило вполне объяснимо, поскольку оба продукта широко используются для приготовления растительных блюд и часто покупаются вместе.

Рассмотрим ассоциацию (конфеты) \rightarrow (помидоры), в которой содержатся, в общем-то, слабо совместимые в гастрономическом плане продукты (тот, кто планирует сделать растительное блюдо, вряд ли станет покупать конфеты, а покупатель, желающий приобрести что-нибудь к чаю, скорее всего, не станет покупать помидоры). Поддержка данной ассоциации $S = 3/10 = 0,3$, а достоверность $C = 3/7 = 0,43$. Таким образом, сравнительно невысокая достоверность данной ассоциации дает повод усомниться в том, что она является правилом.

Аналитики могут отдавать предпочтение правилам, которые имеют только высокую поддержку или только высокую достоверность, либо, что является наиболее частым, оба эти показателя. Правила, для которых значения поддержки или достоверности превышают некоторый, заданный пользователем порог, называются сильными правилами (*strongrules*). Например, аналитика может интересоваться, какие товары в супермаркете, покупаемые вместе, образуют ассоциации с минимальной поддержкой 20% и минимальной достоверностью 70%. С другой стороны, при анализе с целью обнаружения мошенничества, аналитику может потребоваться уменьшение поддержки до 1%, поскольку сравнительно небольшое число транзакций являются связанными с мошенничеством.

Значимость ассоциативных правил

Методики поиска ассоциативных правил обнаруживают все ассоциации, которые удовлетворяют ограничениям на поддержку и достоверность, наложенные пользователем. Это часто приводит к необходимости рассмотреть десятки и сотни тысяч ассоциаций, что делает невозможным «ручную» обработку такого большого количества данных. Очевидно, что желательно уменьшить число правил таким образом, чтобы проанализировать только наиболее значимые правила. Часто значимость связана с разностью между поддержкой правила в целом и произведением поддержки только условия и поддержки только следствия.

Выделяют объективные и субъективные меры значимости правил. Объективными являются такие меры, как поддержка и достоверность, которые могут применяться независимо от конкретного приложения. Субъективные меры связаны со специальной информацией, определяемой пользователем в контексте решаемой задачи. Такими субъективными мерами являются **лифт** (англ: lift) и **левередж**(от англ. leverage- плечо, рычаг).

Лифт – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом. Лифт (оригинальное название - интерес) определяется следующим образом:

$$L(A \rightarrow B) = C(A \rightarrow B) / S(B).$$
 Значения лифта большие, чем единица показывают, что условие более часто появляется в транзакциях, содержащих и следствие, чем в остальных. Можно сказать, что лифт является обобщенной мерой связи двух предметных наборов: при значениях лифта >1 связь положительная, при 1 она отсутствует, а при значениях <1 -отрицательная. Другой мерой значимости правила является **левередж**(англ: leverage; предложена Г. Пятецким-Шапиро):

Левередж – это разность между наблюдаемой частотой, с которой условие и следствие появляются совместно (т.е., поддержкой ассоциации), и произведением частот появления (поддержек) условия и следствия по отдельности.

$$L(A \rightarrow B) = S(A \rightarrow B) - S(A)S(B).$$

Такие меры, как лифт и левередж могут использоваться для последующего ограничения набора рассматриваемых ассоциаций путем установки порога значимости, ниже которого ассоциации отбрасываются.

Генерация ассоциативных правил

В DeductorStudio для решения задач ассоциации используется обработчик **Ассоциативные правила**. В нем реализован алгоритм apriori. Обработчик требует на входе два поля: идентификатор транзакции и элемент транзакции.

Например, идентификатор транзакции – это номер чека или код клиента. А элемент - это наименование товара в чеке или услуга, заказанная клиентом.

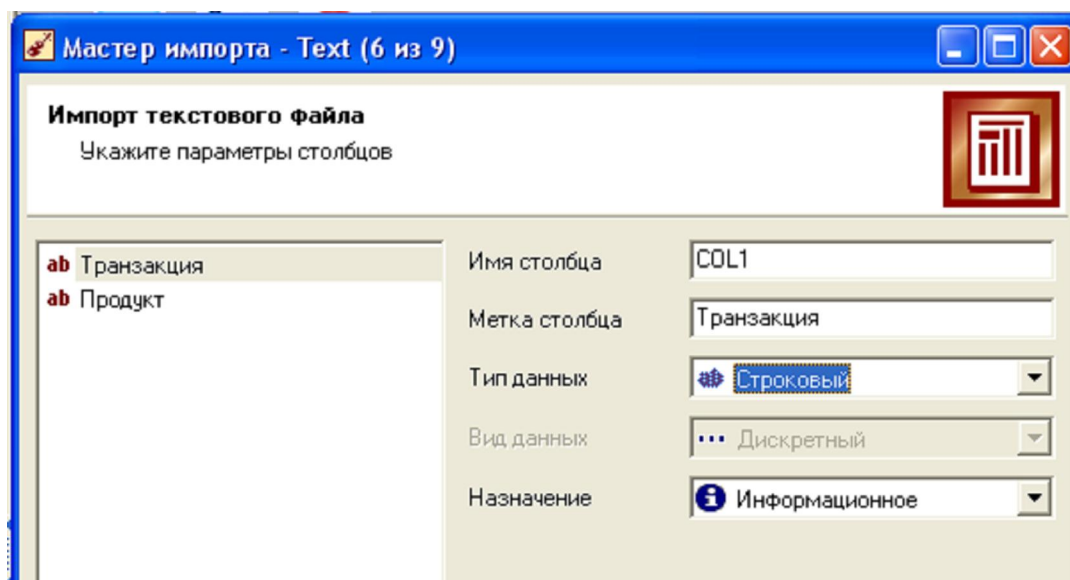
Оба поля (идентификатор и элемент транзакции) должны быть дискретного вида.

Пример решения конкретной задачи ассоциации из области розничной торговли:

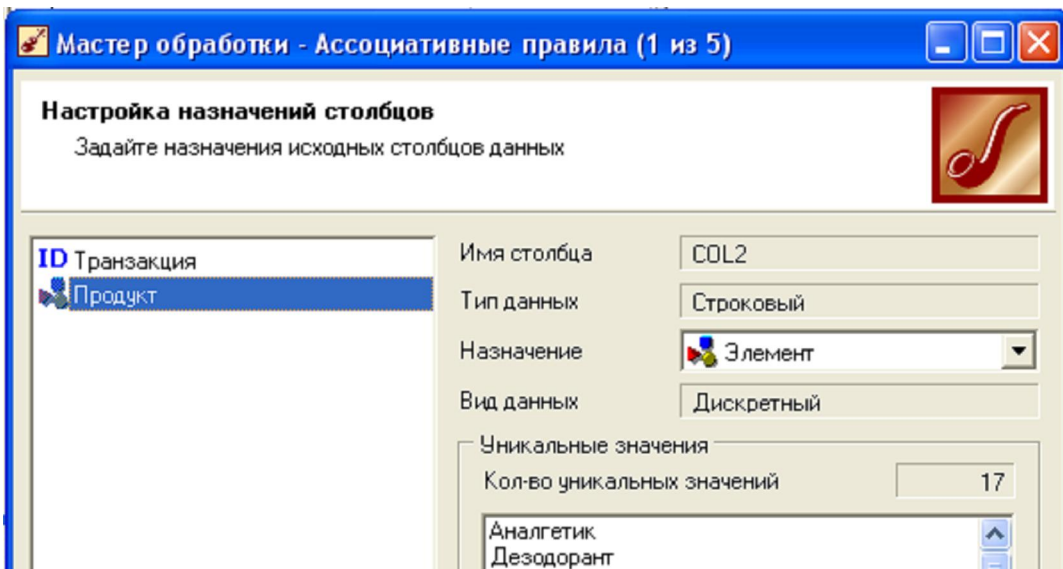
- предсказать то, какие товары покупатели могут выбрать в зависимости от того, что уже есть в их корзинах;
- предложить рекламные акции типа «Каждому купившему товары А и В, товар С в подарок».

Откройте программу DeductorStudio  Deductor Studio , используя ярлык на рабочем столе или через кнопку Пуск.

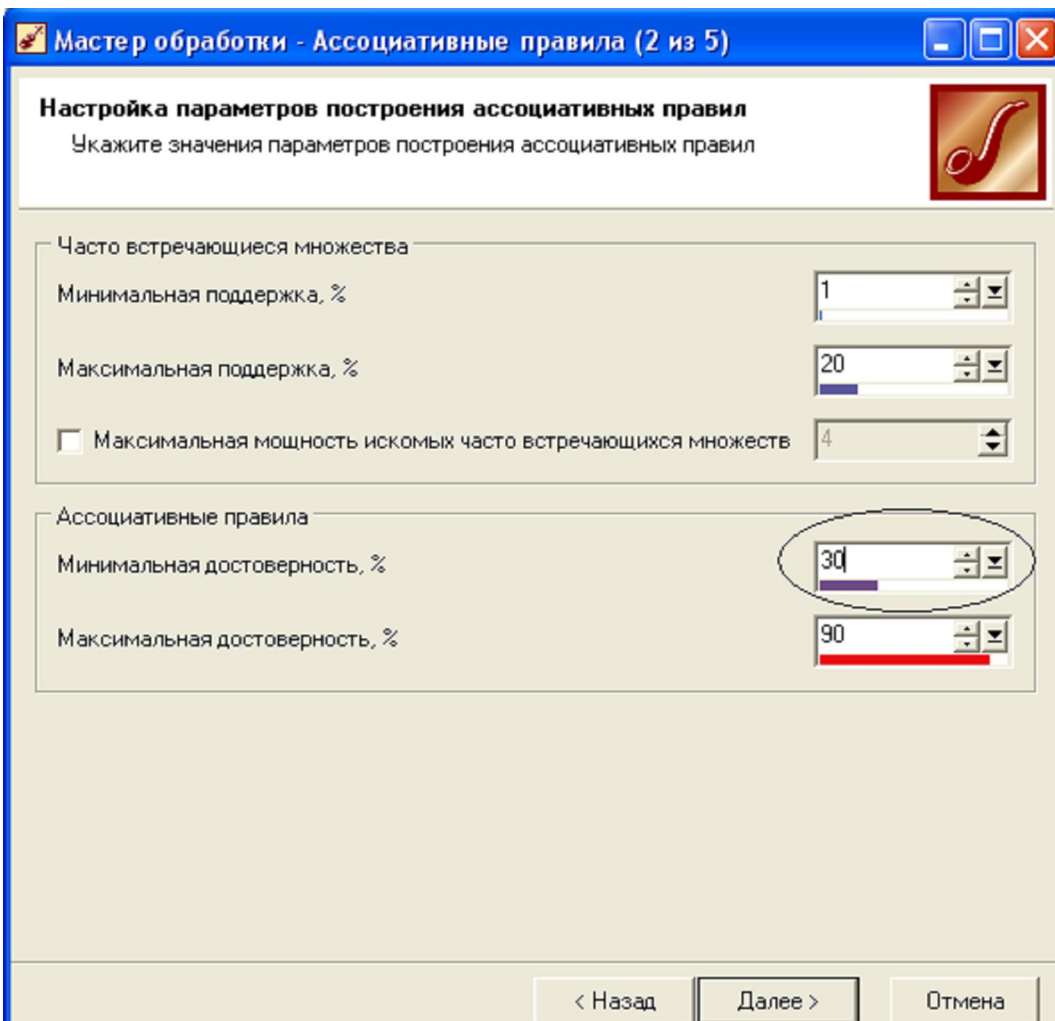
Импортируйте данные из текстового файла transactions.txt в DeductorStudio. В файле данных имеются два столбца Транзакция и Продукт, для которых нужно Тип поля нужно установить строковый.



После импорта к данному загруженному файлу применим обработчик **Ассоциативные правила**. Столбец Транзакция сделаем идентификатором транзакции, а столбецПродукт – ее элементом:



На следующем шаге мастера настроим параметры построения ассоциативных правил, что, по сути, есть



параметры алгоритма apriori:

Здесь для изменения доступны следующие параметры.

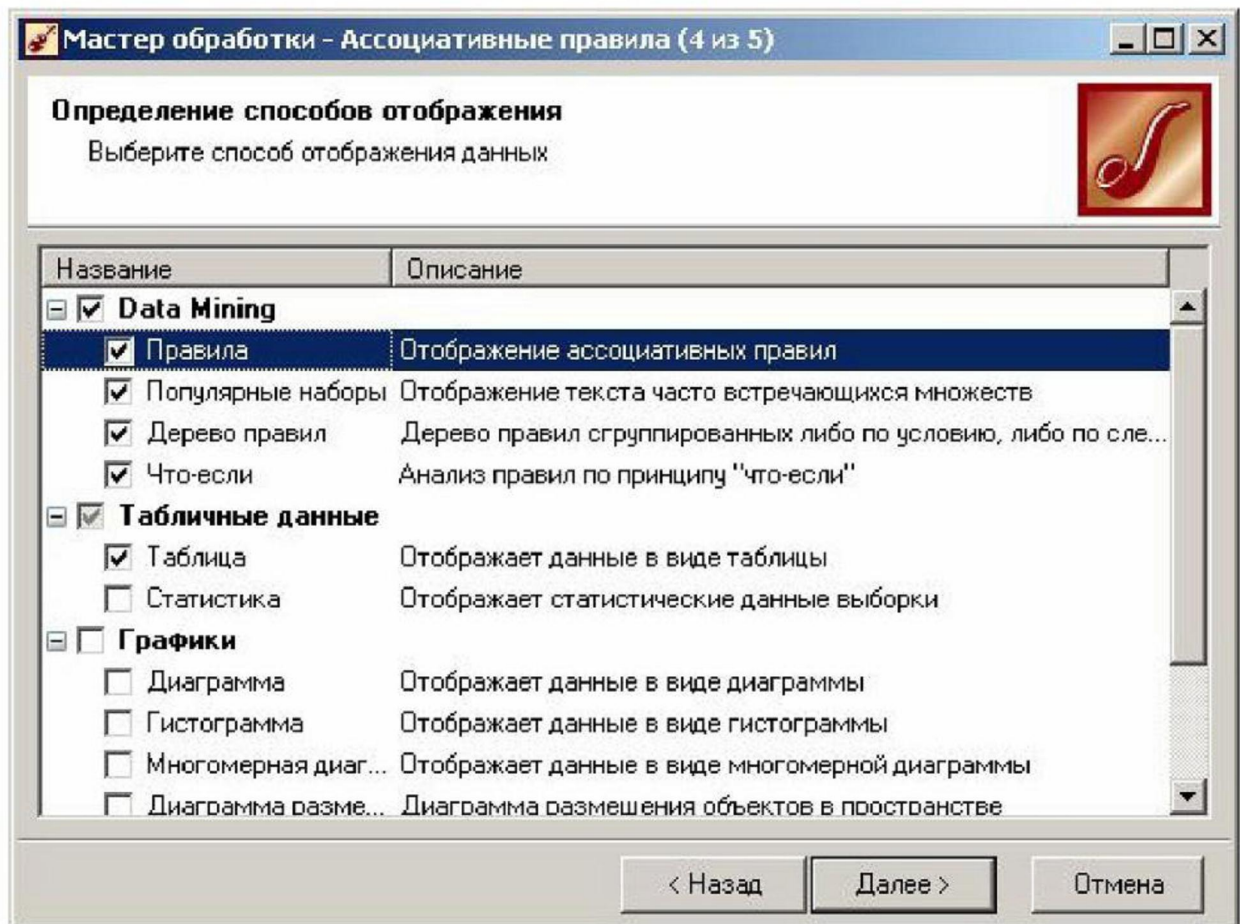
Минимальная и максимальная поддержка в % – ограничивают пространство поиска часто встречающихся предметных наборов. Эти границы определяют множество популярных наборов, из которых и будут создаваться ассоциативные правила.

Минимальная и максимальная достоверность в % – в результирующий набор попадут только те ассоциативные правила, которые удовлетворяют условиям минимальной и максимальной достоверности.

Максимальная мощность искомым часто встречающихся множеств – параметр ограничивает длину k-предметного набора. Например, при установке значения 4 шаг генерации популярных наборов будет остановлен после получения множества 4-предметных наборов. В конечном итоге это позволяет избежать появления длинных ассоциативных правил, которые трудно интерпретируются.

Нажмите на кнопку **Пуск**, что приведет к работе алгоритма поиска ассоциативных правил. По окончании его работы справа в полях появится следующая информация:

Далее выбираем все доступные специализированные визуализаторы DataMining и визуализатор Таблица:




Все эти визуализаторы, кроме **Что-если**, отображают результаты работы алгоритма в различных формах.

На вкладке **Правила** помимо самих ассоциативных правил приводятся их основные расчетные характеристики:

Правила X Популярные наборы X Дерево правил X Что-если X Таблица X								
Правил: 6 из 6 Фильтр: Без фильтрации								
№	Номер правила	Условие	Следствие	Поддержка		Достоверность	Лифт	
				Кол-во	%			
1	1	Зубная щетка	Парфюм	446	2,23	35,20	3,952	
2	2	Зубная паста	Конфеты	218	1,09	45,04	2,618	
3	3	Карандаши	Зубная паста	218	1,09	33,90	2,125	
4	4	Зубная паста	Поздравительная открыт	272	1,36	34,61	2,288	
5	5	Зубная паста	Конфеты	272	1,36	40,60	2,360	
6	6	Конфеты	Зубная паста	272	1,36	30,63	1,920	

поддержка, достоверность и лифт.

На вкладке **Популярные наборы** отображается множество найденных популярных предметных наборов в виде списка. Кнопка  предлагает на выбор несколько вариантов сортировки списка, а кнопка * вызывает окно настройки фильтра множеств. Например, задав в фильтре минимальное значение поддержки 3% и отсортировав их по убыванию поддержки, получим 17 популярных наборов (на картинке изображено только 12):

Популярные наборы X					
Множеств: 17 из 26 Фильтр: Минимальная поддержка = 3,00					
№	ab. Номер множества	ab. Элементы	Поддержка		S Мощность
			Кол-во	%	
1	6	Конфеты	3446	17,20	1
2	2	Зубная паста	3196	15,96	1
3	13	Поздравительная открытка	3029	15,12	1
4	10	Набор ручек	2922	14,59	1
5	4	Карандаши	2714	13,55	1
6	12	Парфюм	1784	8,91	1
7	3	Зубная щетка	1267	6,33	1
8	5	Карта флеш-памяти	1198	5,98	1
9	7	Лекало	1093	5,46	1
10	11	Оберточная бумага	1025	5,12	1
11	24	Конфеты	888	4,43	2
		Поздравительная открытка			
12	9	Мыло	832	4,15	1

выявить наиболее популярные товарные наборы, состоящие из более, чем 1 предмета;

На вкладке **Дерево правил** предлагается еще один удобный способ отображения множества ассоциативных правил, которое строится либо по условию, либо по следствию. При построении дерева правил по условию, на первом (верхнем) уровне находятся узлы с условиями, а на втором уровне – узлы со следствием. В дереве, построенном по следствию, наоборот, на первом уровне располагаются узлы со следствием.

Справа от дерева расположен список правил, построенный по выбранному узлу дерева, например по правилу №5:

предложить рекламные акции типа «Каждому купившему товары А и В, товар С в подарок».

Правило №5; Следствие: Конфеты				
Условие	Поддержка		Достоверность, %	Лифт
	Кол-во	%		
Зубная паста И Поздра...	272	1,36	40,60	2,36
Зубная паста	786	3,92	24,60	1,43
Поздравительная откр...	888	4,43	29,30	1,704

Для каждого правила отображаются поддержка и достоверность и лифт. Если дерево построено по условию, то сверху списка отображается условие правила, а список состоит из его следствий. Тогда правила отвечают на вопрос, что будет при таком условии. Если же дерево построено по следствию, то сверху списка отображается следствие правила, а список состоит из его условий. Эти правила отвечают на вопросы, что нужно, чтобы было заданное следствие или какие товары нужно продать для того, чтобы продать товар из следствия.

Сохраните сценарий под именем **assosiation.ded**.

СОДЕРЖАНИЕ ОТЧЕТА

1. Цель работы.
2. Краткое описание хода работы.
3. Исходные данные
4. Выявленные ассоциативные правила
5. Ответы на вопросы.
6. Заключение.

Вопросы:

1. Какой алгоритм генерации ассоциативных правил имеется в Deductor?
2. Какие входные поля набора данных необходимы для запуска обработчика Ассоциативные правила в Deductor?
3. Какие специализированные визуализаторы предлагаются к узлу-обработчику Ассоциативные правила?

Лабораторная работа №4. Распознавание образов данных (Сеть Хемминга)

Цель работы. Изучение функционирования нейроподобных элементов в виде сети Хемминга. Разработка программы для распознавания образов при преобразовании информации.

Общие сведения

В настоящее время дальнейшее повышение производительности компонентов связывает с системами, обладающими свойствами массового параллелизма.

Одна из таких систем – это нейрокомпьютер, использующий искусственную нейросеть. Искусственная нейросеть (ИНС) – это параллельная структура, которая естественным образом реализует принцип потока данных. Обычно под ИНС понимается набор элементарных нейроподобных преобразователей информации – нейронов, соединенных друг с другом каналами обмена информацией для их совместной работы.

Сформировались две ветви исследований. Первая, нейробиологическая, основывается на моделировании работы живого мозга, имея цель объяснить, каким образом в нем отображаются сложные объекты и связи между ними, как устанавливается соответствие между хранящейся и поступающей извне информацией, и другие вопросы, касающиеся функционирования мозга. Второе направление исследований направлено на решение с помощью ИНС задач переработки информации в различных областях знаний, особенно плохо формализованных, где существующие модели субъективны и неадекватны. Впечатляющие результаты использования ИНС достигнуты при распознавании образов, при построении ассоциативной памяти, при создании самообучающихся Экспертных систем, при решении оптимизационных задач большой размерности.

Предложено и изучено большое количество моделей нейросетей. основными являются три типа сетей, которые соответствуют трем известным методам обучения: самоорганизации, последовательному подкреплению знаний, обучению с учителем.

Теоретическая часть. Сеть Хемминга (СХ) представляет сеть с двухслойной топологией, прямой связью между слоями и с обучением с супервизором. Число нейронов N на входном слое равно размерности векторов памяти, а число нейронов в выходном слое равно числу M векторов памяти.

Сеть состоит из двух слоев. Первый и второй слои имеют по m нейронов, где m – число образцов. Нейроны первого слоя имеют по n синапсов, соединенных со входами сети (образующими фиктивный нулевой слой). Нейроны второго слоя связаны между собой ингибиторными

(отрицательными обратными) синаптическими связями. Единственный синапс с положительной обратной связью для каждого нейрона соединен с его же аксоном.

Работы сети заключается в нахождении расстояния Хэмминга от тестируемого образа до всех образцов. Расстоянием Хэмминга называется число отличающихся битов в двух бинарных векторах. Сеть должна выбрать образец с минимальным расстоянием Хэмминга до неизвестного входного сигнала, в результате чего будет активизирован только один выход сети, соответствующий этому образцу.

На стадии инициализации весовым коэффициентам первого слоя и порогу активационной функции присваиваются следующие значения:

$$w_{ik} = \frac{x_i^k}{2}, i=0..n-1, k=0..m-1 \quad (5)$$

$$T_k = n/2, k = 0..m-1 \quad (6)$$

Здесь x_i^k – i -ый элемент k -ого образца.

Весовые коэффициенты тормозящих синапсов во втором слое берут равными некоторой величине $0 < \square < 1/m$. Синапс нейрона, связанный с его же аксоном имеет вес $+1$.

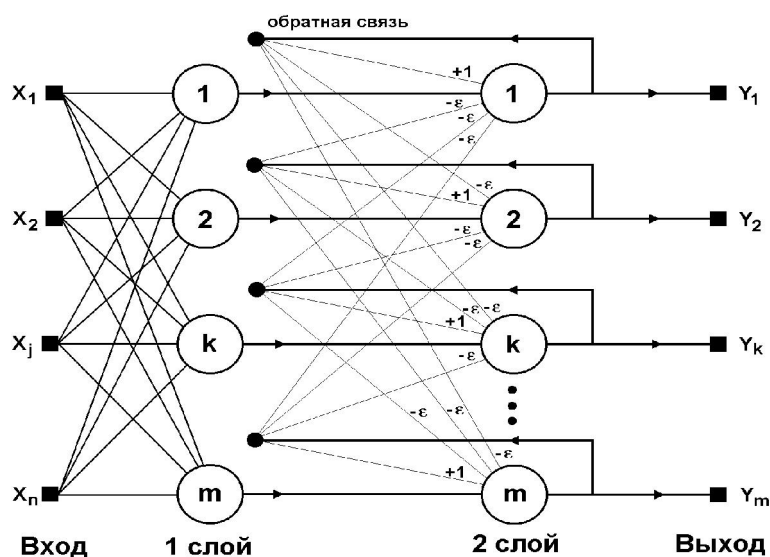


Рис.1 . Схема сети Хемминга

Ход работы. Алгоритм функционирования сети Хэмминга следующий:

1. На входы сети подается неизвестный вектор $X = \{x_i; i=0..n-1\}$, исходя из которого рассчитываются состояния нейронов первого слоя (верхний индекс в скобках указывает номер слоя):

$$y_j^{(1)} = s_j^{(1)} = \sum_{i=0}^{n-1} w_{ij} x_i + T_j, j=0..m-1 \quad (7)$$

После этого полученными значениями инициализируются значения аксонов второго слоя: $y_j^{(2)} = y_j^{(1)}, j = 0..m-1$

2. Вычислить новые состояния нейронов второго слоя:

$$s_j^{(2)}(p+1) = y_j(p) - \varepsilon \sum_{k=0}^{m-1} y_k^{(2)}(p), k \neq j, j = 0 \dots m-1$$

и значения их аксонов:

$$y_j^{(2)}(p+1) = f[s_j^{(2)}(p+1)], j = 0 \dots m-1$$

Активационная функция f имеет вид порога (рис. 2б), причем величина F должна быть достаточно большой, чтобы любые возможные значения аргумента не приводили к насыщению.

3. Проверить, изменились ли выходы нейронов второго слоя за последнюю итерацию. Если да – перейди к шагу 2. Иначе – конец.

Из оценки алгоритма видно, что роль первого слоя весьма условна: воспользовавшись один раз на шаге 1 значениями его весовых коэффициентов, сеть больше не обращается к нему, поэтому первый слой может быть вообще исключен из сети (заменен на матрицу весовых коэффициентов), поэтому так можно сделать в ее конкретной реализации,

Сеть классифицирует произвольные бинарные или аналоговые образы $x = (x_1 \dots x_n)$ в один из M классов. При этом начальное значение $y^j(0)$ нейронов в выходном слое определяется двумя способами в зависимости от характера векторов памяти. Но в обоих случаях вектор стимула x с начало нормируется.

Если векторы памяти являются бинарными, то $y^j(0)$ соответствует перекрытиям нормированного вектора стимула с нормированными векторами памяти. Если векторы памяти являются аналоговыми, то $y^j(0)$ выбирается в соответствии с величиной расстояния Хемминга между нормированными векторами памяти и стимула, с помощью пороговой функции F .

После формирования начальных значений нейронов выходного слоя (в этом слое все нейроны связаны между собой) выполняются итерации, про которых самодействие каждого нейрона является положительным, а вклад остальных нейронов этого слоя отрицателен. С помощью итераций выделяется тот нейрон, у которого значение $y^j(0)$ было максимальным, т. е. итерации прекращаются, когда только один из нейронов имеет ненулевое значение, а номер этого нейрона и определяет результат классификации.

4. Алгоритм программы

1. Инициализация весов (можно взять из файла “obraz.txt”).
2. Ввод распознаваемого образа.
3. Определение кодовое расстояние $d[j]$.
4. Определение максимального сходство искомого образа с одним из исходных.

$$y[j] = \text{porog} - d[j].$$

5. Вывод результата.

Пример результата работы сети Хемминга при распознавании образа цифры.

1. Введите образ:

1111

1111

0011

0011

0011

0011

Результат: 7

Введите образ:

1111

0011

1111

1111

1100

1111

2. Результат: 2

Содержание отчета

1. Цель работы.
2. Краткое описание хода работы.
3. Ответы на вопросы.
4. Листинг программы
5. Заключение.

ВОПРОСЫ

1. Приведите примеры использования сети Хемминга.
2. Сколько слоев имеет сеть Хемминга?
3. Какую роль играют обратные связи?
4. Каким образом определяется распознаваемый образ?
5. Какой вид имеет активационная функция для сети Хемминга?
6. Совпадает ли количество входов и выходов в сети Хемминга?

Приложение:

1. Варианты интерфейса с результатами работы программы

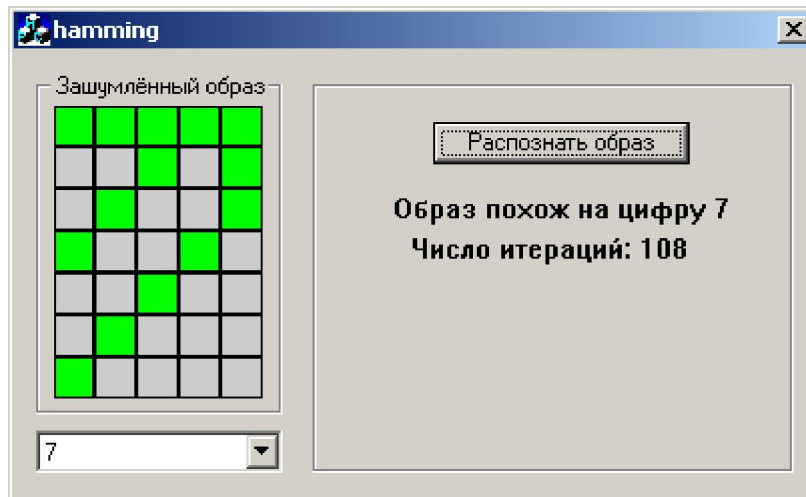


Рис. 2. Зашумлённая цифра 7.

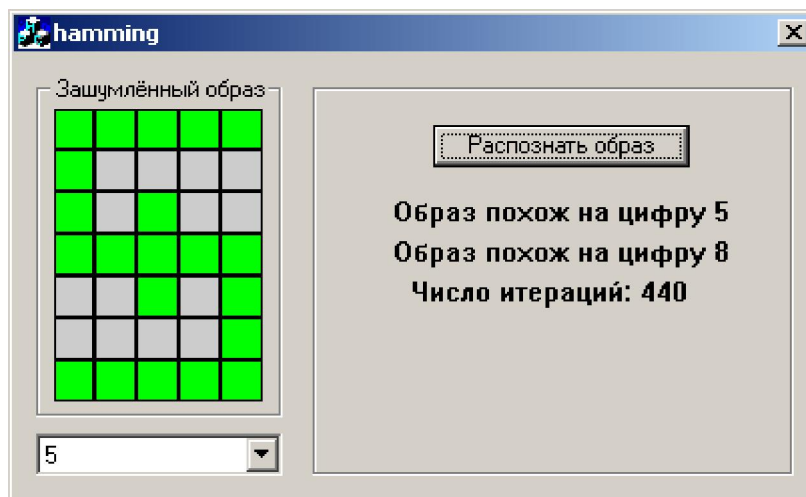
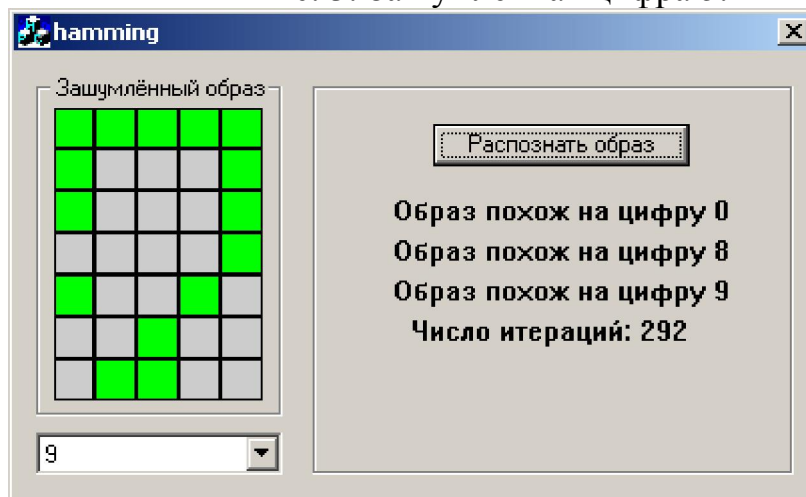


Рис. 3. Зашумлённая цифра 5.



1. Листинг процедур

Процедура распознавания зашумлённого образа
 void CHammingDlg::OnRecognition()

```

{
    // TODO: Add your control notification handler code here
    Invalidate();
    UpdateWindow();

    int k, i, j, iter=0;
    for(k=0; k<m_N; k++) {y1[k]=y2[k]=y2new[k]=tmp[k]=0.0;}

    //Инициализация сети
    T = m_m*m_n/2.0; srand((unsigned)time(NULL));
    double e = rand()%((m_m*m_n))/(m_m*m_n*100.0);

    //Вычисление состояния нейронов первого слоя
    for(k=0; k<m_N; k++) {
        for(i=0; i<m_m; i++)
            for(j=0; j<m_n; j++)
                y1[k] += m_image[i][j] * m_w[k][i*m_n+j];
        y1[k]+=T;
    }

    //Инициализация значений аксонов второго слоя полученными
    значениями
    for(k=0; k<m_N; k++) y2new[k] = y1[k];

    do
    {
        for(k=0; k<m_N; k++) y2[k] = y2new[k];

        //Вычислить новые состояния нейронов второго слоя
        for(k=0; k<m_N; k++) {
            double sum=0;
            for(j=0; j<m_N; j++) if(j!=k) sum += y2[j];
            y2new[k] = y2[k] - e * sum;
        }

        //Вычислить значения аксонов второго слоя
        for(k=0; k<m_N; k++) {
            if(y2new[k]<0) y2new[k]=0;
            else if(y2new[k]>=m_m*m_n) y2new[k]=m_m*m_n;
        }
        iter++;
    }
    while(Change());

    //Определение схожих образов
    double etalon=-100; int index[10], l=0;
    for(k=0; k<m_N; k++) index[k] = -1;
    for(k=0; k<m_N; k++)
        if(y2new[k]>etalon)
            {etalon = y2new[k]; index[0] = k;}

    for(k=0; k<m_N; k++)
        if(fabs(y2new[k]-etalon)<0.001 && k!=index[0])
            index[++l] = k;

    CDC *pdc2 = m_results.GetDC();
    pdc2->SetBkMode(TRANSPARENT);
    char str[50]={0}; int s=0;
    for(k=0; k<m_N; k++) {
        if(index[k]>-1) {
            sprintf(str,"Образ похож на цифру %d", index[k]);
            pdc2->TextOut(40, 60+20*s, str); s++;
        }
    }
}

```



```

    }
    sprintf(str, "Число итераций: %d", iter);
    pdc2->TextOut(50, 60+20*s, str); s++;
    m_results.ReleaseDC(pdc2);
}

Процедура проверки изменения выходов нейронов второго слоя
int CHammingDlg::Change()
{
    for(int k=0; k<m_N; k++)
        if(fabs(y2new[k]-y2[k])>0.001) return 1;
    return 0;
}

Считывание эталонных образов из файла
FILE *f;
if((f=fopen("cifir.txt","r"))==NULL){printf("Невозможно открыть
файл!\n"); return FALSE;}
for(int k=0; k<m_N; k++)
    for(i=0; i<m_m; i++)
        for(int j=0; j<m_n; j++)
            fscanf(f, "%d", &m_etalon[k][i*m_n+j]);
fclose(f);

```

Лабораторная работа № 5. Кластерная обработка данных (карты Кохонена)

Цель работы. Научиться использовать метод кластерной обработки данных в виде самоорганизующихся карт Кохонена».

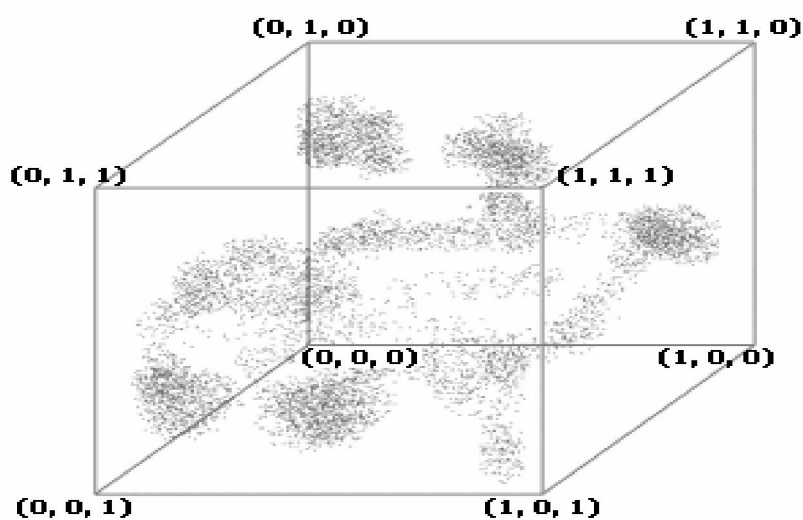
Теоретическая часть. Существуют задачи анализа данных, которые затруднительно представить в числовой форме. При этом нужно извлечь данные, принципы отбора которых заданы нечетко: выделить надежных партнеров, определить перспективный товар и т.п. Также необходимо на основании имеющихся априорных данных получить прогноз на дальнейший период. Существует метод, позволяющий автоматизировать все действия по поиску закономерностей – метод анализа с использованием самоорганизующихся карт Кохонена.

Самоорганизующаяся карта Кохонена (англ. Self-organizing map — SOM) — нейронная сеть с обучением без учителя, выполняющая задачу визуализации и кластеризации. Является методом проецирования многомерного пространства в пространство с более низкой размерностью (чаще всего

двумерное), применяется также для решения задач моделирования, прогнозирования и др.

Каждый объект характеризуется набором различных *параметров*, описывающих его состояние. Например, параметрами будут данные из финансовых отчетов. Эти параметры часто имеют числовую форму или могут быть приведены к ней. Таким образом, нам надо на основании анализа параметров объектов выделить схожие объекты и представить результат в форме, удобной для восприятия. Эти задачи решаются самоорганизующимися картами Кохонена.

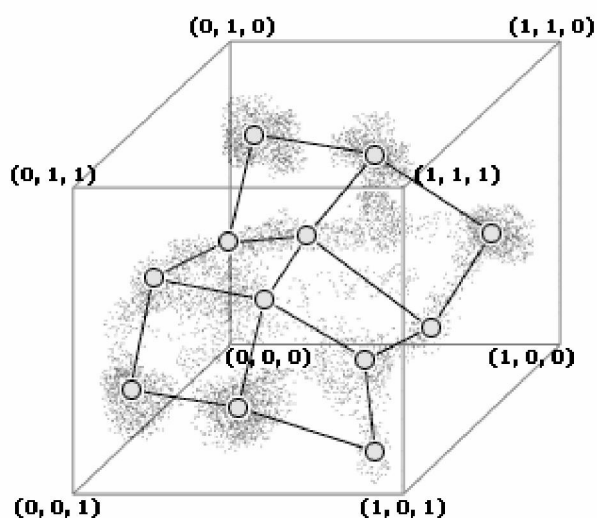
Пусть объект расположен в трехмерном пространстве. Тогда каждый объект с признаками можно представить в виде точки в данном пространстве, и пронормировать эти признаки в интервал $[0,1]$, в результате чего все точки попадут в куб единичного размера. Отобразим эти точки.



Расположение объектов в пространстве

На рисунке видно, как расположены объекты в пространстве, причем легко заметить участки, где объекты группируются, т.е. у них схожи параметры, значит, и сами эти объекты, скорее всего, принадлежат одной группе. Но так можно поступить только в случае, когда признаков немного. Значит, надо найти способ, преобразующий данную систему в простую для восприятия, желательно двумерную систему (потому что уже трехмерную картинку

невозможно корректно отобразить на плоскости) так, чтобы соседние в искомом пространстве объекты оказались рядом и на полученной картинке. Для этого используем самоорганизующуюся карту Кохонена. В первом приближении ее можно представить в виде «гибкой» сети. Предварительно «скомкав», бросаем сеть в пространство признаков, где уже имеются объекты, и далее поступаем следующим образом: берем один объект (точку в этом пространстве) и находим ближайший к нему узел сети. После этого узел подтягивается к объекту (т.к. сетка «гибкая», то вместе с этим узлом так же, но с меньшей силой подтягиваются и соседние узлы). Затем выбирается другой объект (точка), и процедура повторяется. В результате получится карта, расположение узлов которой совпадает с расположением основных скоплений объектов в исходном пространстве. Полученная карта обладает следующим замечательным свойством – узлы ее расположились таким образом, что объектам, похожим между собой, соответствуют соседние узлы карты. Теперь находим, какие объекты попали в какие узлы карты. Это также определяется ближайшим узлом – объект попадает в тот узел, который находится ближе к нему.



Вид пространства после наложения карты

В результате данных операций объекты со схожими параметрами попадут в один узел или в соседние узлы. Хотя задача поиска похожих объектов и их группировки решена, но на этом возможности карт Кохонена не заканчиваются. Они позволяют также представить полученную информацию

в простой и наглядной форме путем нанесения раскраски полученной карты (точнее ее узлы) цветами, соответствующими интересующим нас признакам объектов.

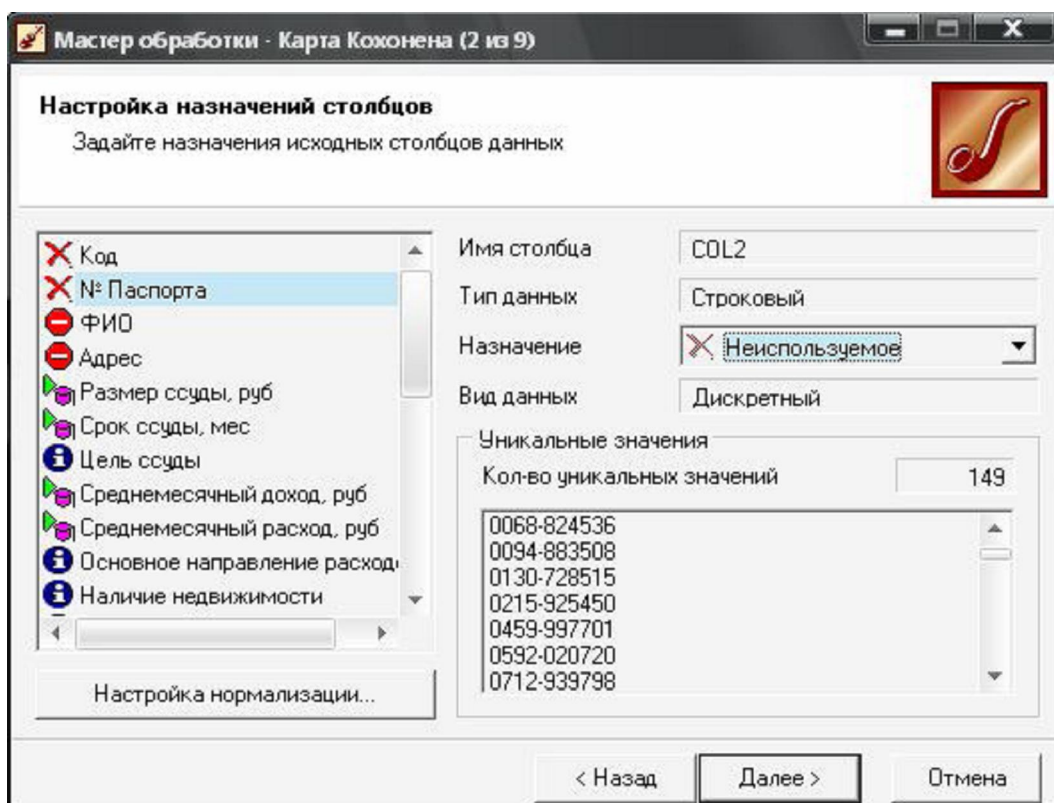
Также можно получить информацию о зависимостях между параметрами. Нанеся на карту раскраску, соответствующую различным статьям отчетов, можно получить так называемый атлас, хранящий в себе информацию о состоянии рынка. Можно анализировать, сравнивать расположение цветов на раскрасках, порожденных различными параметрами, тем самым получая все новую информацию.

При всем этом описанная технология является универсальным методом анализа. С ее помощью можно анализировать различные стратегии деятельности, производить анализ результатов маркетинговых исследований, проверять кредитоспособность клиентов и т.д.

Ход работы

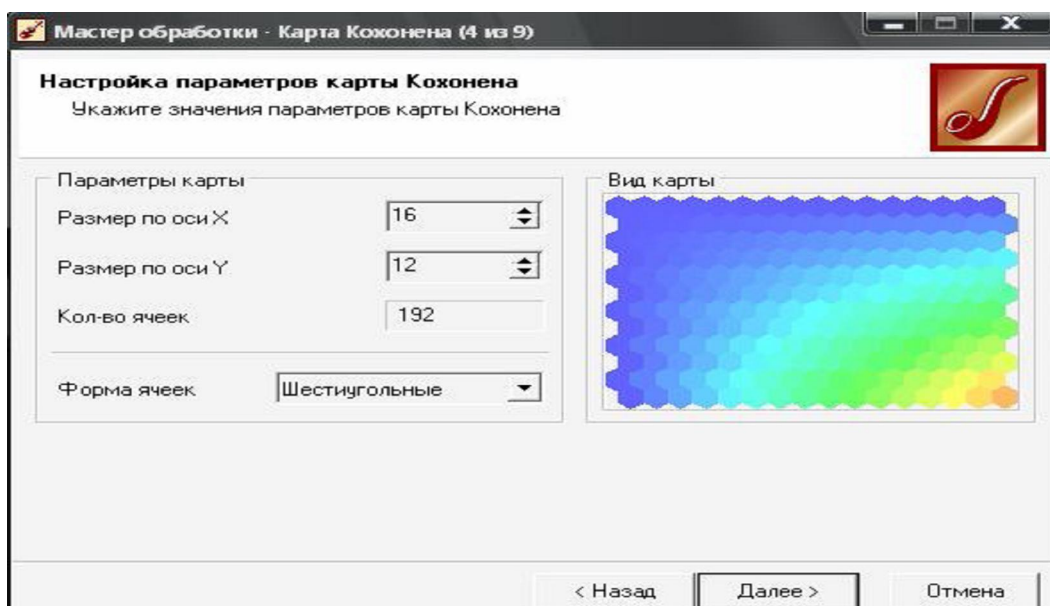
Импортируйте в АП «Deductor» исходные данные из файла C:\Program\Files\BaseGroup\Deductor\Samples\CreditSample.txt. Процесс построения карты Кохонена состоит из 10 этапов.

Запустите *мастер обработки*, в котором в разделе «Data Mining» выберете способ обработки данных «Карта Кохонена», нажмите «Далее». В окне настройки назначения столбцов необходимо обозначить столбцы «Код» и «№ паспорта» как «Неиспользуемые» (так как значения этих столбцов уникальны, а это не позволит их классифицировать по общим признакам). Определите поле «Давать кредит» как «Выходное».



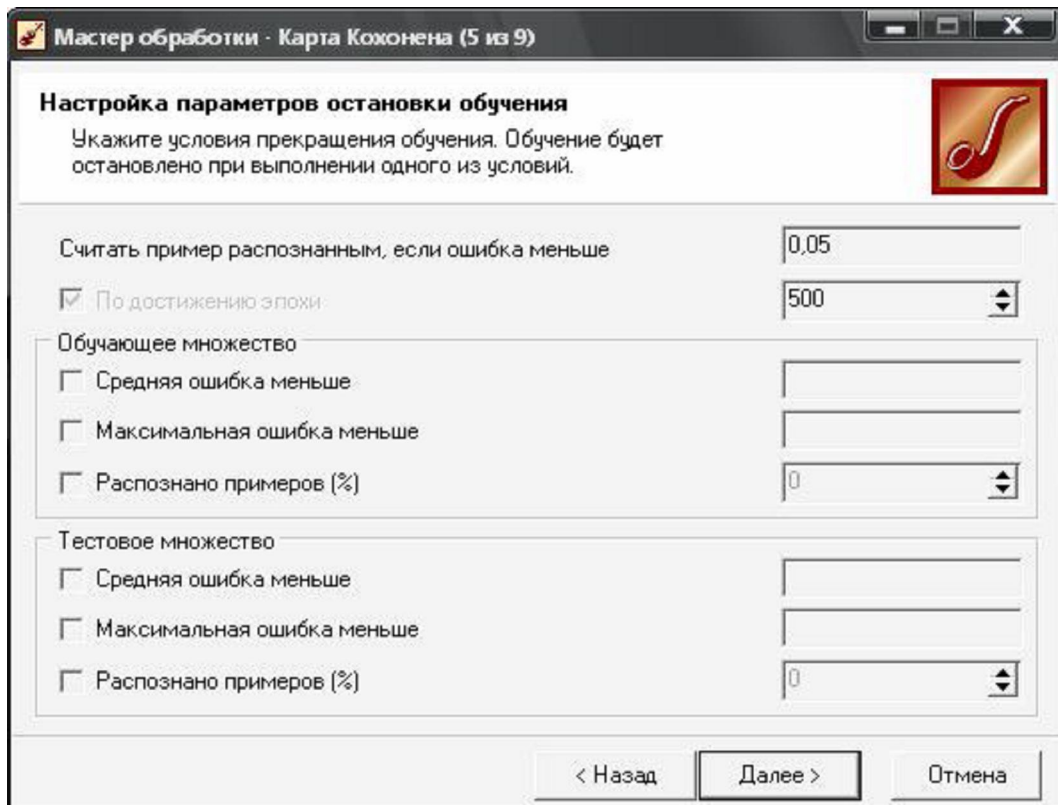
Пример настройки назначений столбцов

Настройку обучающей выборки и параметров карты Кохонена можно оставить без изменений.



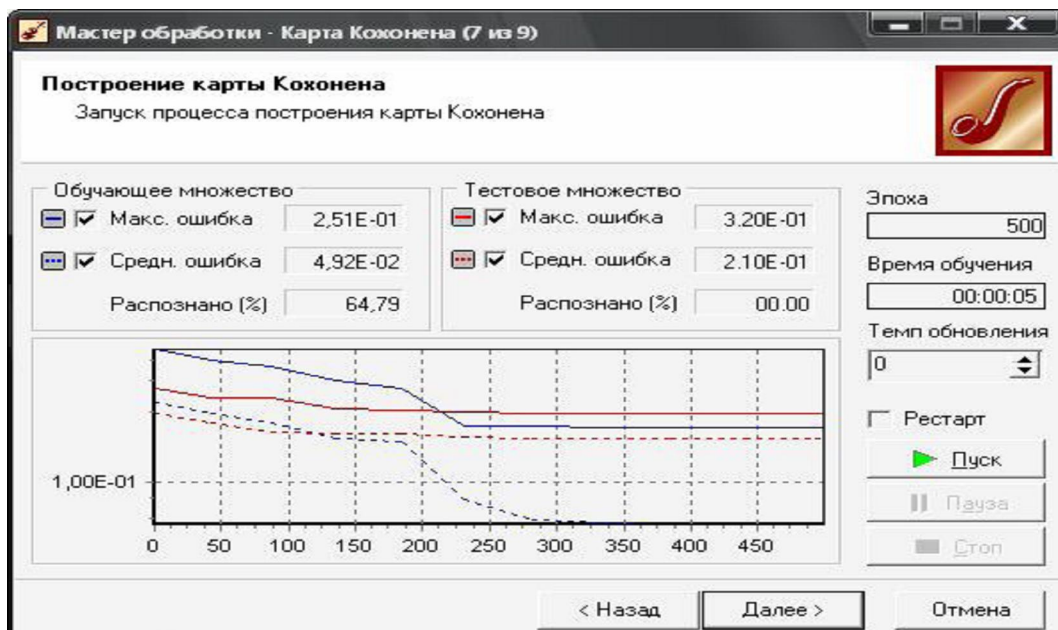
Настройка параметров карты Кохонена

Настройте параметры остановки обучения, указав уровень допустимой погрешности, если он будет превышен, анализ данного множества будет прекращен. Можно оставить значения «по умолчанию».



Настройка параметров остановки обучения.

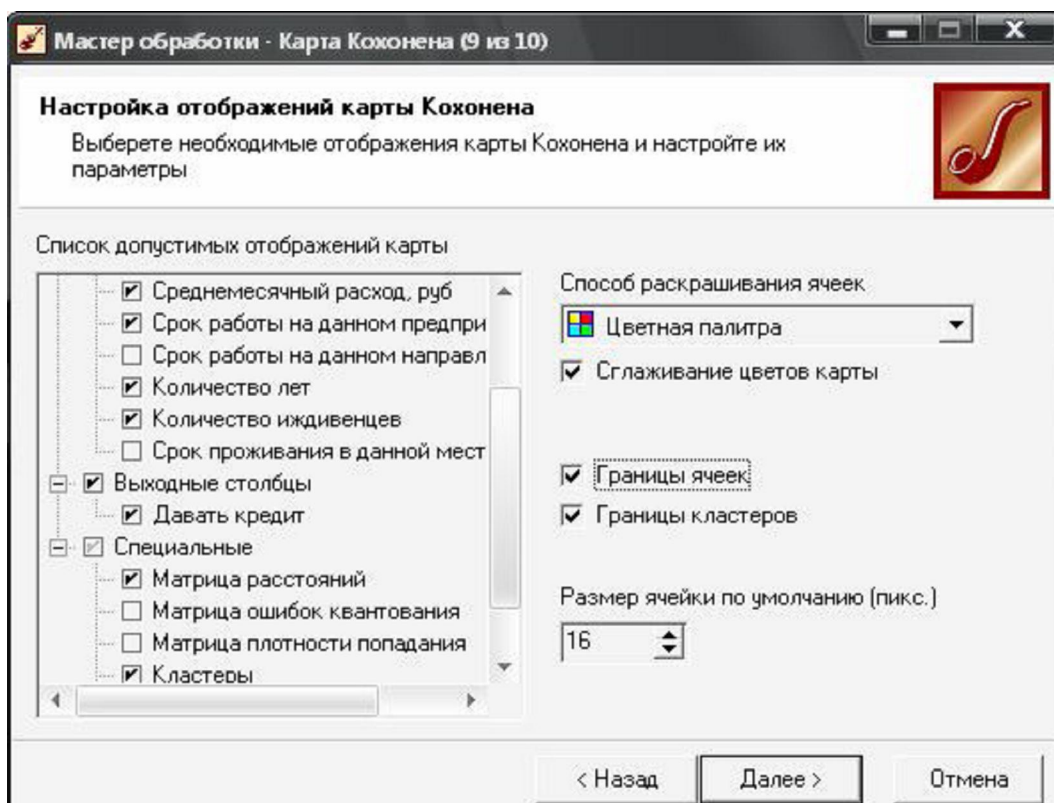
Настройку параметров обучения также оставьте без изменений. Далее запустите процесс построения карты Кохонена, нажав кнопку «Пуск».



Итог построения карты Кохонена

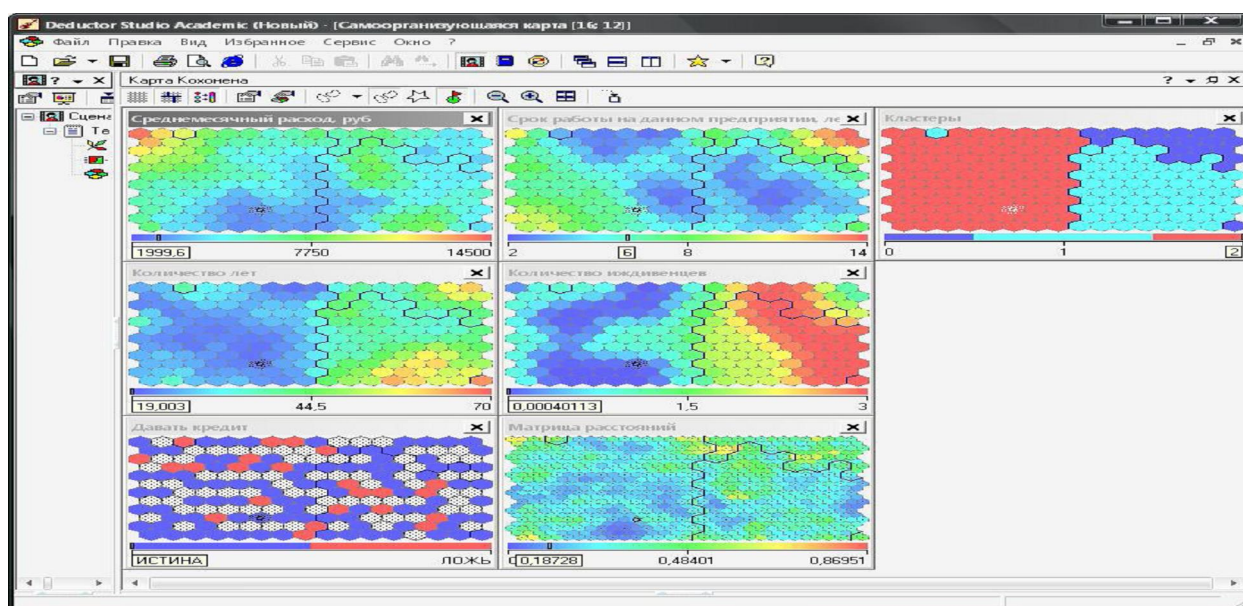
На вкладке «Выбор способа отображения данных» поставьте галочку напротив пункта «Самоорганизующаяся карта Кохонена». Теперь необходимо провести настройку отображения карты:

отметьте разделы «Давать кредит» и «Кластеры» и другие разделы – по желанию.



Настройка отображений карты Кохонена

Далее задайте имя, метку и описание карты (по желанию). В результате получатся карты Кохонена, подобные изображенным на рисунке.



Примеры карт Кохонена

Щелкнув левой клавишей мыши по любому шестиугольнику на любой карте, выделяются соответствующие ему ячейки на остальных картах, в том

числе на картах «Давать кредит» и «Кластеры». При этом на шкалах в нижней части карт отобразятся значения соответствующих параметров.

Задание

1. Выполните описанные выше действия по построению карт Кохонена. Проанализируйте результаты, что можно сказать о вероятности возврата кредита для групп 2, 3 и 4?
2. Используя различные отображения карты Кохонена, постройте 3-4 правила выдачи кредитов.

Содержание отчета

1. Цель работы.
2. Краткое описание хода работы
3. Вид карт Кохонена
4. Ответы на вопросы.
5. Листинг программы
6. Заключение.

Вопросы:

1. Для чего используются карты Кохонена?
2. По какому принципу происходит перенос многомерного пространства на пространство меньшей размерности?
3. Что обозначает фраза “Победитель берет все”
4. Какие метрики используются при разбиении на кластеры
5. Целесообразность применения карт Кохонена при кластеризации данных

Лабораторная работа №6. Классификация данных

Цель работы. Целью данной лабораторной работы является исследование способности нейронной сети решать задачи классификации. Сеть необходимо обучить классификации по пяти классам по 10-20 числовым признакам. Используемая модель: одномерная сеть Кохонена.

Теория. Задача классификации заключается в идентификации объекта и отнесении его к одному из нескольких множеств. При этом задача классификации предполагает, что множества попарно не пересекаются. Применительно к нейронным сетям задачу классификации можно поставить

следующим образом: пусть имеется N множеств D_1, D_2, \dots, D_n признаков объектов. Сеть обучается на парах векторов X и Y , где: $X = (x_1, x_2, \dots, x_m)$ – входной вектор признаков; $Y = (y_1, y_2, \dots, y_n) = C(X)$ – выходной вектор, классифицирующий вектор X . При этом возможно несколько случаев:

1. $Y = k$, классификатор имеет скалярный характер. k – порядковый номер множества, к которому относится X .
2. $Y = (y_1, y_2, \dots, y_k, \dots, y_n)$. При этом только $y_k=1$, остальные компоненты вектора равны 0. Таким образом работает звено Кохонена
3. $Y = (y_1, y_2, \dots, y_n)$. При этом каждая компонента y_k характеризует степень принадлежности к множеству D_k . В режиме нормального функционирования сеть по входному вектору X выдает вектор Z по правилам, аналогичным описанным для векторов Y . Точность решения определяется статистикой: сколько раз вектор Z правильно классифицировал объект с признаками X , соотнося его с той или иной группой D_k . Для пункта 3 возможно вычисление погрешности, при наличии функции-скаляризатора степени принадлежности вектора X к множествам D_k . Задача может быть дополнена введением «шума», однако смысл от этого не изменится. Шум лишь изменит границы областей D_1, \dots, D_n .

Для решения задачи классификации с линейной и нелинейной разделимостью классов используются классические модели нейронных сетей: многослойный персептрон и рекуррентные сети на его основе, радиально-базисные сети, сети Кохонена, гибридные сети, рекуррентные самоорганизующиеся сети. При наложении множества объектов друг на друга, то задача классификации становится более общей и предполагает, что объект характеризуется степенью принадлежности к тому или иному множеству, то есть имеет место задача классификации с вероятностной разделимостью классов.

Ход работы:

1. Необходимо выбрать предметную область, отобрать не менее 10 числовых характеристик объектов и задать их диапазоны.

2. Сгенерировать обучающую выборку размерностью от 10 до 20 примеров для каждого класса. Предусмотреть нормализацию входных векторов.
3. Написать программу, имитирующую работу нейронной сети Кохонена
4. Провести обучение сети Кохонена по алгоритму Кохонена с прямоугольным соседством.
5. Исследовать эффективность алгоритмов обучения от значения коэффициента обучения.
6. Исследовать зависимость погрешности классификации от алгоритма обучения.
7. Исследовать зависимость погрешности классификации от объёма обучающей выборки.
8. Исследовать зависимость погрешности классификации от числа итераций обучения.

Содержание отчета

1. Цель лабораторной работы.
2. Краткое описание хода работы.
3. Файл обучающей выборки.
4. Результаты работы нейросетевого классификатора. Выводы по результатам работы.
5. Ответы на вопросы.
6. Ответы на вопросы.

Вопросы

1. В чем суть классификации данных?
2. Отличие классификации от кластеризации.
3. Какие существуют методы классификации кроме нейросетевого?
4. Целесообразность применения нейросетевого метода для классификации данных.
5. Как определяется погрешность классификации?

Цель работы: изучить существующие алгоритмы распознавания образов в виде прецедентов.

Теоретическая часть. Обзор алгоритмов распознавания образов

1. Алгоритм ближайшего соседа Процедура взятая в качестве решающего правила: оставить в памяти машины все реализации обучающей выборки и классифицируемую точку (образ) отнести к тому классифицирующему образу, чья реализация оказалась ближайшей. Это – правило ближайшего соседа. Учитывая, что результаты реальных измерений свойств могут быть «зашумлены», можно использовать правило k ближайших соседей: если больше половины из k ближайших соседей принадлежат образу i , то и объект (точка) q относится к образу i .
2. Метод потенциальных функций Иногда в «голосовании» принимают участие все реализации обучающей выборки, но с разными весами, зависящими от расстояния. Смысл алгоритма заключается в излучении точками каждого класса потенциала, величина убывающего с расстоянием. Характер убывания может быть самым различным. В точке q определяется «притяжение» к каждому из классов. Если окажется, что какая-то точка распознается с ошибкой, то можно изменить картину потенциального поля. Доказана сходимость алгоритма к оптимальному при увеличении обучающей выборки и конечность числа шагов при не слишком сложной («вычурной») картине потенциального поля.
3. Алгоритм STOPL Сложность заключается в комбинаторном характере задачи - вечный поиск компромиса между требуемой скоростью и затратами памяти. Для сокращения перебора выбирают точки пограничные точки наибольшего риска. Пусть r_{in} - расстояние до своей ближайшей точки, r_{out} - до чужой. Тогда отношение $W = r_{in} / r_{out}$ характеризует величину риска быть опознаной в качестве чужого образа. Среди точек каждого образа выбираются в качестве

прецедентов по одной точке с максимальным значением W . После этого распознаются все точки обучающей выборки с опорой на прецеденты по методу ближайшего соседа. Среди опознанных неправильно вновь выбирается точка с максимальным значением W и ею пополняется список прецедентов. Процесс повторяется до тех пор, пока все точки обучающей выборки не будут распознаваться правильно.

4. Метод дробящихся эталонов - алгоритм ДРЭТ Стремимся опять же к безошибочному распознаванию обучающей выборки, но для выбора прецедентов используется метод покрытий обучающей выборки каждого образа простыми фигурами (которые можно усложнять при необходимости). В качестве покрывающих фигур выбираются гиперсферы. Для каждого образа строится гиперсфера минимального радиуса, покрывающая все его точки (реализации). Значения радиусов гиперсфер и расстояния между центрами позволяет определить непересекающиеся гиперсферы. Это - эталоны первого поколения. Если сферы пересекаются, но пересечения пусты, то такие гиперсферы (центры и радиусы) также относятся к эталонам первого поколения. При этом область пересечения считается принадлежащей 21 гиперсфере меньшего радиуса. Эталоны второго поколения строятся только для пересечений, содержащих точки двух или более образов. Если эталоны второго поколения также пересекаются, то процесс продолжается до полной надежности распознавания обучающих выборок. Контрольные образцы классифицируются по попаданию в эталонные гиперсферы. Если контрольная точка не попадает в гиперсферу, то определяется ближайшие и далее используется метод ближайшего соседа.
5. Логические решающие правила. В задачах распознавания образов зачастую требуется, чтобы компактные точки в n -мерном пространстве признаков были компактными и несовпадающими в проекциях на координатные оси (гипотеза локальной компактности). Конечно, это не

всегда так. Например, все трехмерные геометрические фигуры (сферы, пирамиды, параллелепипеды) могут быть синими. Но здесь мы сталкиваемся с проблемой ситуативной информативности признаков. Очевидно, фигуры с геометрической точки зрения не будут классифицированы по цвету. Обычно, проекции на координатные оси пересекаются, но могут выглядеть по-разному и это дает надежду, что комбинация несовпадающих проекций на несколько осей позволит построить эффективное решающее правило за счет сокращения размерности признакового пространства. Решающие правила, учитывающие разницу в проекциях разных образов имеют вид «Если-условие-то-следствие» и получили наименование логических решающих правил (ЛРП).

6. Алгоритм CORAL Выделяется подмножество значений $X_{jv} \in X_j$

признака X_j . Для сильных признаков – это интервал значений, для шкал порядка – ряд соседних порядковых позиций, для шкал наименований – одно или несколько имен. Обозначается факт, что некоторое значение признака X_j объекта a_i принадлежит подмножеству X_{jv} как $J(a_i, X_{jv})$, факт попадания объекта a в область v , образованную границами подмножеств X_{jv} , т.е. в гиперпараллелепипед, запишется в виде логического высказывания: $n' \leq n$, где n – размерность признакового пространства. Число n' называется длиной высказывания. Логической закономерностью называется высказывание, удовлетворяющее двум условиям: где w – индекс объектов своего образа, w^- – индекс всех чужих объектов, m_w – число всех своих объектов, m_{w^-} – число чужих объектов, m_{wS} – число своих, удовлетворяющих высказыванию S , m_{wS^-} – число чужих,

удовлетворяющих тому же высказыванию S , $23 \alpha, \beta$ - некоторые величины в диапазоне от 0 до 1. Желательно, чтобы высказывание S отбирало больше своих объектов и поменьше чужих, т.е. чтобы α было как можно большим, а β - как можно меньшим. Набор закономерностей называется покрывающим для образа w , если для любой его реализации выполняется хотя бы одна закономерность из этого набора. Желательно, чтобы число закономерностей в наборе было минимальным. Поиск закономерностей начинается с больших значений α (например, $\alpha = 1$) и малых значений $\beta \approx 0.02$. Просматриваются все подмножества значений первого случайно выбранного признака и находится высказывание S , удовлетворяющее условиям 1 и 2. Если такое не находится, то процесс поиска продолжается при более низком пороге α вплоть до $\alpha = 0.5$. Если и в этом случае нет результата, то увеличивается β в предельно допустимой доле «чужих среди своих». Если условия 1 и 2 не выполняются и при $\alpha = \beta = 0.5$., то делается переход к рассмотрению второго признака, случайным образом выбранным из оставшихся. Если условия 1 и 2 на каком-то шаге выполняются, то объекты своего образа, удовлетворяющие высказыванию S , из дальнейшего рассмотрения исключаются. Для оставшихся объектов образа w длина высказывания увеличивается на единицу. Процесс продолжается до получения покрывающего набора закономерностей для всех объектов образа w . Аналогично строятся покрывающие наборы и для всех других распознаваемых образов. Можно потребовать, чтобы алгоритм делал для каждого образа не по одному, а по несколько покрывающих наборов, что соответствует высказыванию Р. Фейнмана о том, что можно говорить о понимании явления, если в состоянии объяснить его несколькими способами. С этой целью после получения первого покрытия исключают из рассмотрения первый признак, включенный в это покрытие, и процесс поиска закономерностей начинается с другого случайно выбираемого

признака. Распознавание контрольного объекта q с помощью покрывающих наборов закономерностей сводится к проверке того, каким высказываниям удовлетворяют его характеристики. Если таких высказываний одно или несколько и все они находятся в списке образа w , то объект q распознается в качестве реализации образа w . Если же объект q удовлетворяет закономерностям нескольких образов, то решение принимается в пользу того образа, которому принадлежит закономерность с наибольшим значением величины R_{ws} . Анализ общего списка закономерностей может показать, что некоторые признаки из исходной системы X в них отсутствуют. Это означает, что они оказались неинформативными и в процессе принятия решения на них можно не обращать внимания. Для каждого i -го образа подмножество информативных признаков может оказаться разным. А это значит, что при проверке гипотезы о принадлежности объекта к тому или иному образу, нужно анализировать не все пространство признаков, а его информативное подпространство, что хорошо согласуется с интуитивными методами неформального распознавания.

7. Метод случайного поиска с адаптацией (алгоритм СПА) Единичный отрезок разбивается на g участков одинаковой длины ($1/g$). Каждому участку сопоставляется свой признак: первому - первый, второму - второй и т.д. Запускается датчик случайных чисел с равномерным распределением в диапазоне $0..1$. После n шагов работы выбирается n признаков. По числу ошибок оценивается качество распознавания. Такая процедура проделывается r раз. В итоге будет получен список оценок $L = (\alpha_1, \dots, \alpha_r)$. Теперь можно упорядочить список L по возрастанию α , т.е. по убыванию качества распознавания, и ввести систему поощрений и наказаний. Участки, соответствующие признакам, дающим лучший результат, увеличиваются, «худшие» - уменьшаются, но так, чтобы суммарная длина по-прежнему была равна. Испытывают r новых признаковых подсистем, но теперь вероятность

попадания на «лучшие» участки выше, чем на плохие. Продолжают процесс адаптации таким образом, что длина участков признаков, регулярно попадающих в самые информативные подсистемы, увеличивается на величину $h < 1/g$, а для самых неинформативных длины их участков уменьшаются. После некоторого количества циклов поиска и адаптации, процесс стабилизируется. Алгоритм СПА был протестирован для систем, в которых возможен полный перебор сочетаний признаков.

Задание. Составить программу, реализующую один из предложенных алгоритмов.

Вариант 1. Написать программу, выполняющую поиск слова в строке при помощи алгоритма ближайших соседей.

Вариант 2. Написать программу, реализующую поиск слова в текстовом файле с помощью алгоритма ближайших соседей.

Вариант 3. Написать программу, реализующую поиск слова в текстовом файле с помощью алгоритма STOPL.

Вариант 4. Написать программу, реализующую поиск слова в текстовом файле с помощью алгоритма CORAL.

Вариант 5. Написать программу, реализующую поиск слова в текстовом файле с помощью метода случайного поиска с адаптацией.

Вариант 6. Написать программу, реализующую поиск слова в текстовом файле с помощью алгоритма ДРЭТ.

Вариант 7. Написать программу, реализующую поиск слова в текстовом файле с помощью алгоритма потенциальных функций.

Содержание отчета

1. Цель лабораторной работы
2. Краткое описание хода работы
3. Схема алгоритма метода распознавания прецедентов
4. Результаты работы программы распознавания прецедентов
5. Выводы по результатам работы

6. Ответы на вопросы

Вопросы

1. Перечислите существующие алгоритмы распознавания образов.
2. Описание алгоритма ближайших соседей.
3. Описание алгоритма потенциальных функций.
4. Описание алгоритма STOPL.
5. Описание алгоритма CORAL.
6. Описание метода случайного поиска с адаптацией. 7. Описание алгоритма ДРЭТ.

Лабораторная работа № 8. Фильтрация данных

Цель работы. Ознакомиться с методом фильтрации данных фильтром Калмана. Оценить возможность применения фильтра Калмана на практике.

Теоретическая часть. Фильтр Калмана применяют в разных областях – от радиотехники до экономики. В экономике, например, измеряемой величиной могут быть курсы валют, колебания цен. Каждый день курс валют разный, т.е. каждый день “его измерения” дают нам разную величину. Измерения всегда идут с некоторой ошибкой. В простейшем случае описанное можно свести к следующему выражению: $z=x+y$, где x – истинное значение, которое нужно измерить, y – ошибка измерения, вносимая измерительным прибором, а z – измеряемая величина. Задача фильтра Калмана состоит в том, чтобы по измеренной z определить истинное значение x , т.е. необходимо отфильтровать (отсеять) из z истинное значение x – убрать из z искажающий шум y .

Для прогнозирования цены отдельного вида продукции предлагается рассмотреть следующую стохастическую нестационарную модель, описывающую процесс изменения цены:

$$p(t+\Delta t) - p(t) = f(t, p(t), z(t))\Delta t + F(t, p(t))\Delta\omega + c(v,)v(\Delta t, \Delta v).$$

или в виде уравнения в дифференциальной форме:

$$p(t) = p_0 + \int_0^t f(t, p(t), z(t)) dt + \int_0^t F(t, p(t)) d\omega(t) + \int_0^t \sum_i c(v_i) \delta(t - t_i) dt$$

где $p(t)$ – значение цены в момент времени t ;

$f(t, p(t), z(t))$ – скорость изменения цен во времени – непрерывная функция по переменной t , определяющая тренд цены и восстанавливаемая по статистической информации;

$F(t, p(t))$ – среднеквадратичное отклонение;

ω – винеровский процесс;

v – случайная пуассоновская мера с параметром $\lambda(t)$;

v_t – случайная величина, вызывающая приращение цены согласно закону $P(c(v)), P(c(v)) = \delta(v-a), a > 0$.

Первый элемент в правой части представленных выше уравнений, это тренд цены, зависящий от величины предложения и от потребности покупателей в данном товаре.

Кроме того, цены постоянно колеблются около своего тренда вследствие аддитивного воздействия на них случайных факторов, и будем считать, что приращения цены - независимые величины. Можно утверждать, что в краткосрочном периоде закон распределения приращений процесса изменения цен близок к нормальному. Также будем считать, что трендовая составляющая не оказывает влияния на случайную, и наоборот, случайные изменения не влияют на характер тренда. Второй элемент в правой части уравнений отражает именно случайную составляющую – незначительные колебания цен около своего тренда.

На практике часто приходится наблюдать резкие изменения цен - скачки. Их появление свидетельствует о нестабильности рынка или экономики в целом или об ее зависимости от внешних факторов: политических, спекулятивных и других. Причем появление таких скачков не связано с чисто сезонными колебаниями цены. Третий элемент в правой части уравнений отражает эту случайную составляющую, причем

предполагается, что величины скачков будут всюду положительными. Время появления скачков можно моделировать с помощью пуассоновского случайного процесса. Амплитуды скачков случайны и закон распределения их вероятности может быть восстановлен по тем же статистическим данным. Отсюда как следствие имеем, что промежутки между скачками характеризуются показательным распределением.

Любая реализация цены есть решение представленного выше дифференциального уравнения и имеет общий вид:

$$p(t) = p_0 + \int_0^t f(t, p(t), z(t)) dt + \int_0^t F(t, p(t)) d\omega(t) + \int_0^t \sum_{i=1}^n c(v_i) \delta(t - t_i) dt$$

Прогнозное значение, полученное с помощью модифицированного фильтра Калмана-Бьюси, будет оптимально с точки зрения принципа максимума апостериорной вероятности. Ниже изложена структура алгоритма для решения поставленной задачи.

2. Структура алгоритма прогнозирования на основе реализации модифицированного фильтра Калмана-Бьюси

Модифицированное уравнение для оптимальной оценки записывается в виде

$$\frac{dp^*(t)}{dt} = f(t, p^*(t), z(t)) + K(t)[y(t) - p^*(t)] + \int_{-\infty}^{\infty} \frac{M[\varphi(t)y(\tau_2)]}{M[y(t)y(\tau_1)]} H(t, \tau_2) y(\tau_2) d\tau_2$$

где $p^*(t)$ – значение оценки; $K(t)$ – коэффициент усиления фильтра, определяемый из следующего соотношения

$$K(t) = f(t, p(t), z(t)) R(t) [R\varphi(t) + R(t)]^{-1},$$

в этом соотношении $R(t) = M[p^*(t) - p(t)] [p^*(t) - p(t)]$ – дисперсия ошибок оценки, для которой справедливо рекуррентное соотношение

$$R(t+1) = [f(t, p(t), z(t)) - K(t)]^2 R(t) + R_{\eta}(t) K^2(t) + R_{\varphi}(t+1) + 2M \left[\int_{-\infty}^{\infty} \frac{M[\varphi(t)y(\tau_2)]}{M[y(t)y(\tau_1)]} H(t, \tau_2) y(\tau_2) d\tau_2 \right] [p^*(t) - p(t)].$$

функции $R_{\eta}(t) = M[\eta(t) \eta(\tau)] = Q(t)\delta(t-\tau)$ и $R_{\varphi}(t+1)$ – известны.

Модификация связана с представлением уравнения состояния как

$$dp(t) = f(t, p(t), z(t))dt + F(t, p(t))d\omega(t) + \int c(v)v(dt, dv)$$

и уравнения наблюдения в виде

$$y(t) = p(t) + \eta(t),$$

где $\eta(t)$ – гауссов белый шум; $y(t)$ – наблюдаемое значение.

Структура алгоритма прогнозирования:

1. Задать начальные приближения из условия $p(\tau) = p^*(\tau)$; $p^*(\tau) = p_0$;
 $R(\tau) = 0$;
2. Вычислить коэффициент усиления $K(t)$;
3. Вычислить по формуле значение $R(t+1)$;
4. Получить прогноз цены $p^*(t+1)$ как численное решение дифференциального уравнения.

Производя операции 2-4, при $t = \tau, \tau + 1, \dots$ получим траекторию предсказанных на шаг вперед значений цены.

Модель и алгоритм прогнозирования цены

Для практической реализации непрерывное время заменено дискретным;

уравнения модифицированной модели:

$$\text{уравнение состояния: } x(k+1) = \Phi(k+1, k)x(k) + \Gamma(k+1, k)\omega(k),$$

$$\text{уравнение измерения: } z(k+1) = H(k+1)x(k+1) + v(k+1),$$

где Φ – переходная матрица состояния размера $n \times n$;

Γ – переходная матрица возмущения размера $n \times p$;

ω – p -вектор возмущения;

H – матрица измерения размера $m \times n$;

v – m -вектор ошибки измерения.

Процесс $\{\omega(k), k = 0, 1, 2, \dots\}$ является p -мерной гауссовской белой последовательностью с неотрицательно определенной корреляционной матрицей $Q(k)$ размером $p \times p$. Процесс $\{v(k), k = 0, 1, 2, \dots\}$ является m -

мерной гауссовской белой последовательностью с неотрицательно определенной корреляционной матрицей $R(k+1)$ размером $m \times m$. Оба эти процесса взаимно независимы

Начальное состояние $x(0)$ есть гауссовский случайный n -вектор с неотрицательно определенной корреляционной матрицей $P(0)$ размера $n \times n$.

Тогда алгоритм фильтра Калмана будет в виде:

- 1) $x(k+1|k) = \Phi(k+1,k)x(k|k)$;
- 2) $P(k+1|k) = \Phi(k+1,k)P(k|k)\Phi'(k+1,k) + \Gamma(k+1,k)Q(k)\Gamma'(k+1,k)$;
- 3) $K(k+1) = P(k+1|k)H'(k+1)[H(k+1)P(k+1|k)H'(k+1) + R(k+1)]^{-1}$;
- 4) $P(k+1|k+1) = [I - K(k+1)H(k+1)]P(k+1|k)$;
- 5) $x(k+1|k+1) = x(k+1|k) + K(k+1)[z(k+1) - H(k+1)x(k+1|k)]$.

Можно еще более упростить модель, взяв вместо векторов x , ω и v одиночные переменные, тогда матрицы Φ , Γ и H превратятся в скаляр.

Ход работы

Из статистической информации выбираем временной ряд, отражающий динамику цен.

Таблица 1

Динамика цен

Дата	Цена
01.06.2016	3,7
02.06.2016	3,65
03.06.2016	3,7
04.06.2016	3,59
05.06.2016	3,7
06.06.2016	3,7
07.06.2016	3,58
08.06.2016	3,52
09.06.2016	3,52
10.06.2016	3,34

11.06.2016	3,52
12.06.2016	3,52
13.06.2016	3,45
14.06.2016	3,47
15.06.2016	3,52
16.06.2016	3,08
17.06.2016	3,09
18.06.2016	3,15
19.06.2016	2,95
20.06.2016	3,05
21.06.2016	3,11
22.06.2016	2,99
23.06.2016	3,1
24.06.2016	3,15
25.06.2016	3,09
26.06.2016	3,15
27.06.2016	3,15
28.06.2016	3,08
29.06.2016	3,37
30.06.2016	3,28

При анализе этого временного ряда формируем тренд изменения цены:

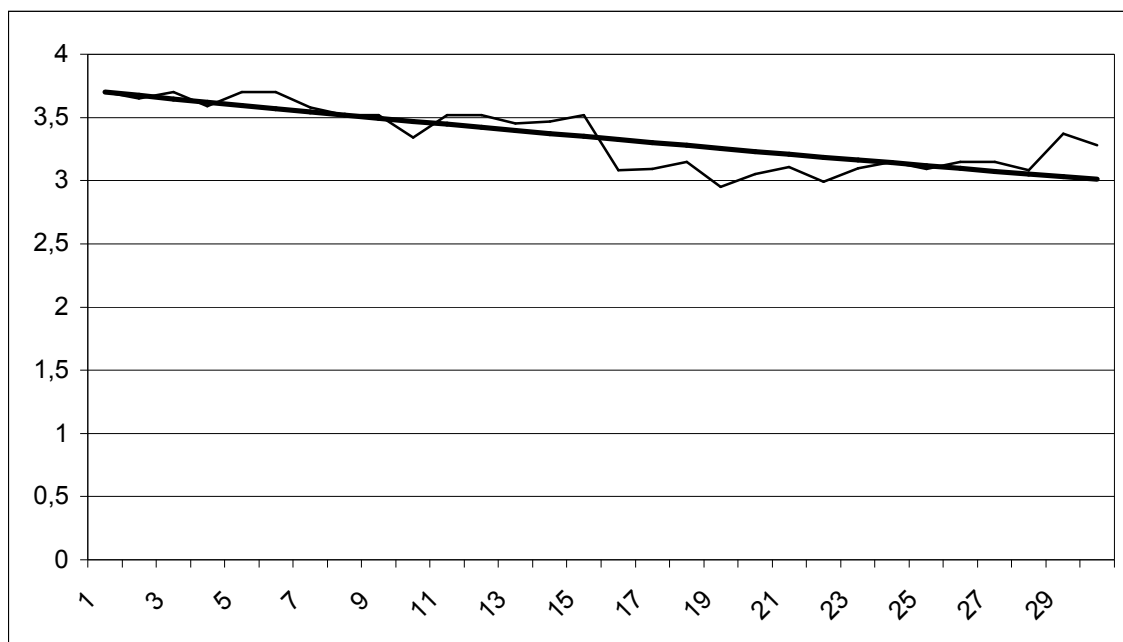


Рис. 1. Изменение цен (исходные данные и тренд)

Уравнение тренда получили в виде: $x(k+1) = 0,9929x(k)$.

Затем формируем следующую модель:

уравнение состояния: $x(k+1) = 0,9929x(k) + \omega(k)$,

уравнение измерения: $z(k+1) = x(k+1) + v(k+1)$,

где дисперсия $\omega(k) = Q$; дисперсия $v(k+1) = R$.

Тогда уравнения фильтра примут вид:

- 1) $x(k+1|k) = 0,9929x(k)$;
- 2) $P(k+1|k) = P(k|k) + Q$;
- 3) $K(k+1) = P(k+1|k)[P(k+1|k) + R]^{-1}$;
- 4) $P(k+1|k+1) = R * K(k+1)$;
- 5) $x(k+1|k+1) = x(k+1|k) + K(k+1)[z(k+1) - x(k+1|k)]$.

при начальном условии $P(0|0) = 0$; $x(0|0) = z(0|0)$.

Исходный код программа, реализующей фильтр Калмана, на языке Си приведен в Приложении 1.

Таблица 2

Результаты работы алгоритма фильтра Калмана ($Q=0,1;R=0,15$)

Дата	Цена	Тренд	$P(k+1 k+1)$	$x(k+1 k)$	$x(k+1 k+1)$
------	------	-------	--------------	------------	--------------

01.06.2016	3,7	3,700000	0,060000	3,673730	3,664238
02.06.2016	3,65	3,673730	0,077419	3,638222	3,670108
03.06.2016	3,7	3,647647	0,081281	3,644050	3,614762
04.06.2016	3,59	3,621748	0,082082	3,589097	3,649785
05.06.2016	3,7	3,596034	0,082246	3,623871	3,665613
06.06.2016	3,7	3,570502	0,082279	3,639587	3,606902
07.06.2016	3,58	3,545151	0,082286	3,581293	3,547669
08.06.2016	3,52	3,519981	0,082287	3,522481	3,521120
09.06.2016	3,52	3,494989	0,082287	3,496120	3,410475
10.06.2016	3,34	3,470175	0,082288	3,386261	3,459628
11.06.2016	3,52	3,445536	0,082288	3,435065	3,481659
12.06.2016	3,52	3,421073	0,082288	3,456939	3,453132
13.06.2016	3,45	3,396783	0,082288	3,428615	3,451318
14.06.2016	3,47	3,372666	0,082288	3,426814	3,477934
15.06.2016	3,52	3,348720	0,082288	3,453241	3,248487
16.06.2016	3,08	3,324944	0,082288	3,225423	3,151132
17.06.2016	3,09	3,301337	0,082288	3,128759	3,140412
18.06.2016	3,15	3,277898	0,082288	3,118115	3,025890
19.06.2016	2,95	3,254625	0,082288	3,004406	3,029418
20.06.2016	3,05	3,231517	0,082288	3,007909	3,063915
21.06.2016	3,11	3,208573	0,082288	3,042161	3,013546
22.06.2016	2,99	3,185792	0,082288	2,992150	3,051315
23.06.2016	3,1	3,163173	0,082288	3,029650	3,095672
24.06.2016	3,15	3,140715	0,082288	3,073693	3,082639
25.06.2016	3,09	3,118416	0,082288	3,060752	3,109712
26.06.2016	3,15	3,096275	0,082288	3,087633	3,121847
27.06.2016	3,15	3,074291	0,082288	3,099682	3,088885
28.06.2016	3,08	3,052464	0,082288	3,066954	3,233200
29.06.2016	3,37	3,030791	0,082288	3,210244	3,248511

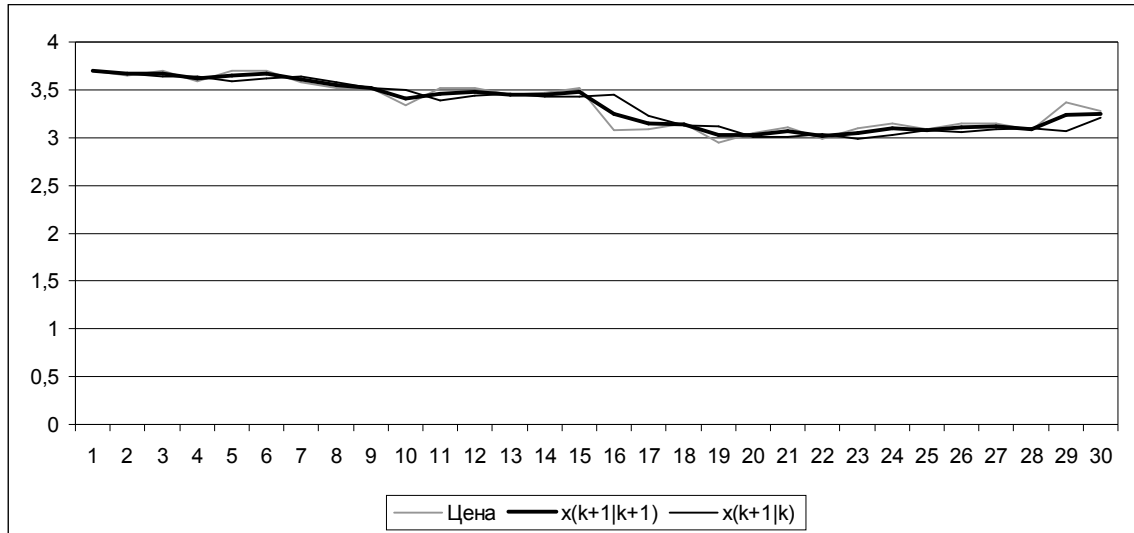


Рис. 2. Результат работы фильтра ($Q=0,1$; $R=0,15$)

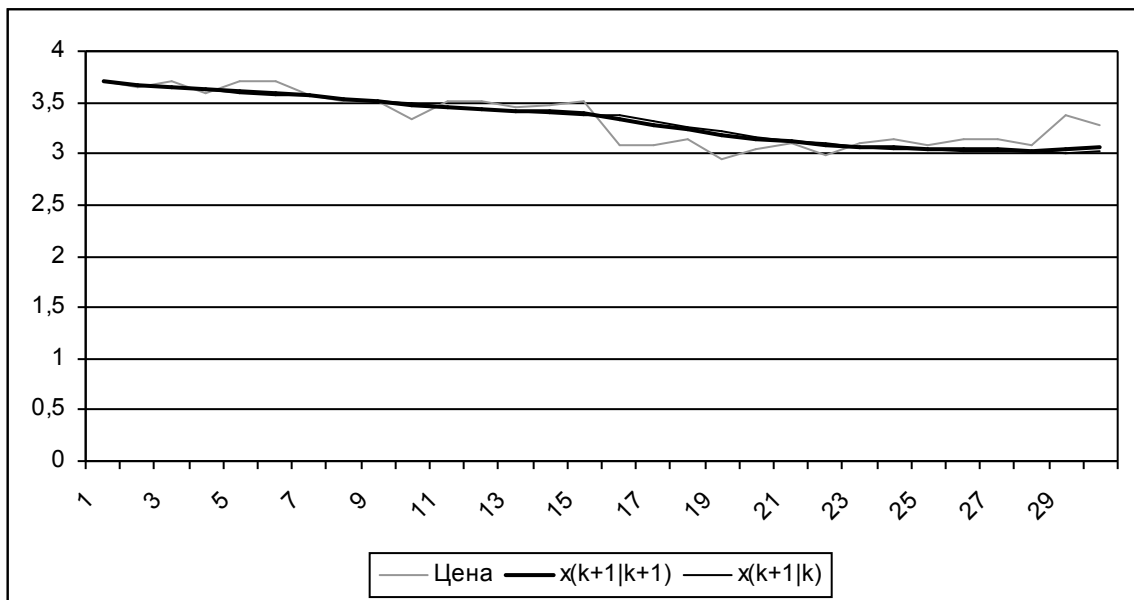


Рис. 3. Результат работы фильтра ($Q=0,01$; $R=0,5$)

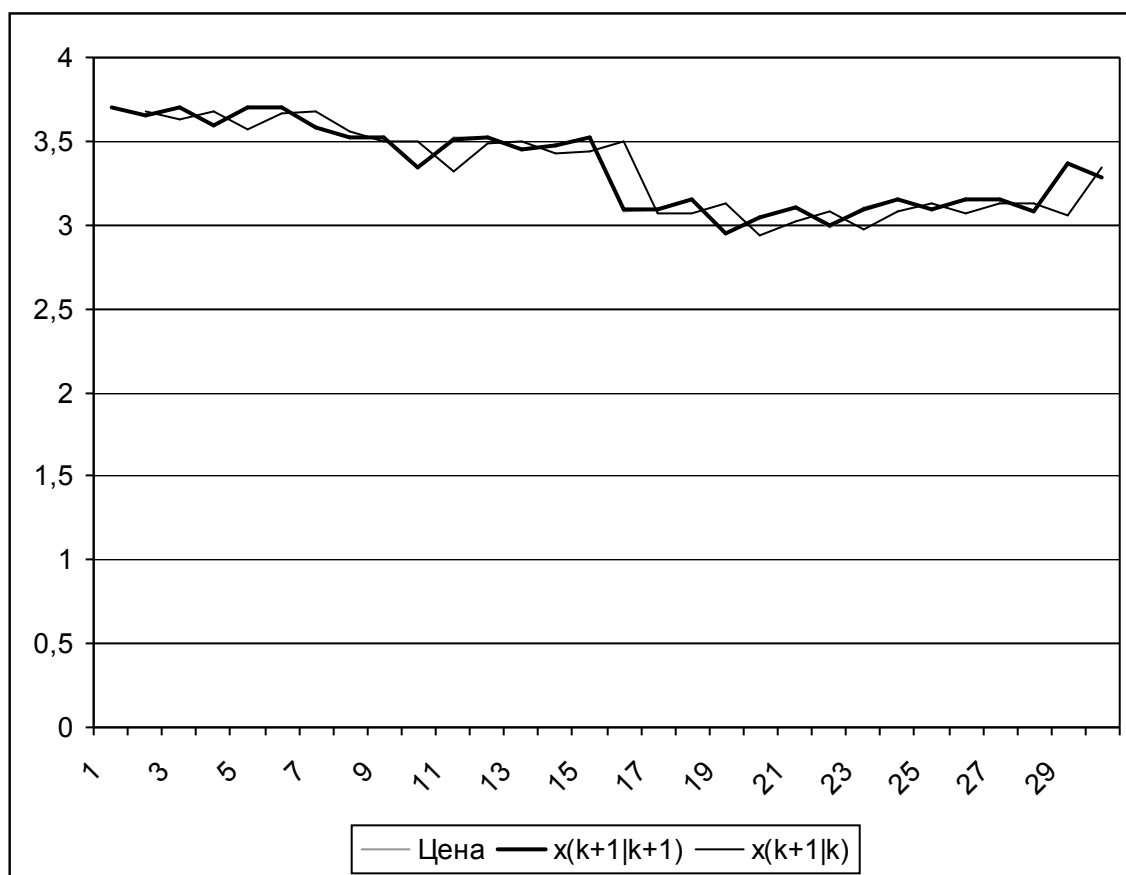


Рис. 4. Результат работы фильтра (Q=0,5; R=0,01)

Как видно из результатов работы алгоритма фильтра Калмана, если дисперсия $\omega(k)$ больше дисперсии $v(k+1)$, то фильтр приближает значение оценки к измерениям. Если же дисперсия $\omega(k)$ меньше дисперсии $v(k+1)$, то наоборот, скорректированная оценка будет ближе к модели. Также видно, что при начальном значении $P(0|0) = 0$, фильтр уже после обработки седьмого измерения, находится в установившемся состоянии. Кроме того, алгоритм был протестирован и при других значениях $P(0|0)$, все равно после седьмого измерения фильтр приходил в устойчивое состояние.

С помощью фильтра Калмана, зная дисперсии случайных процессов $\omega(k)$ и $v(k+1)$, можно получать достаточно точные оценки значений цены отдельного продукта.

Приложение 1

1. `#include <stdio.h>`
2. `#include <string.h>`

```

3. #include <ctype.h>
4. int main(int argc, char *argv[]){
5. FILE *input,*out;
6. char buf[257];
7. int k;
8. float res,z,K,P,P1,Q,R,F,x;//косметика
9. if(argc!=2){
10. printf("Синтаксис: %s <file>\n",argv[0]);
11. printf("      где <file> - имя файла с результатами измерений...\n");
12. printf("В результате работы программы создается файл
      \"result.txt\"\n");
13. return(1);
14. } //есть ли параметр

```

Задание

Для различных объектов (продуктов или изданий по желанию студента) составить статистический ряд и используя фильтр Калмана выполнить прогноз цены. Выполнить исследования полученных результатов при различных значениях Q и R . Сделать выводы.

Содержание отчета

1. Цель лабораторной работы
2. Краткое описание хода работы
3. Файл статистического ряда исследуемого объекта
4. Результаты работы фильтра после прогноза в виде
5. Выводы по результатам работы
6. Ответы на вопросы

Вопросы:

1. В чем заключается суть работы фильтра Калмана?
2. Пояснить, что происходит с фильтром, если дисперсия $\omega(k)$ больше дисперсии $v(k+1)$ и почему?

3. В чем смысл уравнение состояния?
4. Что представляет уравнение измерения?
5. Поясните возможность применения фильтра Калмана для нестационарных процессов.

Лабораторная работа № 9. Парциальная обработка данных

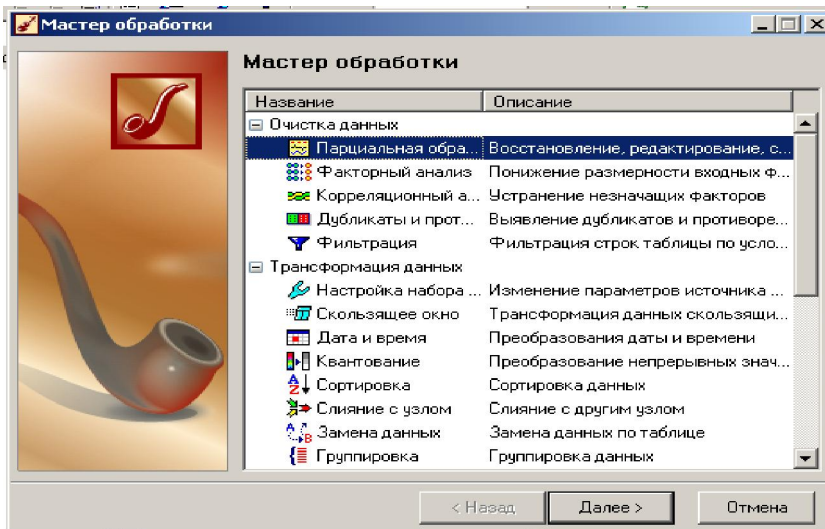
Цель работы. Изучить возможности АП процедур обработки данных.ых выполнить обработку данных выбранной предметной области.

Теоретическая часть. В процессе парциальной обработки восстанавливаются пропущенные данные, редактируются аномальные значения, проводится спектральная обработка. В Deductor Studio при этом используются алгоритмы, в которых каждое поле анализируемого набора обрабатывается независимо от остальных полей, то есть данные обрабатываются по частям. По этой причине такая предобработка получила название парциальной. В числе процедур предобработки данных, реализованных в Deductor Studio, входят сглаживание, удаление шумов, редактирование аномальных значений, заполнение пропусков в рядах данных.

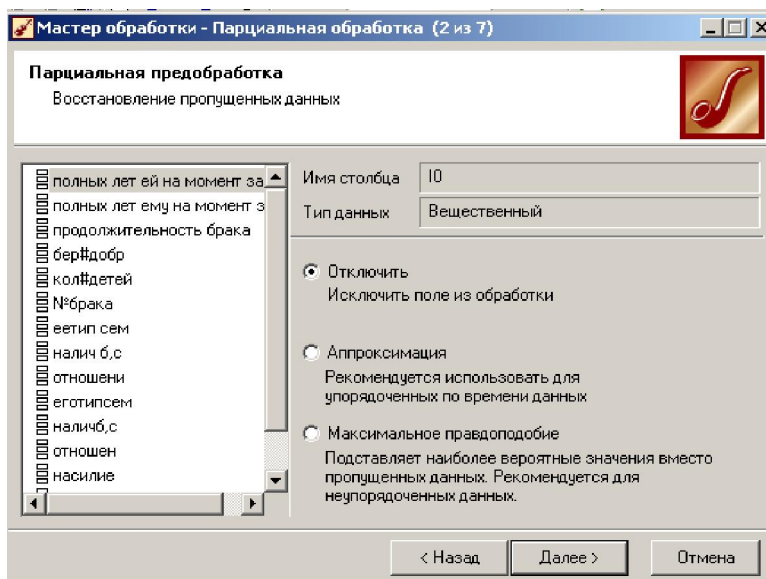
1) таблица с аномальными данными:

полных лет ей на момент заключения брака	полных лет ему на момент	жигител брака	эр#дос	т#де	№брака	еетип сем	налич б,с	ноше	тип	наличб,с	ноше	насилие	алк
100	100	2	нет	0	первый	полна	2	хорош	пол	1	хоро	нет	нет
30	35	5	нет	0	первый	полна	1	хорош	пол	0	хоро	нет	нет
29	31	12	нет	0	первый	полна	1	хорош	пол	1	хоро	нет	нет
28	32	4	да	1	второй	непол	0	хорош	неп	1	хоро	да	нет
27	28	6	да	1	второй	полна	1	хорош	пол	0	хоро	нет	да
27	25	1,5	да	1	первый	непол	2	хорош	пол	2	хоро	нет	нет
27	40	7	нет	0	первый	полна	1	хорош	пол	3	не оч	нет	нет
27	28	2	нет	1	первый	полна	1	хорош	неп	0	хоро	да	нет
26	25	17	нет	2	первый	полна	0	плохи	пол	0	хоро	да	да
26	22	8	да	1	первый	полна	0	хорош	пол	0	не оч	да	нет
26	30	1	да	2	первый	полна	0	хорош	пол	0	плохи	да	да
26	27	3	нет	1	первый	полна	2	хорош	пол	1	не оч	да	нет
26	26	11	нет	1	первый	полна	1	не оч	пол	1	хоро	да	нет
26	30	11	нет	0	первый	полна	1	хорош	пол	1	хоро	нет	нет

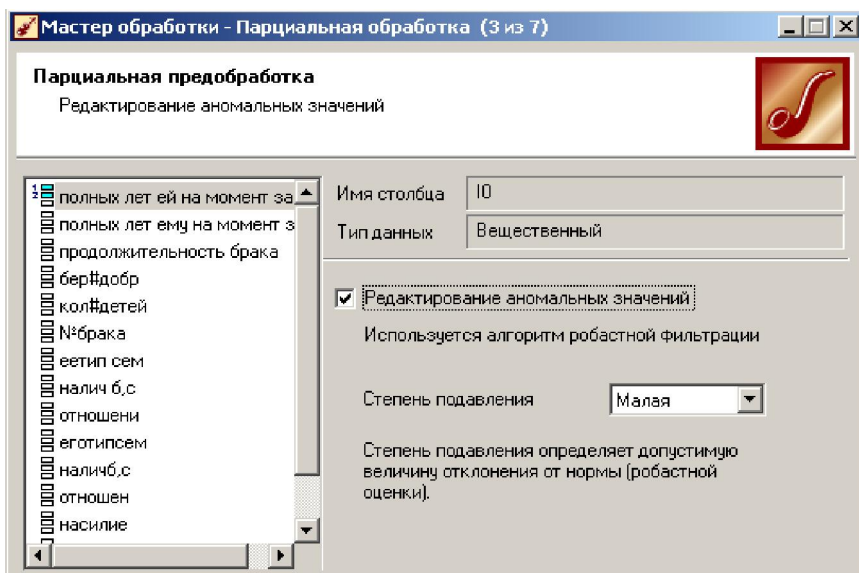
2) открываем мастер обработки и выбираем парциальную обработку:



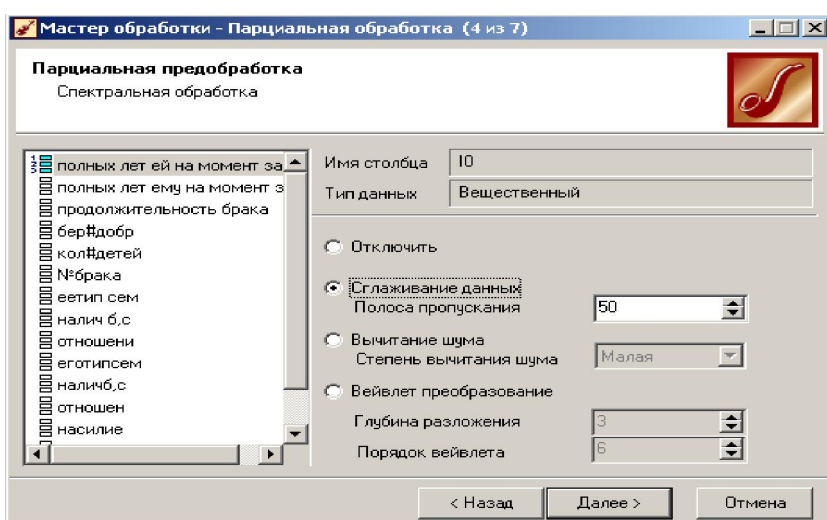
3) выбор операции восстановления пропущенных данных:



4) выбор степени подавления:



5) сглаживание данных возможно выполнить с помощью вейвлет-преобразования и вычитания шума:



6) полученная таблица:

полных лет ей на момент заключения брака	полных лет ему на момент заключения брака	продолжительность брака	бер#добр	кол#детей	№брака	еетип сем	налич б,с	отношени	еготипсем	наличб,с	отношен	насилие	алкоголизм
19,8514571837075	30	13	нет	1	первый дл	полная	0	не очень	полная	0	плохие	да	да
20,7632477464631	23	17	да	2	первый дл	полная	0	хорошие	неполн	1	не очень	да	да
21,3719146851208	24	11	нет	1	первый дл	полная	0	хорошие	полная	0	хорошие	да	да
20,3821882356354	23	12	нет	1	первый дл	полная	0	хорошие	полная	0	хорошие	нет	нет
19,5736272606974	22	14	нет	1	первый дл	полная до	0	хорошие	полная	3	плохие	нет	да
20,6027766202498	19	10	нет	1	первый дл	полная	0	хорошие	полная	0	хорошие	нет	нет
22,5543322481671	22	15	да	2	первый дл	полная	0	хорошие	полная	0	хорошие	нет	нет
24,009369451793	25	17	нет	2	первый дл	полная	0	плохие	полная	0	хорошие	да	да
24,8590407431192	24	9	да	1	первый дл	полная	0	не очень	полная	0	не очень	да	да
24,9933531213713	23	5	нет	1	первый дл	полная	0	не очень	полная	0	хорошие	да	нет
23,9055282093373	22	8	да	1	первый дл	полная	0	хорошие	полная	0	не очень	да	нет
22,6445951391282	40	3	нет	1	первый дл	полная	1	плохие	полная	3	плохие	да	нет
23,3919878087563	22	4	да	1	первый дл	полная	0	не очень	полная	0	хорошие	да	да
25,8477646579785	30	1	да	2	первый дл	полная	0	хорошие	полная	0	плохие	да	да
26,6551971393871	28	6	да	1	второй дл	полная	1	хорошие	полная	0	хорошие	нет	да
23,91919511190294	37	4	нет	1	первый дл	полная	0	хорошие	полная	0	хорошие	нет	нет
20,085655484726	19	4	нет	1	первый дл	полная	2	хорошие	полная	0	хорошие	нет	нет
18,7333854739885	19	11	нет	1	первый дл	полная	1	хорошие	полная	1	хорошие	да	да
20,1864738812745	23	4	нет	1	первый дл	полная	1	хорошие	полная	0	не очень	нет	нет

Полученная таблица отличается от первоначальной.

Задание: Извлечь данные выбранной предметной области из хранилища данных и выполнить парциальную обработку.

Содержание отчета

1. Цель лабораторной работы
2. Краткое описание хода работы
3. Файл первоначальных данных
4. Файл данных после парциальной обработки
5. Выводы по результатам работы
6. Ответы на вопросы

Вопросы:

1. Что такое парциальная обработка данных?
2. Зачем нужен спектральный анализ при обработке данных?
3. Приведите основные методики восстановления данных
4. Основное назначение вейвлет-преобразования при обработке данных
5. Укажите каким образом выбирается степень подавления шума

10. Список рекомендуемой литературы

1. Методы и модели анализа данных: OLAP и Data Mining. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод – СПб.: БХВ-Петербург, 2004.- 336 с.: ил.
2. Аналитика: методология, технология и организация информационно - аналитической работы /П.Ю. Конотопов, Ю.В. Курносков - М.: РУСАКИ, 2004. - 512 с.
3. Официальный сайт компании «BaseGroup Labs» [Электрон. ресурс]. Рязань, 1995-2010.- Режим доступа: <http://www.basegroup.ru/>
4. Data Mining и аналитическая платформа Deductor [Электрон. ресурс] : [статья]. М., 2008.- Режим доступа: http://sttc.ru/index.php?option=com_content&task=view&id=56&Itemid=90

5. Г.И. Просветов (МГУ им. М.В. Ломоносова). Дерево решений [Электрон. ресурс] : [статья] / Г.И. Просветов.- СПб, 2008.- Режим доступа: http://www.elitarium.ru/2008/04/09/derevo_reshenijj.html
6. Дервянко В.А. Поиск ассоциативных правил при интеллектуальном анализе данных [Электрон. ресурс] : [статья] / В.А. Дервянко.- 2009.- Режим доступа: http://www.rammus.ru/products/arda/article_lam_translation