Методы анализа данных

Примеры задач. Иллюстрации

Ганелина Наталья Давидовна

Кафедра АСУ

12657@211.ru

Структура курса

- Задачи и методы анализа данных
- Корреляционный анализ данных
- Регрессионный анализ данных
- Поиск ассоциативных взаимосвязей
- Кластеризация
- Классификация
- Снижение размерности многомерного признака. Отбор наиболее информативных показателей. Факторный анализ
- Исследование и прогнозирование временных рядов

Структура курса

- Генетические алгоритмы и эволюционное моделирование задач анализа данных
- Statistica
- PolyAnalyst
- SPSS
- Deductor
- Excel

БРС

- Лабораторные работы: 40 баллов
- РГР: 40 баллов
- Зачет: 20 баллов

- «Автомат»: от 77 баллов

Рекомендуемая литература

- Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности.- М.: Финансы и статистика, 1989.
- Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. – М.: «Финансы и статистика», 1983. – 471 с.
- Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998. – 1022 с.
- Альсова О.К. Решение задач интеллектуального анализа данных на основе вариативного моделирования./Методические указания к лабораторным работам; составитель Альсова О.К. – Новосибирск: Изд-во НГТУ, 2005. – 75 с.
- Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. – Спб.: БХВ-Петербург, 2004. – 336 с.
- Боровиков В.П. Statistica. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.
- Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы. – М.: ФИЗМАТЛИТ, 2006. – 320 с.

Рекомендуемая литература

http://archive.ics.uci.edu/ml/
http://www.ics.uci.edu/~MLearn/MLRepository.html
Базы данных с реальными данными из разных предметных областей для оценки эффективности работы методов ИАД.

http://www.statsoft.ru/
Описание интегрированной системы Statistica, электронный учебник по статистике, Data Mining, примеры реальных задач.

http://exponenta.ru/soft/statist/statist.asp
Демо-версия программ. Ссылка на электронный учебник.

http://www.r-project.org/
http://cran.gis-lab.info/
R is a free software environment for statistical computing and graphics.

Рекомендуемая литература

- Бериков В.Б. Анализ статистических данных с использованием деревьев решений: Учебное пособие. – Новосибирск. Изд-во НГТУ, 2002. – 60 с.
- Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, вып. 1, 1974. – 406 с.; вып. 2 – 197 с.
- Боровиков В.П., Ивченко Г.И. Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика на компьютере. Учеб. Пособие. – М.: Финансы и статистика, 1999. – 384 с.
- Губарев В.В. Интеллектуальный анализ данных и вариативное моделирование в экспериментальных исследованиях.//Информационные системы и технологии. ИСТ, 2001: Сб. научн. статей. – Новосибирск: НГТУ, 2001. – С. 5-25.
- Губарев В.В. Вероятностные модели / Новосиб. электротехн. ин-т. – Новосибирск, 1992. – Ч.1. – 198 с; Ч.2. – 188 с.
- Губарев В.В., Альсова О.К. Вариативное моделирование на примере решения прикладной задачи.// ИСТ-2000: Матер. междун. науч.-техн. конф. – Новосибирск, НГТУ, 2000, том 2, С. 285-286.
- Губарев В.В., Альсова О.К., Швайкова И.Н. Интеллектуальный анализ «данных» и вариативное моделирование с системных позиций.// SCM’2000: International Conference on Soft Computing and Measurements. – Санкт-Петербург, СПб-ГЭТУ, 2000, С. 65-68.

Рекомендуемая литература

- Дюк В.А., Самойленко А.П. Data Mining: учебный курс. — СПб.: Питер, 2001. – 368 с.
- Елманова Н. Введение в Data Mining.// Компьютер Пресс 8, 2003, С. 28-39.
- Кендэл М. Временные ряды. – М.: Финансы и статистика, 1981. – 199 с.
- Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В. Базы данных. Интеллектуальная обработка информации. – М.: Изд-во Нолидж, 2001. – 496 с.
- Курейчик В.М., Родзин С.И. Эволюционные алгоритмы: генетическое программирование. Обзор // Известия РАН. ТиСУ. 2002. №1. С. 127-137.
- Струнков Т. Что такое генетические алгоритмы.//PC Week RE, №19, 1999.
- Факторный, дискриминантный и кластерный анализ/Пер. с англ. А.М. Хотинского. Под ред. И.С. Енюкова. -М.: Финансы и статистика, 1989.
- Четыркин Е.М. Статистические методы прогнозирования. – М.: Статистика, 1977. – 199с.
- Шапот М. Интеллектуальный анализ данных в системах поддержки принятия решений.//Открытые системы, №1, 1998, С. 30-35.
- Шапот М., Рощупкина В. Интеллектуальный анализ данных и управление процессами.//Открытые системы №4-5, 1998, С. 40-44.
- Щавелев Л.В. Способы аналитической обработки данных для поддержки принятия решений.// СУБД. - 1998. - № 4-5.
- Эвоинформатика: Теория и практика эволюционного моделирования./И.Л. Букатова, Ю.И. Михасев, А.М. Шаров. – М.: Наука, 1991. – 206 с.

Рекомендуемая литература

- Гайдышев И. Анализ и обработка данных: специальных справочник. – Спб.: Питер, 2001. – 752 с.
- И.Гайдышев. Решение научных и инженерных задач средствами Excel, VBA и C/C++.- СПб.: БХВ-Петербург, 2004. – 504 с.

Иллюстрации

- Большинство примеров и иллюстраций заимствованы из учебных пособий, представленных в списке рекомендованной литературы.
- На лекции в обязательном порядке указывается источник.

Признаки

Действия над признаками, измеренными в различных шкалах

Шкала	Допустимые действия	Пример применения
Номинальная	Различение	Наличие или отсутствие симптома
Порядковая	Различение, сравнение	Школьная оценка
Количественная	Различение, сравнение, сложение, вычитание, умножение, деление	Температура, масса, время, длина

Шкала	**Математические и статистические величины, вычисление которых допустимо на данном уровне**
Номинальная	Мода, процентные частоты, доли, корреляция
Порядковая	Мода, медиана, квартили, коэффициент корреляции, дисперсионный анализ
Интервальная	Мода, медиана, квартили, коэффициент корреляции, ранговые критерии, средняя, дисперсия, стандартное отклонение, коэффициент корреляции
отношений	Все арифметические операции, все понятия и методы математической статистики

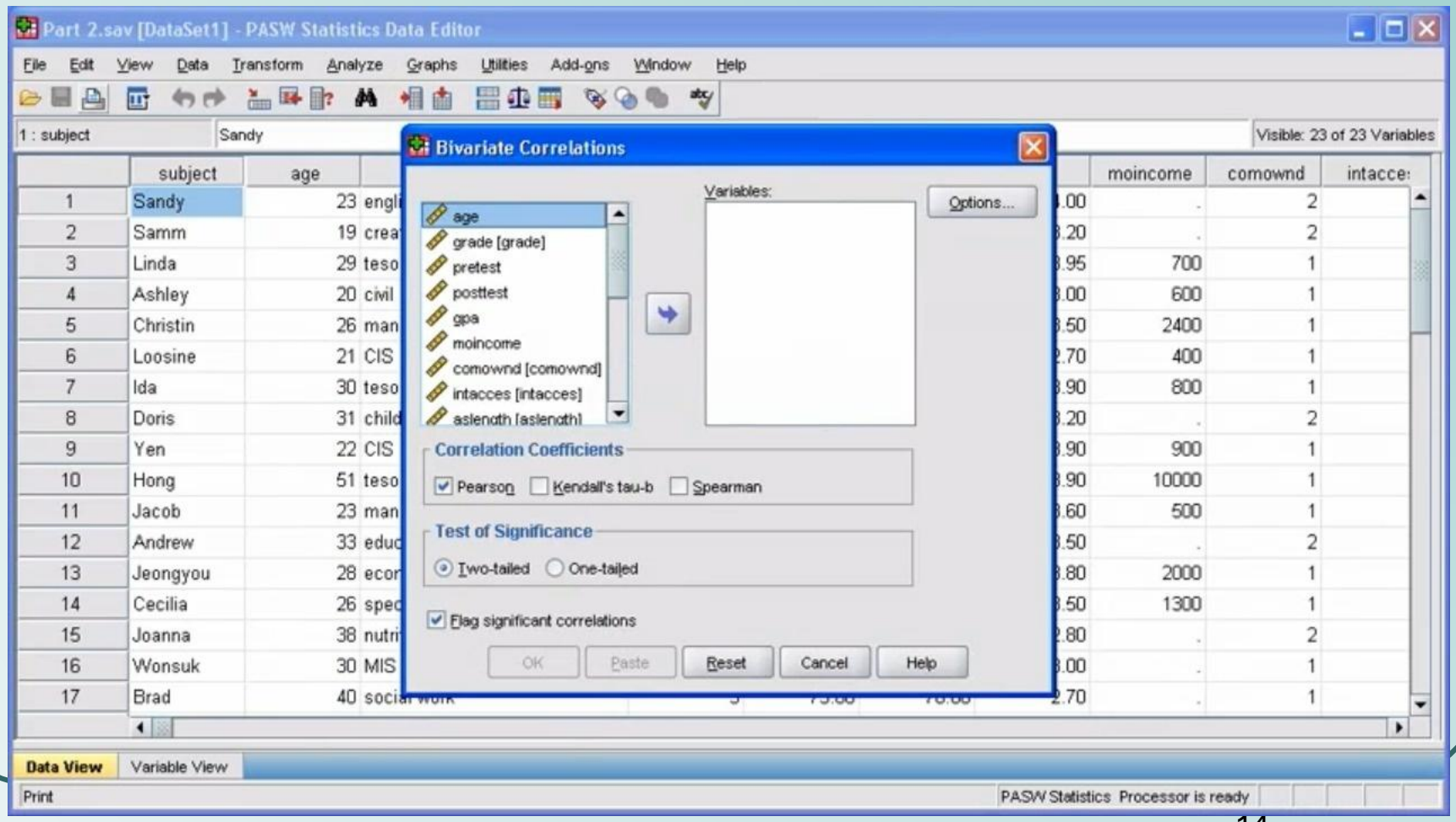
Методы **DM**

Data Mining — мультидисциплинарная область

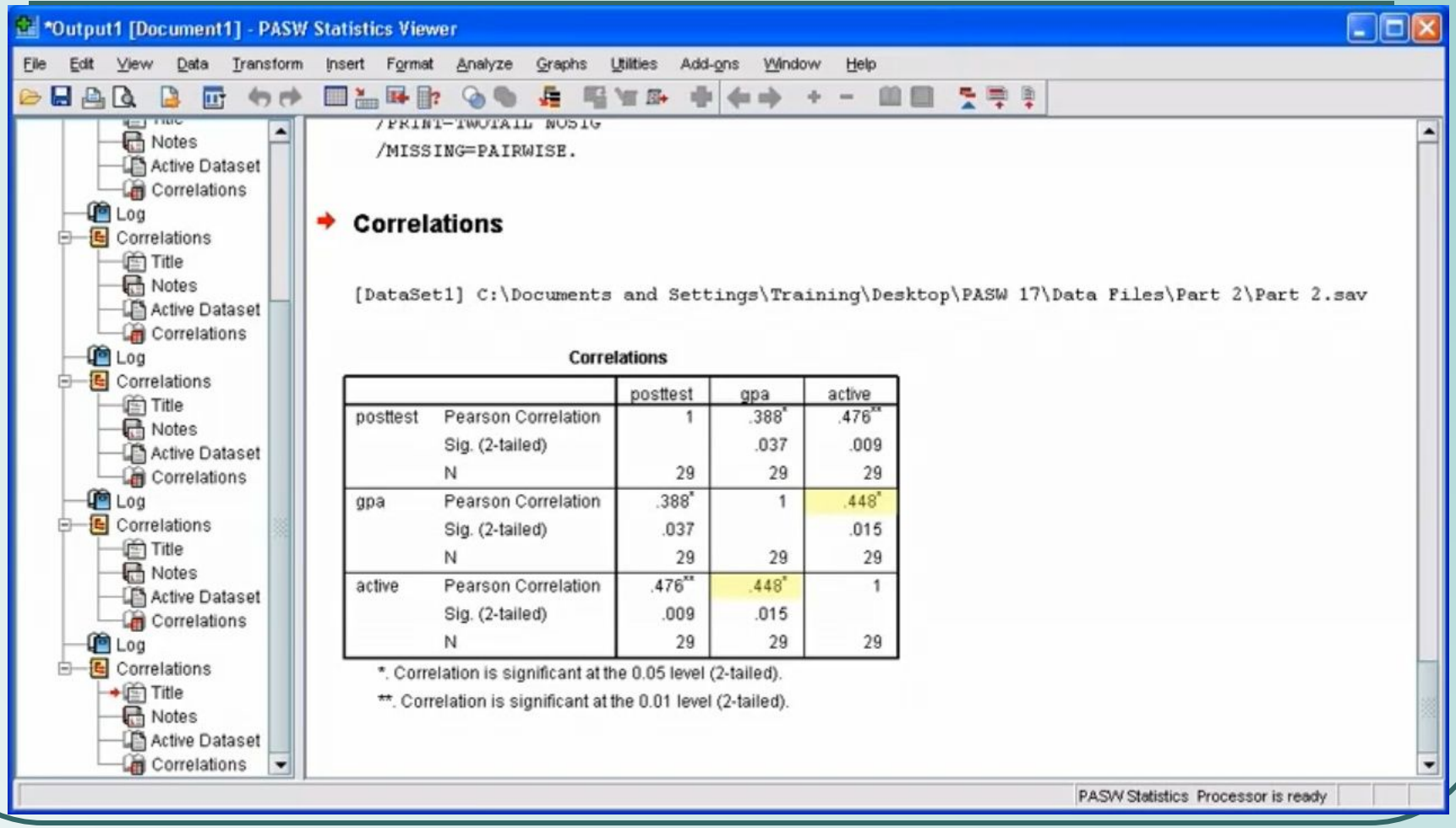
Системы **DM**


Технологии
Data Mining и OLAP
Сиситемы Data Mining
Этапы
Исследование баз знаний
Построение модели или структуры данных
Проверка модели
Системы OLAP
OLAP - технологии
Предметно-ориентированные аналитические системы
Статистические пакеты
Нейронные сети
Системы рассуждений на основе аналогичных случаев
Другие системы

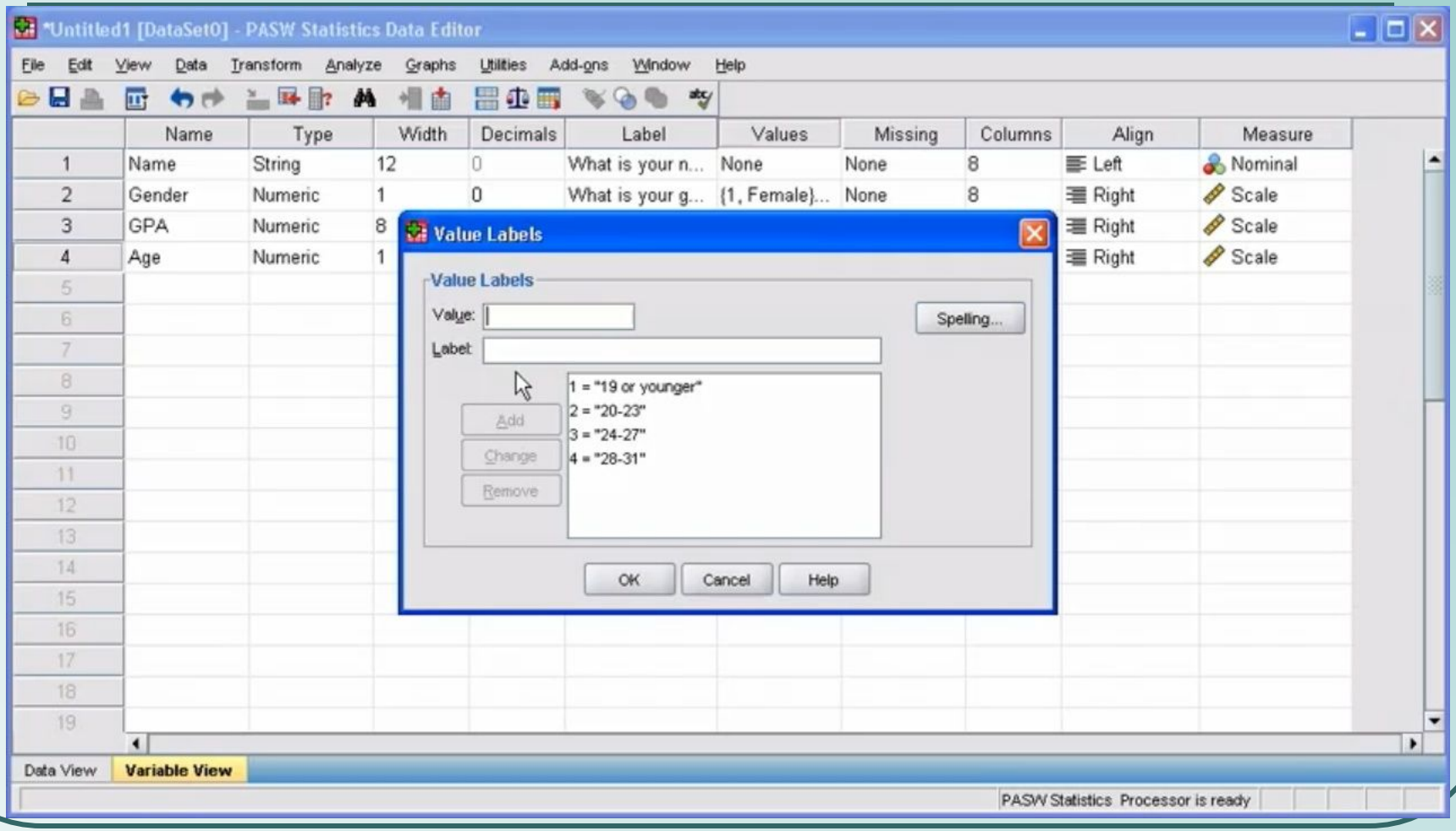
Программное обеспечение анализа данных



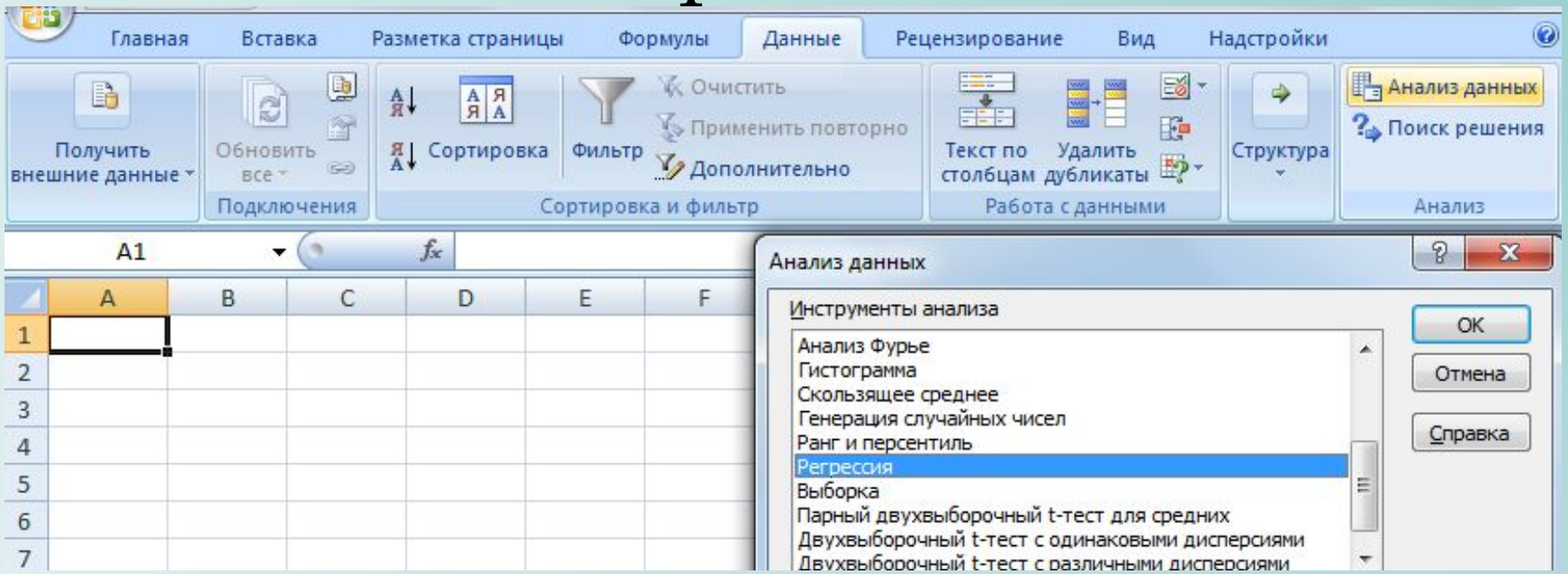
Программное обеспечение анализа данных



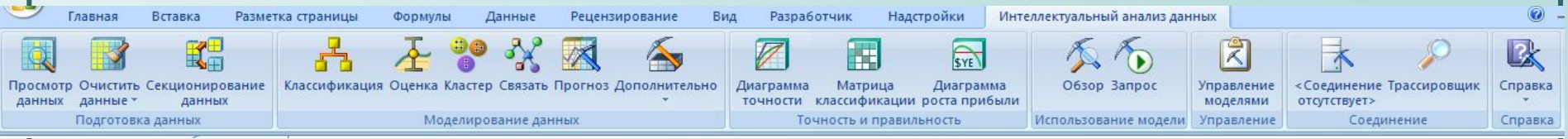
Пакеты

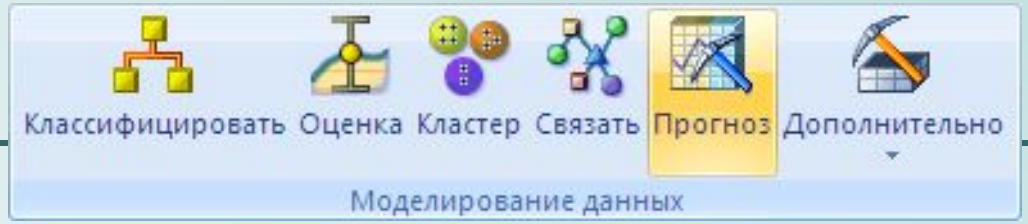

*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure
1 Name String 12 0 What is your n... None None 8 Left Nominal
2 Gender Numeric 1 0 What is your g... {1, Female}... None 8 Right Scale
3 GPA Numeric 8 Right Scale
4 Age Numeric 1 Right Scale
Value Labels
Value Labels
Value:
Label:
Spelling...
1 = "19 or younger"
2 = "20-23"
3 = "24-27"
4 = "28-31"
Add
Change
Remove
OK
Cancel
Help
Data View
Variable View
PASW Statistics Processor is ready

Надстройки Excel



Надстройки Data Mining к приложению Microsoft Office Excel 2007 для извлечения и обработки данных





Дисперсионный анализ

Пример 5.1. Удобрения для комнатных растений фасуются в пакеты весом по 0,5 кг. Из партии пакетов, расфасованных в течение суток, случайным образом отобрали 30 пакетов. Они были распределены по трем различным условиям хранения. После хранения в течение одной недели определялось содержание влаги в продукте, хранящемся в каждом пакете.

Данные о содержании влаги приводятся ниже.

Условия хранения	Содержание влаги, %
1	10,1 7,3 5,6 6,2 8,4 8,1 8,0 7,6 5,3 7,2
2	11,7 12,2 11,8 7,8 8,9 9,9 12,4 11,0 10,3 13,8 10,5 9,8 9,1
3	10,2 12,0 8,8 8,7 10,5 11,0 9,1

Дисперсионный анализ

На уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что условия хранения продукта не оказывают влияния на содержание влаги.

Предполагается, что выборки получены из независимых нормально распределенных совокупностей с одной и той же дисперсией.

Решение. Задача состоит в проверке гипотезы $H_0 : m_1 = m_2 = m_3$, где m_k — математическое ожидание случайной величины — содержание влаги в продукте с k-м условием хранения, $k = 1, 2, 3$. В нашем случае число уровней фактора «условия хранения продукта», $l = 3$, общий объем всей выборки: $n = 10 + 13 + 7 = 30$.

Вычисления удобно проводить в такой последовательности.

Вычислим суммы элементов выборок для каждого уровня фактора, по группам $x_{.k} = \sum_{i=1}^{n_k} x_{ik}$; $x_{.1} = 73{,}8$; $x_{.2} = 139{,}2$; $x_{.3} = 70{,}3$.

Дисперсионный анализ

Сумма всех элементов выборки равна

$$x_{..} = \sum_{k=1}^{l} x_{.k} = 73{,}8 + 139{,}2 + 70{,}3 = 283{,}3,$$

а сумма их квадратов будет

$$\sum_{k=1}^{l} \sum_{i=1}^{n_k} x_{ik}^2 \approx 2801{,}61.$$

Далее получаем:

$$Q = \sum_{k=1}^{l} \sum_{i=1}^{n_k} (x_{ik} - \bar{x})^2 = \sum_{k=1}^{l} \sum_{i=1}^{n_k} (x_{ik})^2 - \frac{1}{n}(x_{..})^2 = 2801{,}61 - \frac{1}{30}(283{,}3)^2 \approx 126{,}31,$$

$$Q_1 = \sum_{k=1}^{l} n_k (\bar{x}_k - \bar{x})^2 = \sum_{k=1}^{l} \frac{1}{n_k} x_{.k}^2 - \frac{1}{n}(x_{..})^2 = \frac{1}{10} 73{,}8^2 + \frac{1}{13} 139{,}2^2 +$$

$$+ \frac{1}{7} 70{,}3^2 - \frac{1}{30}(283{,}3)^2 \approx 65{,}869,$$

$$Q_2 = Q - Q_1 = 126{,}31 - 65{,}869 = 60{,}441.$$

Дисперсионный анализ

Вычисляем выборочное значение статистики F:

$$F_{в} = \frac{Q_1/(l-1)}{Q_2/(n-l)} = \frac{65{,}869/(3-1)}{60{,}441/(30-3)} \approx 14{,}712.$$

Используя вероятностный калькулятор или таблицы квантилей распределения Фишера, находим $F_{0,95}(2,27) = 3{,}35$. Так как $F_{в} = 14{,}712 > 3{,}35$, то на уровне значимости $\alpha = 0{,}05$ гипотеза о равенстве средних отклоняется: условия хранения продукта оказывают значимое влияние на содержание влаги.

Дисперсионный анализ

Пример 5.2. В условиях примера 5.1 при двусторонних альтернативных гипотезах проверить гипотезы $H_0^{(1)}: m_1 = m_2$; $H_0^{(2)}: m_1 = m_3$; $H_0^{(3)}: m_2 = m_3$; $H_0^{(4)}: \frac{1}{2}(m_1 + m_3) = m_2$.

Решение. В соответствии с проверяемыми гипотезами $H_0^{(i)}$, $i = 1, 2, 3, 4$, определяются линейные контрасты:

$Lk_1 = m_1 - m_2$;	$c_1 = 1$,	$c_2 = -1$,	$c_3 = 0$;
$Lk_2 = m_1 - m_3$;	$c_1 = 1$,	$c_2 = 0$,	$c_3 = -1$;
$Lk_3 = m_2 - m_3$;	$c_1 = 0$,	$c_2 = 1$,	$c_3 = -1$;
$Lk_4 = 1/2(m_1 + m_2) - m_3$;	$c_1 = 1/2$,	$c_2 = 1/2$,	$c_3 = -1$.

Найдем границы доверительных интервалов для линейных контрастов Lk_i, $i = 1, 2, 3, 4$.

Предварительно вычислим оценки линейных контрастов и их дисперсий. Выборочные средние по группам равны: $\bar{x}_1 = 7{,}38$, $\bar{x}_2 \approx 10{,}71$, $\bar{x}_3 \approx 10{,}04$.

Оценка дисперсии ошибок наблюдений:

$$\tilde{\sigma}^2 = \frac{Q_2}{n - l} = \frac{60{,}441}{30 - 3} \approx 2{,}239.$$

Дисперсионный анализ

Вычислим оценки контрастов и их дисперсий:

$$\tilde{L}k_1 = 7{,}38 - 10{,}71 = -3{,}33,\ s^2_{Lk_1} = 2{,}239 \cdot \left(\frac{1}{10} + \frac{1}{13}\right) \approx 0{,}396;$$

$$\tilde{L}k_2 = 7{,}38 - 10{,}04 = -2{,}66,\ s^2_{Lk_2} = 2{,}239 \cdot \left(\frac{1}{10} + \frac{1}{7}\right) \approx 0{,}544;$$

$$\tilde{L}k_3 = 10{,}71 - 10{,}04 = 0{,}67,\ s^2_{Lk_3} = 2{,}239 \cdot \left(\frac{1}{13} + \frac{1}{7}\right) \approx 0{,}492;$$

Дисперсионный анализ

$$\tilde{L}k_4 = \frac{1}{2}(7{,}38 + 10{,}71) - 10{,}04 = -0{,}995,$$

$$s^2_{Lk_4} = 2{,}239 \cdot \left(\frac{(1/2)^2}{10} + \frac{(1/2)^2}{13} + \frac{1}{7} \right) \approx 0{,}419.$$

По таблице (см. [1], либо воспользуйтесь вероятностным калькулятором) находим квантиль распределения Фишера $F_{1-\alpha}(l-1, n-l) = F_{0,95}(2,27) = 3{,}35$. Чтобы определить доверительные интервалы для линейных контрастов, предварительно вычислим

$$\sqrt{(l-1)F_{1-\alpha}(l-1, n-l)} = \sqrt{(3-1)\cdot 3{,}35} \approx 2{,}59.$$

Таким образом, доверительные границы для контрастов Lk_i, $i = 1, 2, 3, 4$, равны соответственно $-3{,}33 \pm 1{,}63$; $-2{,}66 \pm 1{,}91$; $0{,}67 \pm 1{,}82$; $-0{,}995 \pm 1{,}68$.

Так как нулевое значение накрывается доверительными интервалами для Lk_3 и Lk_4, то гипотезы $H_0^{(3)}$ и $H_0^{(4)}$ принимаются, гипотезы $H_0^{(1)}$ и $H_0^{(2)}$ отклоняются. Таким образом, значимо различны средние первой и второй групп, а также средние первой и третьей групп.

Дисперсионный анализ

- **Однофакторный дисперсионный анализ для несвязанных выборок**
- Последовательность операций

T_c	суммы индивидуальных значений по каждому из условий
$\Sigma(T^2_c)$	сумма квадратов суммарных значений по каждому из условий
c	количество условий (градаций фактора)
n	количество значений в каждом комплексе (испытуемых в каждой группе)
N	общее количество индивидуальных значений
$(\Sigma x_i)^2$	квадрат общей суммы индивидуальных значений
$\Sigma(x_i)^2 / N$	константа, необходимая для вычитания из каждой суммы квадратов
x_i	каждое индивидуальное значение
$\Sigma(x_i)^2$	сумма квадратов индивидуальных значений

НЕ ОДНО И ТО ЖЕ !

Дисперсионный анализ

- **Однофакторный дисперсионный анализ для несвязанных выборок**
- Обозначения
- СК или SS – сумма квадратов
- SSфакт. – вариативность, обусловленная действием исследуемого фактора
- SSобщ. – общая вариативность
- SSсл. – случайная вариативность
- MS – «средний квадрат» (математическое ожидание суммы квадратов, усредненная величина соответствующих SS)
- df – число степеней свободы.

Дисперсионный анализ

- **Однофакторный дисперсионный анализ для несвязанных выборок**
- Последовательность операций

Подсчитать SSфакт.	$SS_{факт.} = 1/n\ \Sigma T^2_c - 1/n\ (\Sigma x_i)^2$
Подсчитать SSобщ.	$SS_{общ.} = \Sigma x^2_i - 1/N\ (\Sigma x_i)^2$
Подсчитать случайную остаточную величину SSсл.	$SS_{сл.} = SS_{общ.} - SS_{факт.}$
Определить число степеней свободы	$df_{факт.} = c - 1$ $df_{общ.} = N - 1$ $df_{сл.} = df_{общ.} - df_{факт.}$
Разделить каждую SS на соответствующее число степеней свободы	$MS_{факт.} = SS_{факт.} / df_{факт.}$ $MS_{сл.} = SS_{сл.} / df_{сл.}$
Подсчитать значение Fэмп.	$F_{эмп.} = MS_{факт.} / MS_{сл.}$
Определить по таблицам критические значения F и сопоставить с ним полученное эмпирическое значение	При $F_{эмп.} >= F_{кр.}$ H_0 отклоняется.

Дисперсионный анализ

T_c	Суммы индивидуальных значений по каждому из условий
ΣT^2_c	Сумма квадратов суммарных значений по каждому из условий
c	Количество значений у каждого респондента, то есть – количество условий
n	Количество респондентов
N	общее количество значений
T_n	Суммы индивидуальных значений по каждому респонденту
ΣT^2_n	Сумма квадратов сумм индивидуальных значений по респондентам
x_i	каждое индивидуальное значение
$(\Sigma x_i)^2$	квадрат общей суммы индивидуальных значений
$1/N(\Sigma x_i)^2$	константа, необходимая для вычитания из каждой суммы квадратов
$\Sigma(x_i)^2$	сумма квадратов индивидуальных значений

Дисперсионный анализ

Подсчитать SSфакт.	$SS_{факт.} = 1/n\ \Sigma T^2 c - 1/n\ (\Sigma xi)^2$
Подсчитать SSресп.	$SS_{респ.} = 1/c\ \Sigma T^2 n - 1/N\ (\Sigma xi)^2$
Подсчитать SSобщ.	$SS_{общ.} = \Sigma x^2 i - 1/N\ (\Sigma xi)^2$
Подсчитать случайную остаточную величину SSсл.	$SS_{сл.} = SS_{общ.} - SS_{факт.} - SS_{респ.}$
Определить число степеней свободы	$df_{факт.} = c - 1$ $df_{респ.} = n - 1$ $df_{общ.} = N - 1$ $df_{сл.} = df_{общ.} - df_{факт.} - df_{респ.}$
Разделить каждую SS на соответствующее число степеней свободы	$MS_{факт.} = SS_{факт.} / df_{факт.}$ $MS_{респ.} = SS_{респ.} / df_{респ.}$ $MS_{сл.} = SS_{сл.} / df_{сл.}$
Подсчитать значения F	$F_{факт.} = MS_{факт.} / MS_{сл.}$ $F_{респ.} = MS_{респ.} / MS_{сл.}$
Определить по таблицам критические значения F и сопоставить с ними полученные эмпирические значения	При $F_{эмп.} >= F_{кр.}$ H_0 отклоняется.

Корреляционный анализ

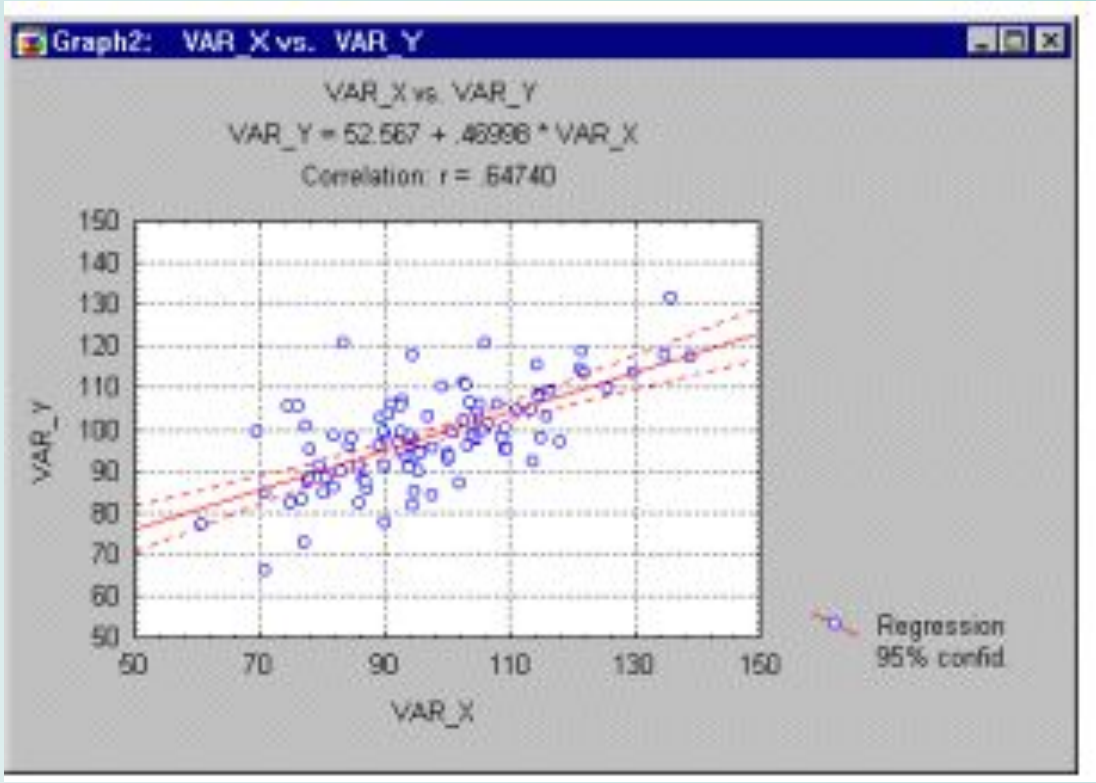
Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая
		средняя		сильная	

Корреляционный анализ

- Коэффициенты корреляции в зависимости от типа переменных

Тип шкалы		**Мера связи**
Переменная X	**Переменная У**	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент «ϕ»
Дихотомическая	Ранговая	Рангово-бисериальный
Дихотомическая	Интервальная или отношений	Бисериальный

Линия регрессии

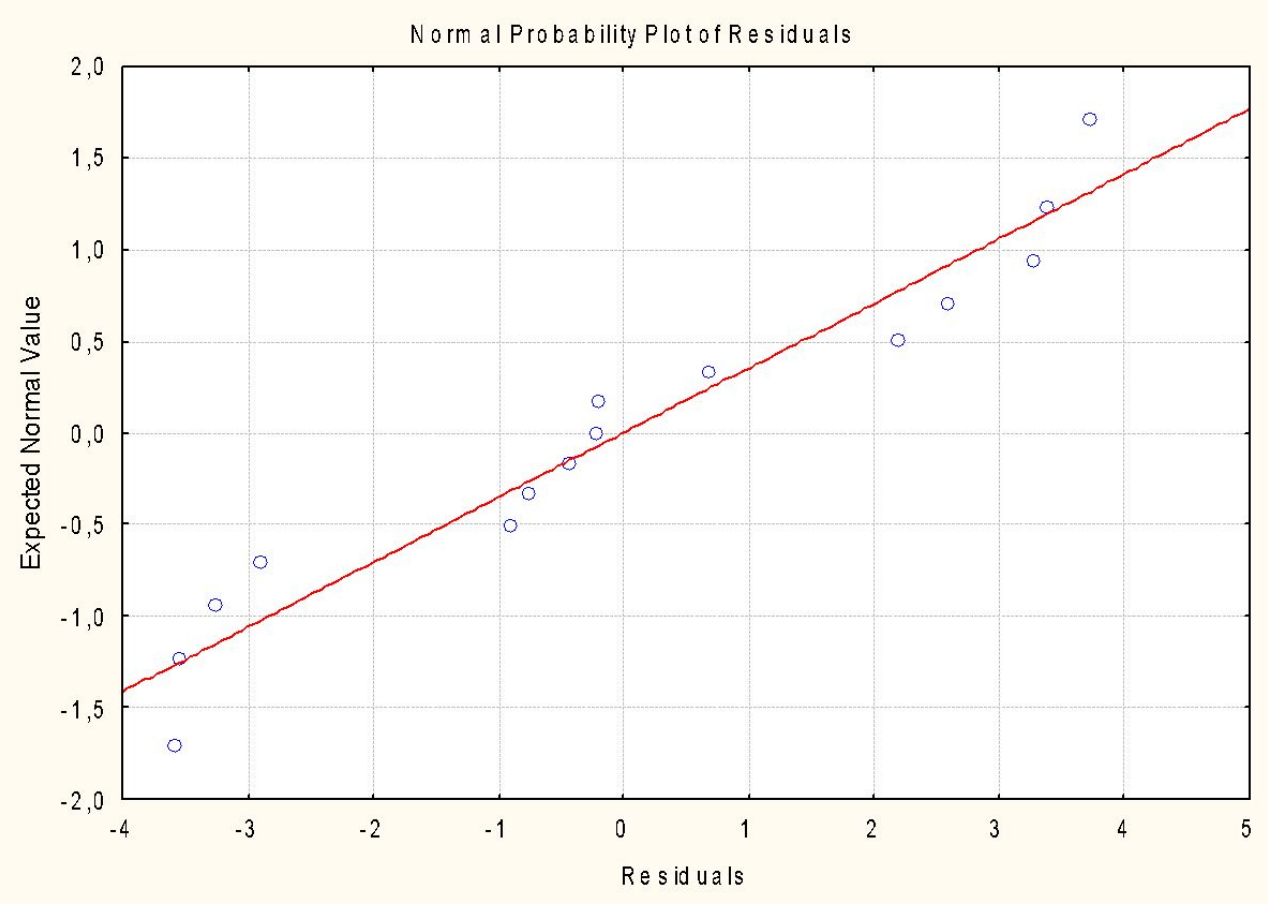


Регрессионный анализ

Анализ остатков

Case No.	Predicted & Residual Values (Product.sta) Dependent variable: Product Include condition: z=1 Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	18,30000	21,89366	-3,59366	-1,01506	-1,33665	1,006910	1,030354	-4,17996	0,169520
2	31,10000	27,37918	3,72082	-0,29810	1,38395	0,726478	0,088866	4,01389	0,081371
3	27,00000	23,62593	3,37407	-0,78866	1,25498	0,896114	0,621977	3,79575	0,110718
4	37,90000	35,31874	2,58126	0,73960	0,96009	0,874250	0,547008	2,88647	0,060940
5	20,30000	23,19286	-2,89286	-0,84526	-1,07599	0,922371	0,714460	-3,27877	0,087525
6	32,40000	33,15341	-0,75341	0,45659	-0,28023	0,767805	0,208474	-0,82031	0,003796
7	31,20000	34,74132	-3,54132	0,66413	-1,31719	0,842387	0,441070	-3,92682	0,104713
8	39,70000	42,96960	-3,26960	1,73957	-1,21612	1,429786	3,026102	-4,55894	0,406599
9	46,60000	44,41315	2,18684	1,92824	0,81339	1,549704	3,718121	3,27493	0,246489
10	33,10000	29,83323	3,26677	0,02264	1,21507	0,694372	0,000513	3,50025	0,056530
11	26,90000	26,22433	0,67567	-0,44904	0,25131	0,765504	0,201640	0,73528	0,003032
12	24,00000	24,20335	-0,20335	-0,71319	-0,07564	0,862844	0,508634	-0,22670	0,000366
13	24,20000	24,63642	-0,43642	-0,65658	-0,16233	0,839327	0,431102	-0,48355	0,001576
14	33,70000	34,59697	-0,89697	0,64526	-0,33362	0,834781	0,416365	-0,99266	0,006571
15	18,50000	18,71784	-0,21784	-1,43015	-0,08102	1,240121	2,045315	-0,27671	0,001127
Minimum	18,30000	18,71784	-3,59366	-1,43015	-1,33665	0,694372	0,000513	-4,55894	0,000366
Maximum	46,60000	44,41315	3,72082	1,92824	1,38395	1,549704	3,718121	4,01389	0,406599
Mean	29,66000	29,66000	0,00000	0,00000	0,00000	0,950184	0,933333	-0,03586	0,089392

Регрессионный анализ



Регрессионный анализ

- Пример расчетов

Пример 6.1. Пример простой линейной регрессии Y на x. Исходные данные: результаты наблюдений зависимой переменной (y) и фактора (x) следующие:

y	x
4,0	5,5
5,6	8,1
5,7	8,5
3,6	5,9
4,0	7,8

Решение.

1. По данным примера вычислим суммы квадратов Q_y, Q_x и сумму произведений Q_{xy}; $n = 5$. Предварительно найдем средние значения:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{5}(5{,}5 + 8{,}1 + 8{,}5 + 5{,}9 + 7{,}8) = 7{,}16;$$

Регрессионный анализ

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{5}(4 + 5{,}6 + 5{,}7 + 3{,}6 + 4) = 4{,}58;$$

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2 = (5{,}5^2 + 8{,}1^2 + 8{,}5^2 + 5{,}9^2 + 7{,}8^2) -$$
$$- 5 \cdot (7{,}16)^2 = 263{,}76 - 5 \cdot 51{,}266 = 7{,}432;$$

$$Q_y = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = (4^2 + 5{,}6^2 + 5{,}7^2 + 3{,}6^2 + 4^2) -$$
$$- 5 \cdot (4{,}58)^2 = 108{,}81 - 5 \cdot 20{,}976 = 3{,}928;$$

$$Q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \cdot \bar{x} \cdot \bar{y} = (5{,}5 \cdot 4 + 8{,}1 \cdot 5{,}6 + 8{,}5 \cdot 5{,}7 +$$
$$+ 5{,}9 \cdot 3{,}6 + 7{,}8 \cdot 4) - 5 \cdot 7{,}16 \cdot 4{,}58 = 168{,}25 - 5 \cdot 7{,}16 \cdot 4{,}58 = 4{,}289.$$

Оценки параметров линейной регрессии $y = \beta_0 + \beta_1 x$, по формулам (1) и (2) равны:

$$\tilde{\beta}_1 = \frac{Q_{xy}}{Q_x} = \frac{4{,}289}{7{,}432} \approx 0{,}577;$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \cdot \bar{x} = 4{,}58 - 0{,}577 \cdot 7{,}16 \approx 0{,}451.$$

Таким образом, уравнение линейной регрессии Y на x имеет вид

$$y = 0{,}451 + 0{,}577x.$$

Регрессионный анализ

Аналогично, оценки параметров линейной регрессии X на y:

$$\tilde{\beta}_1' = \frac{Q_{xy}}{Q_y} \approx 1{,}091;\ \tilde{\beta}_0' = \bar{x} - \tilde{\beta}_1' \bar{y} = 7{,}16 - 1{,}091 \cdot 4{,}58 \approx 2{,}163.$$

Уравнение линейной регрессии X на y имеет вид

$$x = 2{,}163 + 1{,}091y.$$

2. Диаграмма рассеяния исходных данных и прямая регрессии Y на x показана на рис. 6.2.

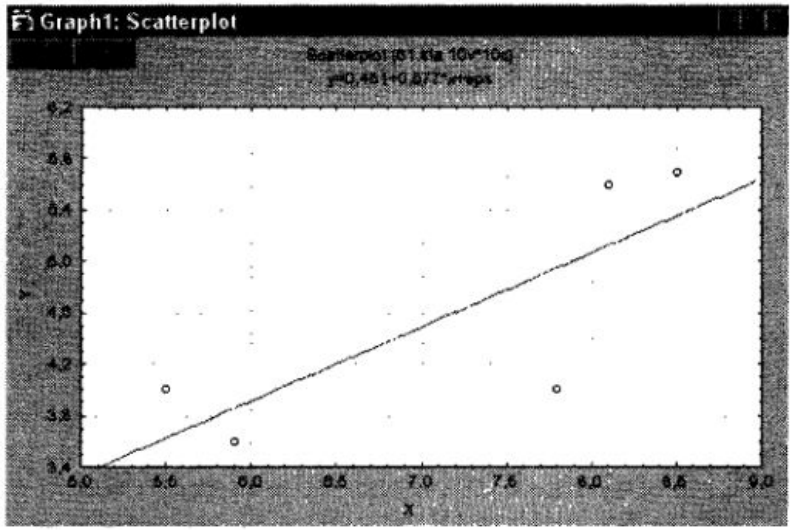


Рис. 6.2. Диаграмма рассеяния и прямая регрессии Y на x

Регрессионный анализ

3. Для линейной регрессии Y на x вычислим остатки:

$$e_i = y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i),\ i = 1, 2, \ldots, 5;$$

$$e_1 = 4 - (0{,}451 + 0{,}577 \cdot 5{,}5) = 0{,}377;$$

$$e_2 = 5{,}6 - (0{,}451 + 0{,}577 \cdot 8{,}1) = 0{,}478;$$

.....

$$e_5 = 4 - (0{,}451 + 0{,}577 \cdot 7{,}8) = -0{,}949.$$

Остаточная сумма квадратов Q_e:

$$Q_e = (0{,}377)^2 + (0{,}478)^2 + (0{,}35)^2 + (-0{,}25)^2 + (-0{,}949)^2 \approx 1{,}457.$$

Оценка дисперсии ошибок наблюдений

$$S^2 = \frac{Q_e}{n-k} = \frac{1{,}457}{5-2} \approx 0{,}486,$$

где k — число оцениваемых параметров; для простой линейной регрессии $k = 2$.

Регрессионный анализ

Коэффициент детерминации R^2:

$$R^2 = 1 - \frac{Q_e}{Q_y} = 1 - \frac{1{,}457}{3{,}928} \approx 0{,}629.$$

Оценка коэффициента корреляции r:

$$r = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}} = \frac{4{,}286}{\sqrt{7{,}438 \cdot 3{,}928}} \approx 0{,}793.$$

4. Вычислим оценки параметров линейной регрессии Y на x в матричном виде, используя формулу (5):

$$\tilde{\beta} = (A^T A)^{-1} A^T Y,$$

где $\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix}$; A — регрессионная матрица: $A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} 1 & 5{,}5 \\ 1 & 8{,}1 \\ 1 & 8{,}5 \\ 1 & 5{,}9 \\ 1 & 7{,}8 \end{pmatrix}$;

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix}.$$

Регрессионный анализ

Последовательно вычисляем:

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix},$$

$$B = A^T A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix} \cdot \begin{pmatrix} 1 & 5{,}5 \\ 1 & 8{,}1 \\ 1 & 8{,}5 \\ 1 & 5{,}9 \\ 1 & 7{,}8 \end{pmatrix} = \begin{pmatrix} 5 & 35{,}8 \\ 35{,}8 & 263{,}76 \end{pmatrix}.$$

Определитель матрицы B:

$$|B| = \det(A^T A) = 37{,}16.$$

Обратная матрица к матрице B:

$$B^{-1} = \frac{1}{|B|} \cdot B^* = \frac{1}{37{,}16} \cdot \begin{pmatrix} 263{,}76 & -35{,}8 \\ -35{,}8 & 5 \end{pmatrix} = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix},$$

где B^* — присоединенная матрица к матрице B, составленная из алгебраических дополнений к элементам матрицы B.

Регрессионный анализ

Далее вычисляем произведения матриц

$$B^{-1} \cdot A^T = \begin{pmatrix} 7{,}098 & -0{,}963 \\ -0{,}963 & 0{,}135 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 5{,}5 & 8{,}1 & 8{,}5 & 5{,}9 & 7{,}8 \end{pmatrix} =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix}.$$

Окончательно

$$\tilde{\beta} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \end{pmatrix} = B^{-1} A^T Y =$$

$$= \begin{pmatrix} 1{,}7992 & -0{,}7056 & -1{,}0910 & 1{,}4139 & -0{,}4166 \\ -0{,}2234 & 0{,}1265 & 0{,}1803 & -0{,}1695 & 0{,}0861 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 5{,}6 \\ 5{,}7 \\ 3{,}6 \\ 4 \end{pmatrix} = \begin{pmatrix} 0{,}4509 \\ 0{,}5767 \end{pmatrix}.$$

Сравнивая полученные значения с результатами в п. 2 видно, что расхождение имеется только в третьем десятичном знаке.

Задание на л/р

- По результатам статистического исследования физического развития мальчиков 5 лет известно, что их средний рост (x) равен 109 см, а средняя масса тела (y) равна 19 кг. Коэффициент корреляции между
- ростом и массой тела составляет + 0,9, средние квадратические отклонения представлены в таблице.
- Требуется:
- 1) рассчитать коэффициент регрессии;
- 2) по уравнению регрессии определить, какой будет ожидаемая масса
- тела мальчиков 5 лет при росте, равном x1 = 100 см, x2 = 110 см,
- x3= 120 см;
- 3) рассчитать сигму регрессии, построить шкалу регрессии и предста-
- вить результаты ее решения в графическом виде;
- 4) сделать соответствующие выводы.

Задание на л/р

Условия задачи					Результаты решения задачи				
					уравнение регрессии		сигма рег-рессии	шкала регрессии (ожидаемая масса тела, кг)	
	M	σ	r_{xy}	$R_{y/x}$	x, см	y, кг	σ $R_{y/x}$	$y - \sigma_{Ry/x}$	$y + \sigma_{Ry/x}$
1	2	3	4	5	6	7	8	9	10
Рост, см (x)	109	± 4,4	+0,9	0,16	100	17,56	± 0,35	17,21	17,91
Масса тела, кг (y)	19	±0,8			110	19,16		18,81	19,51
					120	20,76		20,41	21,11

Решение задачи

- ЭТАПЫ РЕШЕНИЯ ЗАДАЧИ
- 1. Коэффициент регрессии:
- $R_{y/x} = r_{xy} \times (\sigma_y/\sigma_x) = +0,9 \times (0,8/4,4) = 0,16$ кг/см.
- Таким образом, при увеличении роста мальчиков 5 лет на 1 м масса
- тела увеличивается на 0,16 кг.
- 2. Уравнение регрессии:
- $y = M_y + R_{y/x} (x – M_x)$
- $x1 = 100$ см
- $x2 = 110$ см
- $x3 = 120$ см
- $y1 = 19 + 0,16 (100 – 109) = 17,56$ кг
- $y2 = 19 + 0,16 (110 – 109) = 19,16$ кг
- $y3 = 19 + 0,16 (120 – 109) = 20,76$ кг

Решение

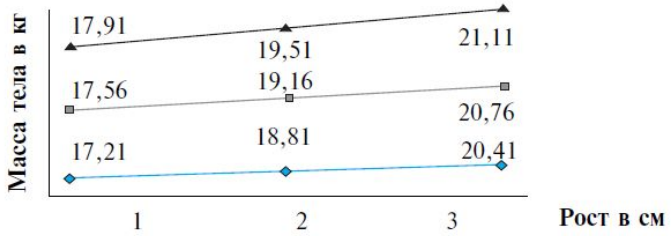
3. Сигма регрессии:

$\sigma_{Ry/x} = \sigma_y \sqrt{1 - r_{xy}^2}$; $\sigma_{Ry/x} = 0{,}8 \sqrt{1 - 0{,}9^2} = \pm 0{,}35$ кг

4. Шкала регрессии:

Рост, см	Среднее значение массы тела, кг	Наименьшее значение массы тела, кг	Наибольшее значение массы тела, кг
x	y	$y - \sigma_{Ry/x}$	$y + \sigma_{Ry/x}$
100	17,56	17,21	17,91
110	19,16	18,81	19,51
120	20,76	20,41	21,11

5. Графическое изображение регрессии:



Шкала регрессии массы тела по росту 5-летних мальчиков

Вывод: таким образом, шкала регрессии в пределах расчетных величин массы тела позволяет определить ее при любом другом значении роста или оценить индивидуальное развитие ребенка. Для этого следует восстановить перпендикуляр к линии регрессии.

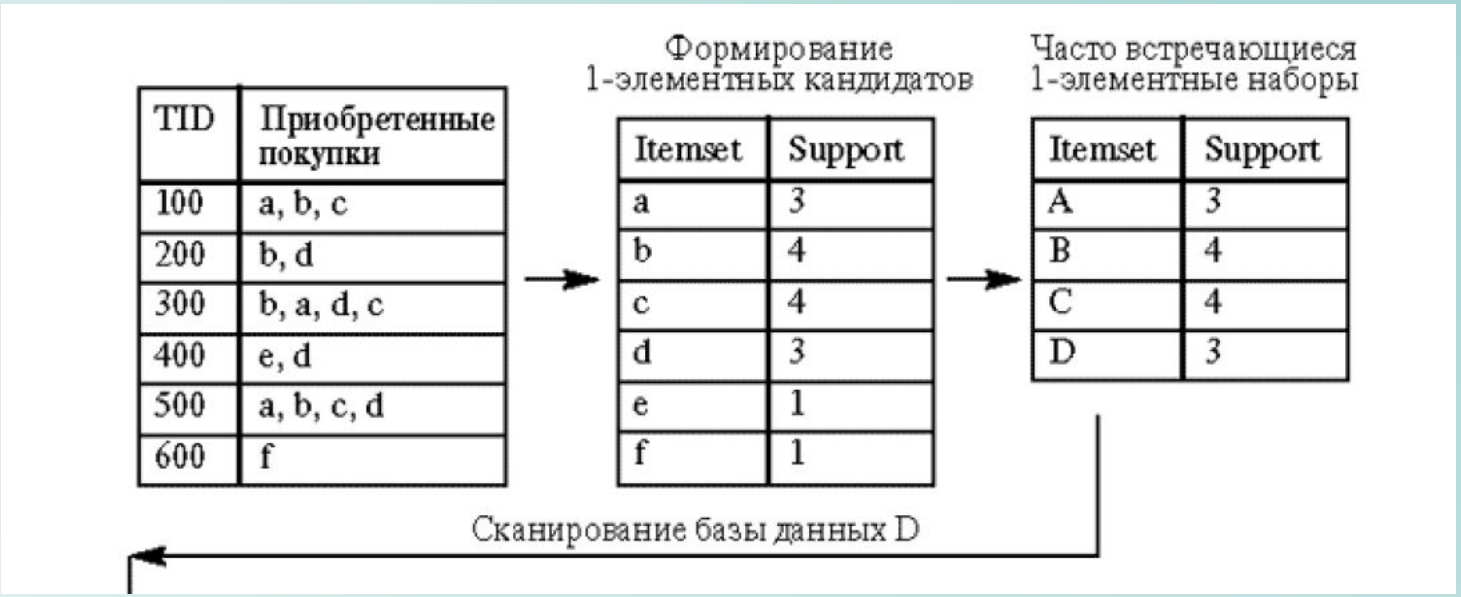
Транзакции

Объектно-признаковая таблица транзакций

Покупатели/товары	Пиво	Пряники	Молоко	Мюсли	Чипсы
C_1	1	0	0	0	1
C_2	0	1	1	1	0
C_3	1	0	1	1	1
C_4	1	1	1	0	1
C_5	0	1	1	1	1

- $supp(\{\text{Пиво, Чипсы}\}) = 3/5$
- $supp(\{\text{Пряники, Мюсли }\} \rightarrow \{\text{ Молоко }\}) = \frac{|(\{\text{Пряники, Мюсли}\} \cup \{\text{Молоко}\})'|}{|G|} = \frac{|\{C2, C5\}|}{5} = 2/5$
- $conf(\{\text{Пряники, Мюсли }\} \rightarrow \{\text{ Молоко }\}) = \frac{|(\{\text{Пряники, Мюсли}\} \cup \{\text{Молоко}\})'|}{|\{\text{Пряники, Мюсли}\}'|} = \frac{|\{C2, C5\}|}{|\{C2, C5\}|} = 1$

Алгоритм (* [Чубукова])

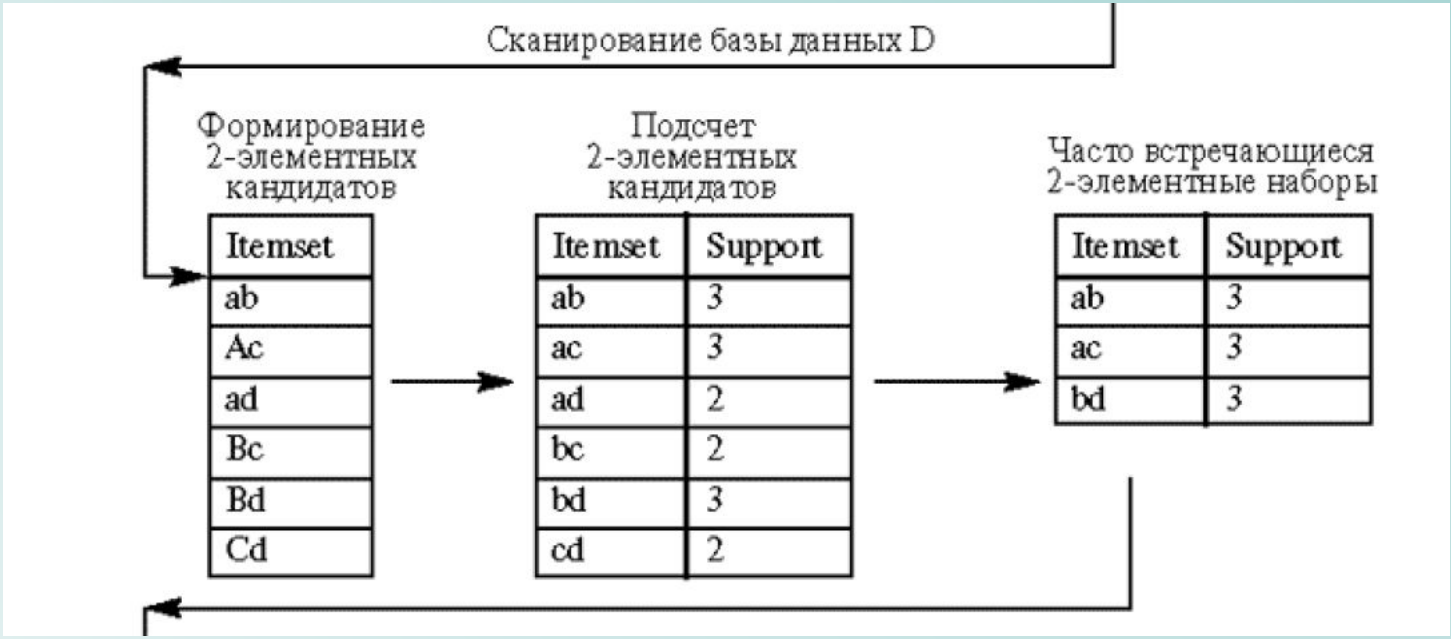


TID	Приобретенные покупки
100	a, b, c
200	b, d
300	b, a, d, c
400	e, d
500	a, b, c, d
600	f

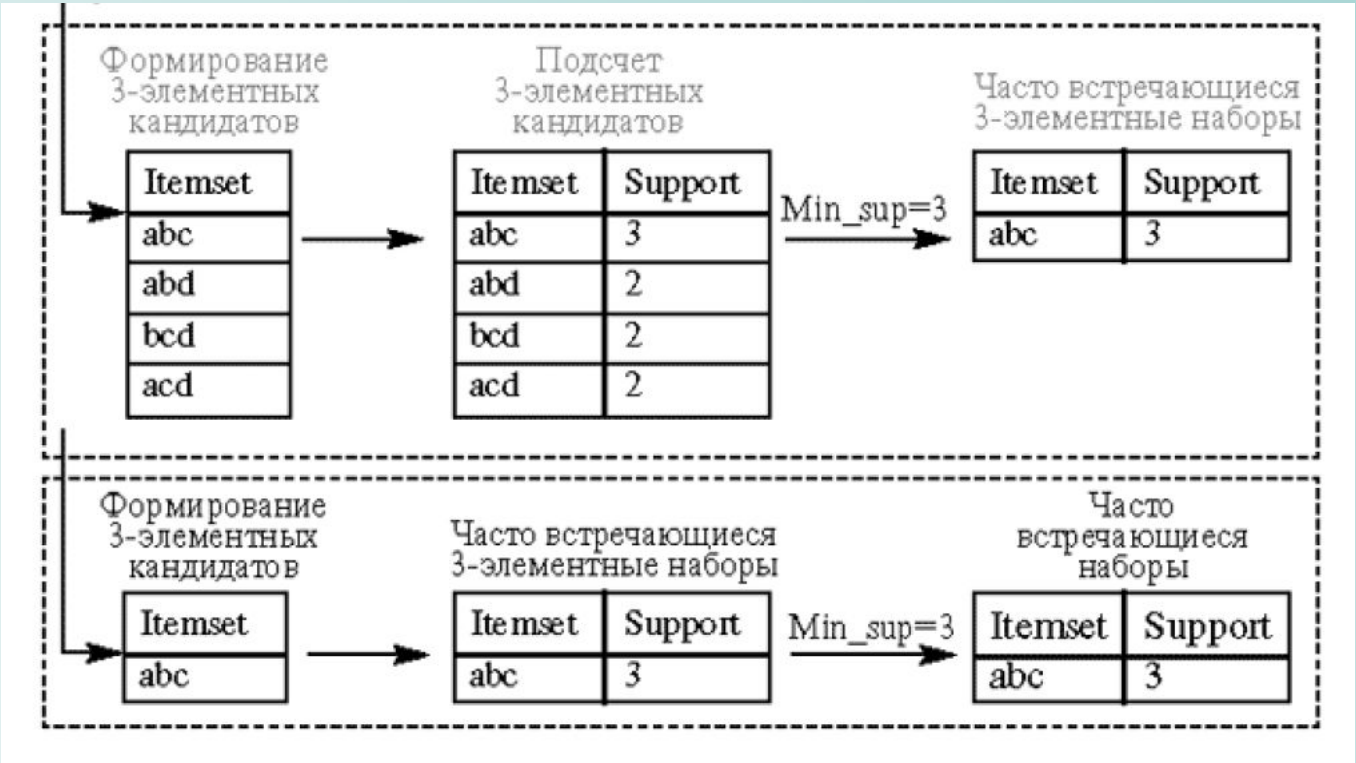
Itemset	Support
a	3
b	4
c	4
d	3
e	1
f	1

Itemset	Support
A	3
B	4
C	4
D	3

Алгоритм (* Чубукова)



Алгоритм (* Чубукова)



Алгоритм

Алгоритм 2.5.1. APRIORI($Context, min\_supp$)

комментарий: Context - набор данных, min_supp - минимальная поддержка,

I_F — все частые множества признаков.

$C_1 \leftarrow \{1\text{-itemsets}\}$
$i \leftarrow 1$
while $(C_i \neq \emptyset)$
do $\begin{cases} SupportCount(C_i) \\ F_i \leftarrow \{f \in C_i \mid f.support \geq min\_supp\} // F - \text{частые множества признаков} \\ C_{i+1} \leftarrow AprioriGen(F_i) // C - \text{кандидаты} \\ i{+}{+} \end{cases}$
$I_F \leftarrow \bigcup F_i$
return (I_F)

Алгоритм 2.5.2. APRIORIGEN(F_i)

комментарий: F_i - частые множества признаков длины i

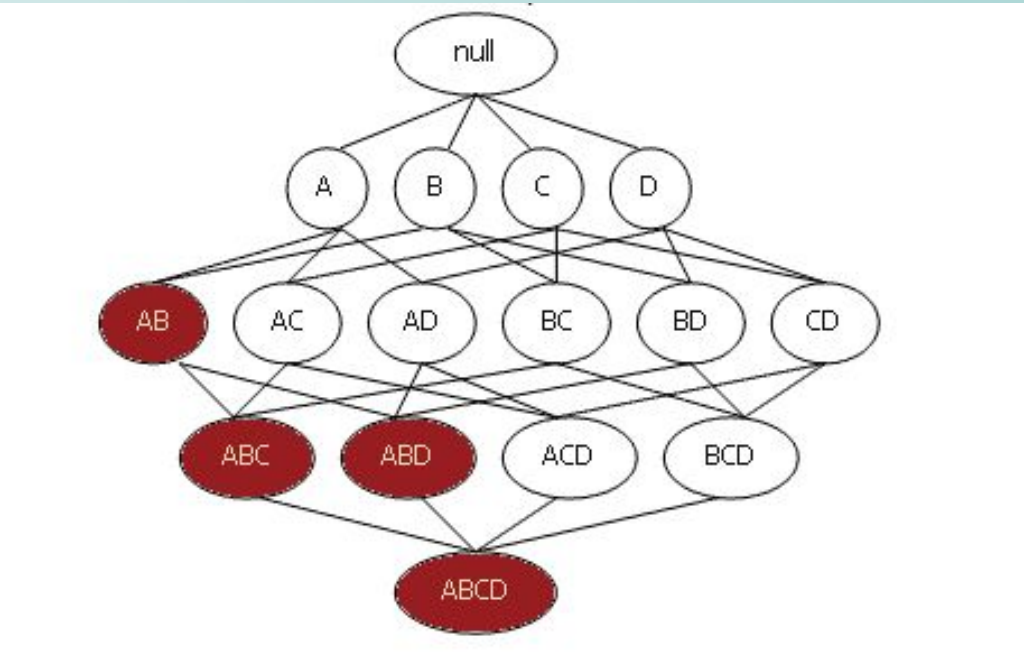
C_{i+1} — потенциальные кандидаты частых множеств признаков.

insert into C_{i+1} // объединение
select $p[1], p[2], ..., p[i], q[i]$
from $F_i p, F_i q$
where $p[1] = q[1], ..., p[i-1] = q[i-1], p[i] < q[i]$
for each $c \in C_{i+1}$ // удаление
do $\begin{cases} S \leftarrow (i-1)\text{-элементные подмножества } c \\ \textbf{for each } s \in S \\ \quad \textbf{do} \begin{cases} \textbf{if } (s \notin F_i) \\ \quad \textbf{then } C_{i+1} \leftarrow C_{i+1} \setminus c \end{cases} \end{cases}$
return (C_{i+1})

Алгоритм

1. F_1 = {часто встречающиеся 1-элементные наборы}
2. для (k=2; $F_{k-1} <> \varnothing$; k++) {
3. C_k = Apriorigen(Fk-1) // генерация кандидатов
4. для всех транзакций $t \in T$ {
5. C_t = subset(C_k, t) // удаление избыточных правил
6. для всех кандидатов $c \in C_t$
7. c.count ++
8. }
9. F_k = { $c \in Ck$ | c.count >= minsupport} // отбор кандидатов
10. }
11. Результат $\cup F_k$

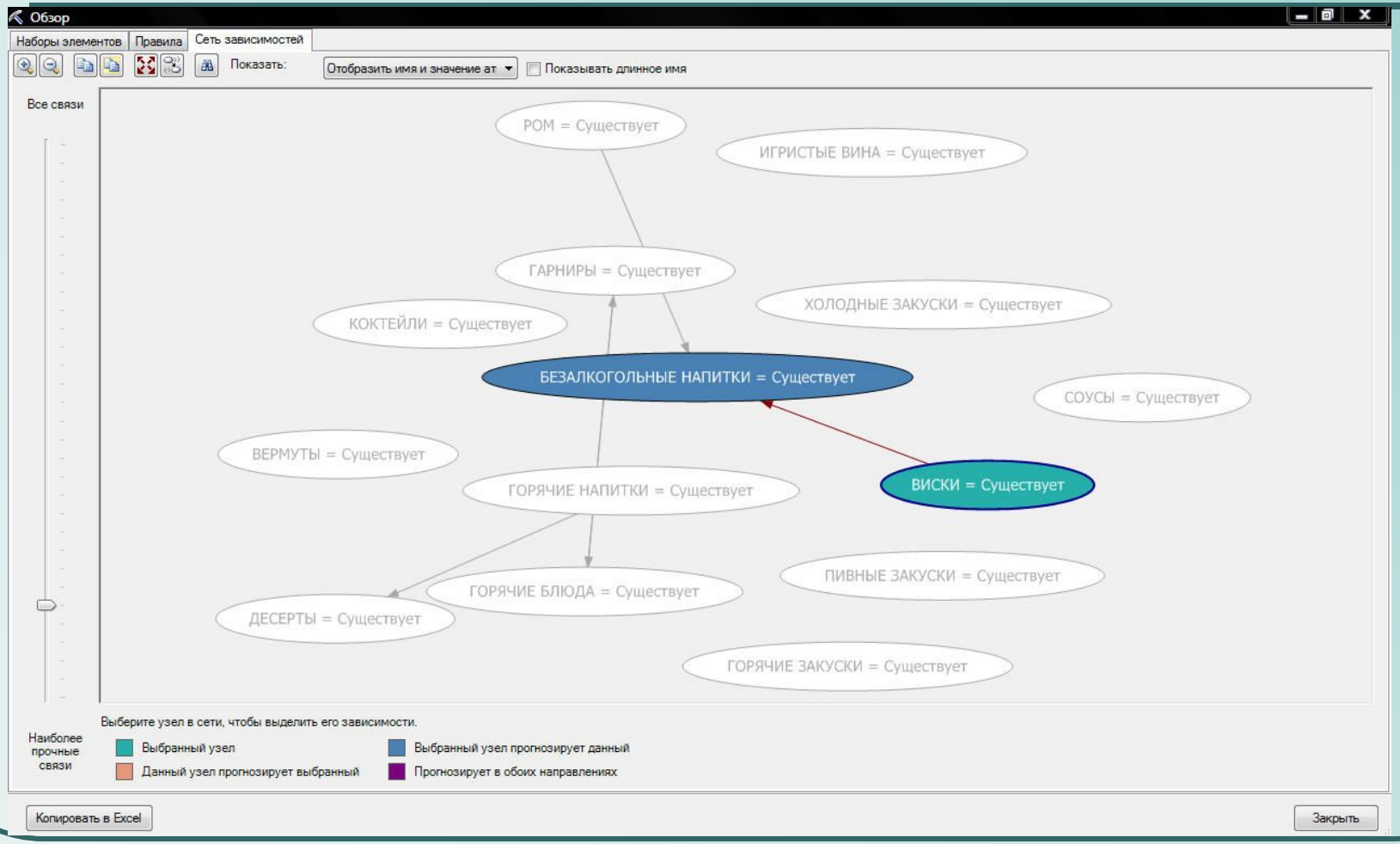
Алгоритм. Свойство антимонотонности



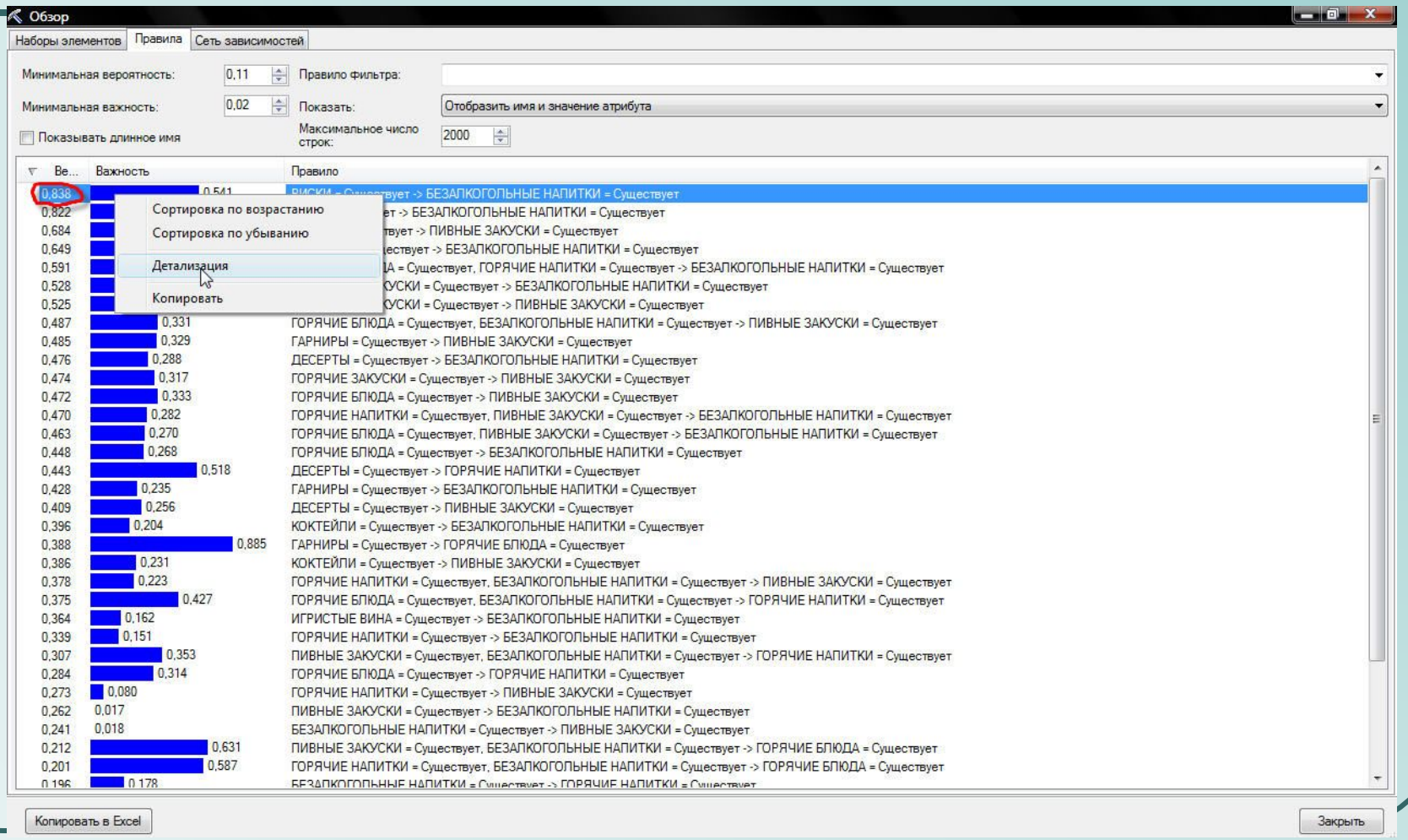
Алгоритм

```
1.  // Оптимизация 1
2.  Вычислить I* множества предков элементов для каждого элемента;
3.  L1 = {Часто встречающиеся множества элементов и групп элементов};
4.  для ( k = 2; Lk-1 <> ∅; k++ )
5.      {
6.          Ck = {Генерация кандидатов мощностью k на основе Lk-1};
7.          // Оптимизация 2
8.          если k = 2 то
9.              удалить те кандидаты из Ck, которые содержат элемент и его предок;
10.         // Оптимизация 3
11.         Пометить как удаленные множества предков элемента, который не содержится в списке кандидатов;
12.         для всех транзакций t ∈D
13.         {
14.             // Оптимизация 3
15.             для каждого элемента x ∈t
16.                 добавить всех предков x из I* к t;
17.             Удалить дубликаты из транзакции t;
18.             // Оптимизация 4,5
19.             если (t не помечена как удаленная) и ( |t| >= k) то
20.             {
21.                 для всех кандидатов c ∈Ck
22.                     если c ⊆t то
23.                         c.count++;
24.                 // Оптимизация 5
25.                 если в транзакцию t не вошел ни один кандидат то
26.                     пометить эту транзакцию как удаленную;
27.             }
28.         }
29.         // Отбор кандидатов
30.         Lk = { c >= Ck | c.count >= minsupp };
31. }
32. Результат = UkLk;
```

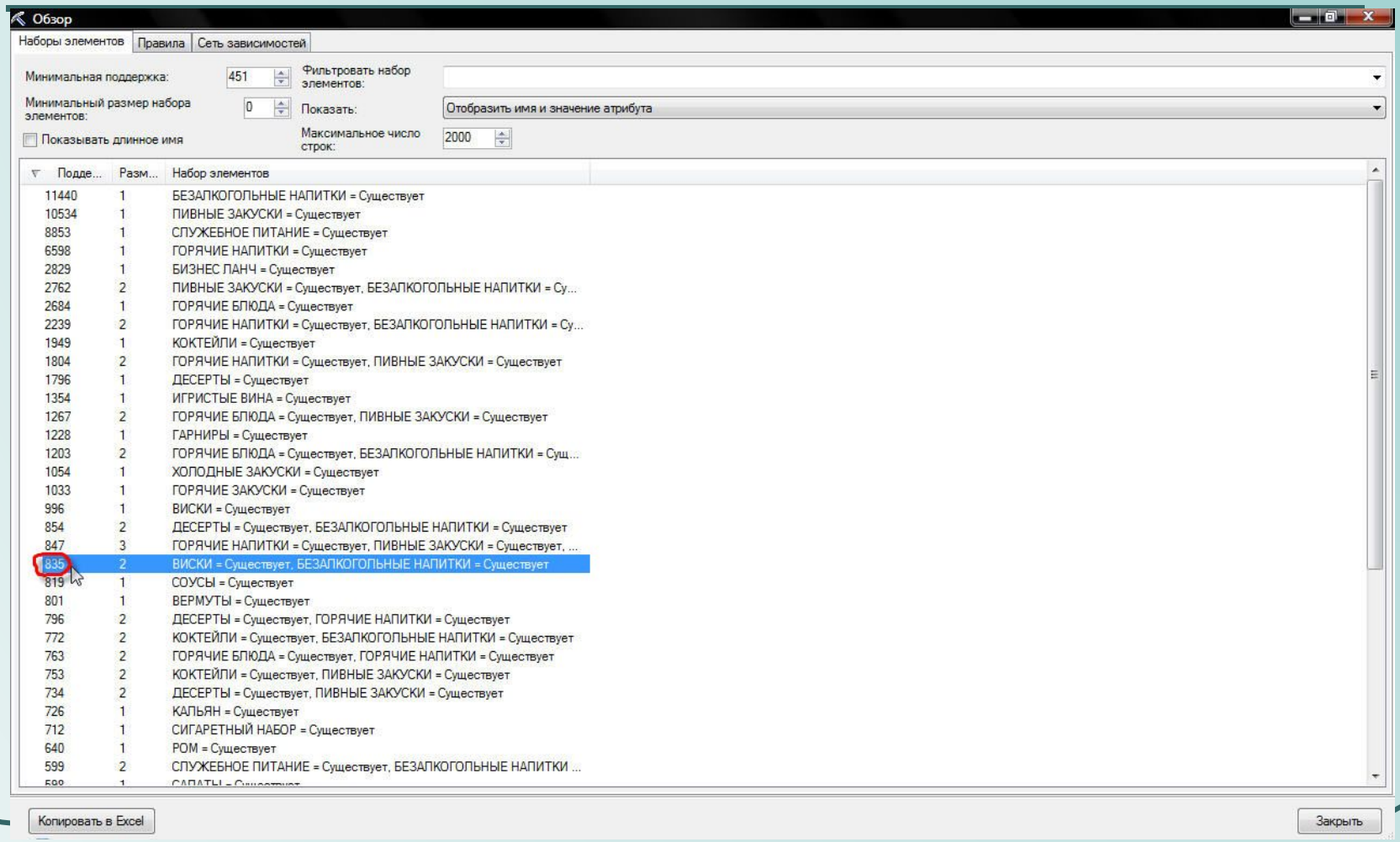
Примеры


Обзор
Наборы элементов
Правила
Сеть зависимостей
Показать:
Отобразить имя и значение ат
Показывать длинное имя
Все связи
РОМ = Существует
ИГРИСТЫЕ ВИНА = Существует
ГАРНИРЫ = Существует
ХОЛОДНЫЕ ЗАКУСКИ = Существует
КОКТЕЙЛИ = Существует
БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Существует
СОУСЫ = Существует
ВЕРМУТЫ = Существует
ГОРЯЧИЕ НАПИТКИ = Существует
ВИСКИ = Существует
ПИВНЫЕ ЗАКУСКИ = Существует
ГОРЯЧИЕ БЛЮДА = Существует
ДЕСЕРТЫ = Существует
ГОРЯЧИЕ ЗАКУСКИ = Существует
Выберите узел в сети, чтобы выделить его зависимости.
Наиболее прочные связи
Выбранный узел
Данный узел прогнозирует выбранный
Выбранный узел прогнозирует данный
Прогнозирует в обоих направлениях
Копировать в Excel
Закрыть

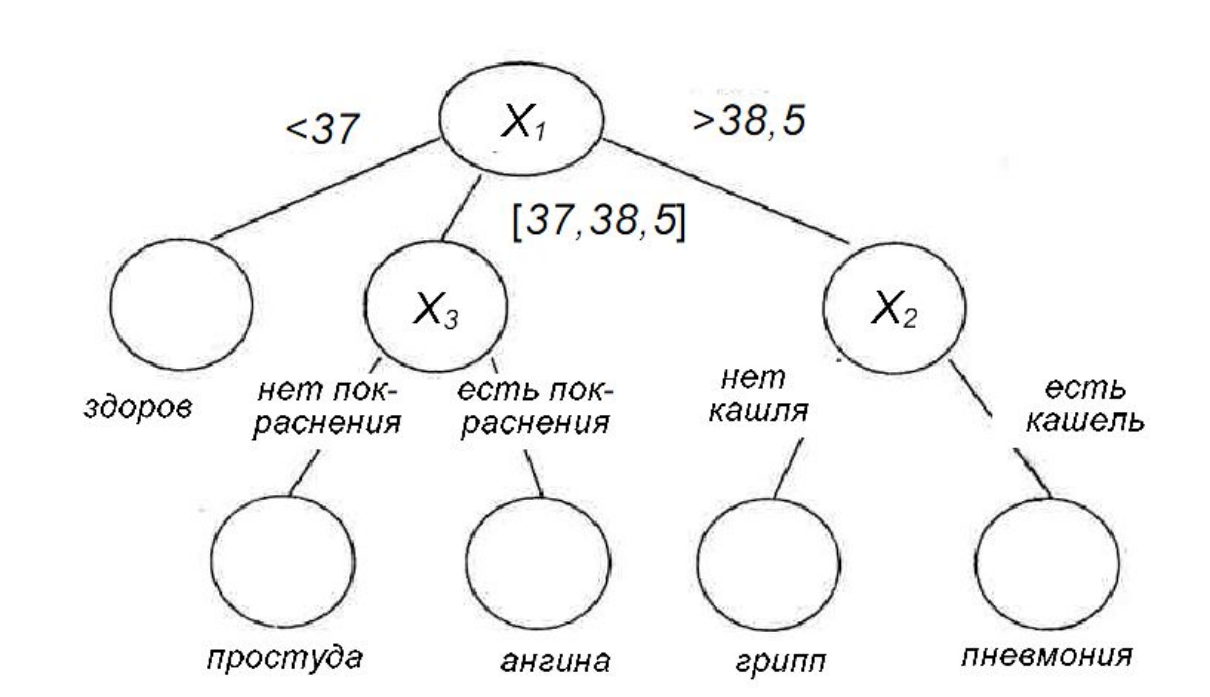
Примеры


Обзор
Наборы элементов
Правила
Сеть зависимостей
Минимальная вероятность:
0,11
Правило фильтра:
Минимальная важность:
0,02
Показать:
Отобразить имя и значение атрибута
Показывать длинное имя
Максимальное число строк:
2000
Ве...
Важность
Правило
Сортировка по возрастанию
Сортировка по убыванию
Детализация
Копировать
Копировать в Excel
Закрыть

Примеры

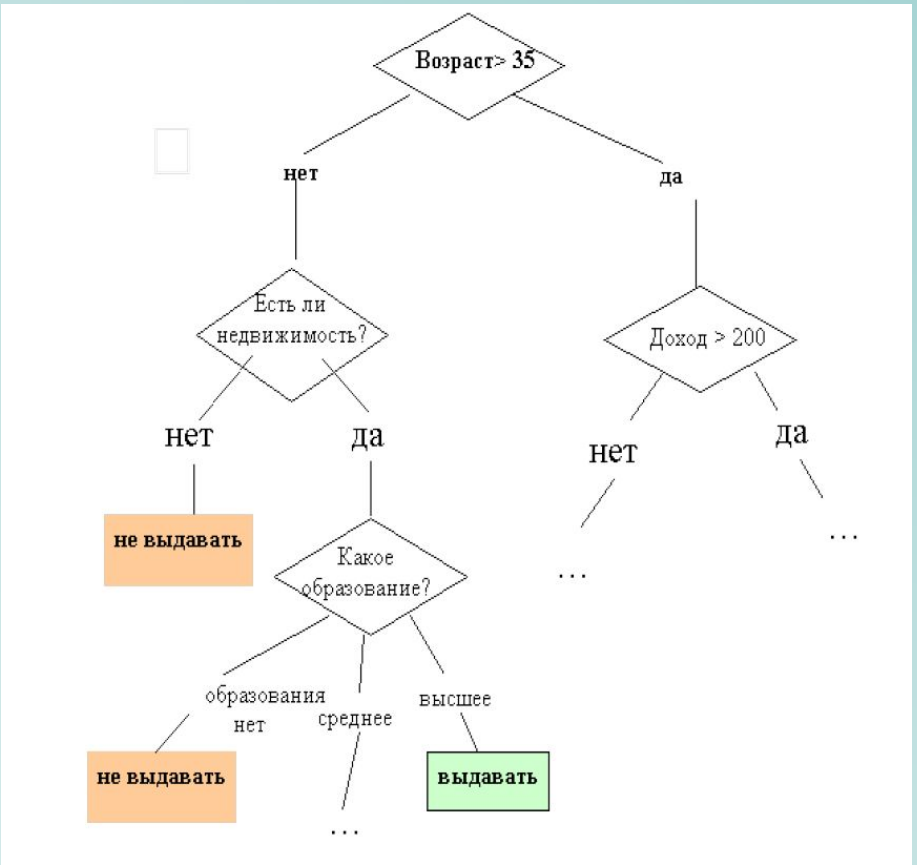

Обзор
Наборы элементов
Правила
Сеть зависимостей
Минимальная поддержка:
451
Фильтровать набор элементов:
Минимальный размер набора элементов:
0
Показать:
Отобразить имя и значение атрибута
Показывать длинное имя
Максимальное число строк:
2000
Подде...
Разм...
Набор элементов
11440 1 БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Существует
10534 1 ПИВНЫЕ ЗАКУСКИ = Существует
8853 1 СЛУЖЕБНОЕ ПИТАНИЕ = Существует
6598 1 ГОРЯЧИЕ НАПИТКИ = Существует
2829 1 БИЗНЕС ЛАНЧ = Существует
2762 2 ПИВНЫЕ ЗАКУСКИ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Су...
2684 1 ГОРЯЧИЕ БЛЮДА = Существует
2239 2 ГОРЯЧИЕ НАПИТКИ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Су...
1949 1 КОКТЕЙЛИ = Существует
1804 2 ГОРЯЧИЕ НАПИТКИ = Существует, ПИВНЫЕ ЗАКУСКИ = Существует
1796 1 ДЕСЕРТЫ = Существует
1354 1 ИГРИСТЫЕ ВИНА = Существует
1267 2 ГОРЯЧИЕ БЛЮДА = Существует, ПИВНЫЕ ЗАКУСКИ = Существует
1228 1 ГАРНИРЫ = Существует
1203 2 ГОРЯЧИЕ БЛЮДА = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Сущ...
1054 1 ХОЛОДНЫЕ ЗАКУСКИ = Существует
1033 1 ГОРЯЧИЕ ЗАКУСКИ = Существует
996 1 ВИСКИ = Существует
854 2 ДЕСЕРТЫ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Существует
847 3 ГОРЯЧИЕ НАПИТКИ = Существует, ПИВНЫЕ ЗАКУСКИ = Существует, ...
835 2 ВИСКИ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Существует
819 1 СОУСЫ = Существует
801 1 ВЕРМУТЫ = Существует
796 2 ДЕСЕРТЫ = Существует, ГОРЯЧИЕ НАПИТКИ = Существует
772 2 КОКТЕЙЛИ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ = Существует
763 2 ГОРЯЧИЕ БЛЮДА = Существует, ГОРЯЧИЕ НАПИТКИ = Существует
753 2 КОКТЕЙЛИ = Существует, ПИВНЫЕ ЗАКУСКИ = Существует
734 2 ДЕСЕРТЫ = Существует, ПИВНЫЕ ЗАКУСКИ = Существует
726 1 КАЛЬЯН = Существует
712 1 СИГАРЕТНЫЙ НАБОР = Существует
640 1 РОМ = Существует
599 2 СЛУЖЕБНОЕ ПИТАНИЕ = Существует, БЕЗАЛКОГОЛЬНЫЕ НАПИТКИ ...
Копировать в Excel
Закрыть

Деревья решений **(decision trees)**



Деревья решений

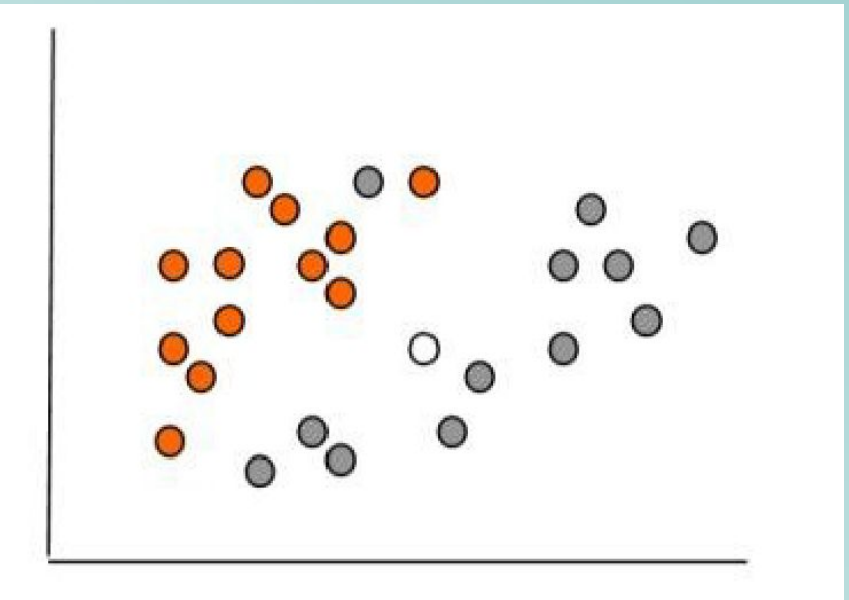
- Дерево решений (выдача кредита)



Классификация

База данных клиентов туристического агентства

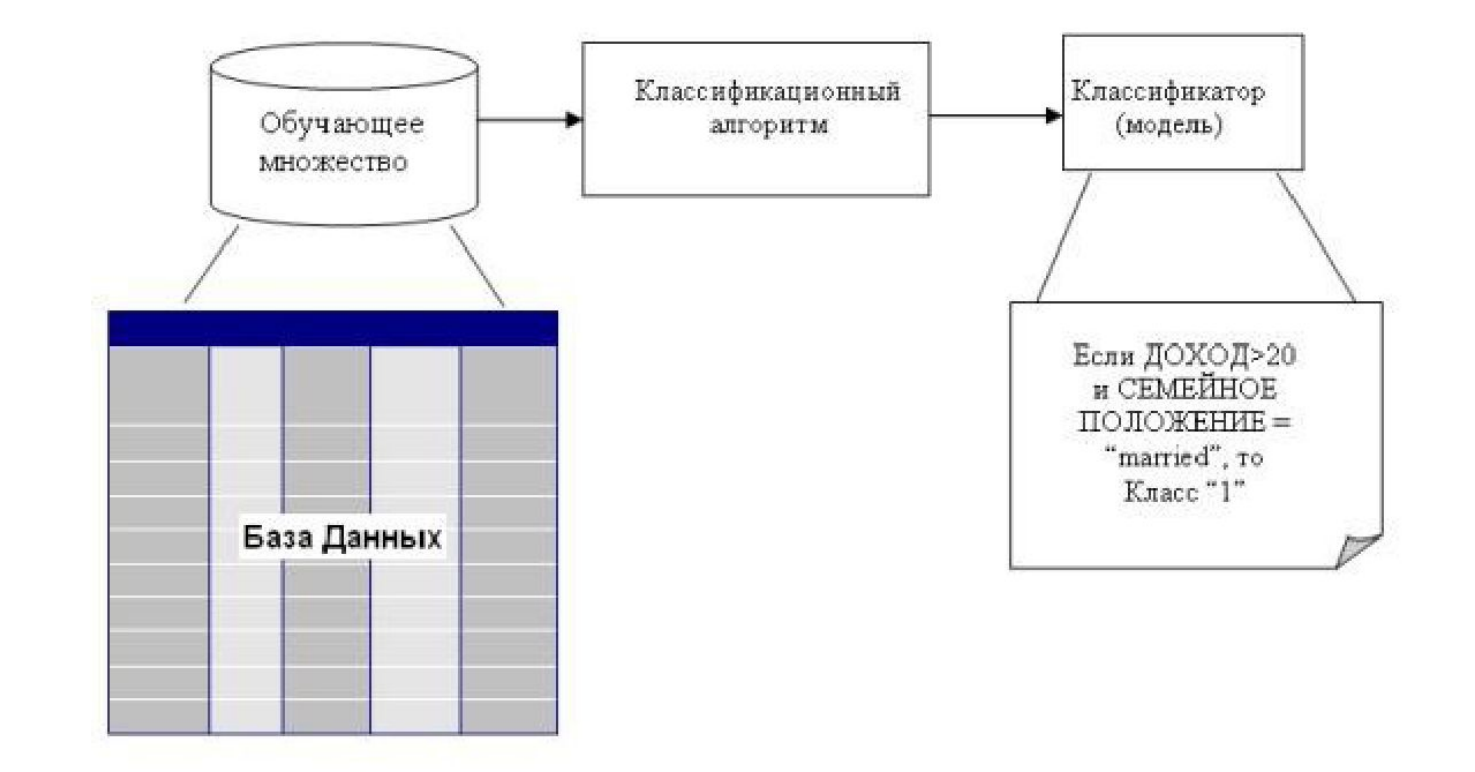
Код клиента	Возраст	Доход	Класс
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2



Множество объектов в двумерном измерении, цвет обозначает класс (оранжевый – класс1, серый – класс2, белый – неизвестный класс, новый объект)

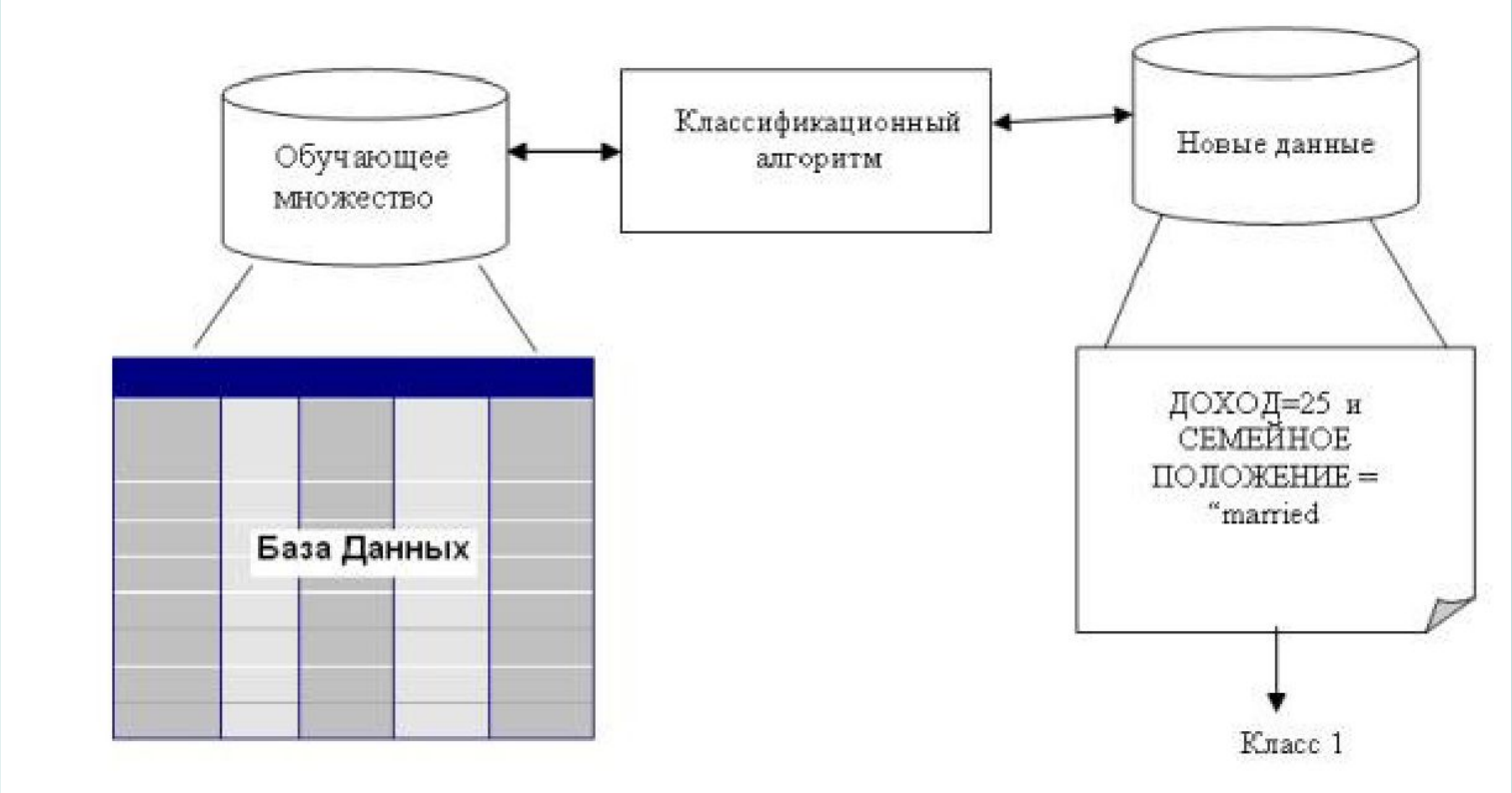
Классификация

- Конструирование модели



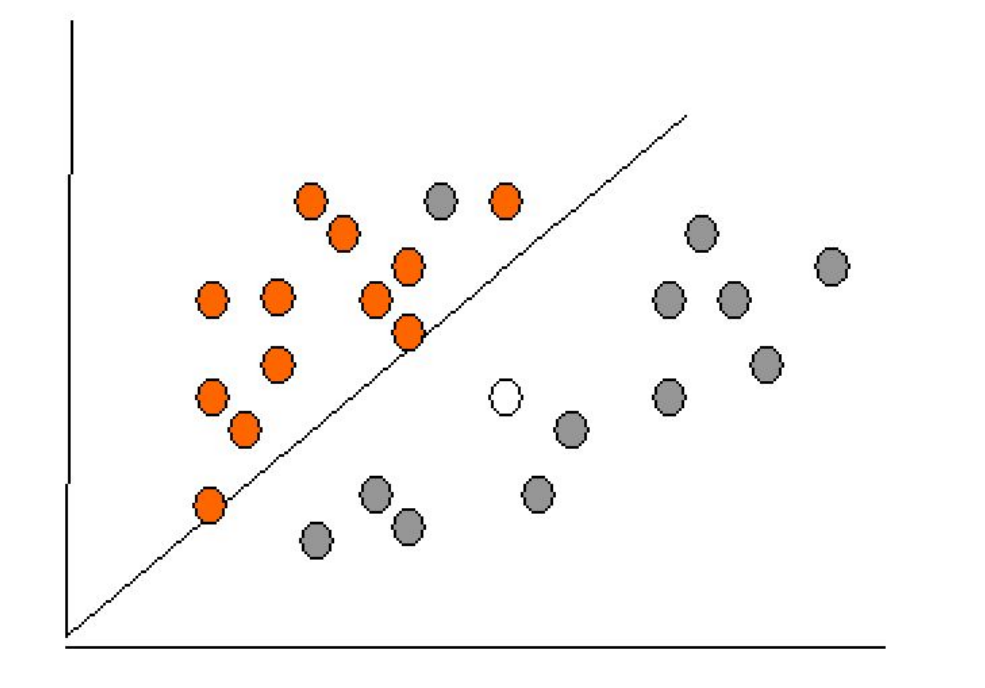
Классификация

- Использование модели



Классификация

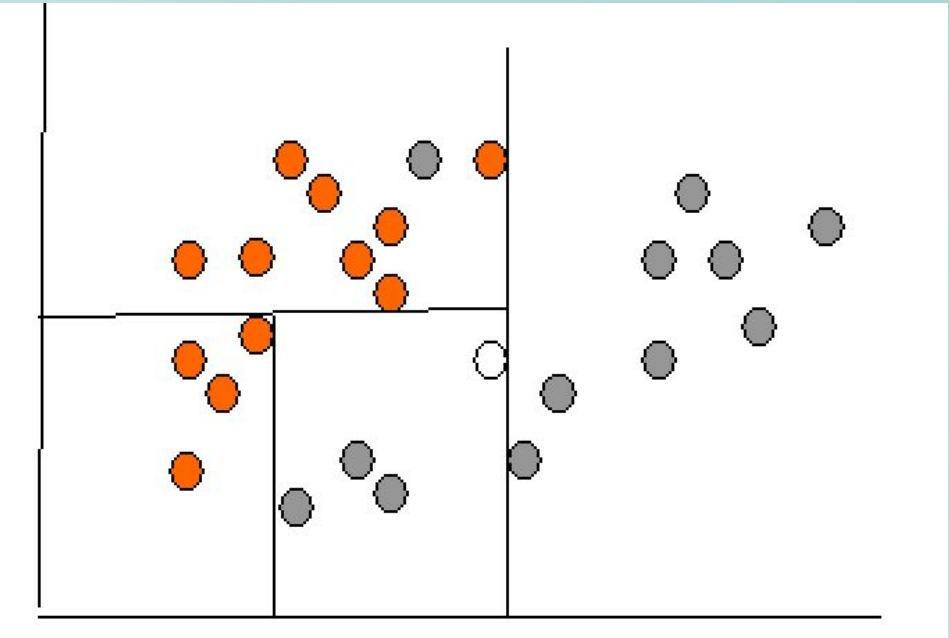
- Пример решения методом линейной регрессии (схематическое решение)



Классификация

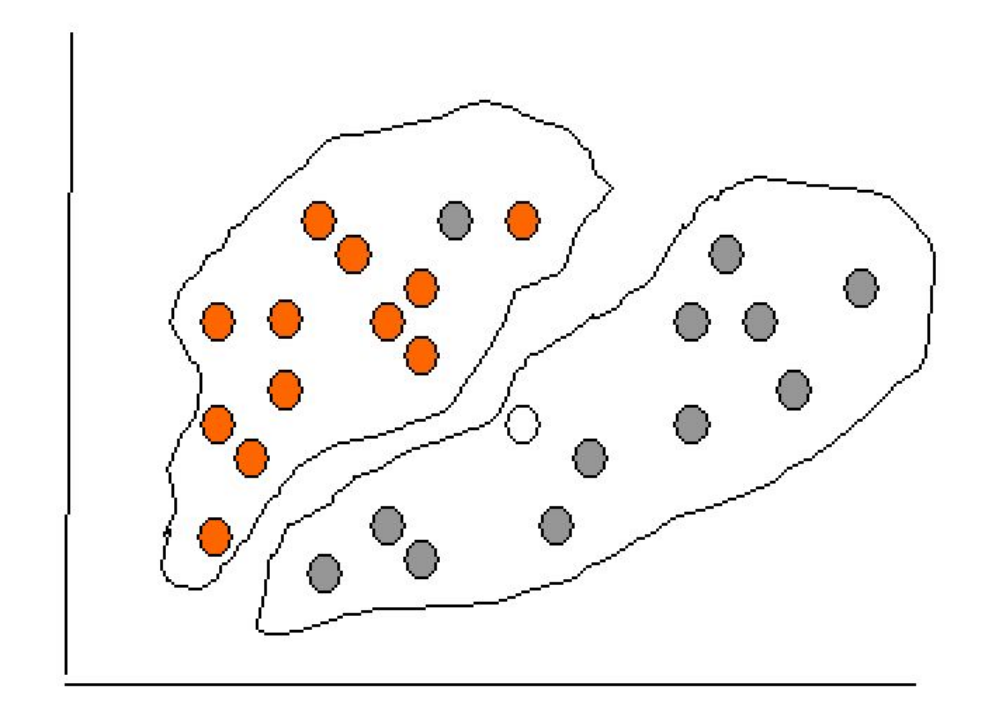
- Пример решения методом деревьев решений

```
if X > 5 then grey
   else if Y > 3 then orange
      else if X > 2 then grey
         else orange
```

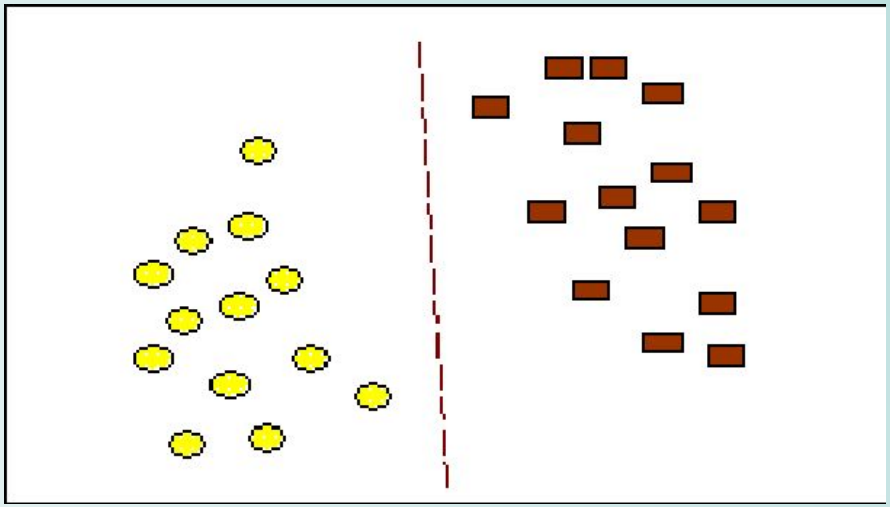


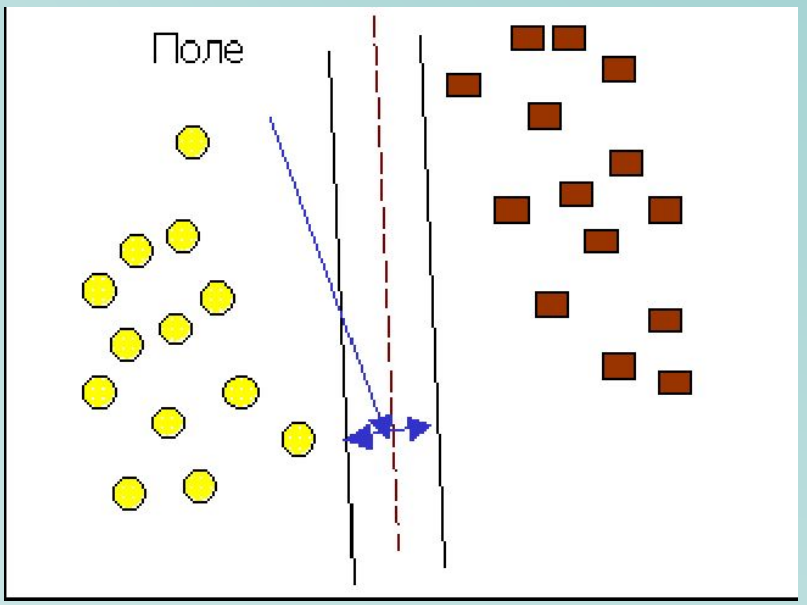
Классификация

- Пример решения методом нейронный сетей



Классификация

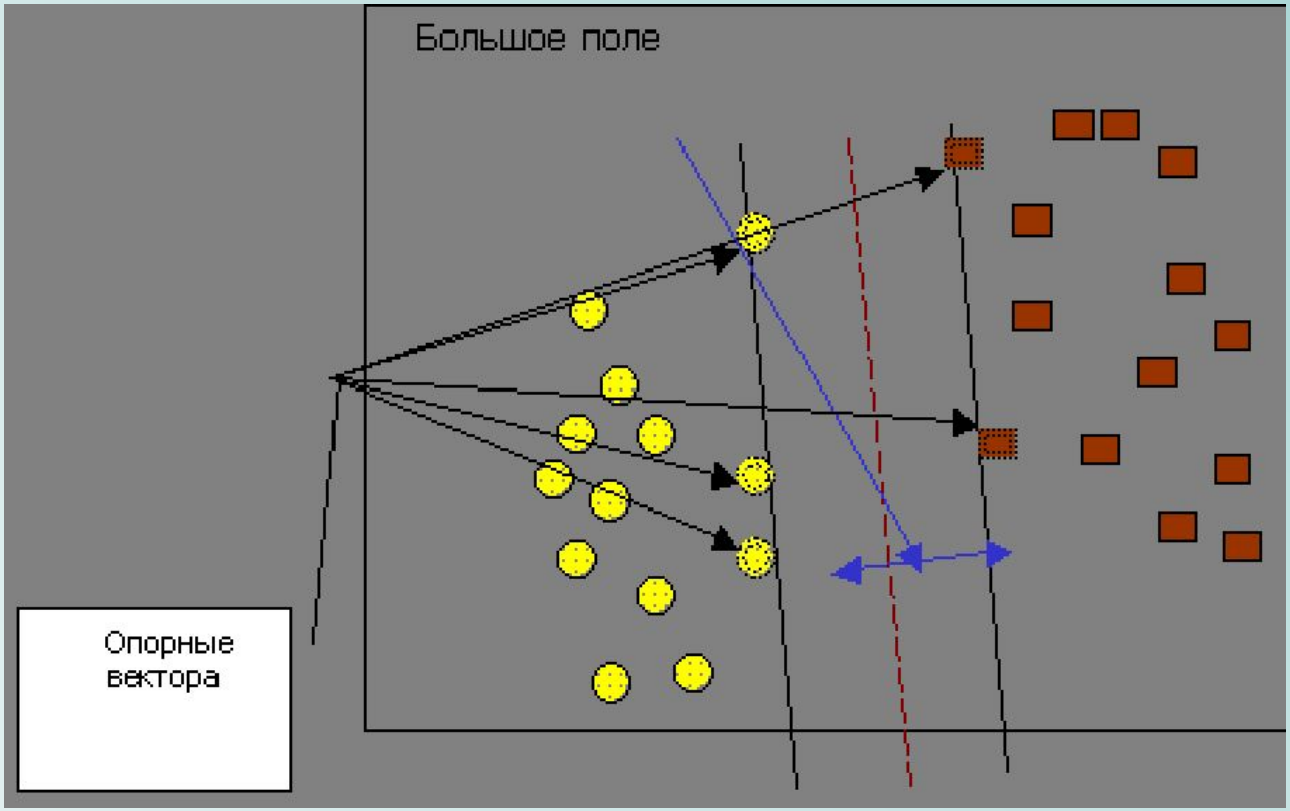




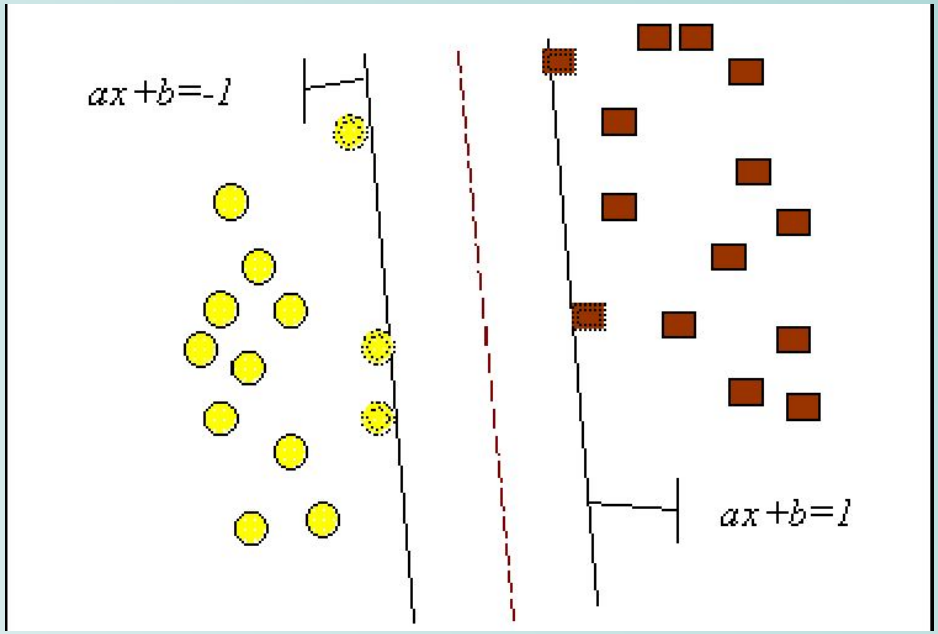
- Метод находит образцы, находящиеся на границах между двумя классами, т.е. опорные вектора.
- **Опорными векторами** называются объекты множества, лежащие на границах областей.

Классификация

- Классификация считается хорошей, если область между границами пуста.

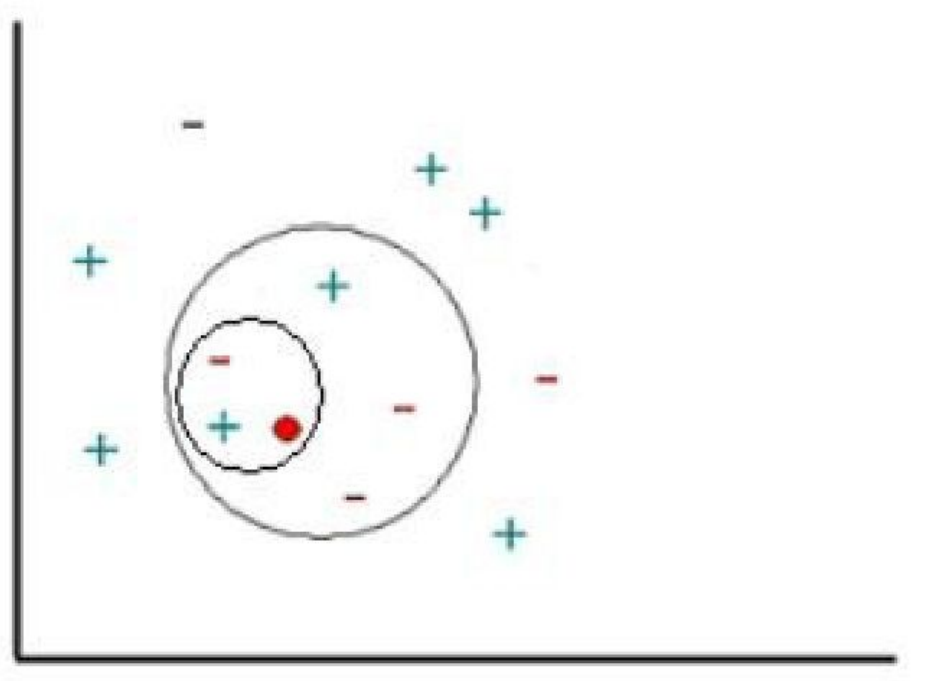


Классификация



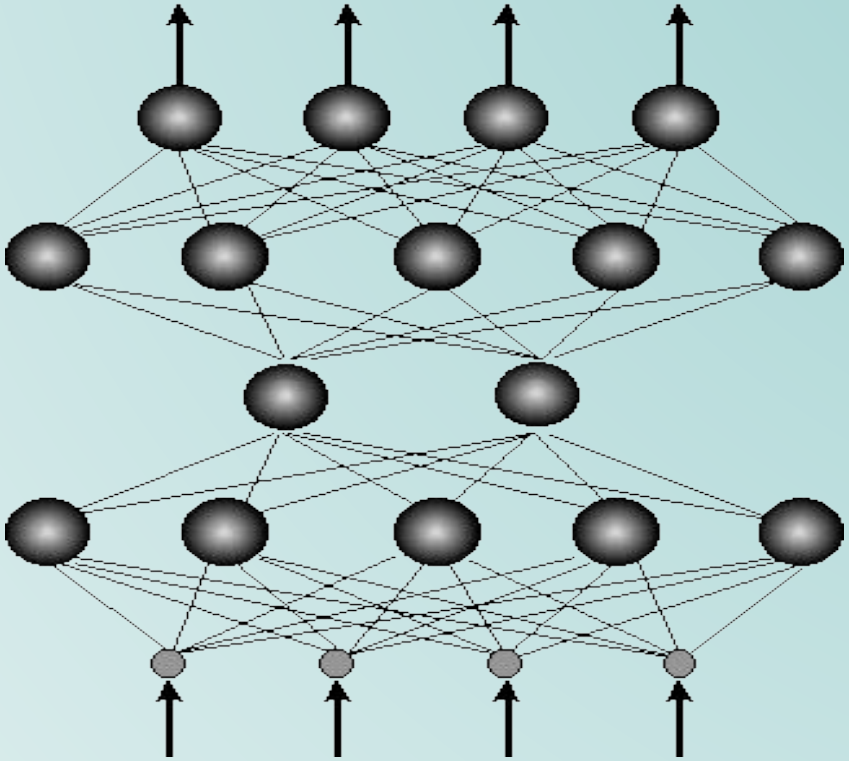
Классификация

- Метод *k*-ближайших соседей для решения задач классификации



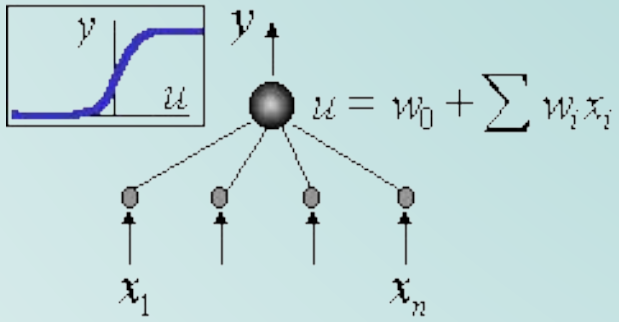
+ известный экземпляр принадлежит классу;
-известный экземпляр не принадлежит классу;
- красный круг – новый объект, для которого нужно определить принадлежность классу.

Классификация

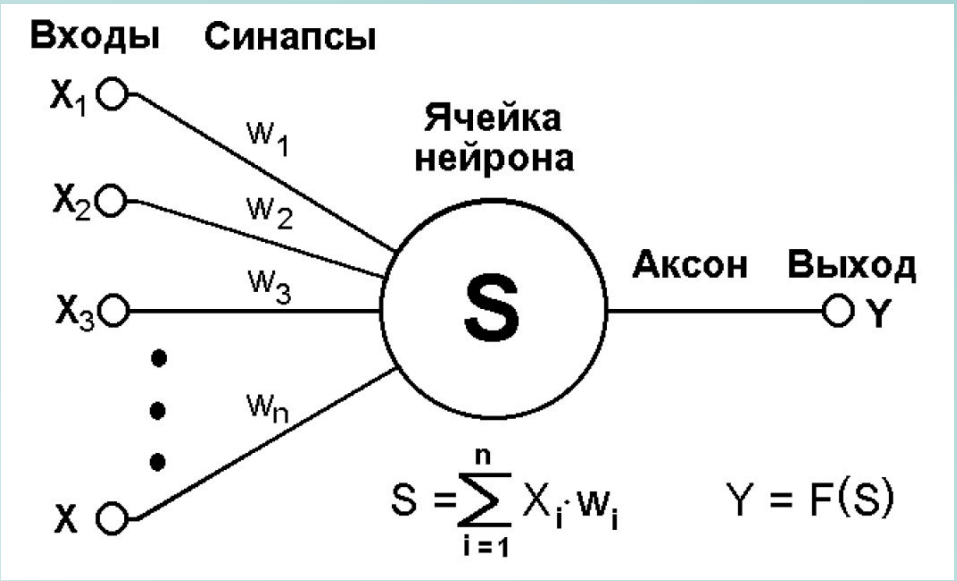


Классификация

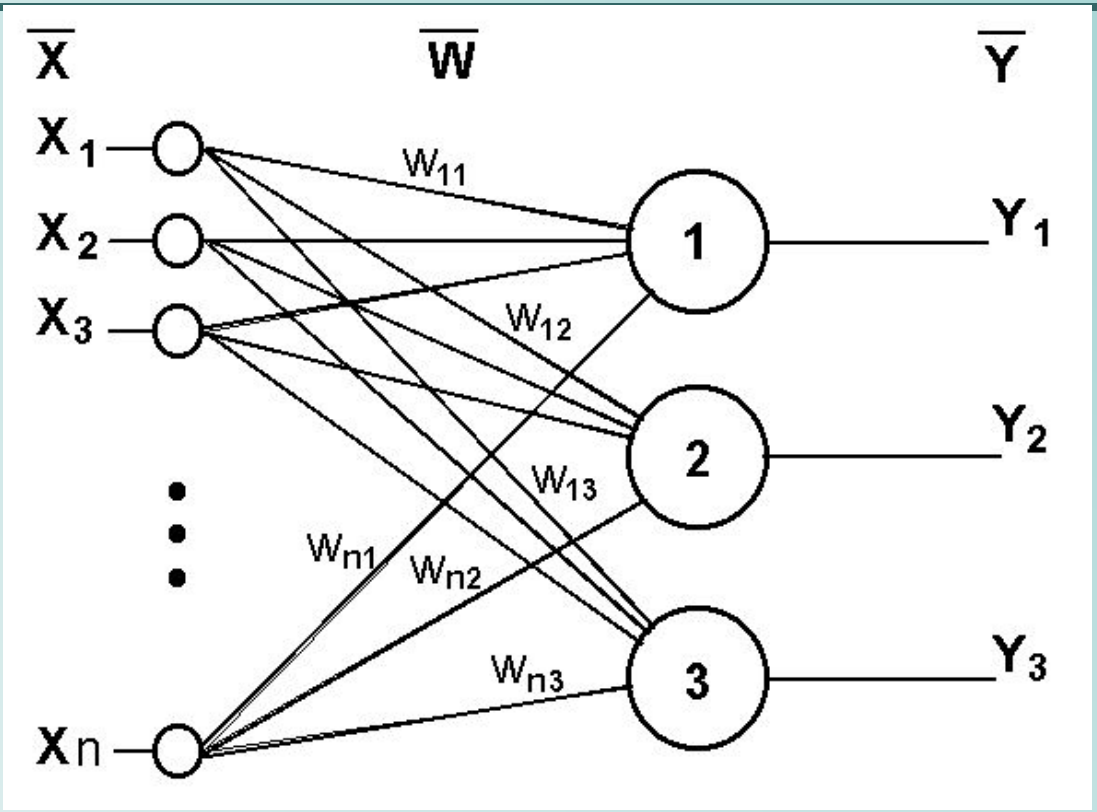
$$y = f(u), \quad u = w_0 + \sum_i w_i x_i$$



Классификация

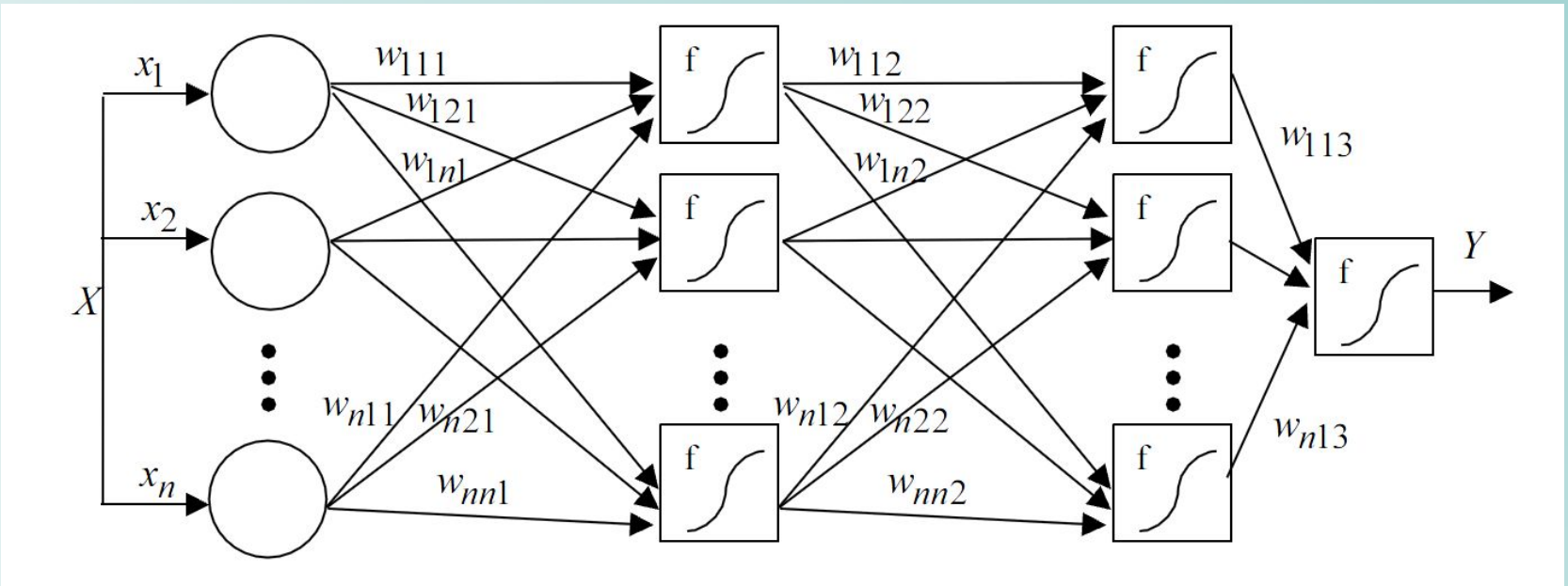


Классификация



n входов, на которые поступают сигналы, идущие по синапсам на 3 нейрона. Эти три нейрона образуют единственный слой данной сети и выдают три выходных сигнала.

Классификация



- Y – вектор выходных сигналов, X – вектор входных сигналов, в выходном слое N_0 нейронов, в каждом скрытом слое – N_H нейронов, входной слой – N_I нейронов.

Классификация

- Результат работы *i*-го слоя (Y_i – вектор выхода *i*-го слоя многослойного перцептрона):

$$Y_i = \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{ni} \end{pmatrix} = F\left\{ \begin{pmatrix} w_{11i} & w_{21i} & \cdots & w_{n1i} \\ w_{12i} & w_{22i} & \cdots & w_{n2i} \\ \cdots & \cdots & \ddots & \cdots \\ w_{n1i} & w_{n2i} & \cdots & w_{nni} \end{pmatrix} \begin{pmatrix} y_{1(i-1)} \\ y_{2(i-1)} \\ \vdots \\ y_{n(i-1)} \end{pmatrix} \right\}.$$

Классификация

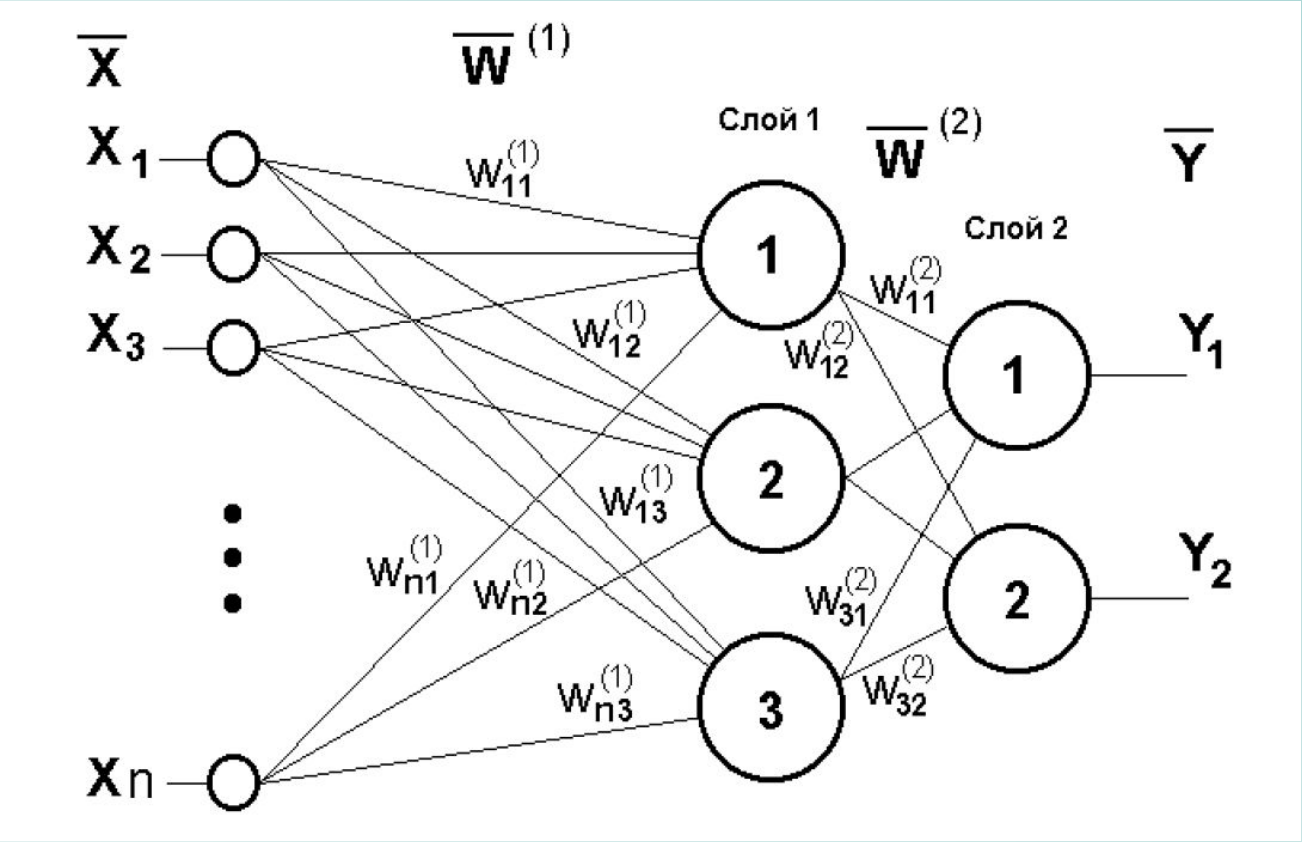
- Если заданы начальные значения Y: $y_{j,0}=x_j$, то результат работы перцептрона

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = F\left\{ (w_{113}\ w_{213} \ldots w_{n13}) \cdot F\left[\begin{pmatrix} w_{112} & w_{212} & \cdots & w_{n12} \\ w_{122} & w_{222} & \cdots & w_{n22} \\ \cdots & \cdots & \ddots & \cdots \\ w_{n12} & w_{n22} & \cdots & w_{nn2} \end{pmatrix} \times \right.\right.$$

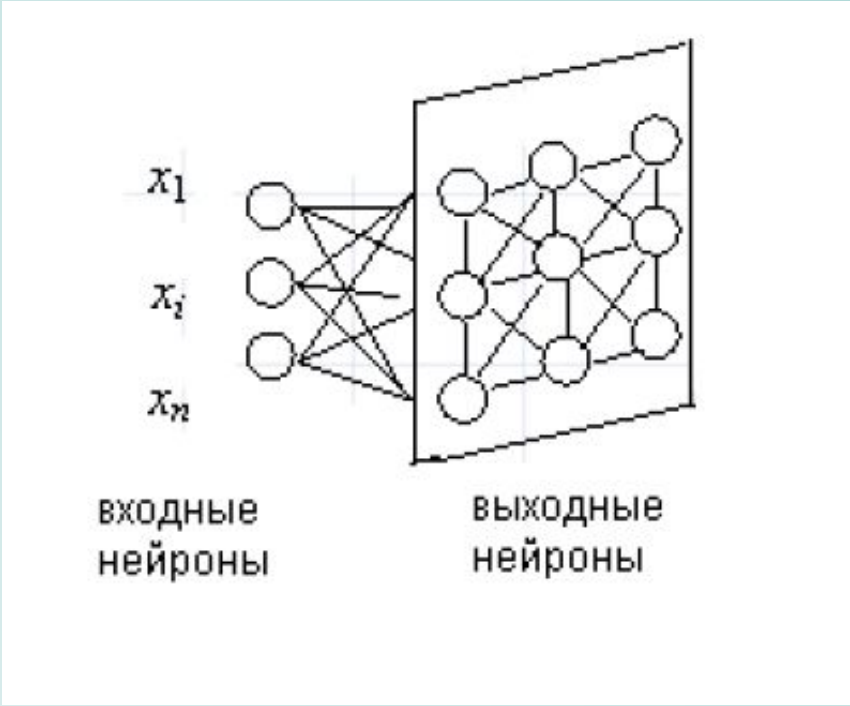
$$\left.\left. \times F\left\langle \begin{pmatrix} w_{111} & w_{211} & \cdots & w_{n11} \\ w_{121} & w_{221} & \cdots & w_{n21} \\ \cdots & \cdots & \ddots & \cdots \\ w_{n11} & w_{n21} & \cdots & w_{nn1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right\rangle \right] \right\}.$$

Классификация

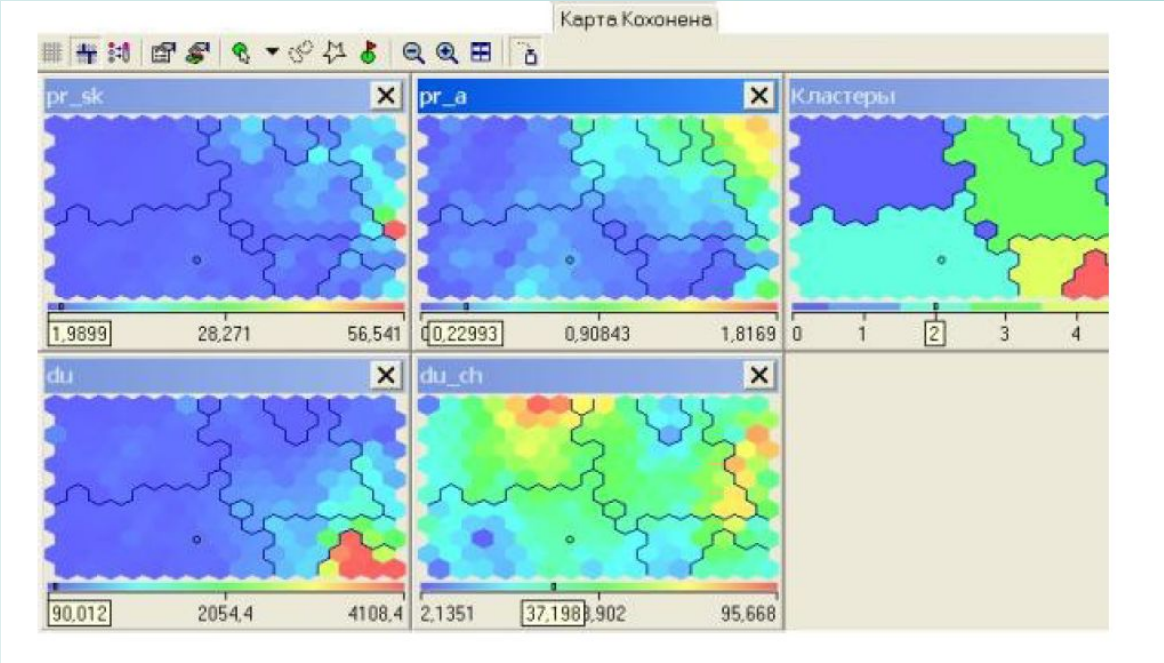
- Двухслойный перцептрон



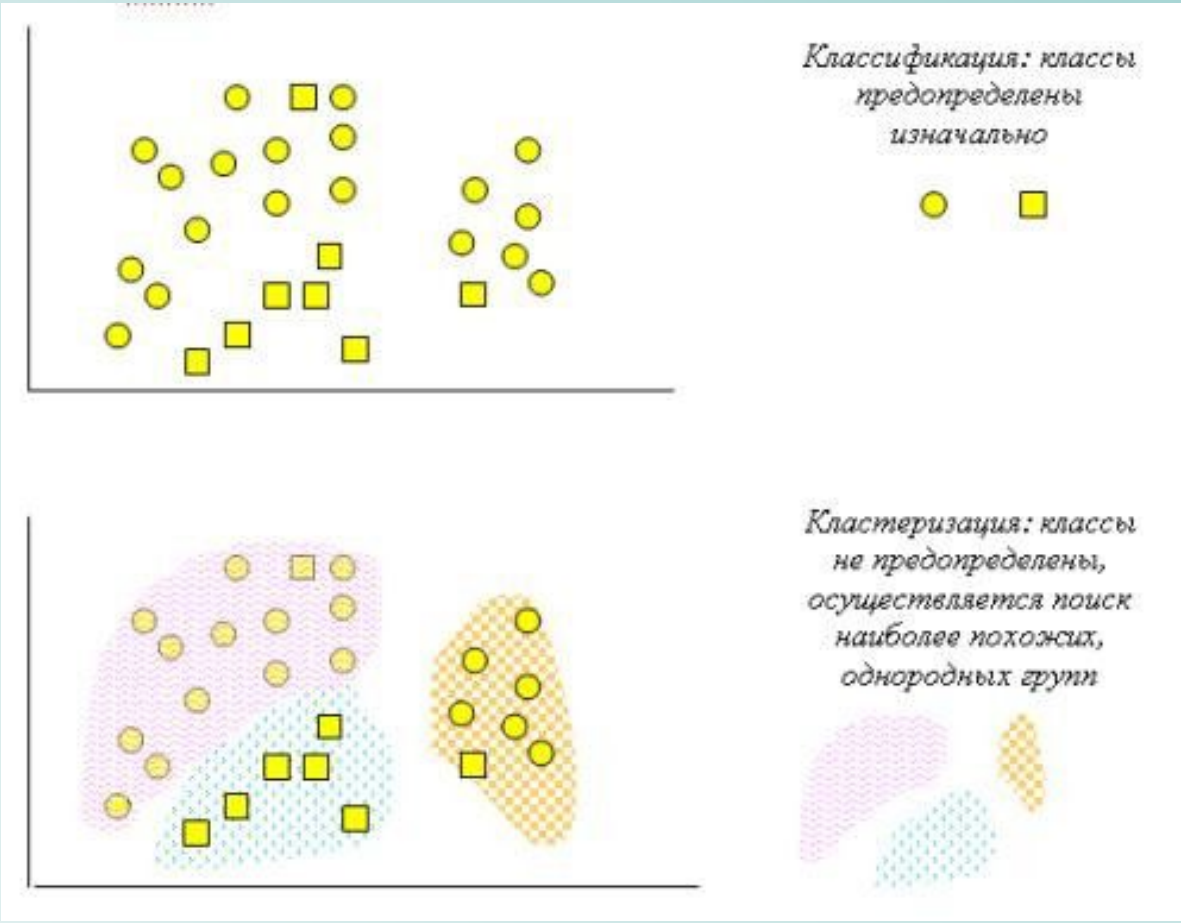
Классификация



Классификация

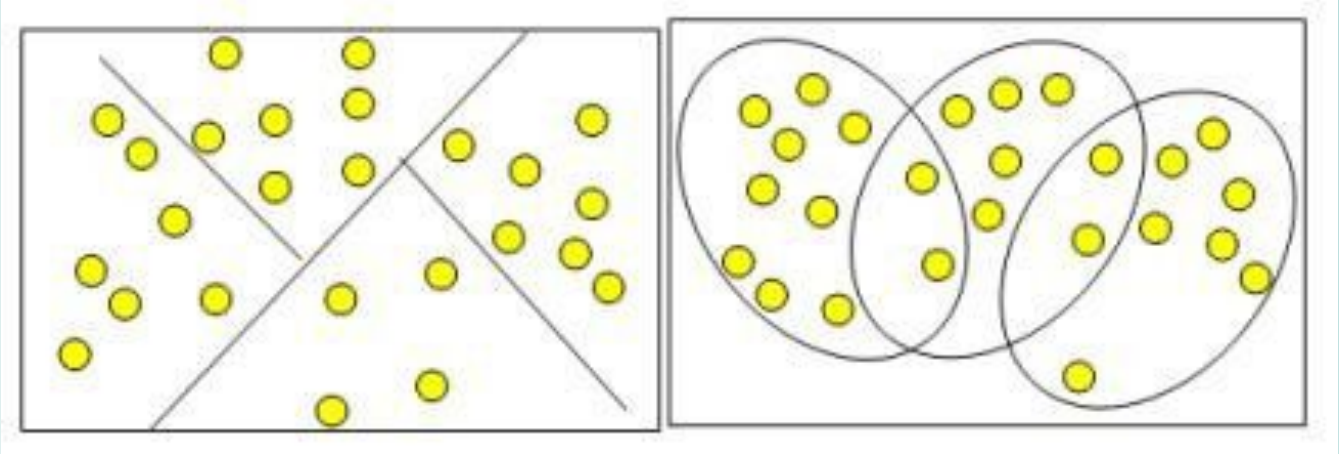


Кластеризация

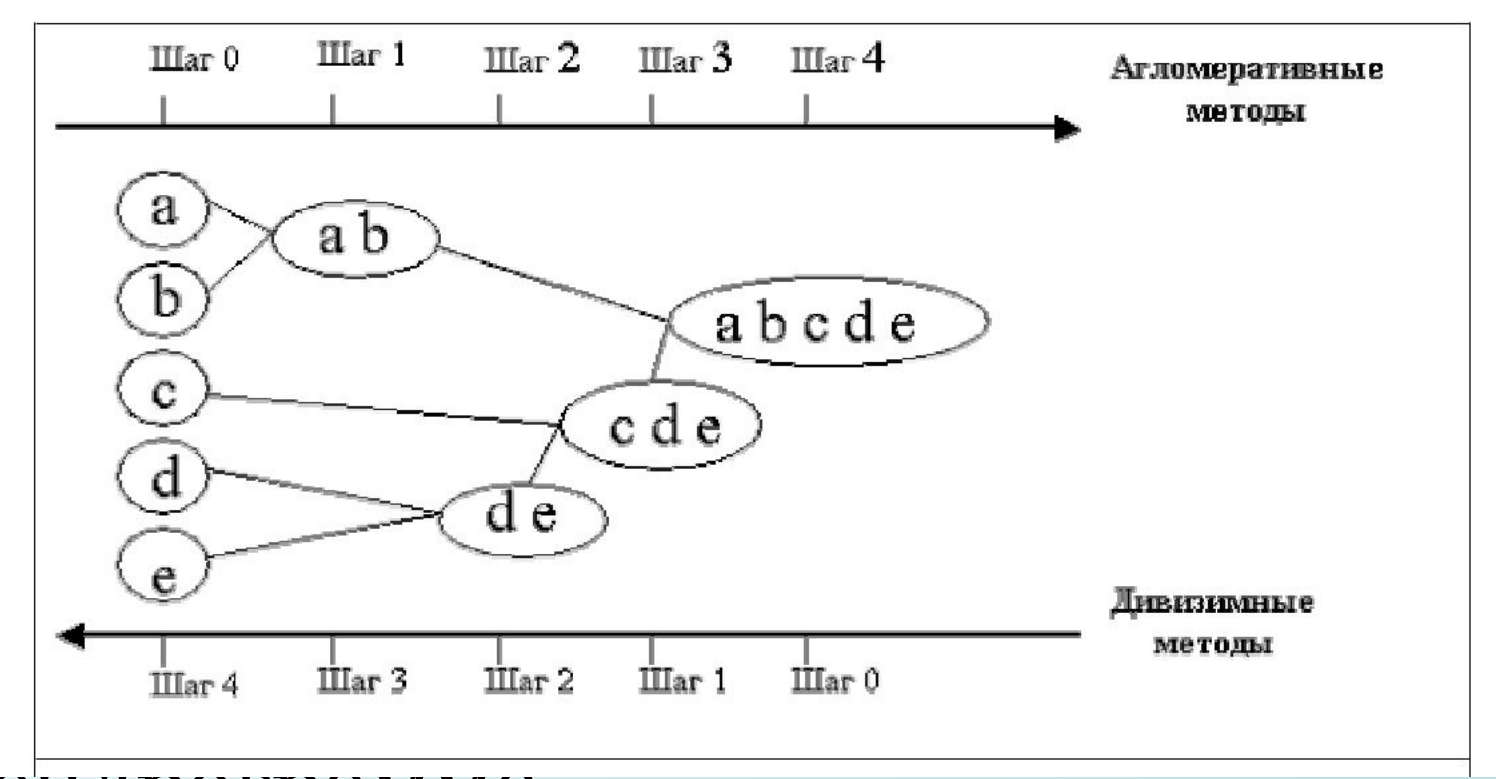


Кластеризация

- Кластеры: пересекающиеся и непересекающиеся



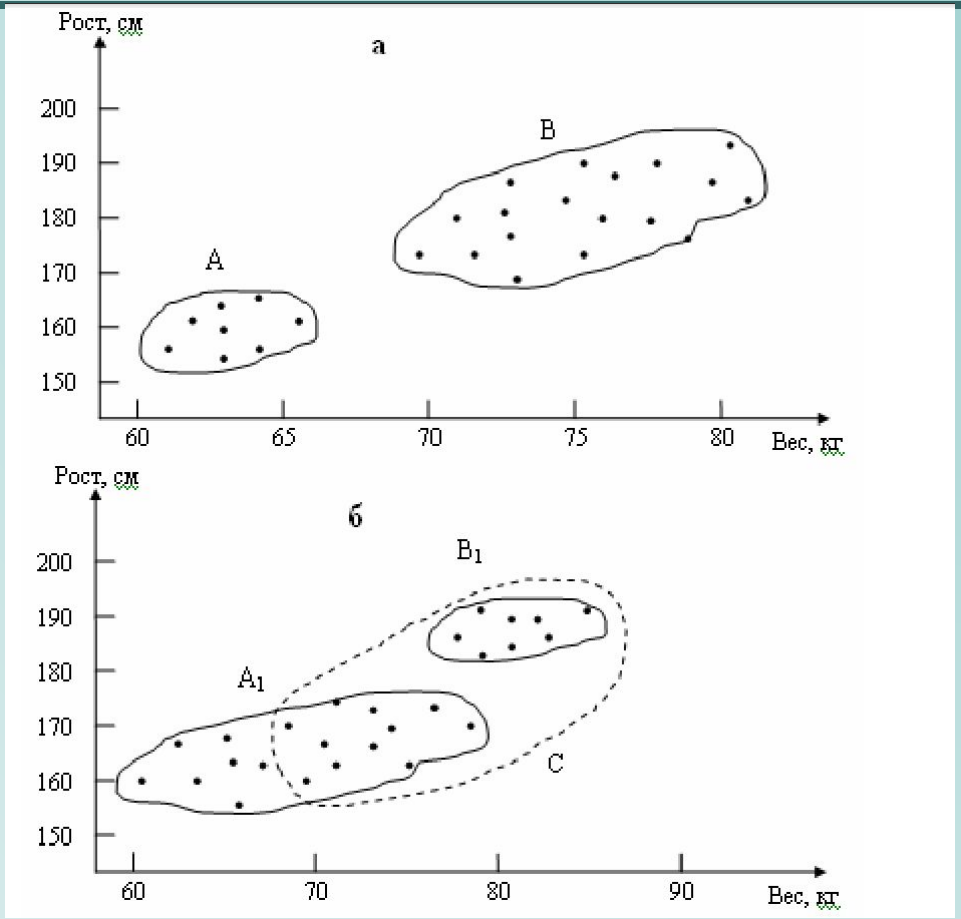
Кластеризация



Дендрограмма

Кластеризация

Необходимость нормировки (разные масштабы → разные классы)



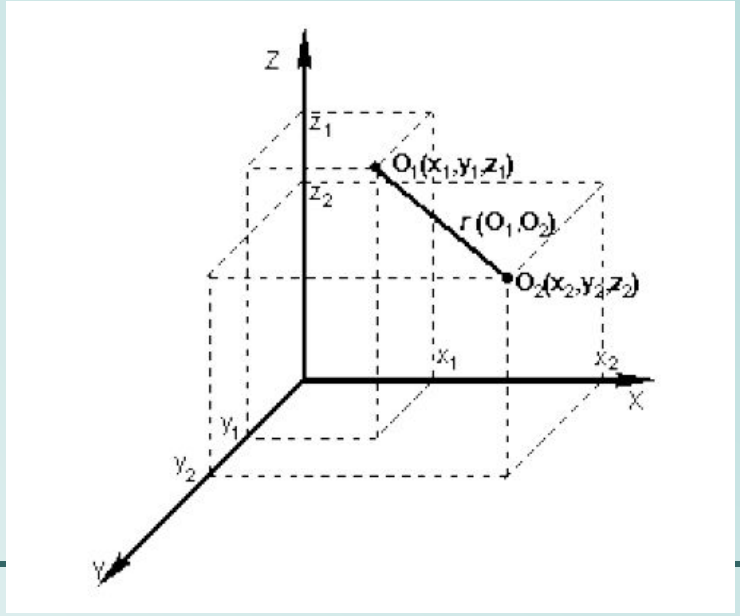
Кластеризация

Показатели	Формулы
Для количественных шкал	
Линейное расстояние	$d_{lij} = \sum_{l=1}^{m} \left| x_i^l - x_j^l \right|$
Евклидово расстояние	$d_{Eij} = \left(\sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^2 \right)^{\frac{1}{2}}$
Квадрат евклидово расстояния	$d^2{}_{Eij} = \sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^2$
Обобщенное степенное расстояние Минковского	$d_{Pij} = \left(\sum_{l=1}^{m} \left(x_i^l - x_j^l \right)^P \right)^{\frac{1}{P}}$
	$d_{ij} = \max_{1 \le i,j \le l} \left| x_i - x_j \right|$
Расстояние Чебышева	$d_H \left(x_i, x_j \right) = \sum_{l=1}^{k} \left| x_i^l - x_j^l \right|$
Расстояние городских кварталов (Манхэттенское расстояние)	

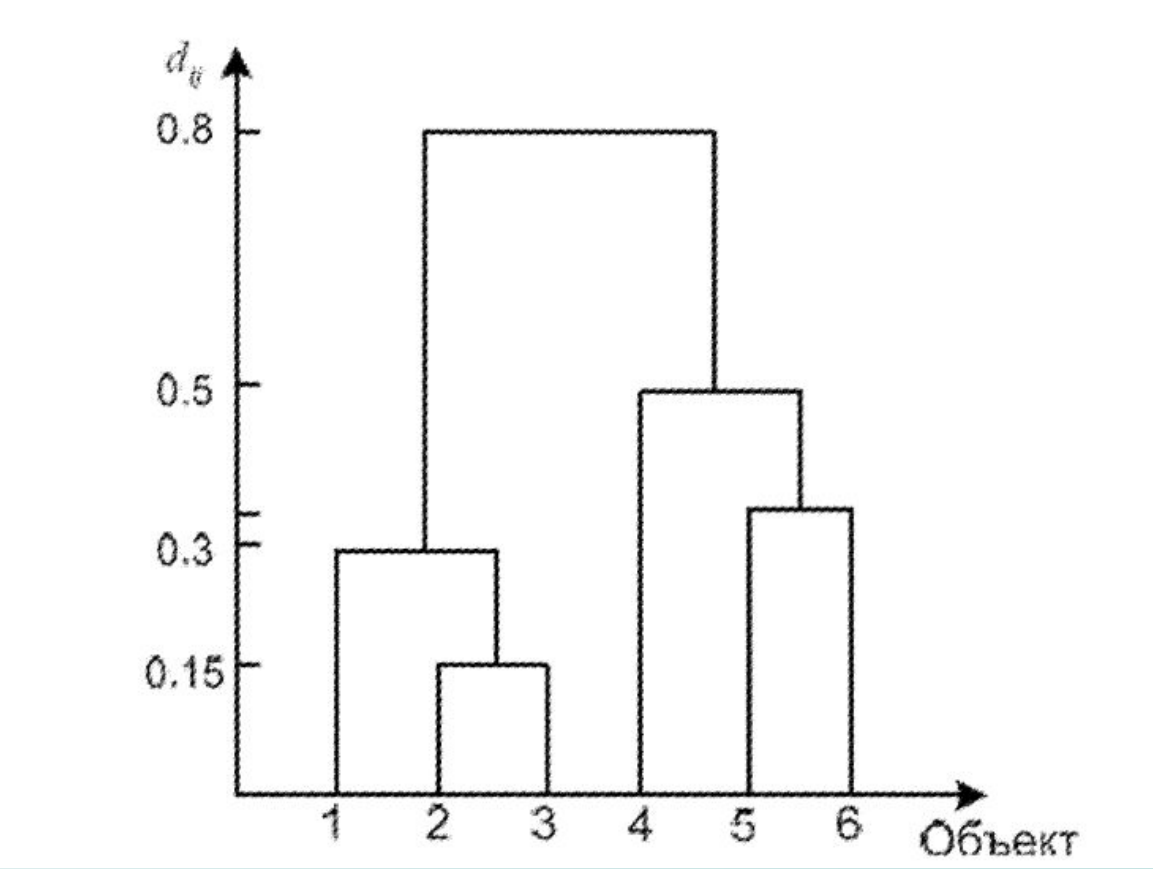
Кластеризация

- Расстояние в пространстве трех измерений

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

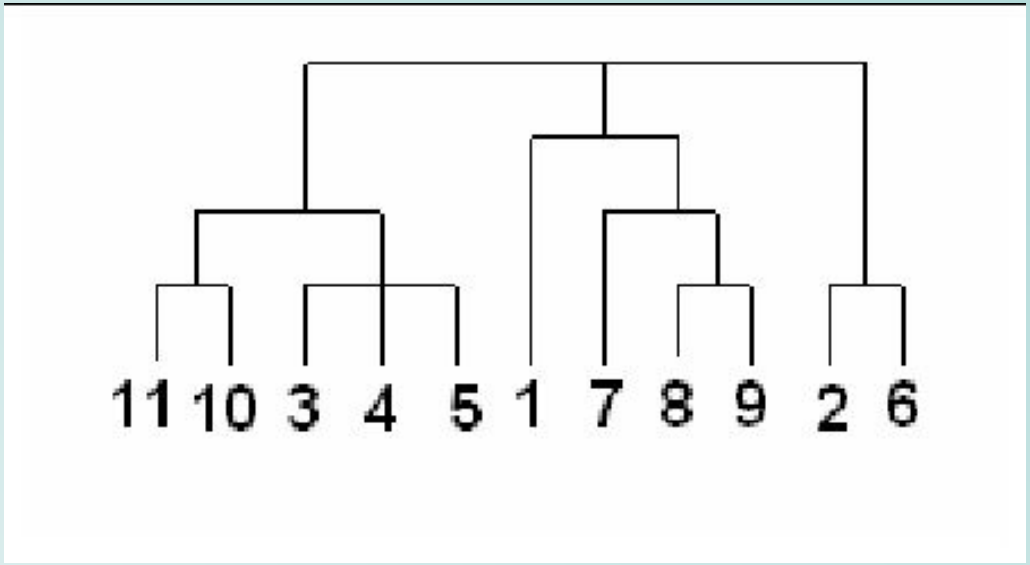


Кластеризация



Кластеризация

- Задание: описать последовательность объединения в классы



Кластеризация

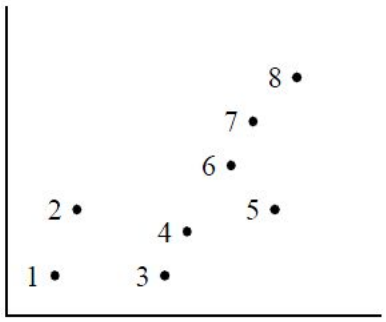


Рис. 1.4.1.

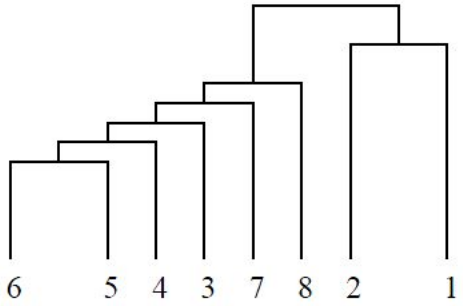


Рис. 1.4.2.

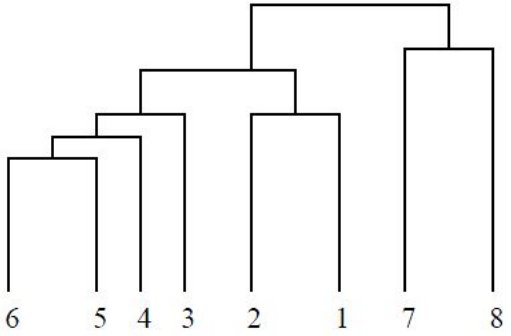
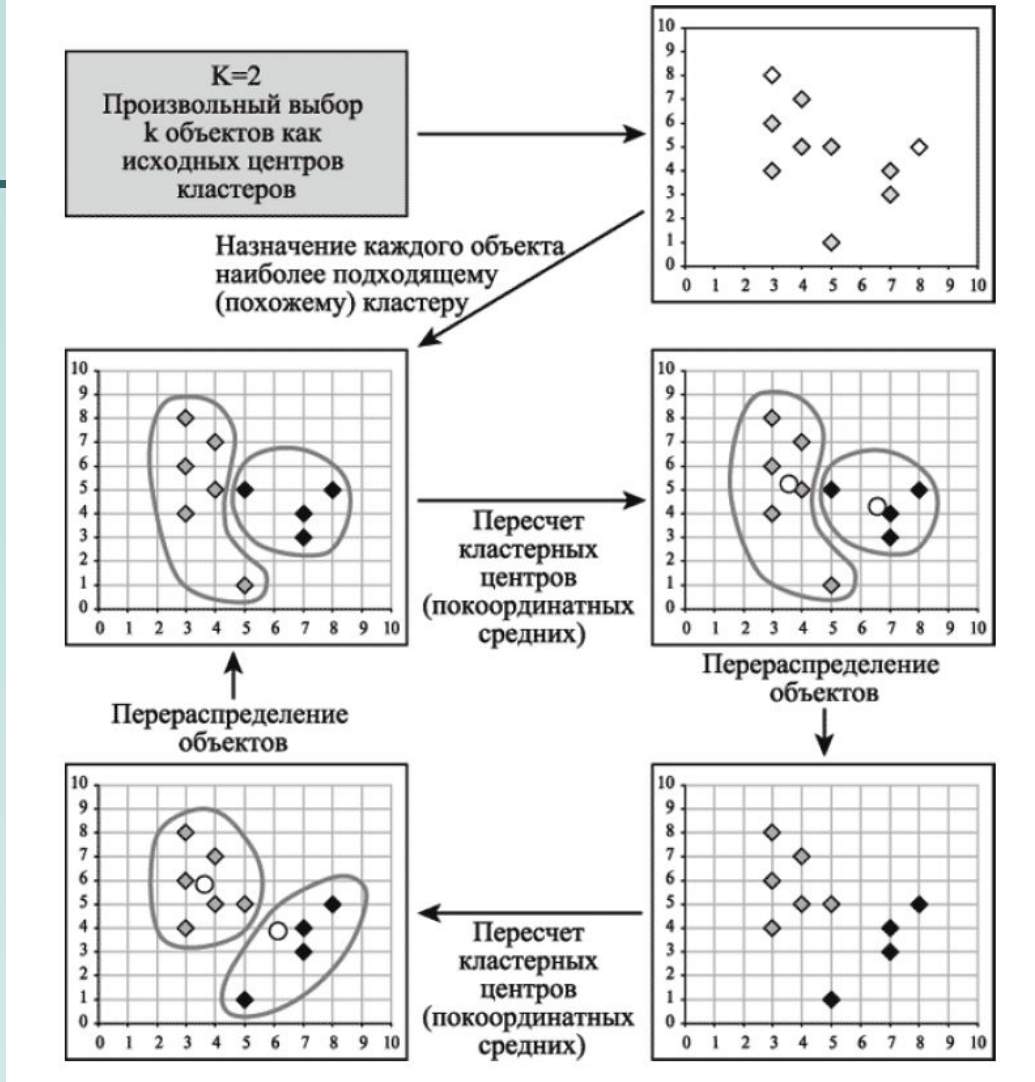


Рис. 1.4.3.

Кластеризация

Метод *k*-средних, *k*=2

Выбор *k*:
Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.



Факторный анализ

Матрица факторных нагрузок

Исходные переменные	Выделенные факторы		
	1	2	3
Счет в уме	**0.804**	–0.056	0.195
Аналогии	**0.769**	0.183	0.074
Числовые ряды	**0.762**	0.217	–0.005
Умозаключения	**0.696**	0.007	0.284
Заучивание слов	–0.057	**0.829**	0.048
Осведомленность	0.094	**0.746**	0.077
Пропущенные слова	0.352	**0.630**	0.202
Понятливость	0.042	0.155	**0.739**
Скрытые фигуры	0.153	0.225	**0.737**
Геометрическое сложение	0.421	–0.224	**0.692**
Исключение изображений	0.087	0.487	0.506

- Жирным выделены значимые нагрузки

Факторный анализ

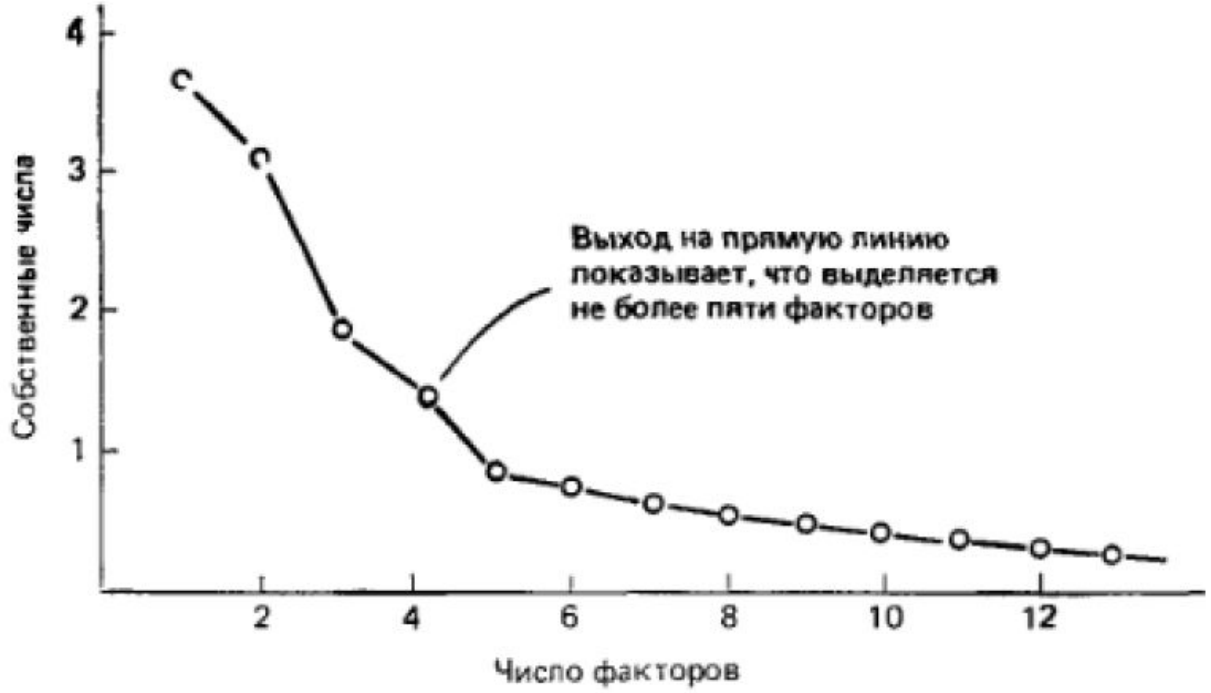


График собственных значений
(диаграмма каменистой осыпи)

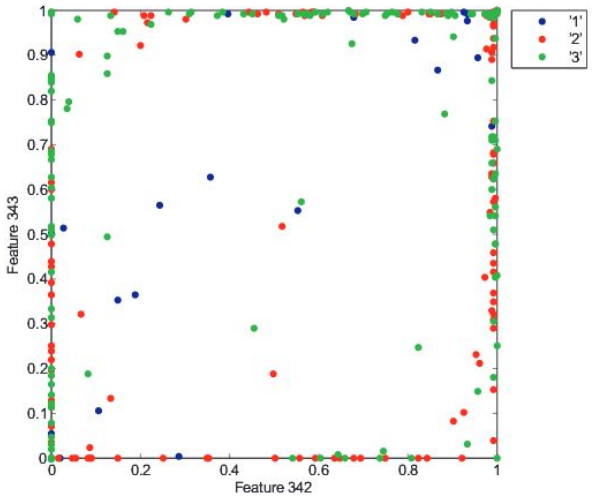
Факторный анализ

Алгоритм 1 Схема метода главных компонент

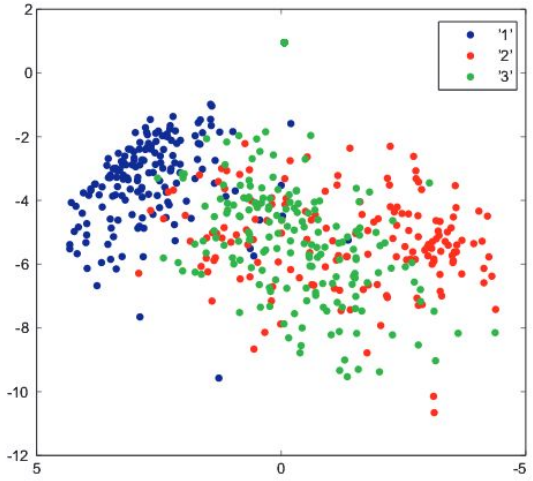
Вход: $X \in \mathbb{R}^{N \times D}$ – исходная выборка данных, d – размерность редуцированного пространства
Выход: $T \in \mathbb{R}^{N \times d}$ – представление выборки в редуцированном пространстве

$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$; *// Вычисляем выборочное среднее*
$\boldsymbol{x}_n \leftarrow \boldsymbol{x}_n - \bar{\boldsymbol{x}}$; *// Переносим начало координат в центр выборки*
если $N > D$ **то**
  $S = \frac{1}{N} X^T X$; *// Вычисляем выборочную матрицу ковариации*
  $S = Q \Lambda Q^T, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы ковариации*
  Выбираем d наибольших собственных значений $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$ и соответствующие им собственные вектора $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;
иначе
  $S = \frac{1}{N} X X^T$;
  $S = Q \Lambda Q^T, \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D), Q^T Q = I, Q = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_D)$; *// Находим собственные вектора и собственные значения матрицы S*
  $Q \leftarrow \frac{1}{\sqrt{N}} X^T Q \mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_D}}\right)$; *// Переходим к нормированным собственным векторам выборочной матрицы ковариации*
  Выбираем собственные вектора, соответствующие d наибольшим собственным значениям $W = (\boldsymbol{q}_1 | \ldots | \boldsymbol{q}_d)$;
$T = XW$; *// Проектируем выборку на выбранные направления*

Факторный анализ



(a)



(b)

. Проекция выборки изображений цифр ’1’, ’2’ и ’3’ на два признака, соответствующих интенсивностям пикселов (a) и на два признака, полученных с помощью метода главных компонент (b).

Анализ временных рядов

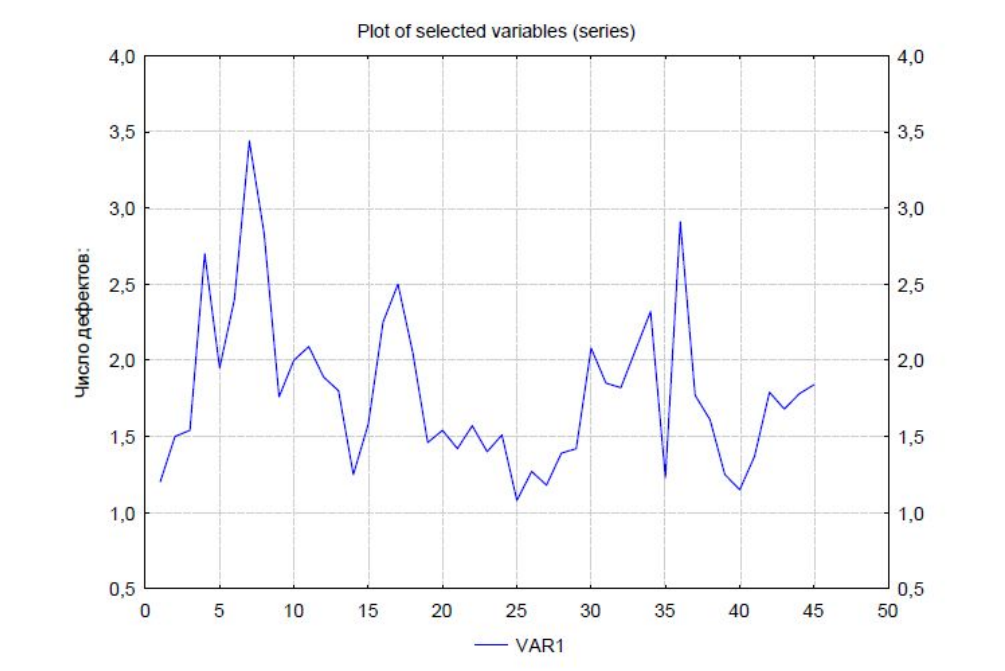
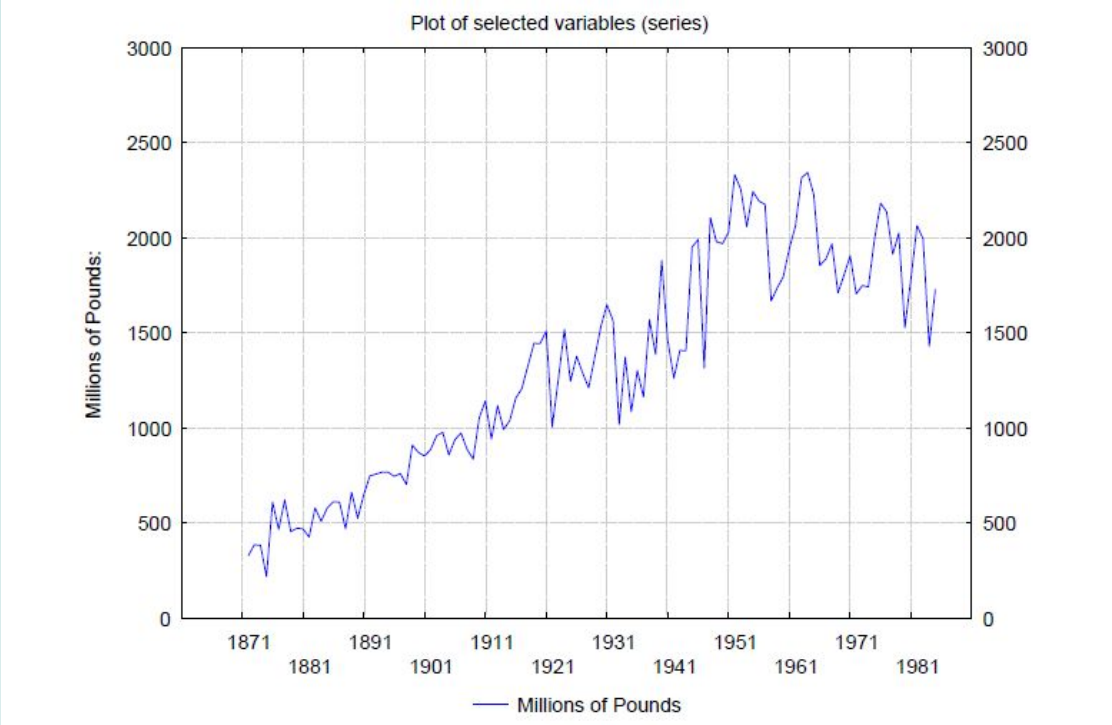


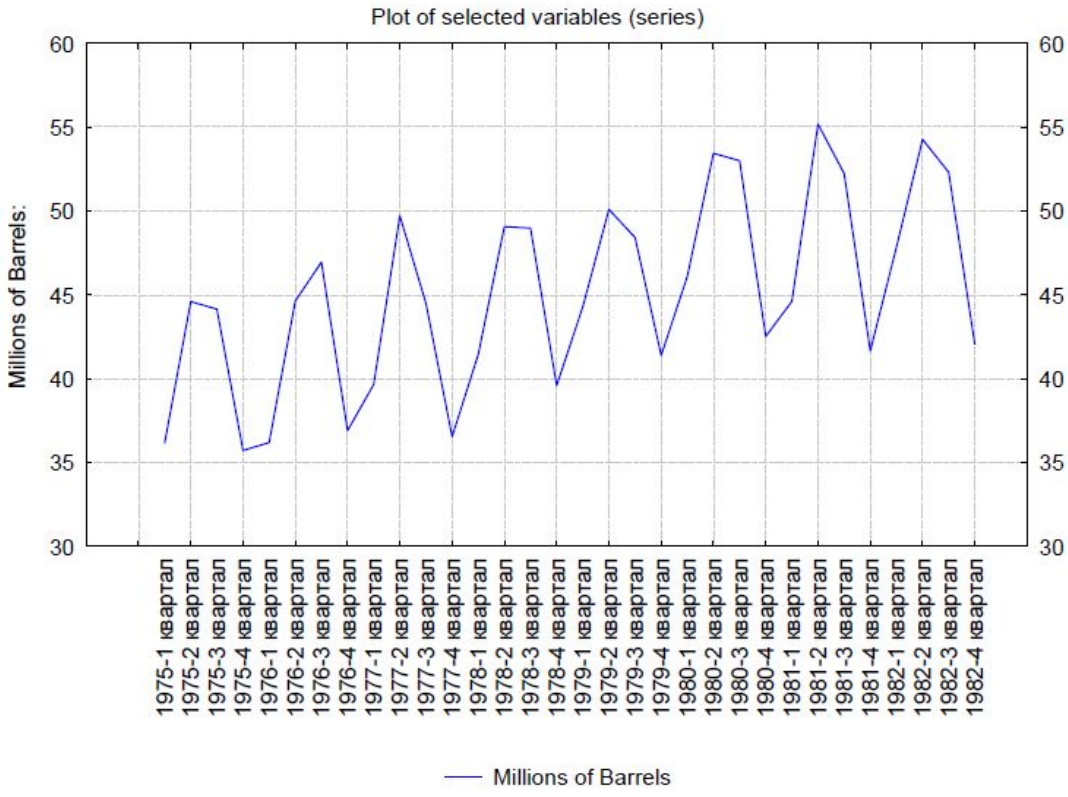
График ежедневных данных о среднем числе дефектов на грузовик в конце сборочного конвейера на предприятии по производству грузовиков. Наблюдения осциллируют на некотором постоянном уровне. Стационарный временной ряд (стационарный в среднем, специальный случай стационарных временных рядов). Ряд может быть описан авторегрессионной моделью скользящего среднего (ARMA), предложенной в методологии Бокса–Дженкинса.

Анализ временных рядов



Данные о производстве (ежегодном) табака в США. Не варьируются около постоянного значения, выявляют предельный, вверх направленный тренд. Дисперсия увеличивается с увеличением времени. Нестационарный по среднему и по дисперсии временной ряд.

Анализ временных рядов



Ежеквартальные данные о производстве пива в США в течение нескольких лет. Сезонный временной ряд, проявляющий ежегодную тенденцию к повторению. Период сезонности, т.е. интервал, через который тенденция повторяется, равен 4.

Для анализа данного ряда может быть предложена модификация модели Бокса–Дженкинса.

Альтернативным способом моделирования является сезонная декомпозиция.

Анализ временных рядов

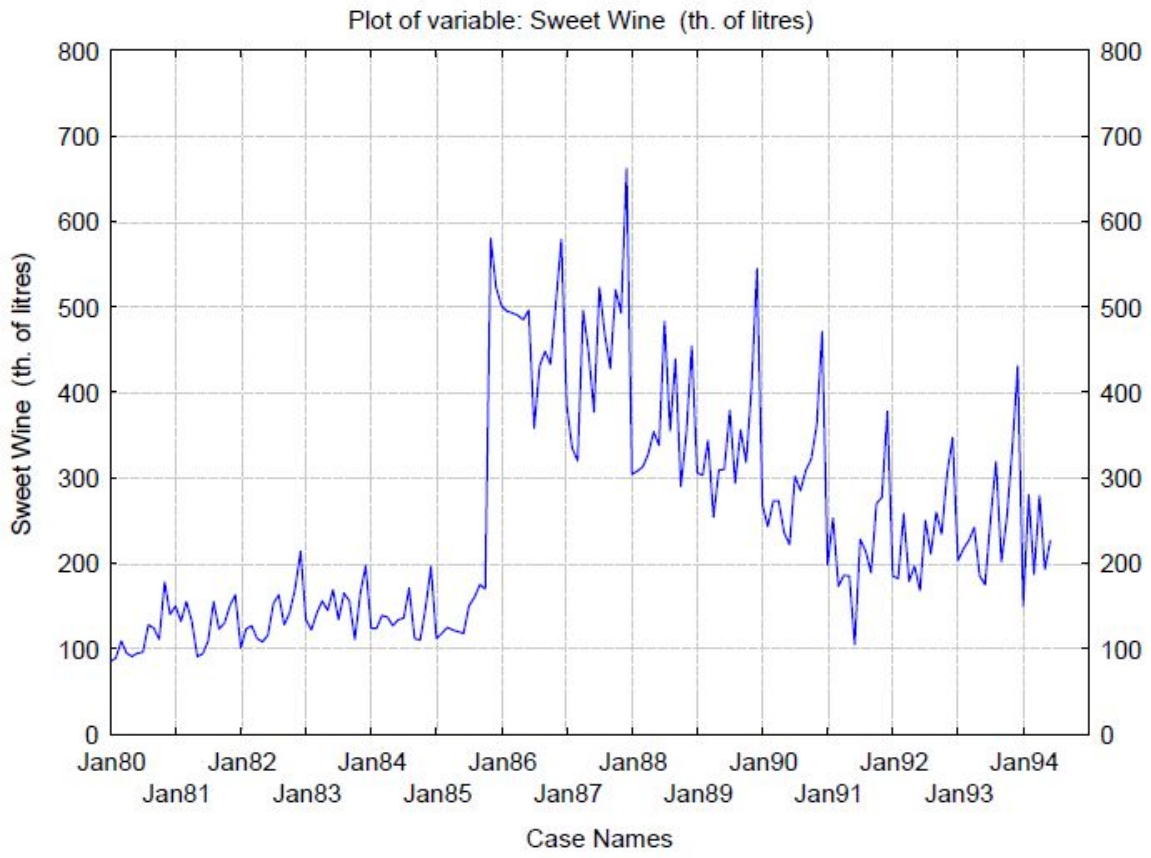
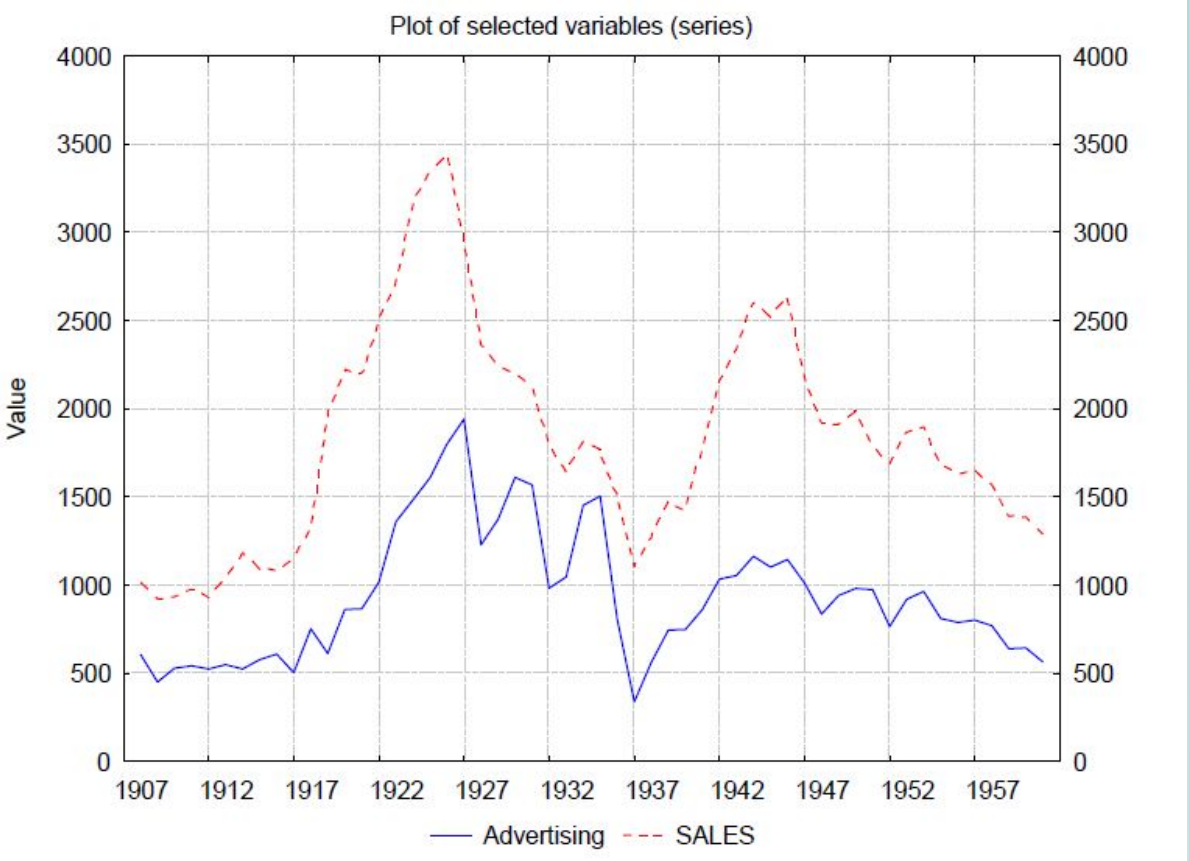


График реализации вина сладкого сорта на территории Австралии с января 1980 по июнь 1994 года. Нестационарный ряд– изменение в структуре ряда, возникшее из-за некоторого внешнего события. Такой тип нестационарности нельзя учесть, применяя то или иное стандартное преобразование.

Анализ временных рядов



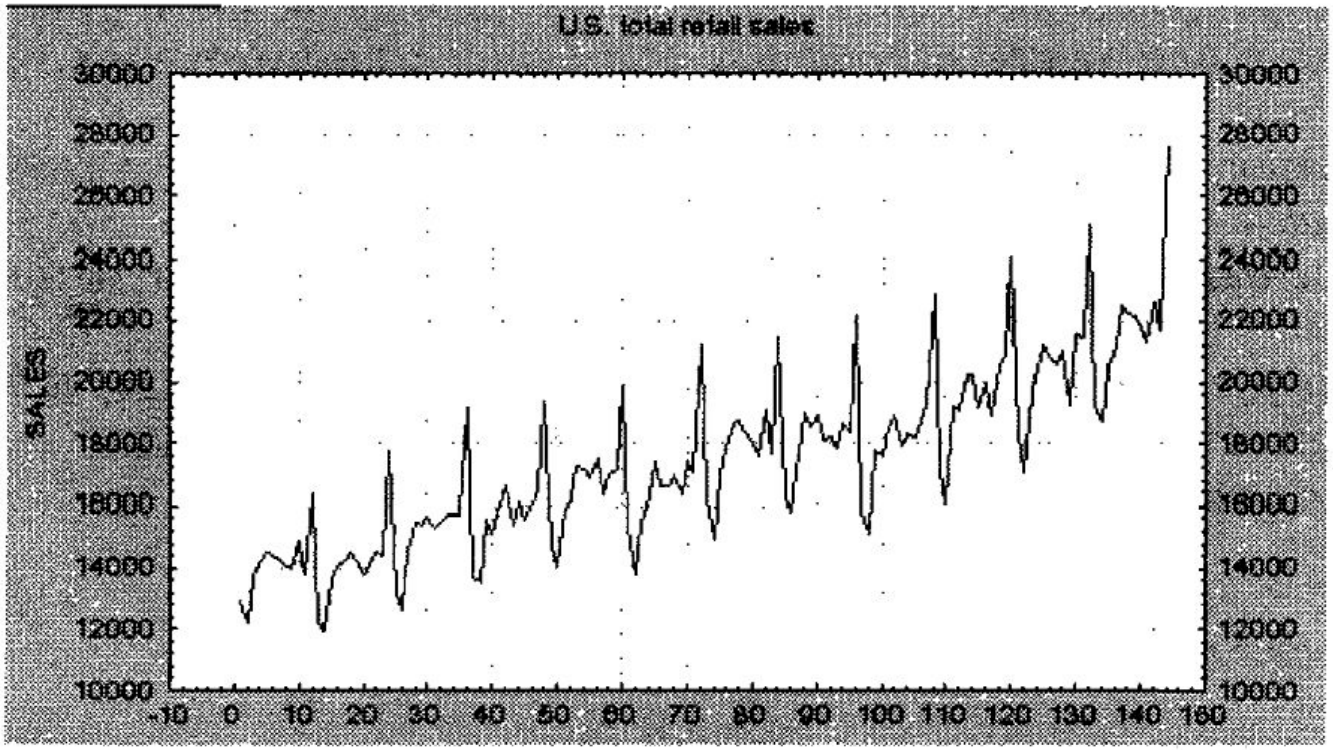
Двумерный временной ряд.

Ряды коррелированны.

Переменные взаимно влияют друг на друга.

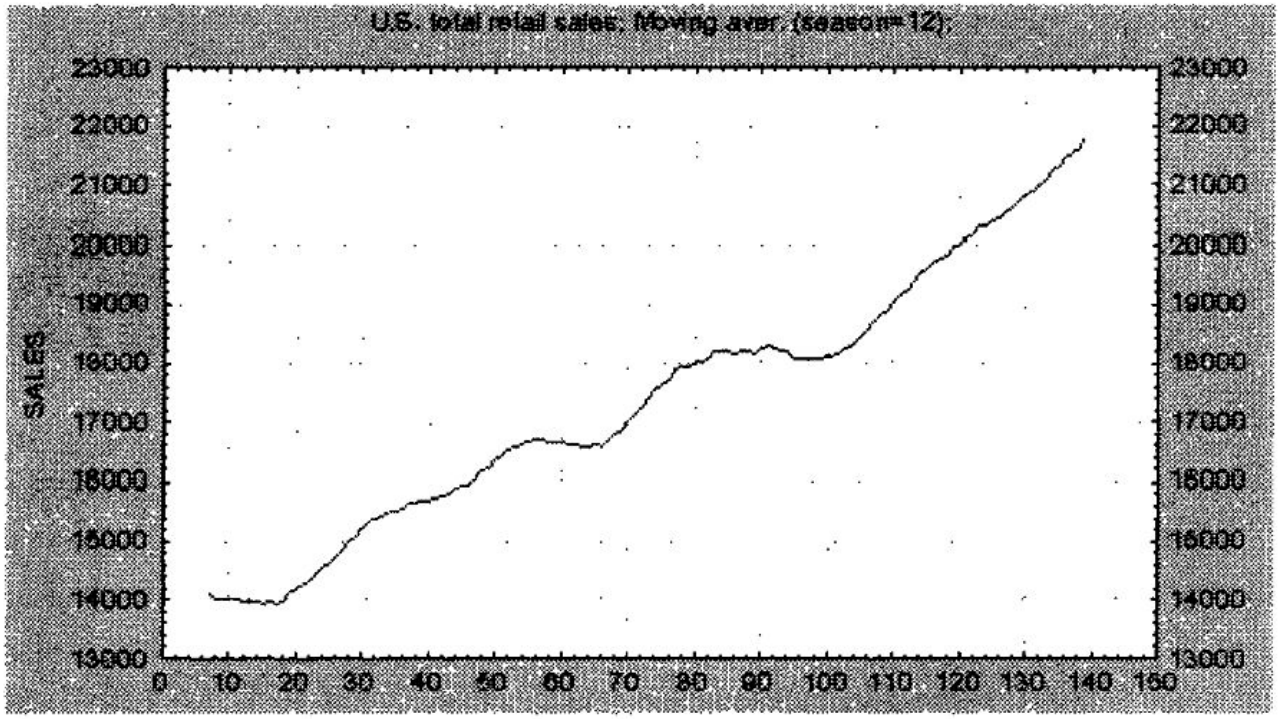
Необходимо использовать сложные методы анализа, например, векторные авторегрессионные модели скользящего среднего.

Анализ временных рядов



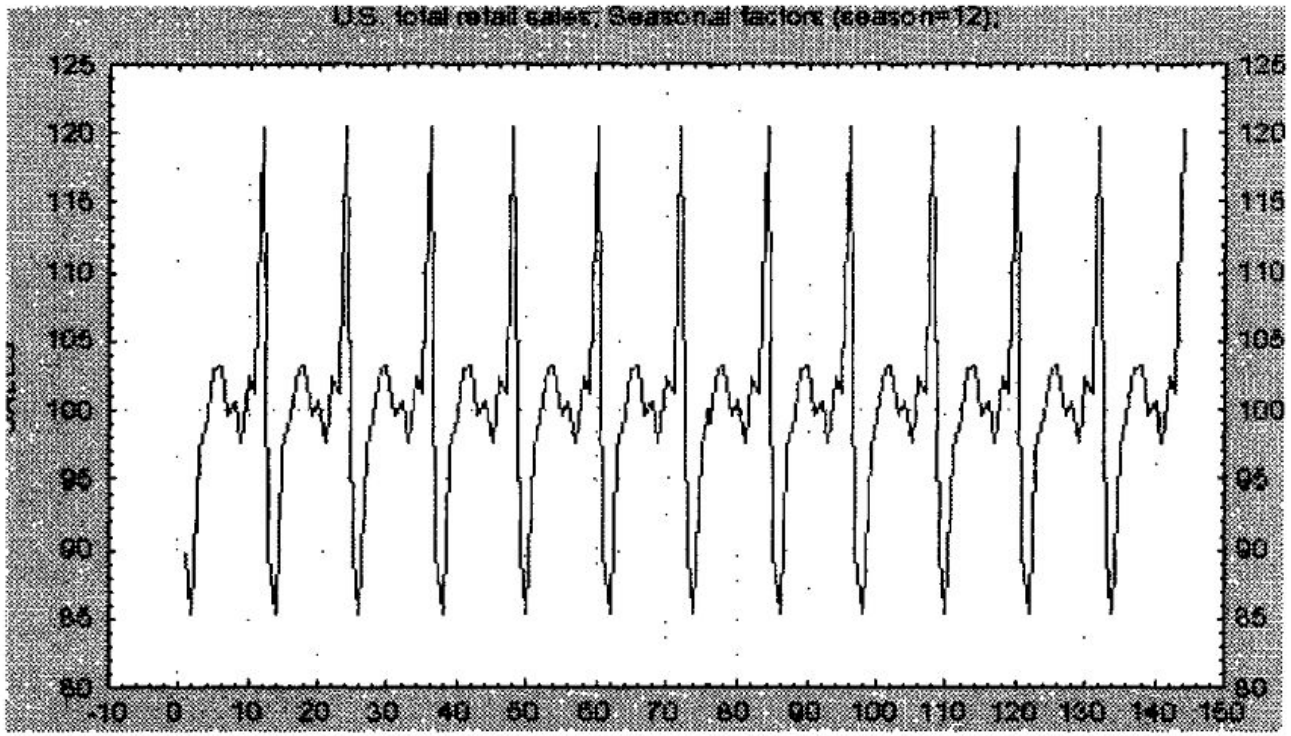
Ежемесячный объем продаж в США с 1953 по 1964 гг.

Анализ временных рядов



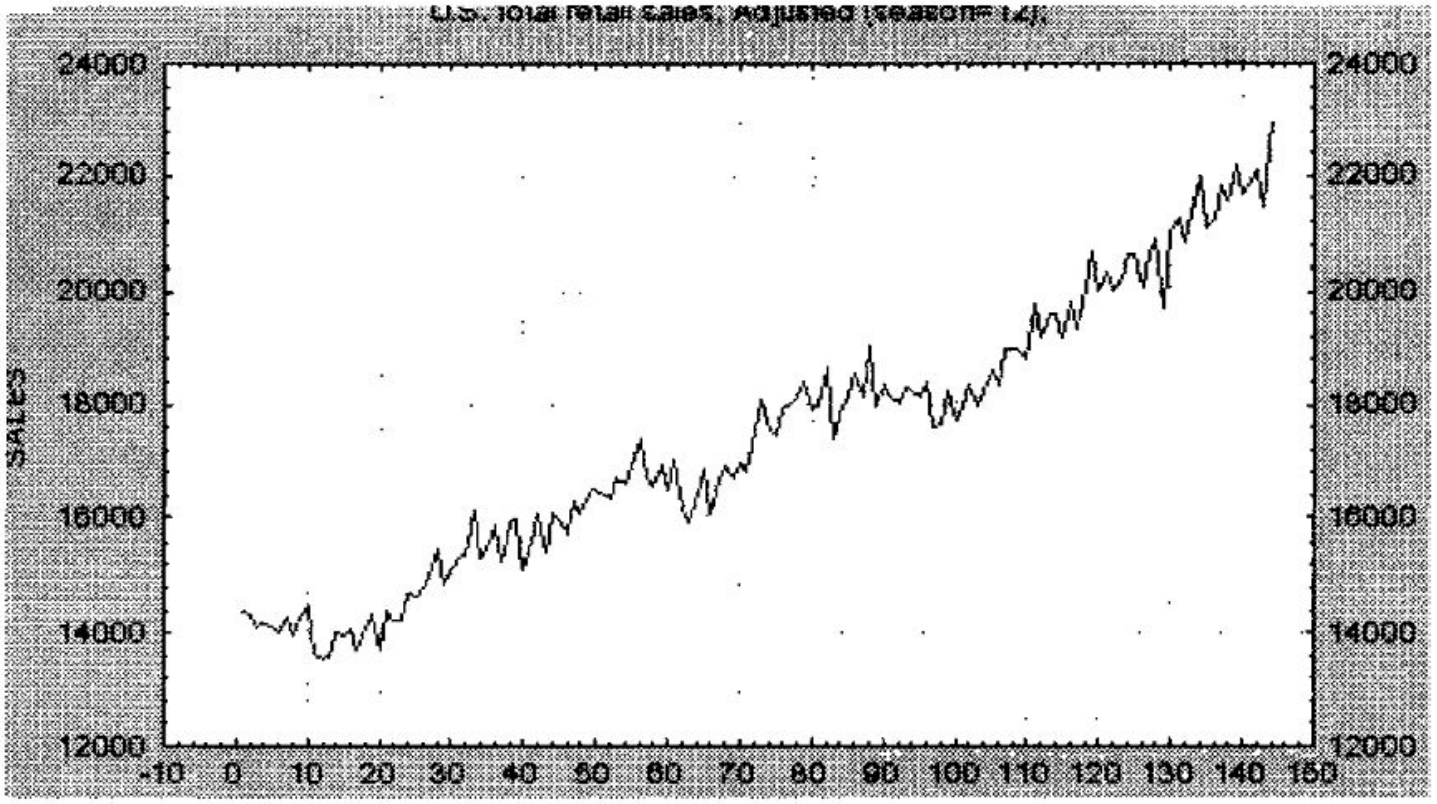
Результаты простого скользящего среднего по 12 точкам

Анализ временных рядов



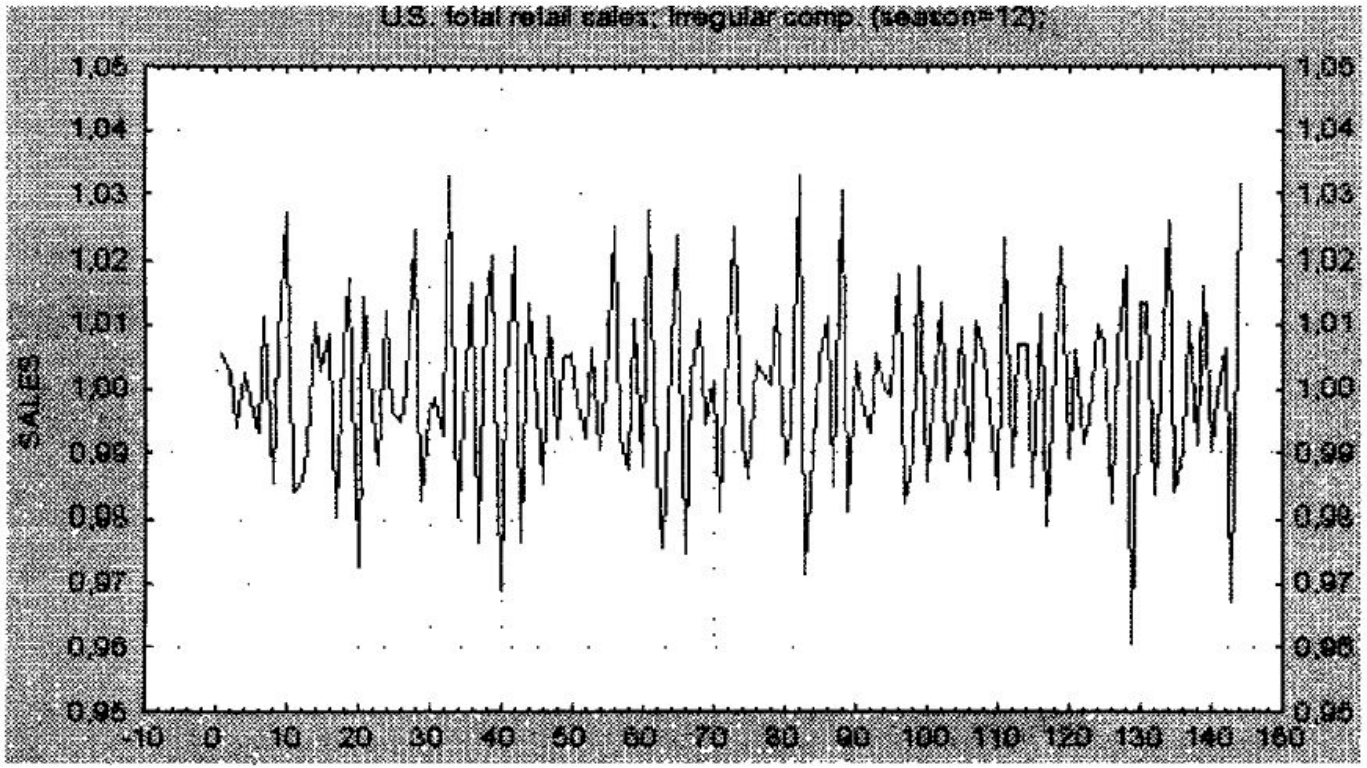
Сезонная компонента

Анализ временных рядов



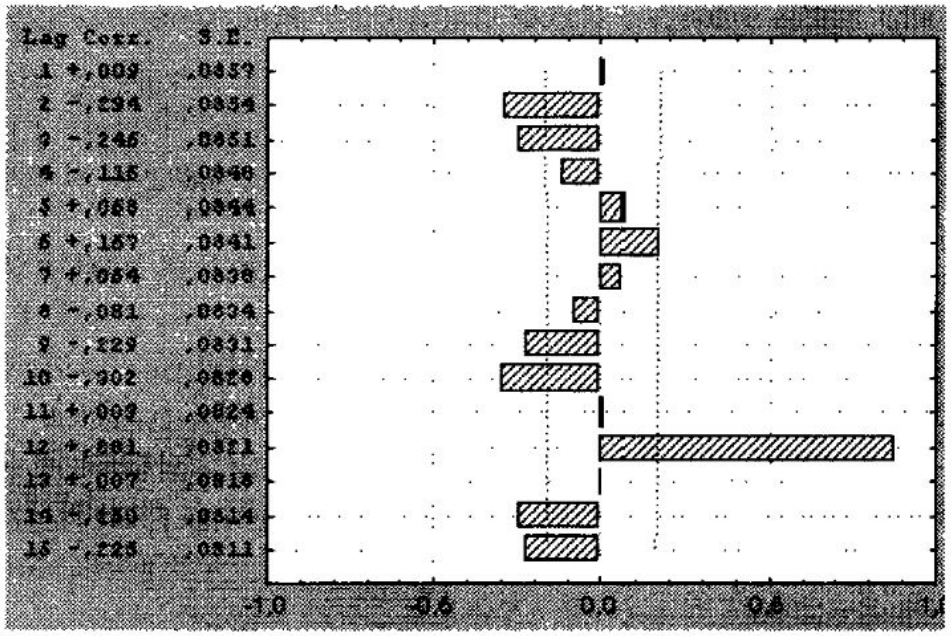
Данные скорректированные на сезонную составляющую

Анализ временных рядов



Случайная остаточная составляющая

Анализ временных рядов



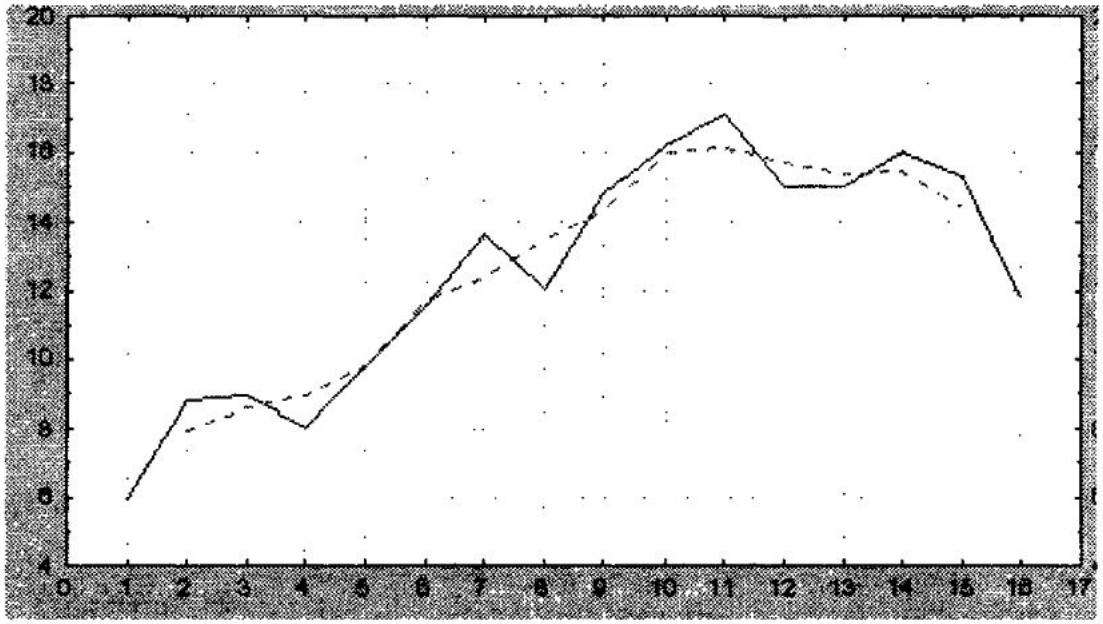
Коррелограмма ряда без тренда

Обнаружена сезонная составляющая с периодом, равным 12 месяцев (r_{12}≈0.9).

Анализ временных рядов

t	y_t	Сглаженный ряд
1	6,00	—
2	8,82	7,92
3	8,94	8,60
4	8,05	8,91
5	9,75	9,77
6	11,51	11,65
7	13,69	12,41
8	12,04	13,50
9	14,76	14,33
10	16,18	16,02
11	17,11	16,09
12	14,99	15,70
13	15,01	15,33
14	16,00	15,42
15	15,26	14,34
16	11,75	—

Анализ временных рядов



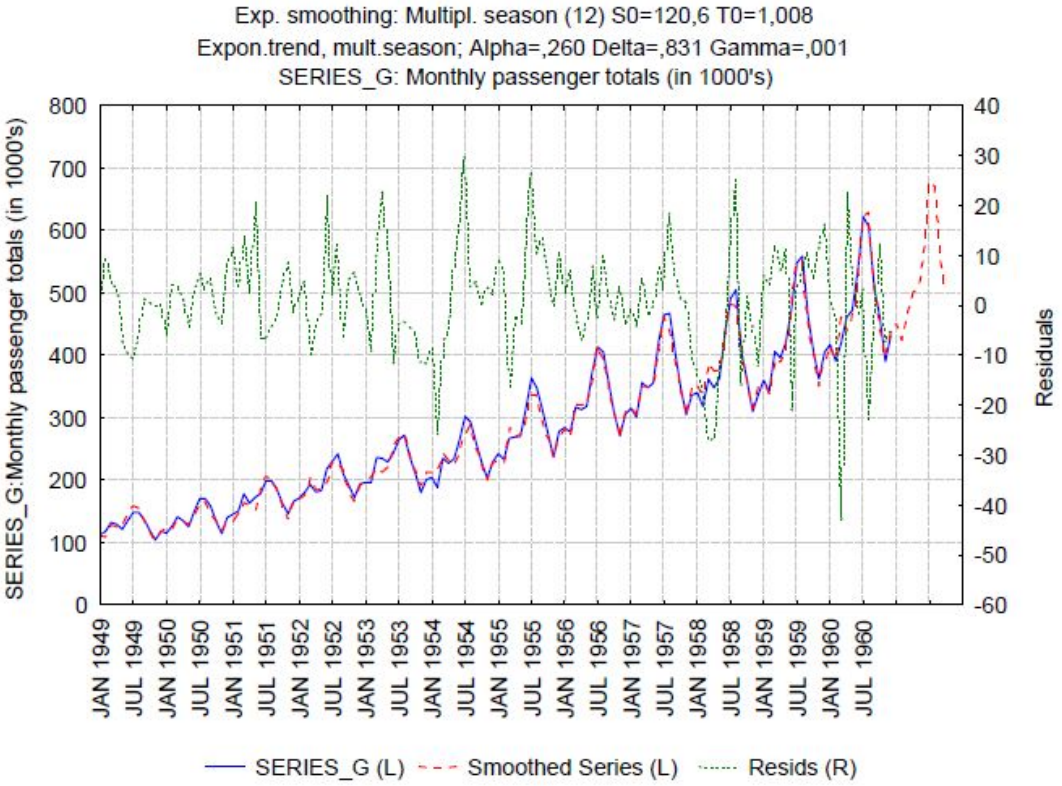
— — исходный ряд; --- — сглаженный ряд, полученный при помощи простого скользящего среднего по трем точкам

Анализ временных рядов

Расчет весов экспоненциального сглаживания

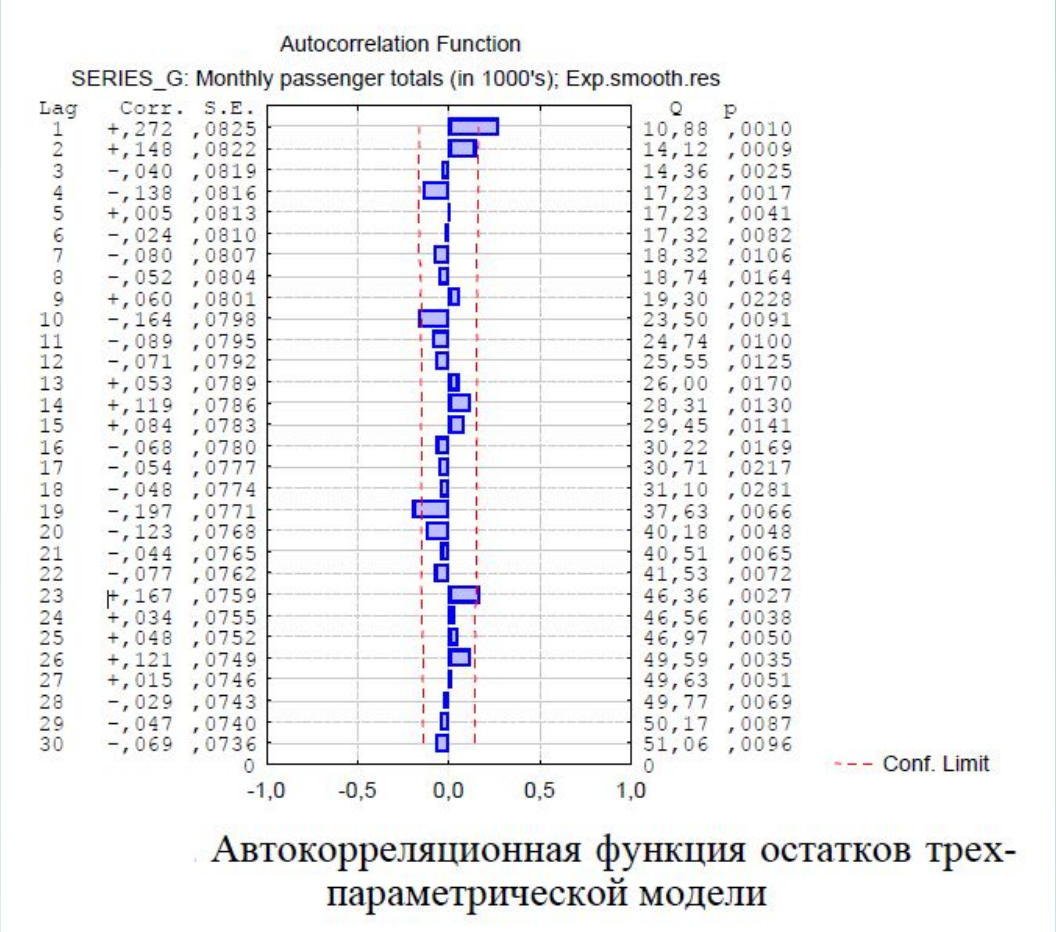
α	$\alpha(1-\alpha)$	$\alpha(1-\alpha)^2$	$\alpha(1-\alpha)^3$	$\alpha(1-\alpha)^4$
0,01	0,0099	0,0098	0,0097	0,0096
0,3	0,21	0,147	0,1029	0,0720
0,7	0,21	0,063	0,0189	0,0057

Анализ временных рядов



Результаты трехпараметрического экспоненциального сглаживания

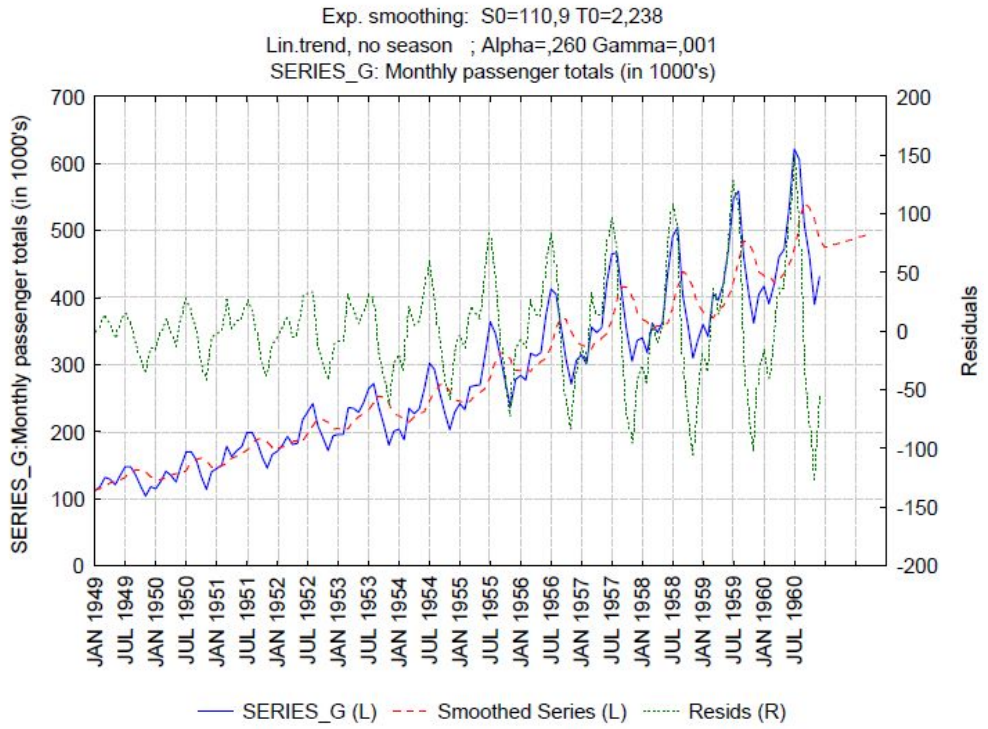
Анализ временных рядов



Автокорреляционная функция остатков трех-параметрической модели

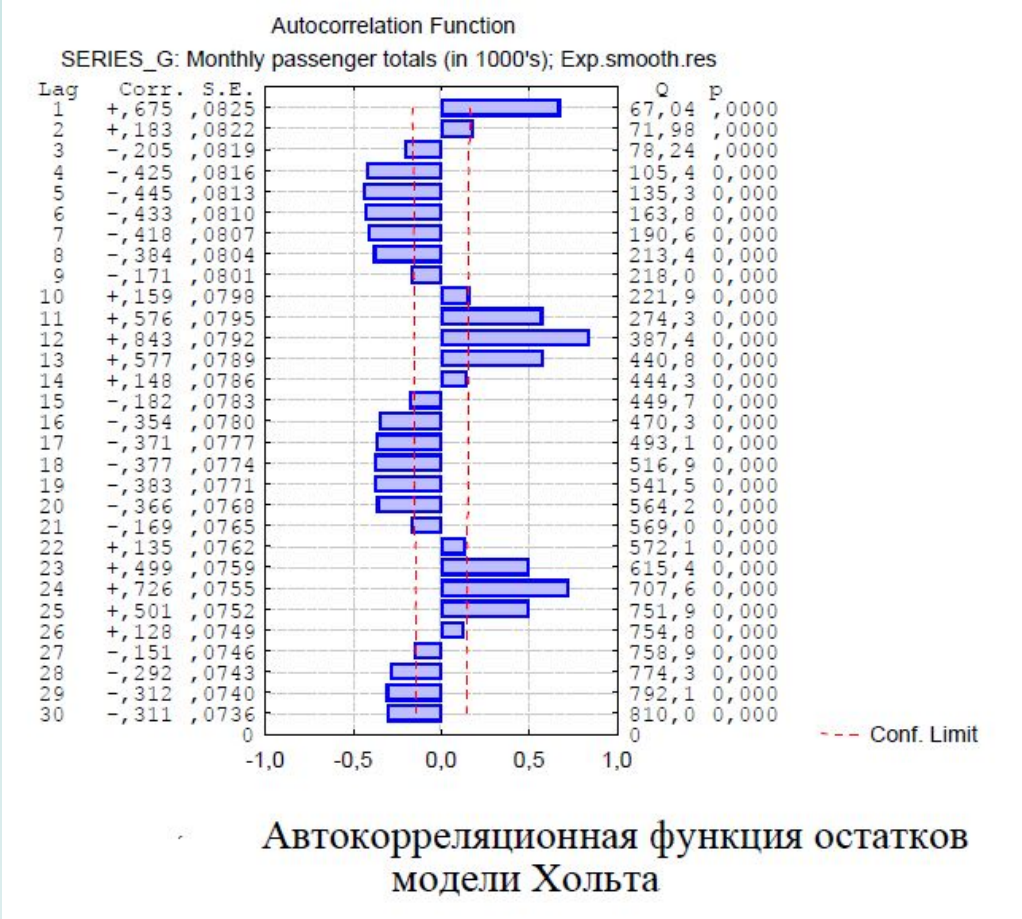
Анализ временных рядов

- Неадекватная модель



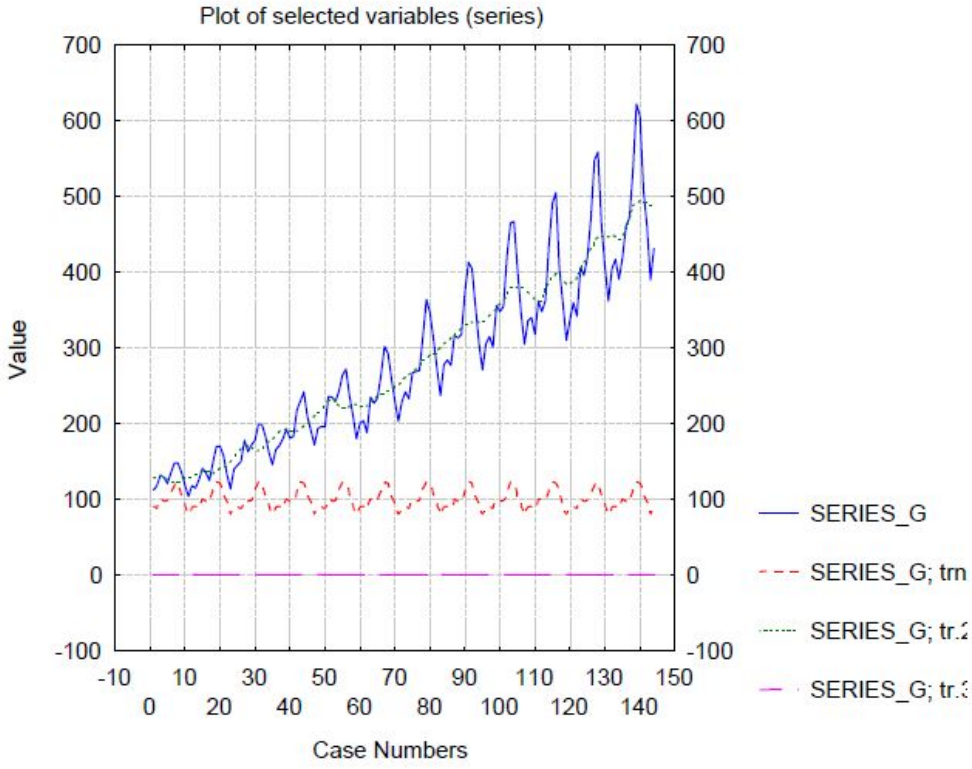
Результаты двухпараметрического экспоненциального сглаживания

Анализ временных рядов



Автокорреляционная функция остатков модели Хольта

Анализ временных рядов



Результаты декомпозиции временного ряда авиаперевозок

Выделены тренд-циклический, сезонный и случайный компоненты

Анализ временных рядов

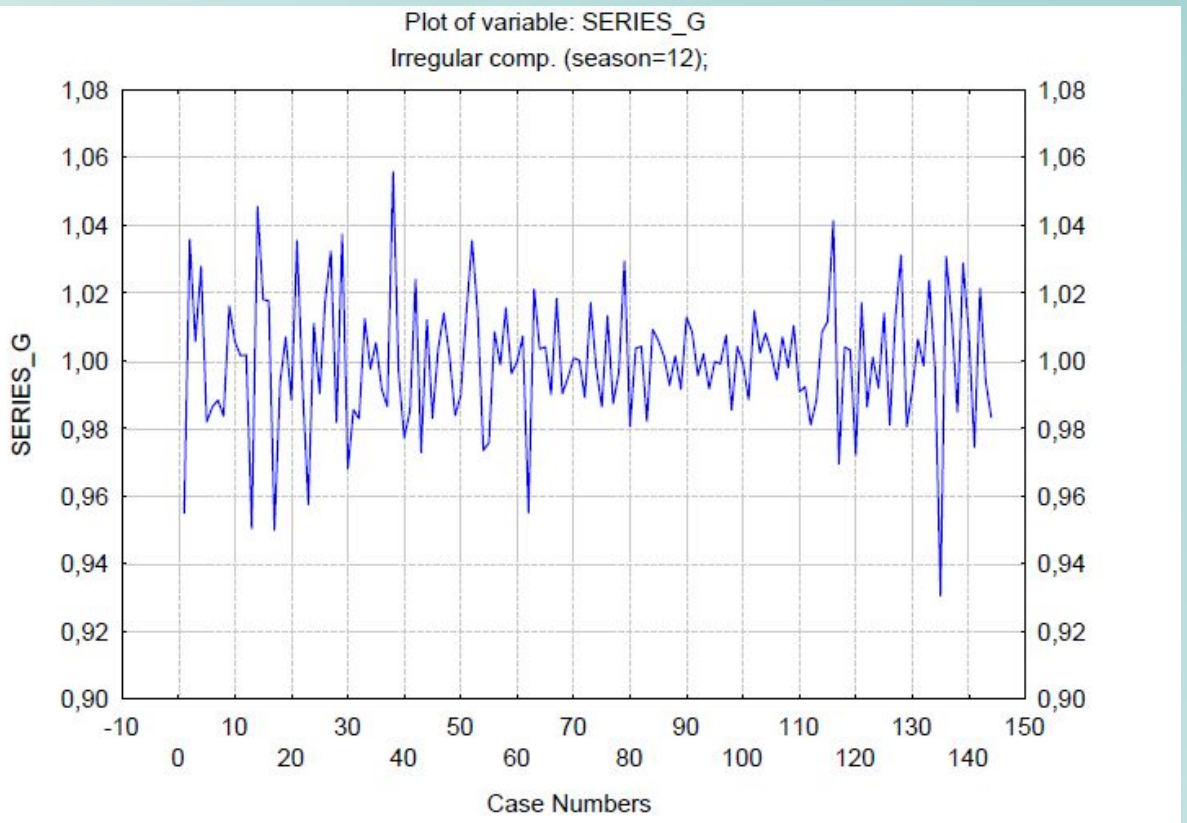
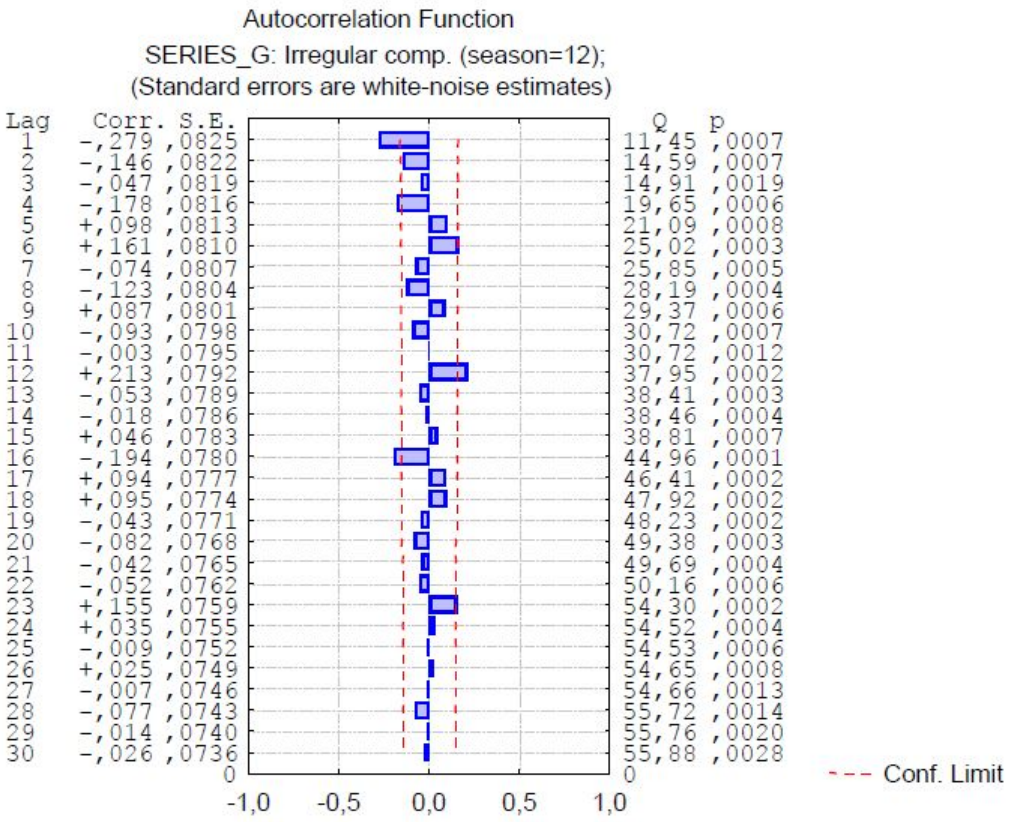


График нерегулярного компонента декомпозиции ряда авиаперевозок

Анализ временных рядов



Автокорреляционная функция нерегулярного компонента

Анализ временных рядов

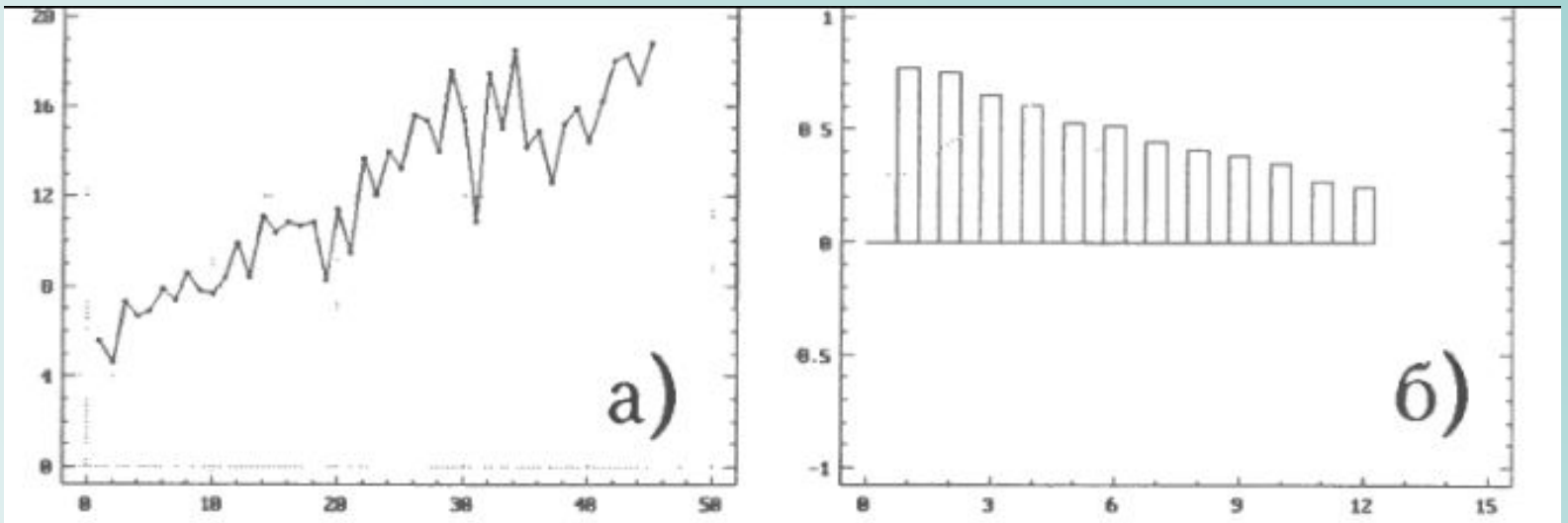


Рис. 12.14. Коррелограмма ряда урожайности зерновых:
а) исходный временной ряд; б) его коррелограмма

- Временной ряд, содержащий тренд: коррелограмма не стремится к 0.

Анализ временных рядов

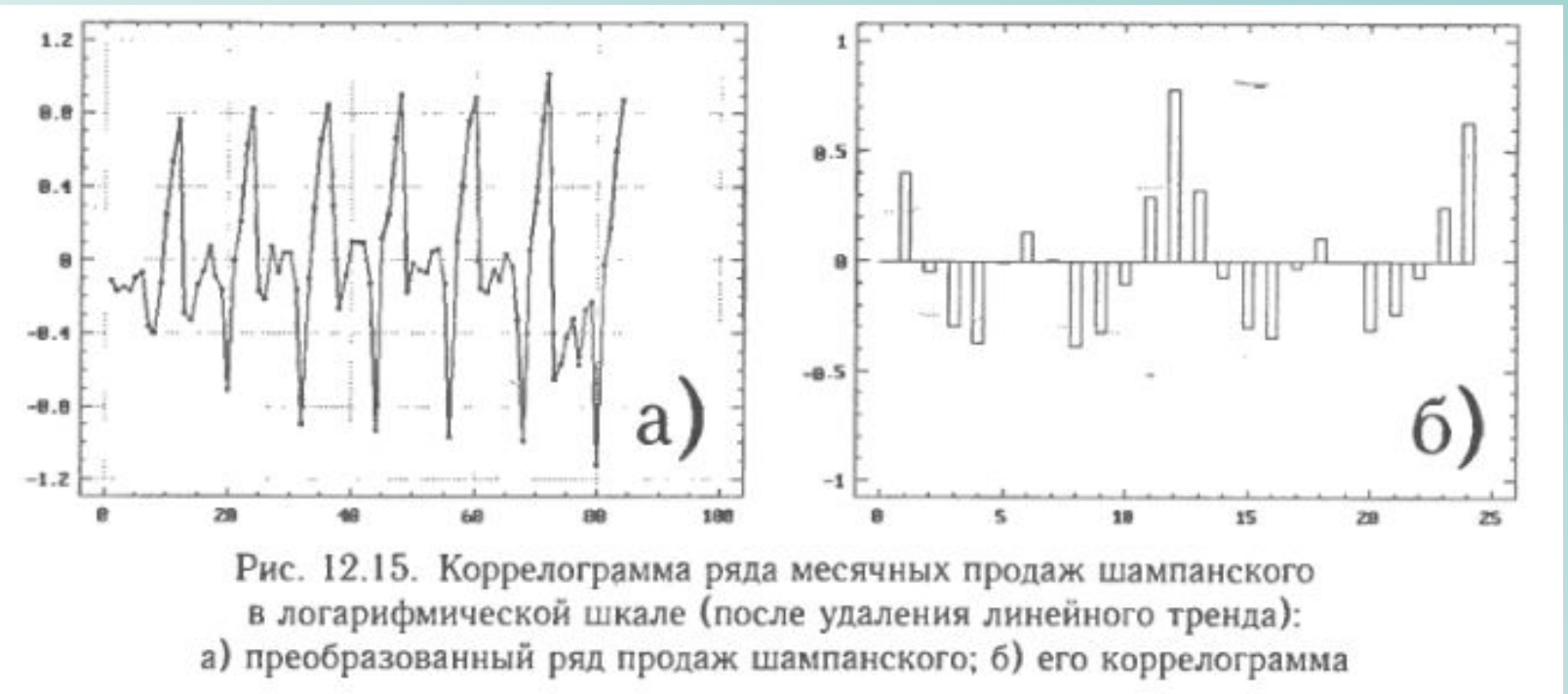


Рис. 12.15. Коррелограмма ряда месячных продаж шампанского в логарифмической шкале (после удаления линейного тренда): а) преобразованный ряд продаж шампанского; б) его коррелограмма

- Ряд с сезонной составляющей, после удаления тренда: коррелограмма показывает наличие сезонной составляющей

Анализ временных рядов

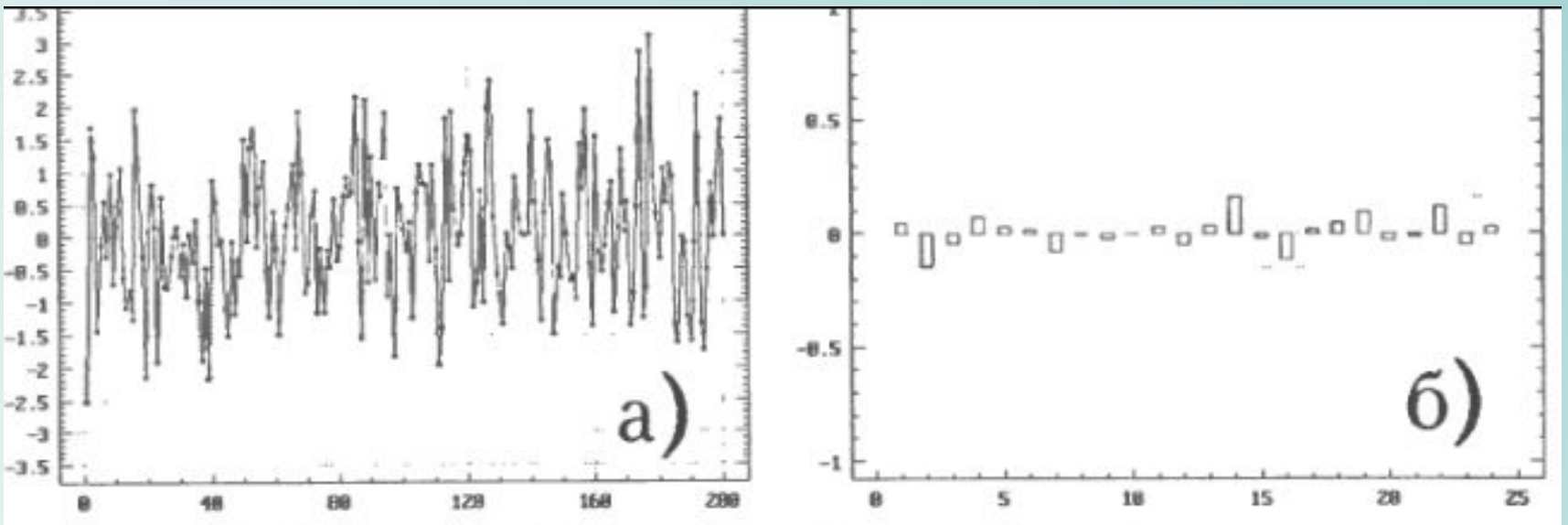
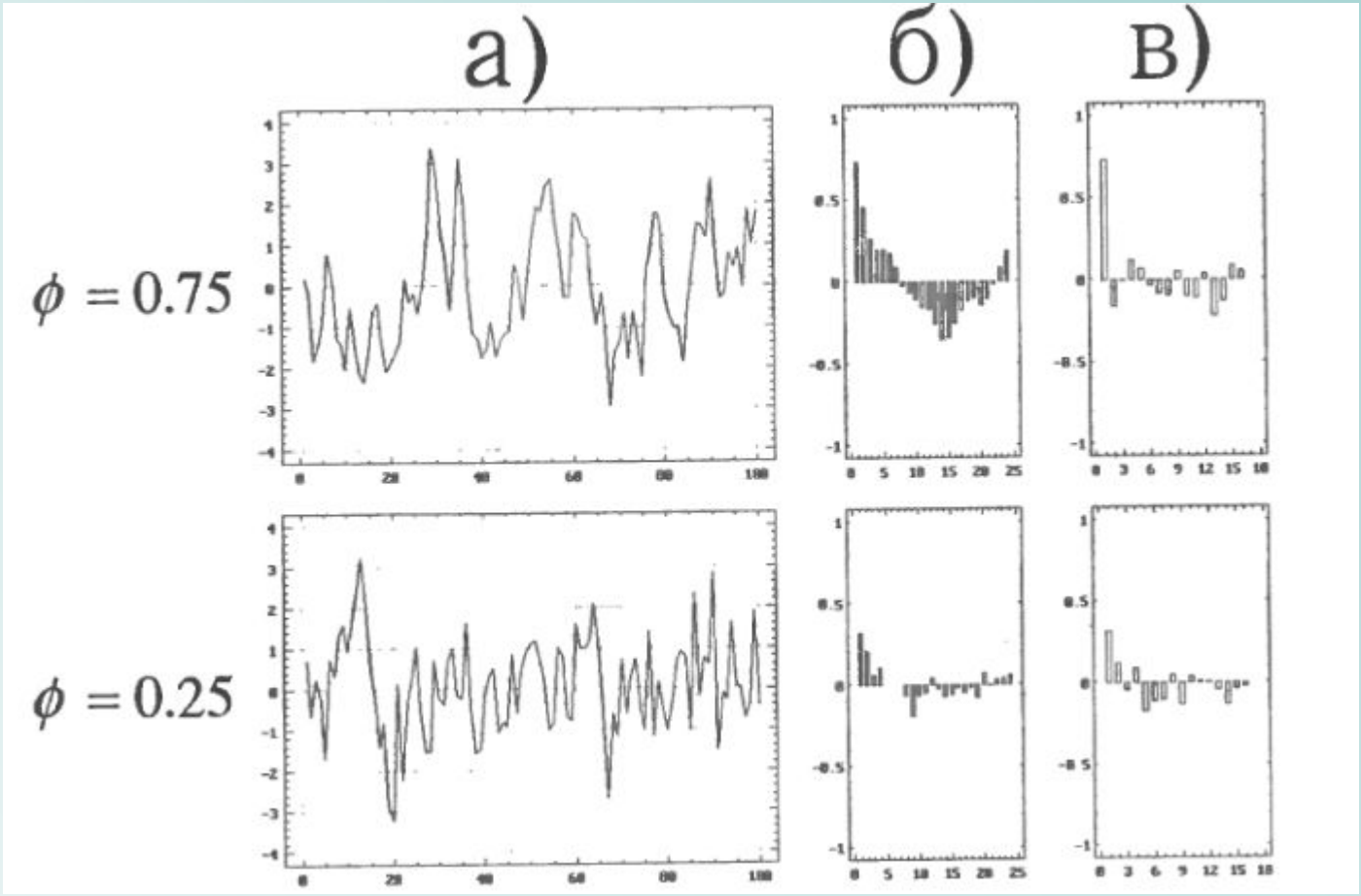


Рис. 12.16. Коррелограмма белого шума: а) исходный ряд; б) его коррелограмма

Анализ временных рядов



$\sigma^2 = 1$

Анализ временных рядов

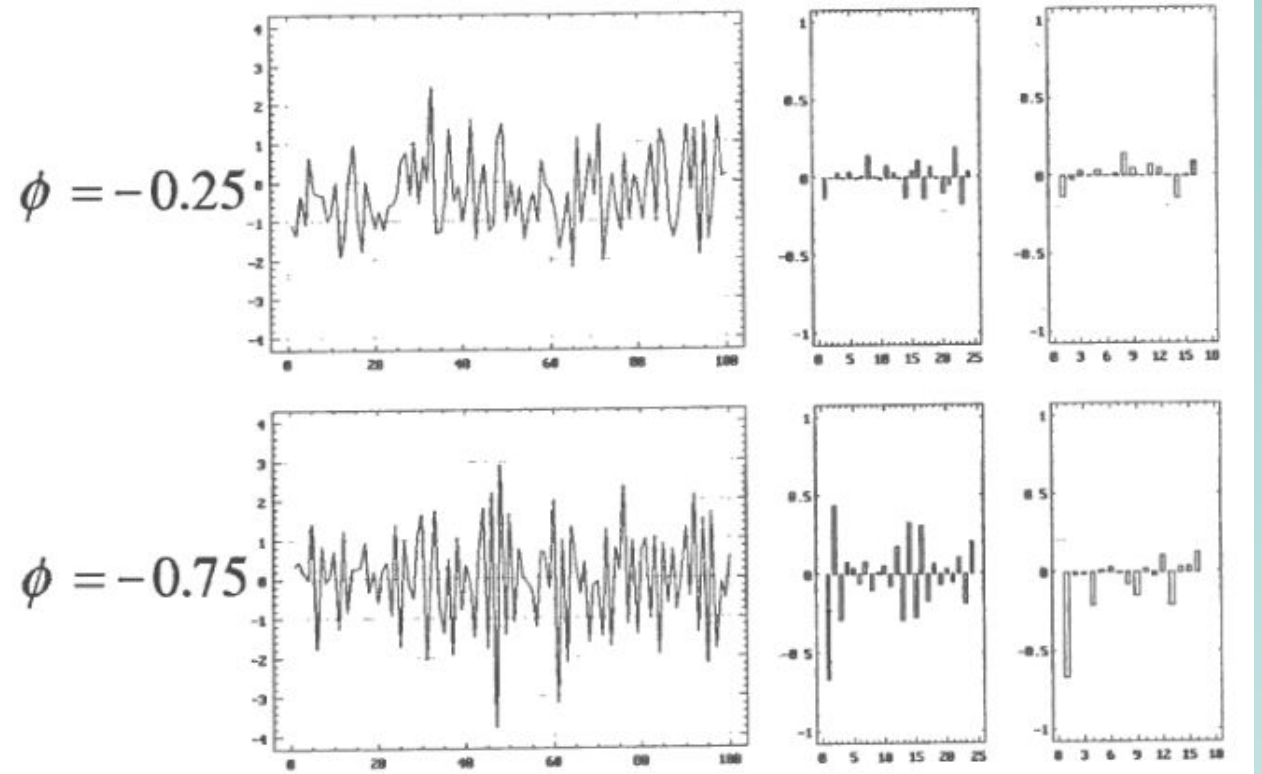


Рис. 14.1. Графики AR(1) процессов и их выборочных автокорреляционных и частных автокорреляционных функций для различных значений коэффициента ϕ а) график исходного ряда; б) график выборочной автокорреляционной функции; в) график частной автокорреляционной функции

Анализ временных рядов

- Автокорреляционные функции авторегрессионных рядов экспоненциально затухают или представляют экспоненциально затухающие синусоидальные волны.

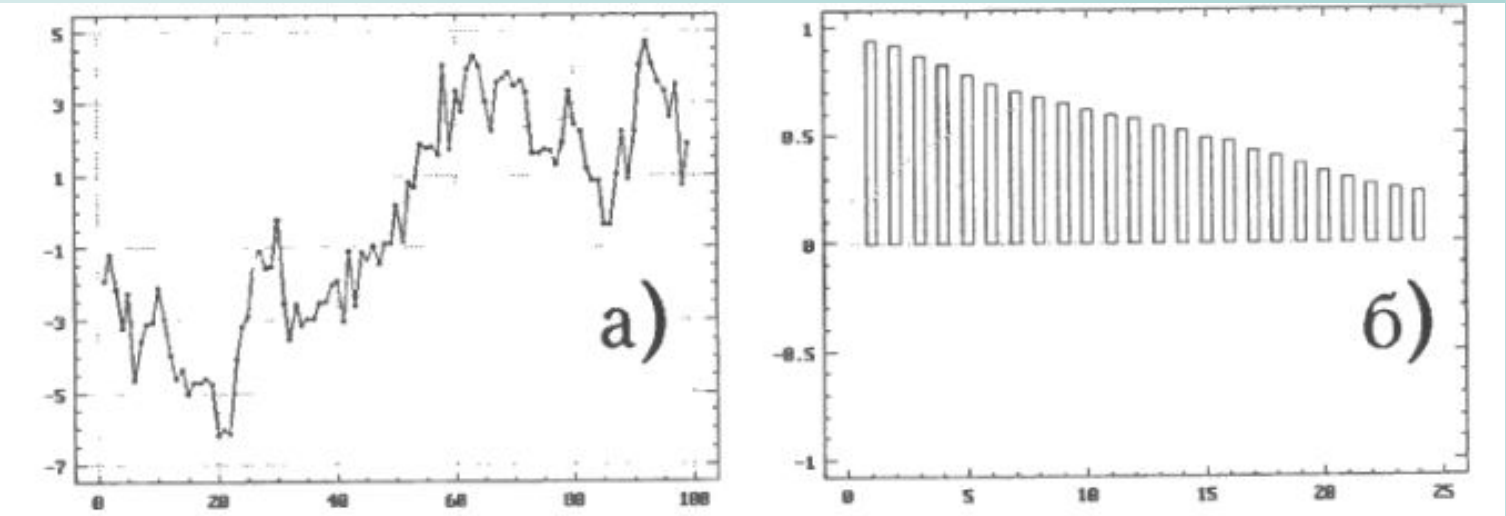


Рис. 12.19. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$, $\phi_2 = 0.25$: а) исходный ряд; б) его коррелограмма

Анализ временных рядов

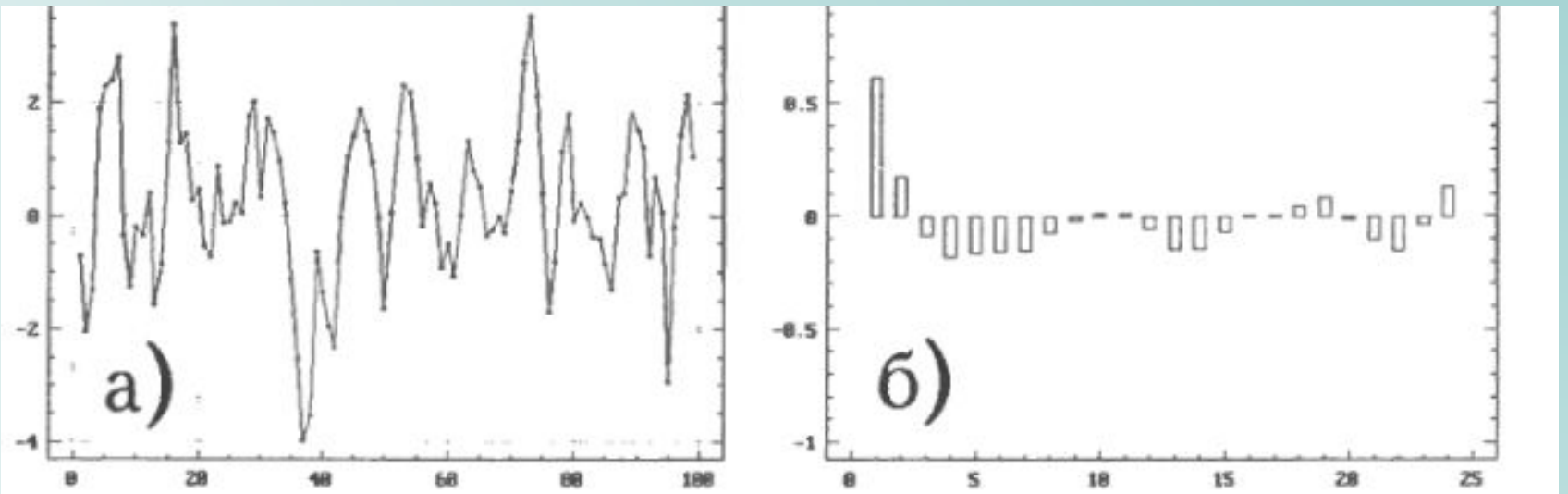
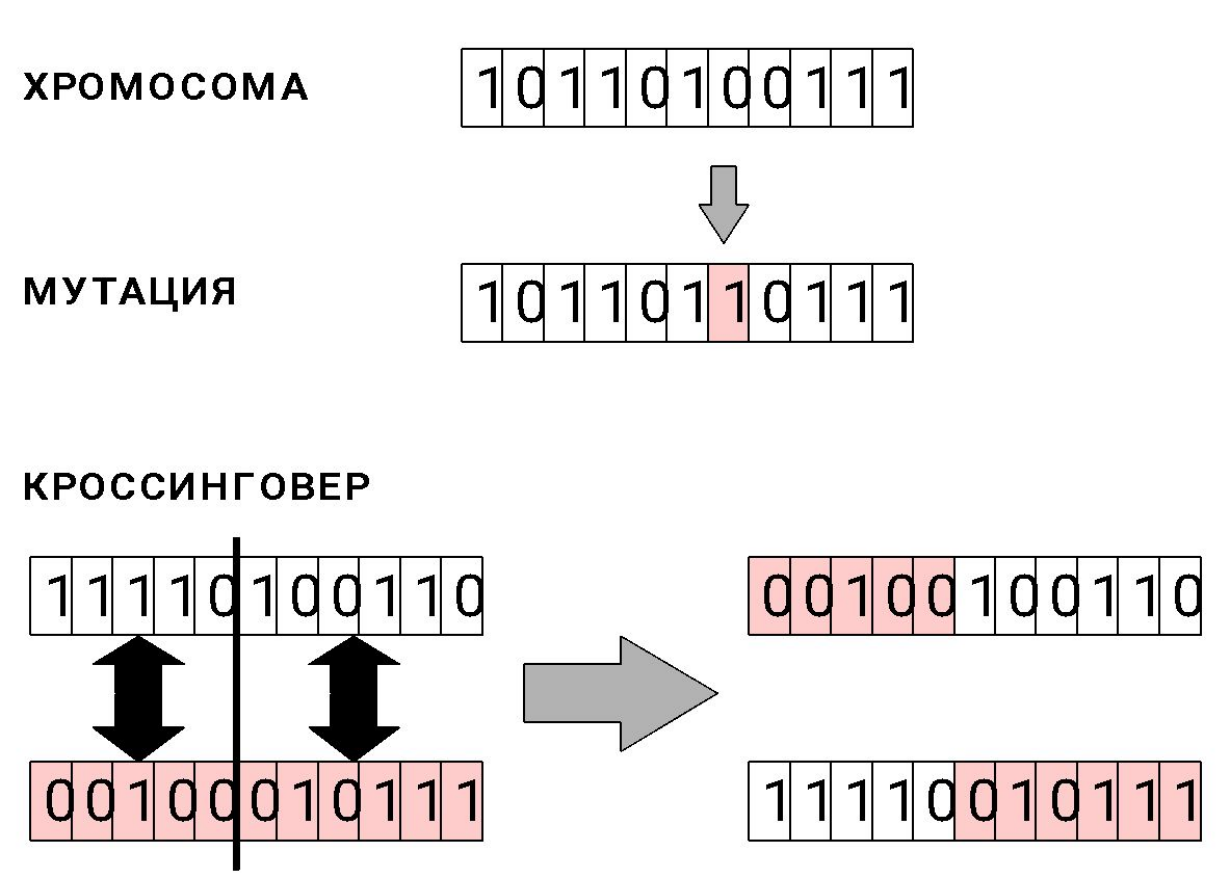
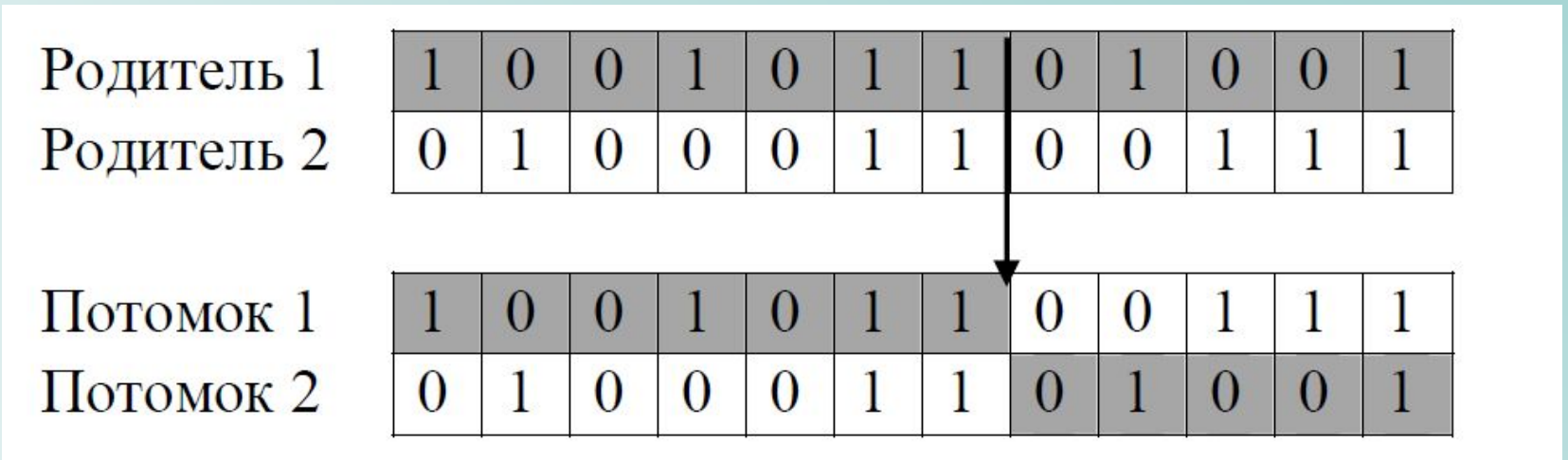


Рис. 12.20. Коррелограмма AR(2) процесса при $\phi_1 = 0.7$ $\phi_2 = -0.25$: а) исходный ряд; б) его коррелограмма

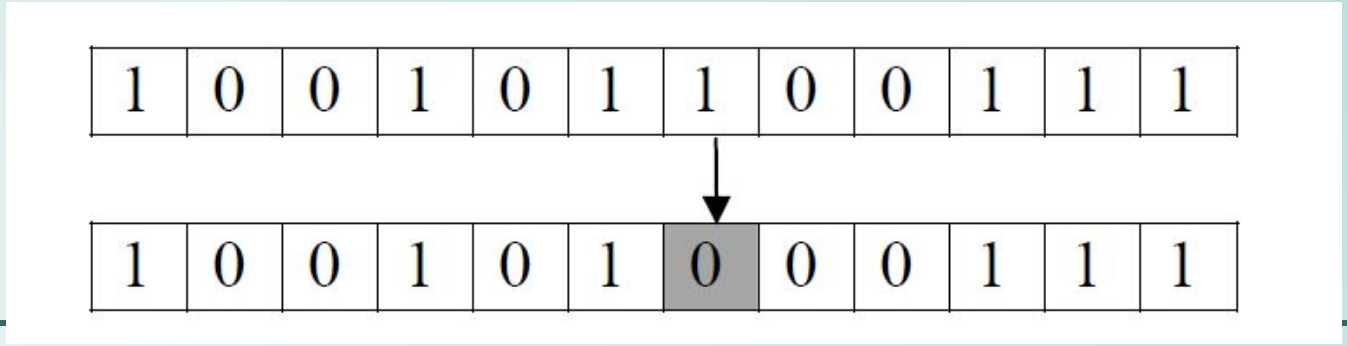
Генетические алгоритмы



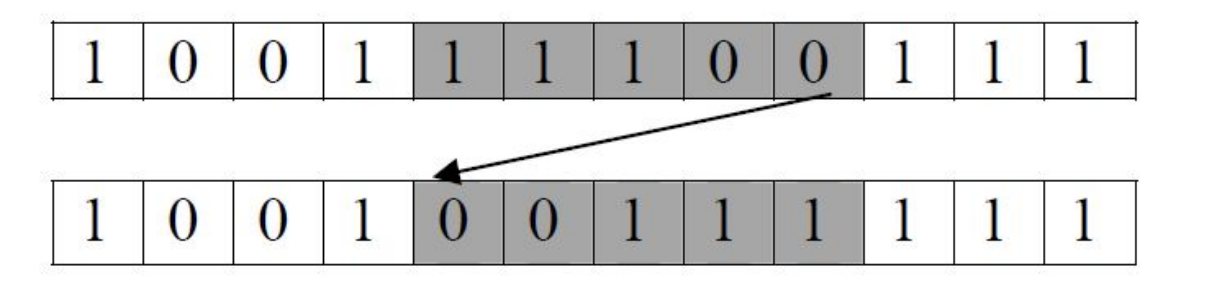
Генетические алгоритмы

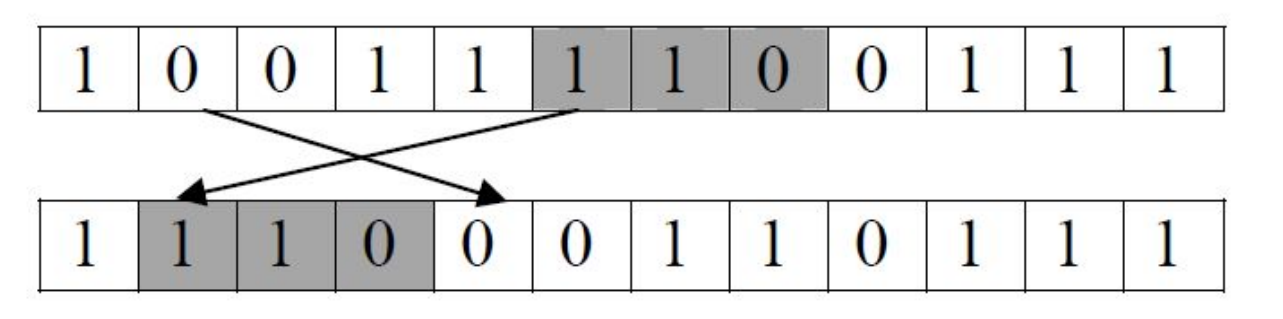


Генетические алгоритмы



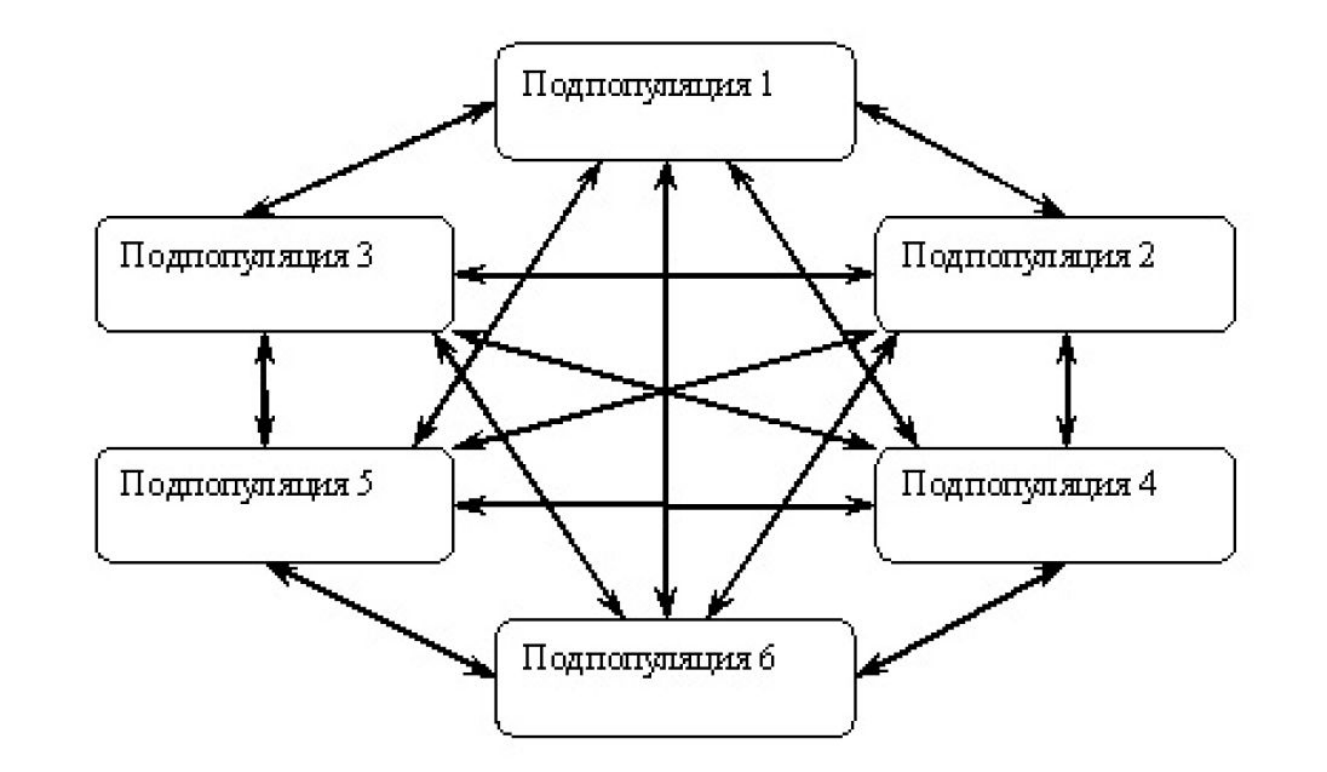
Генетические алгоритмы



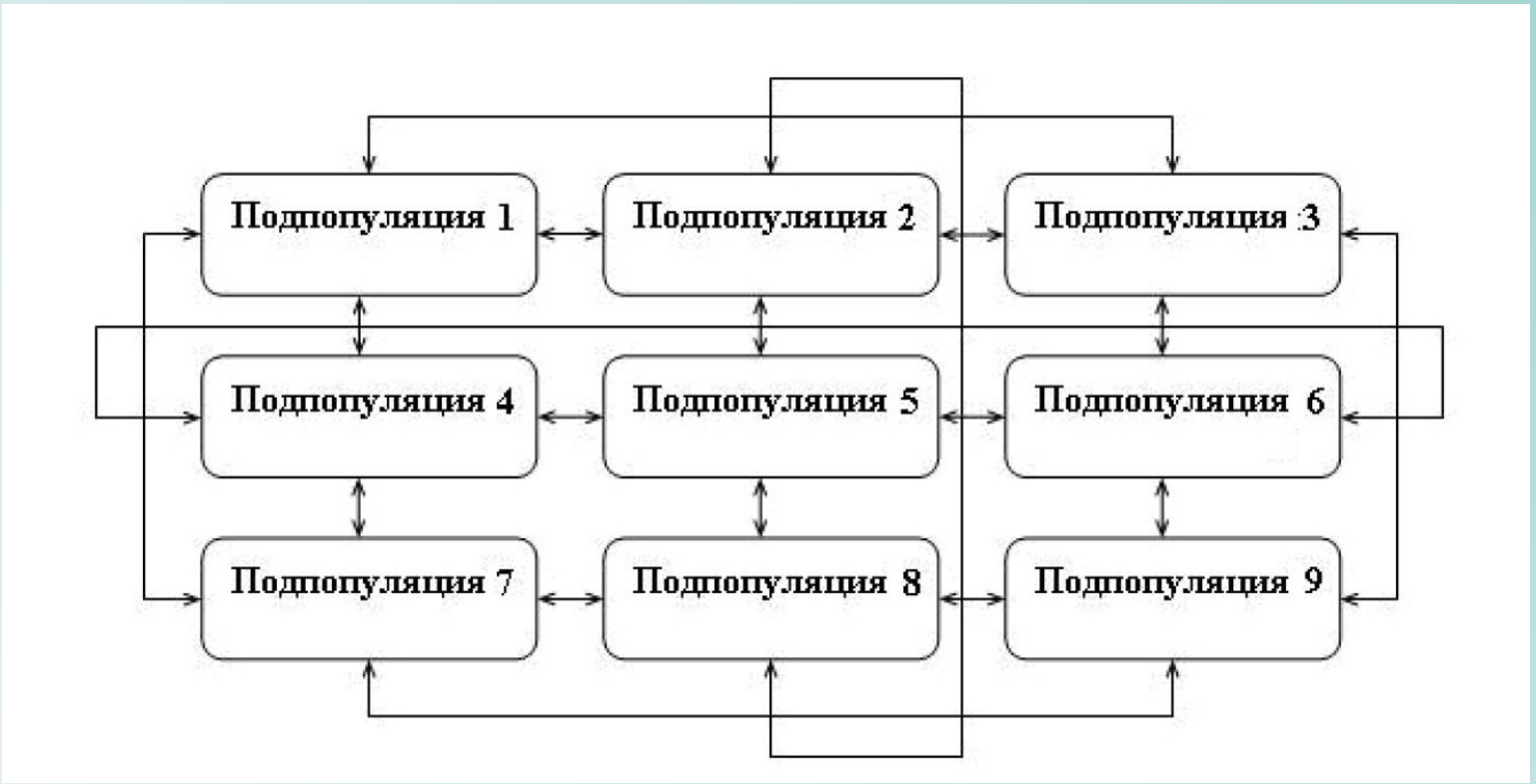


Параллельные ГА

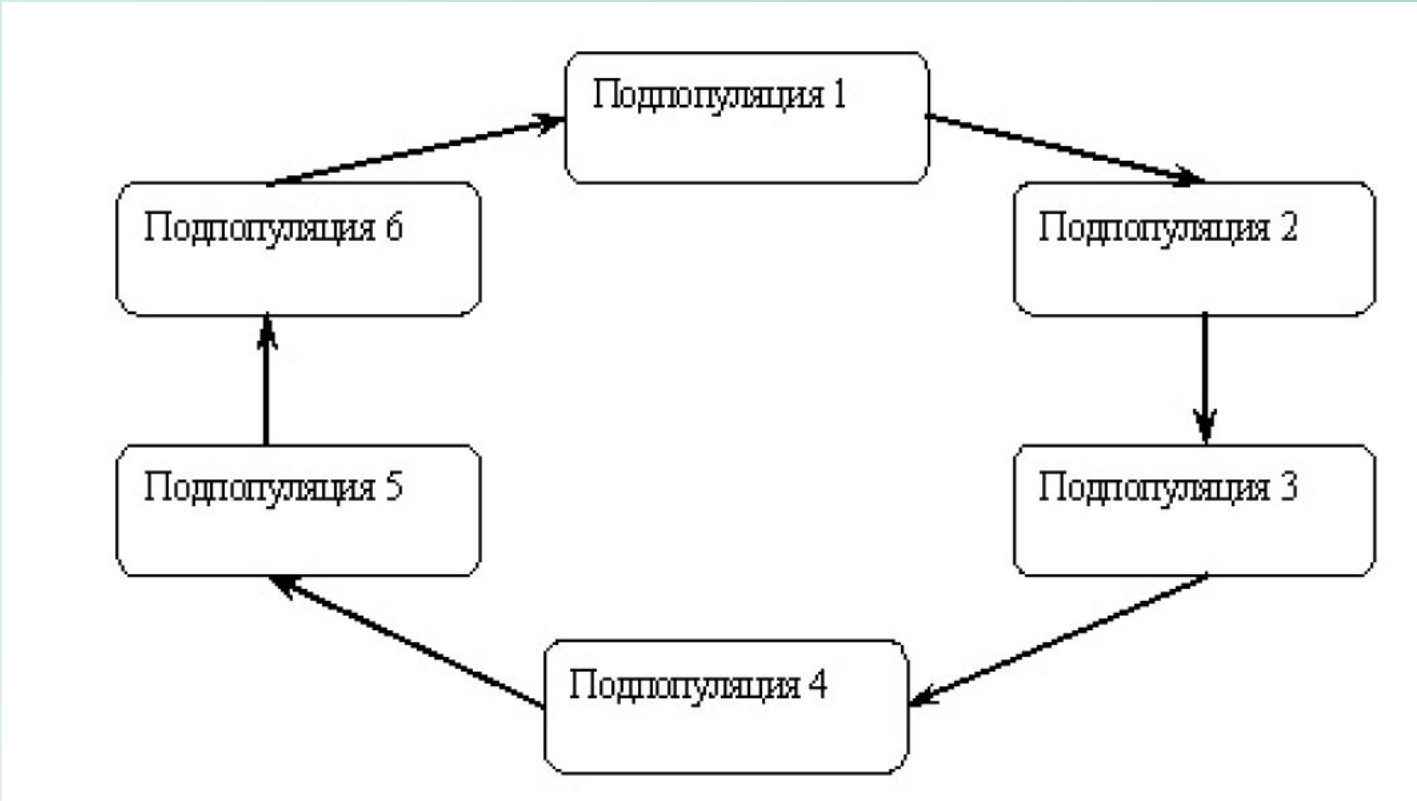
- Модель миграции



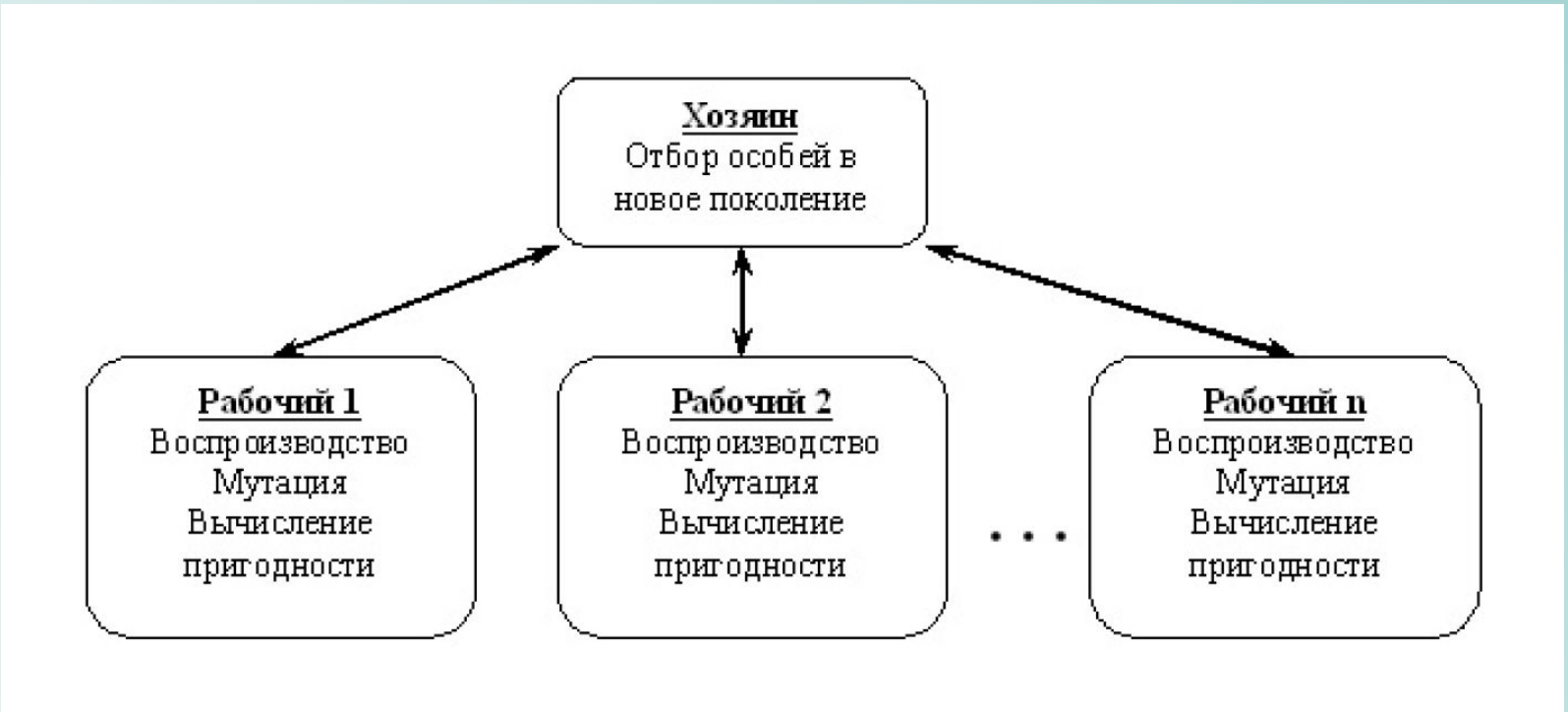
Параллельные ГА



Параллельные ГА



Параллельные ГА



PolyAnalyst

Название модуля	Технология/методы
Find Laws Algorithm (FL)	Symbolic Knowledge Acquisition Technology, Эволюционное программирование
PolyNet Predictor Algorithm (PN)	GMDH-Neural Net hybrid, гибрид метода МГУА и нейронных сетей
Find Dependencies Algorithm (FD)	N-dimensional distribution analysis, N-мерный анализ распределений
Cluster Algorithm (FC)	Localization of Anomalies, N-мерный кластеризатор
PAY Algorithm (MB)	Memory Based Reasoning and Genetic Algorithms hybrid, гибрид метода "ближайших соседей" и генетических алгоритмов
Market Basket Analysis (BA)	Transactional clustering and directed association rules, транзакционный кластеризатор с генерацией направленных ассоциативных правил
Linear Regression (LR)	Stepwise Linear Regression, многопараметрическая линейная регрессия с автоматическим выбором независимых переменных
Classify Algorithm (CL)	Fuzzy logic classification, классификация по булевой целевой переменной, необходимо наличие модуля FL, или PN, или MB, или LR
Disciminate (DS)	Модификация модуля CL, обнаруживает различия между двумя таблицами
Decision Trees (DT)	Модуль "деревья решений", классификация на категории
Summary Statistics (SS)	Модуль общей статистики

Генетические алгоритмы

Хромосома	вектор генов
Генотип ↔ Фенотип	набор хромосом ↔ вариант решения задачи
Отбор	один из методов выбора определённого числа особей, которые будут участвовать в формировании новой популяции
Кроссинговер (кроссовер)	операция при которой две хромосомы обмениваются своими частями
Мутация	Случайное изменение одного или нескольких генов в хромосоме
Фитнес-функция	Функция, определяющая приспособленность особи

Генетические алгоритмы

- **НАЧАЛО** // простой генетический алгоритм
  Создать начальную совокупность структур(популяцию)
  Оценить каждую структуру
  останов := **FALSE**
  ПОКА НЕ останов **ВЫПОЛНЯТЬ**
  НАЧАЛО // новая итерация (поколение)
  Применить оператор отбора
  ПОВТОРИТЬ (размер_популяции/2) **РАЗ**
  НАЧАЛО // цикл воспроизводства
  Выбрать две структуры (родители) из множества предыдущей итерации
  Применить оператор скрещивания с заданной вероятностью к выбранным структурам и получить две новые структуры (потомки)
  Оценить эти новые структуры
  Если оператор скрещивания не применяется, то потомки становятся копиями своих родителей
  Поместить потомков в новое поколение
  КОНЕЦ
  Применить оператор мутации с заданной верояностью
  ЕСЛИ популяция сошлась **ТО** останов := **TRUE**
  КОНЕЦ

КОНЕЦ