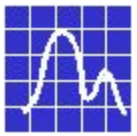
Сегментация в примерах

Полежаев Илья
Владимир Боровиков



StatSoft® Russia

Кластерный анализ

Анализируемый объект представляет собой жилое многоквартирное здание. подвергнутое модернизации с целью экономии потребления ресурсов.

Постановка задачи

Требуется разбить квартиры на три как можно больше отличающиеся друг от друга группы по признаку - количество потребляемого тепла.

Данные

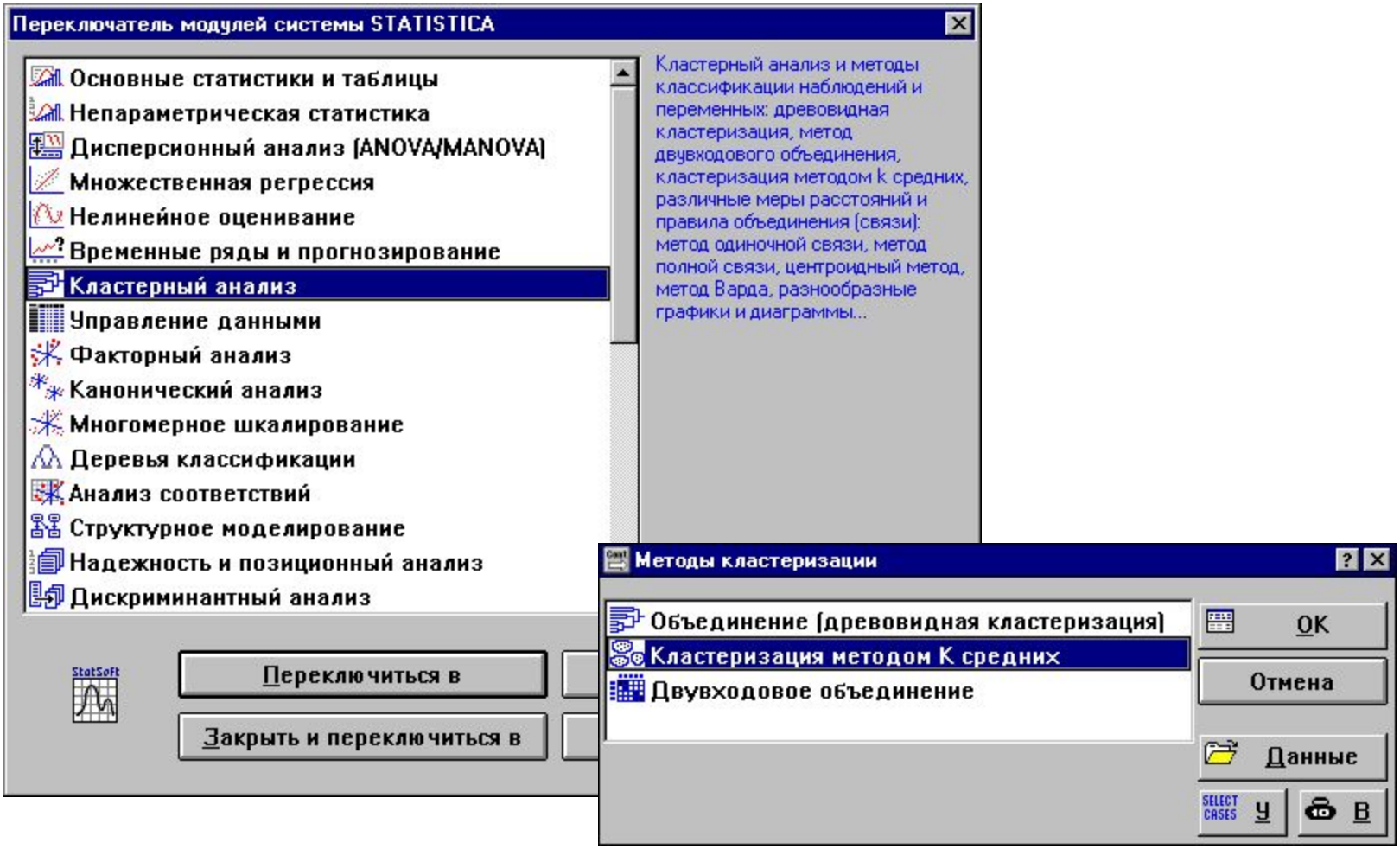
Показания датчиков температуры, размещенных в квартирах. Данные собирались в течение 2-х лет.

Данные: mean.sta 25п * 218н

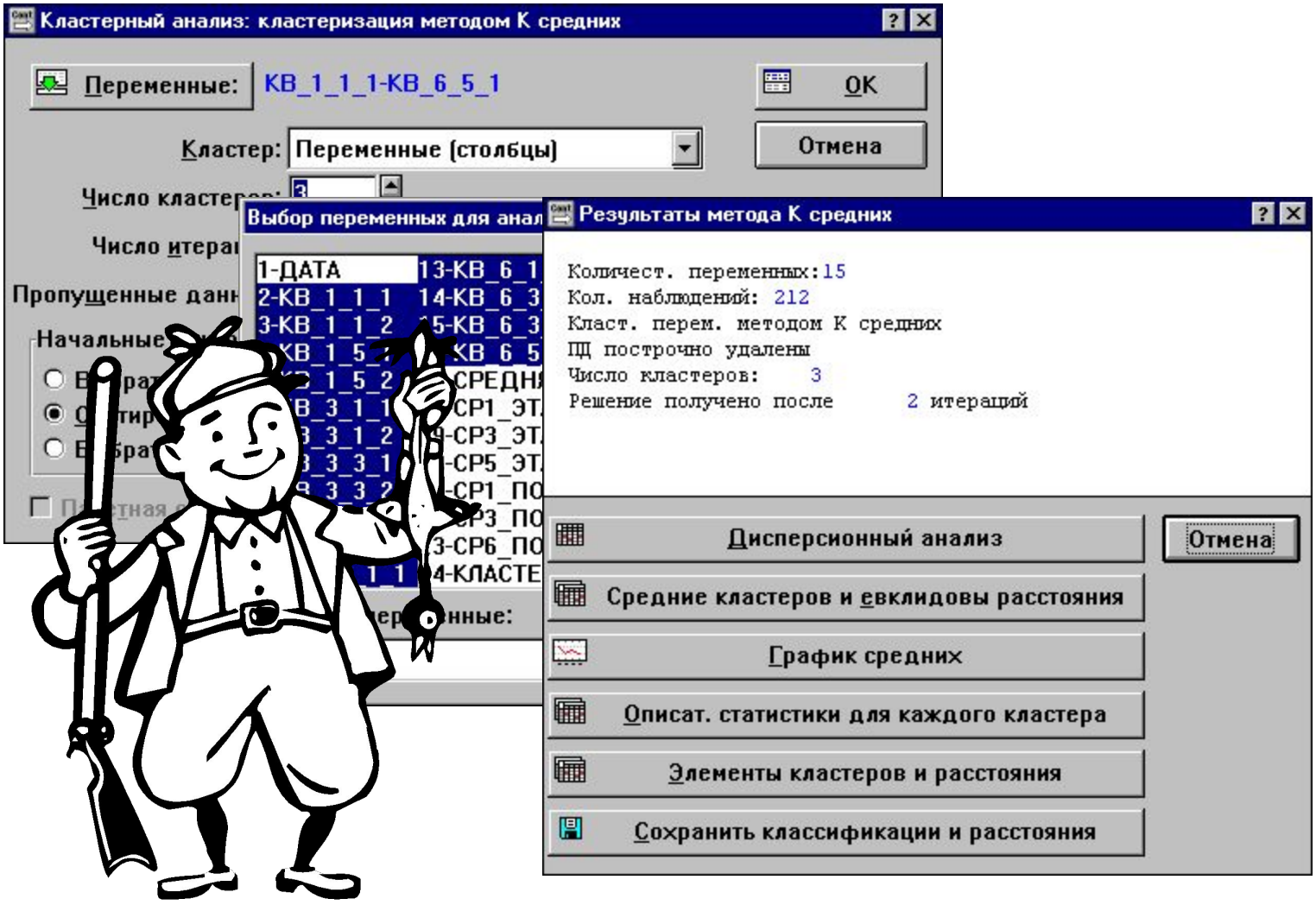
(hc41rt99.sta)

	1 ДАТА	2 KB_1_1_1	3 KB_1_1_2	4 KB_1_5_1	5 KB_1_5_2	6 KB_3_1_1	7 KB_3_1_2	8 KB_3_3_1
1	19/10/98	21.03000	23.97800	22.58400	21.31000	21.06400	23.94800	24.3940
2	20/10/98	21.07000	21.45167	22.49250	22.30250	19.45833	23.14083	24.0258
3	21/10/98	21.05000	20.48167	22.31375	22.02167	19.85458	23.39833	23.9891
4	22/10/98	20.66583	19.48958	21.42833	21.69625	20.33708	22.11792	23.7962
5	23/10/98	20.14000	18.12708	20.02667	19.93042	21.39500	22.82625	23.4716
6	24/10/98	20.04333	18.64958	21.42333	19.63542	22.04917	23.01792	23.7804
7	25/10/98	20.49583	19.79625	22.50083	21.13375	21.86125	22.52250	24.2758
8	26/10/98	20.46292	18.78500	21.65458	21.73667	21.45833	22.49125	23.8795
9	27/10/98	20.39333	19.48458	22.10458	21.88833	19.96708	22.85458	23.7770
10	28/10/98	20.54000	19.32292	21.89292	21.95917	20.87542	22.83083	23.6770
11	29/10/98	20.42000	19.06542	21.08167	20.71583	21.70125	22.24917	24.1016

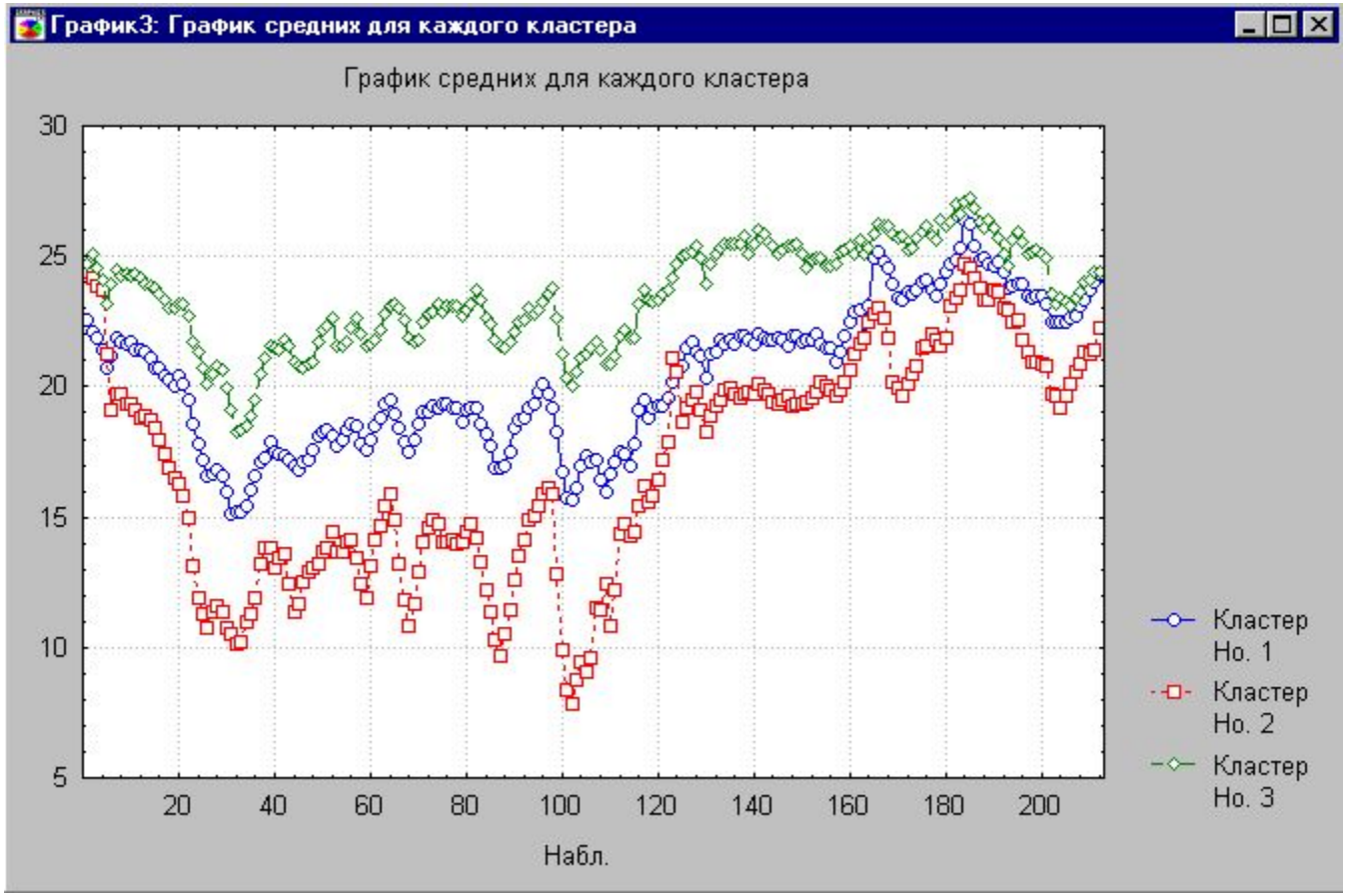
Анализ


Переключатель модулей системы STATISTICA
Основные статистики и таблицы
Непараметрическая статистика
Дисперсионный анализ (ANOVA/MANOVA)
Множественная регрессия
Нелинейное оценивание
Временные ряды и прогнозирование
Кластерный анализ
Управление данными
Факторный анализ
Канонический анализ
Многомерное шкалирование
Деревья классификации
Анализ соответствий
Структурное моделирование
Надежность и позиционный анализ
Дискриминантный анализ
Кластерный анализ и методы классификации наблюдений и переменных: древовидная кластеризация, метод двувходового объединения, кластеризация методом k средних, различные меры расстояний и правила объединения (связи): метод одиночной связи, метод полной связи, центроидный метод, метод Варда, разнообразные графики и диаграммы...
StatSoft
Переключиться в
Закрыть и переключиться в
Методы кластеризации
Объединение (древовидная кластеризация)
Кластеризация методом K средних
Двувходовое объединение
OK
Отмена
Данные
SELECT CASES
У
В

Анализ


Кластерный анализ: кластеризация методом К средних
Переменные:
KB_1_1_1-KB_6_5_1
OK
Кластер:
Переменные (столбцы)
Отмена
Число кластеров:
Число итераций:
Пропущенные данные
Начальные
Выбор переменных для анализа
1-ДАТА
2-KB_1_1_1
3-KB_1_1_2
13-KB_6_1
14-KB_6_3
Результаты метода К средних
Количест. переменных:15
Кол. наблюдений: 212
Класт. перем. методом К средних
ПД построчно удалены
Число кластеров: 3
Решение получено после 2 итераций
Дисперсионный анализ
Отмена
Средние кластеров и евклидовы расстояния
График средних
Описат. статистики для каждого кластера
Элементы кластеров и расстояния
Сохранить классификации и расстояния

Графический анализ



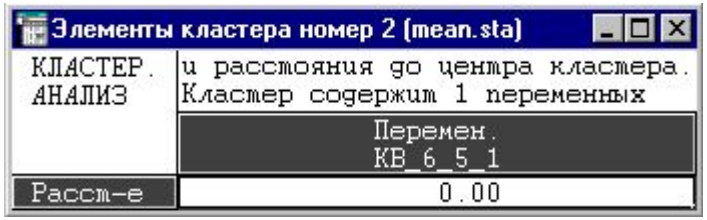
Результаты

Элементы кластера номер 1 (mean.sta)

КЛАСТЕР. АНАЛИЗ

u рассмояния до центра кластера.
Кластер содержит 8 переменных

	Перемен. KB_1_1_1	Перемен. KB_1_1_2	Перемен. KB_1_5_1	Перемен. KB_1_5_2	Перемен. KB_3_1_1	Перемен. KB_3_1_2	Перемен. KB_6_1_1	Перемен. KB_6_1_2
Рассм-е	1.971874	1.506634	1.089310	1.077315	1.598448	1.637242	1.385841	.954671



Элементы кластера номер 2 (mean.sta)

КЛАСТЕР. АНАЛИЗ

u рассмояния до центра кластера.
Кластер содержит 1 переменных

	Перемен. KB_6_5_1
Рассм-е	0.00

Элементы кластера номер 3 (mean.sta)

Далее...

u рассмояния до центра кластера.
Кластер содержит 6 переменных

	Перемен. KB_3_3_1	Перемен. KB_3_3_2	Перемен. KB_3_5_1	Перемен. KB_3_5_2	Перемен. KB_6_3_1	Перемен. KB_6_3_2
Рассм-е	1.058891	.824863	1.405683	.976158	.981679	1.529956

Результаты

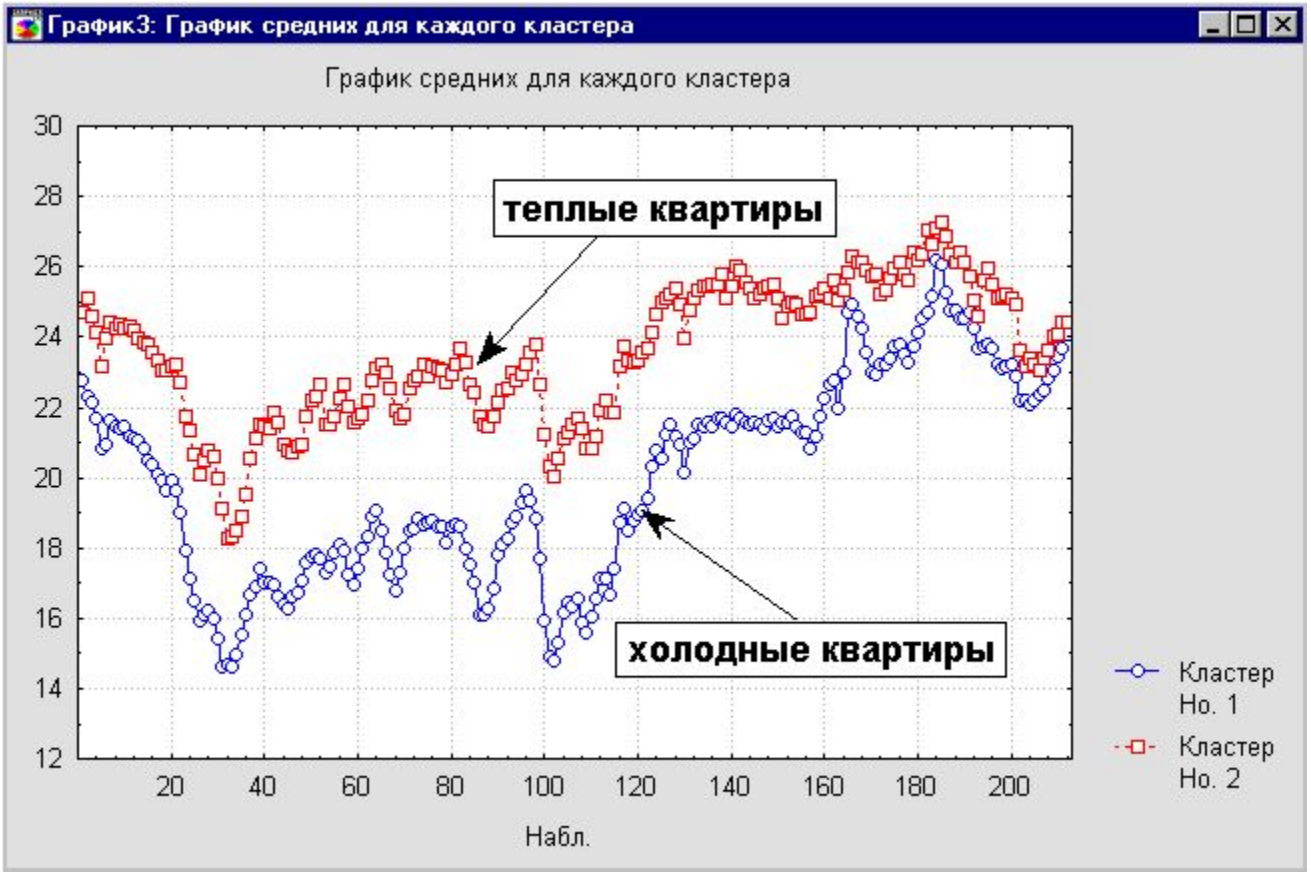
Средн. класт. (mean.sta)

КЛАСТЕР. АНАЛИЗ	Кластер No. 1	Кластер No. 2	Кластер No. 3
1	22.57700	24.37400	24.71400
2	22.07531	24.16000	25.10563
3	21.89677	23.90042	24.59660
4	21.43693	23.69958	24.11257
5	20.75703	21.26333	23.17507
6	21.18417	19.14208	23.94160
7	21.86151	19.70167	24.44437
8	21.70307	19.74208	24.24840
9	21.65292	19.36208	24.36618
10	21.69016	19.35708	24.24667
11	21.40417	19.14667	24.29604
12	21.42771	18.82333	24.17430
13	21.30474	18.87083	23.96694
14	21.12156	18.69917	23.83021
15	20.74688	18.43667	23.79194

Евклид. расстояния между кластерами (mean.sta)

КЛАСТЕР. АНАЛИЗ	Рассояния под диагональю Квадраты рассояний над диагональю		
Кластер номер	No. 1	No. 2	No. 3
No. 1	0.000000	14.66892	11.26418
No. 2	3.830003	0.00000	49.23798
No. 3	3.356215	7.01698	0.00000

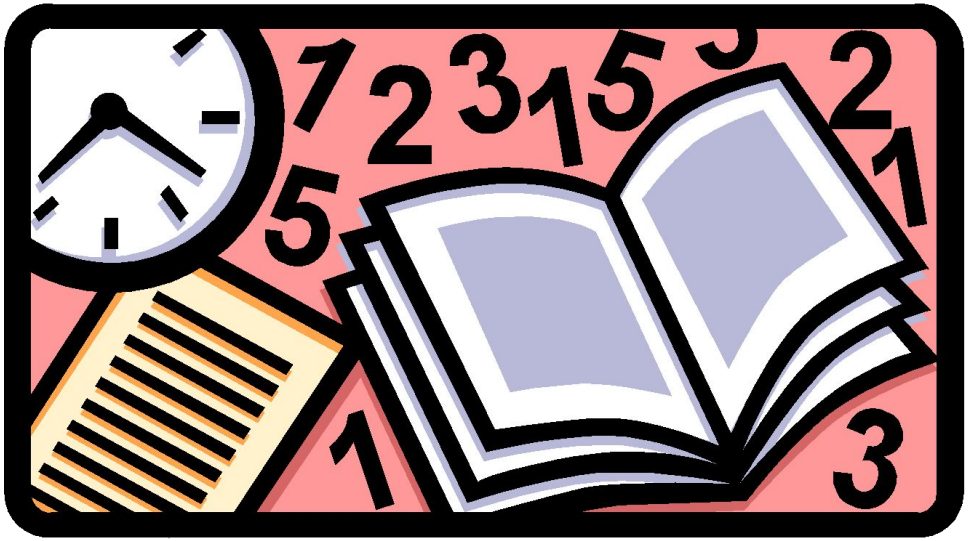
А если 2 кластера?



Выводы

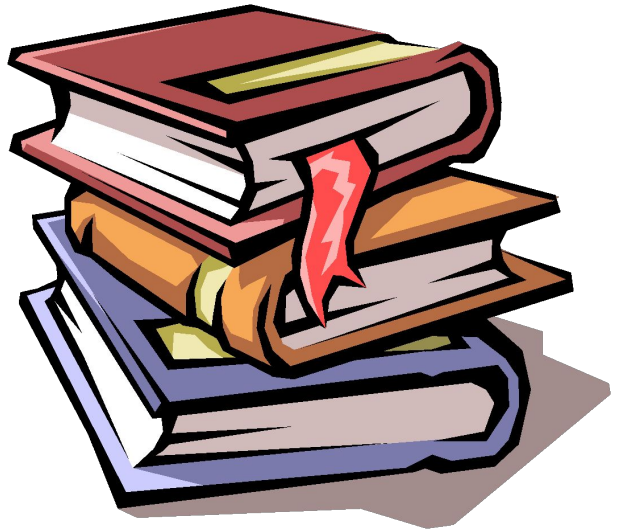
- Кластеризация выполнена
- Самые теплые квартиры расположены в середине дома
- Самые холодные квартиры - угловые

Сегментирование потребителей услуги "Подключение к сети Интернет"



Что необходимо?

- Наличие у фирмы минимум 100 клиентов
- Возможность проведения опроса клиентов фирмы
- Наличие STATISTICA



Этапы методики

- Определение возможных признаков сегментирования
- Проведение опроса
- Определение "пригодных" признаков сегментирования
- Выделение сегментов
- Формулировка сегментов

Признаки сегментирования

- Возраст
- Стаж работы в сети Интернет
- Профессиональная специализация
- Время работы в сети

Опросный лист

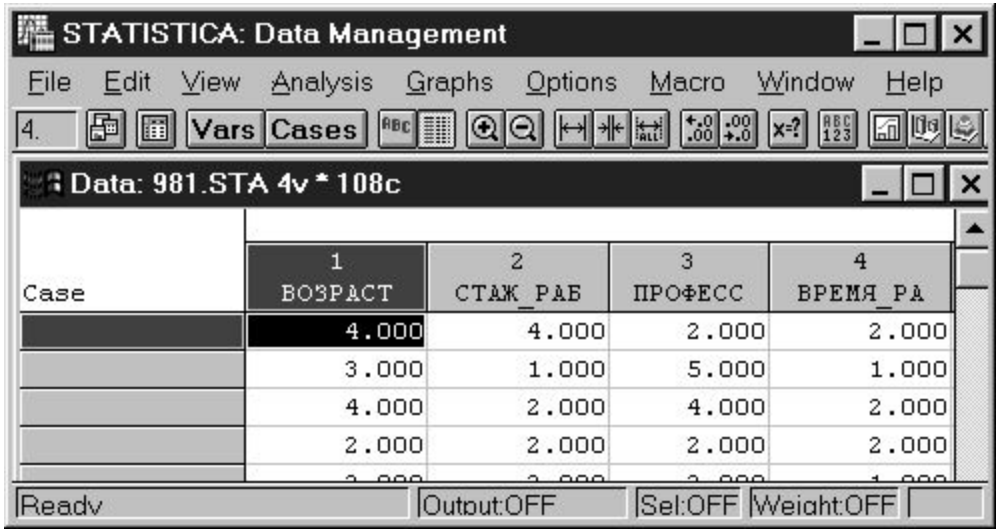
- Укажите, пожалуйста, Ваш возраст:
  - до 20 лет
  - 20-35
  - 35-55
  - старше 55
- Как давно Вы работаете в Интернет?
  - менее 1 года
  - 1-2 года
  - 2-3 года
  - более 3-х лет

Опросный лист

- В какой степени Вы используете возможности сети в вашей работе?
  - не использую
  - использую крайне редко
  - ежедневно обращаюсь к сети
  - интернет - часть моей работы
- Как часто Вы выходите в Интернет
  - несколько раз в месяц
  - несколько раз в неделю
  - один раз в день
  - несколько раз в день

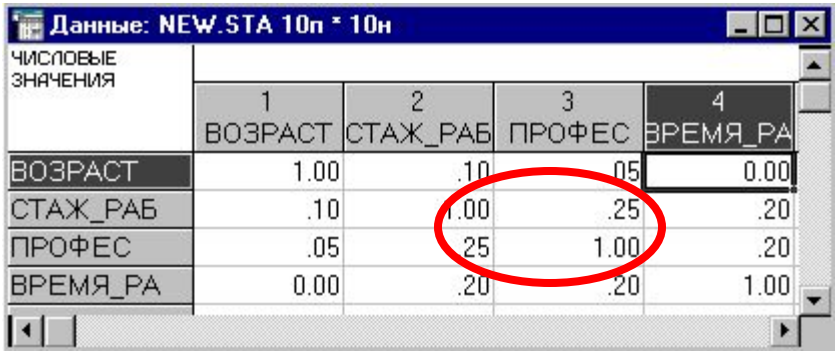
Ответы кодируются

- Не использую == 1
- Крайне редко == 2
- Ежедневно == 3
- Интернет - моя работа == 4

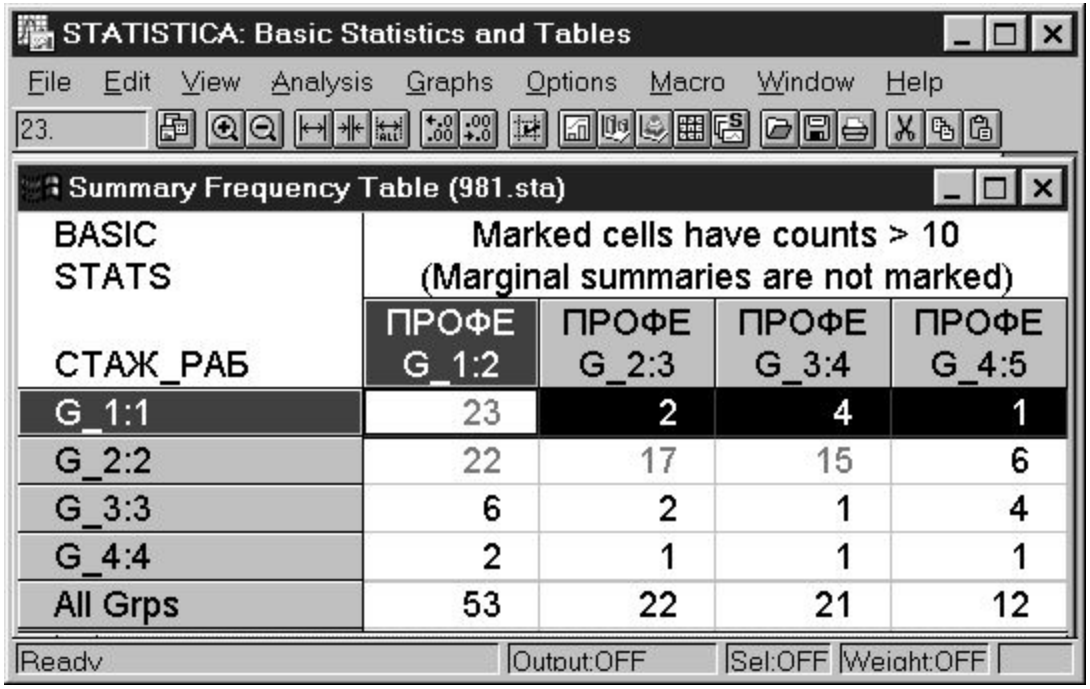


Выделение признаков

Степенью "пригодности" определенного признака сегментирования можно считать наличие определенной математической корреляции между парой предполагаемых признаков

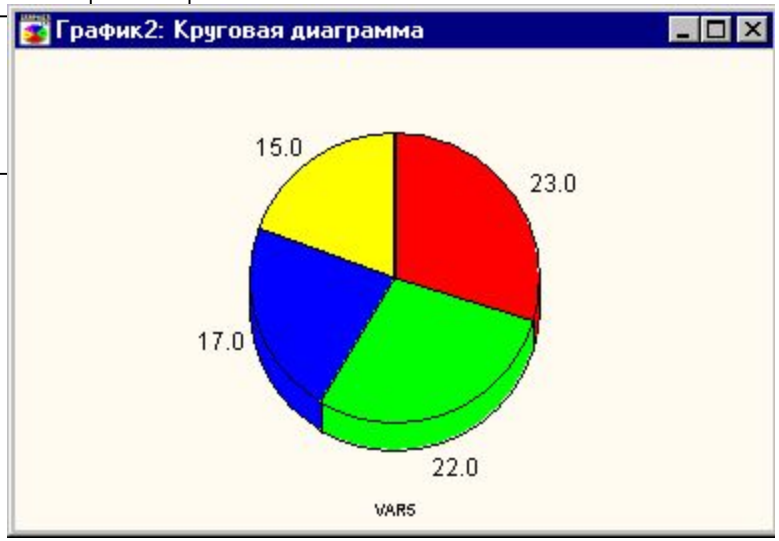


Выделение сегментов



Результат сегментации

Профессиональная специализация"	Стаж работы в сети"	%
Не использую	Менее 1 года	23
Не использую	1-2 года	22
Использую крайне редко	1-2 года	17
Ежедневно обращаюсь к сети	1-2 года	



Формулировка сегментов

Профессиональная специализация	Стаж работы в сети	"Прозвище сегмента"	Описание (характеристики сегмента)
Не использую	Менее 1 года	"Новичок"	Данный сегмент еще не успел основательно освоиться в среде Интернет, использует Интернет в бытовых целях, не применяет его в своей профессиональной деятельности
Не использую	1-2 года	"Бытовой Интернет"	Представители сегмента освоили сеть Интернет и легко в ней ориентируются, но их профессиональная деятельность либо лежит за пределами сети Интернет, либо они не знают возможностей использования Интернет в своей профессиональной деятельности
Использую крайне редко	1-2 года	"Справка на работе"	Представители сегмента представляют себе возможности Интернет и используют его на работе для получения справочной информации в редких случаях. Профессия чаще не связана с информационными потоками
Ежедневно обращаюсь к сети	1-2 года	"Интернет в работе"	Представители сегмента представляют себе возможности Интернет и активно используют его на работе для получения профессионально ориентированной информации

Кластерный анализ

Разбиение на три кластера покупателей продукции со склада

Данные

Сумма, перечисленная покупателем за месяц

Данные: NEW1.STA 10п * 24н

ЧИС ЗНА	1 ДАТА	2 ПОК1	3 ПОК2	4 ПОК3	5 ПОК4	6 ПОК5	7 ПОК6	8 ПОК7
3	3/1/99	3000.000	3250.000	1000.000	40.000	2522.400	50.000	90.000
4	4/1/99	4000.000	4250.000	3000.000	30.000	3353.200	80.000	150.000
5	5/1/99	5000.000	5250.000	4000.000	50.000	4204.000	130.000	250.000
6	6/1/99	5000.000	5250.000	30.000	80.000	4484.000	410.000	810.000
7	7/1/99	1000.000	1250.000	50.000	130.000	1090.800	220.000	430.000
8	8/1/99	3000.000	3250.000	80.000	30.000	2532.400	30.000	110.000
9	9/1/99	4000.000	4250.000	130.000	50.000	3303.200	50.000	50.000
10	10/1/99	5000.000	5250.000	5105.000	80.000	4124.000	80.000	90.000
11	11/1/99	1000.000	5250.000	5105.000	130.000	4154.000	130.000	150.000
12	12/1/99	3000.000	5000.000	1101.000	120.000	1000.800	130.000	250.000
13	1/1/00	4000.000	1000.000	3103.000	400.000	2882.400	410.000	30.000
14	2/1/00	5000.000	3000.000	4104.000	210.000	3493.200	220.000	50.000

Анализ аналогичный