Кластерный анализ

Понятие кластерного анализа

фото

- Трион, 1939 год – появление кластерного анализа
- Кластерный анализ – совокупность различных алгоритмов классификации
- Ключевой вопрос –организация наблюдаемых данных в наглядные структуры (таксономии)

фото

- Отсутствие процедуры проверки статической значимости

Области применения и методы

Области:

- медицина
- психиатрия
- арехеология
- менеджмент

Методы:

- древовидная кластеризация
- двувходовое объединение
- метод K средних

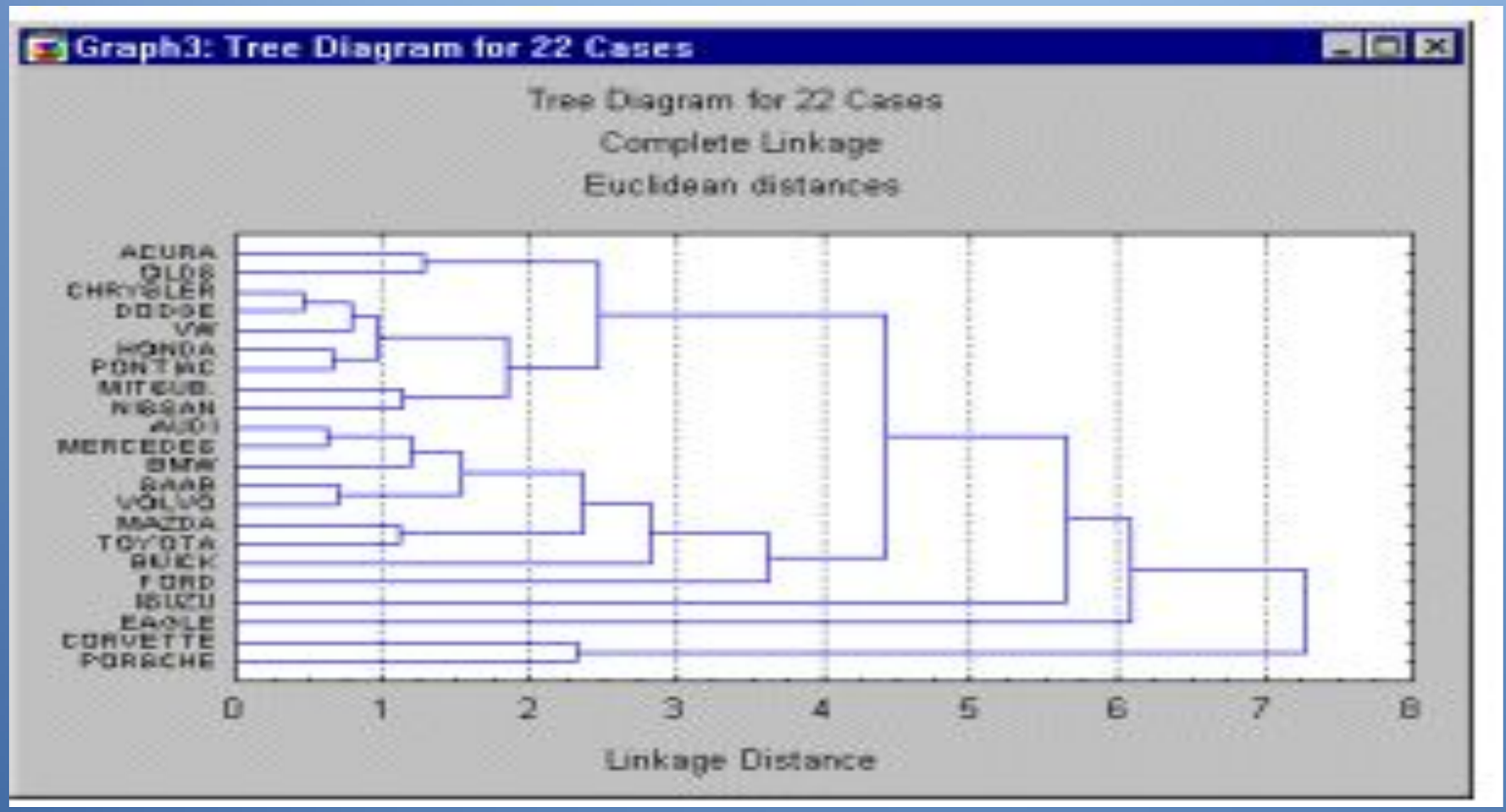
Этапы кластерного анализа

1. Отбор **выборки** для кластеризации
2. Определение **множества** переменных
3. Вычисление **значений** той или иной меры сходства между объектами
4. Применение **метода** кластерного анализа
5. Проверка **достоверности** результатов

Древовидная кластеризация

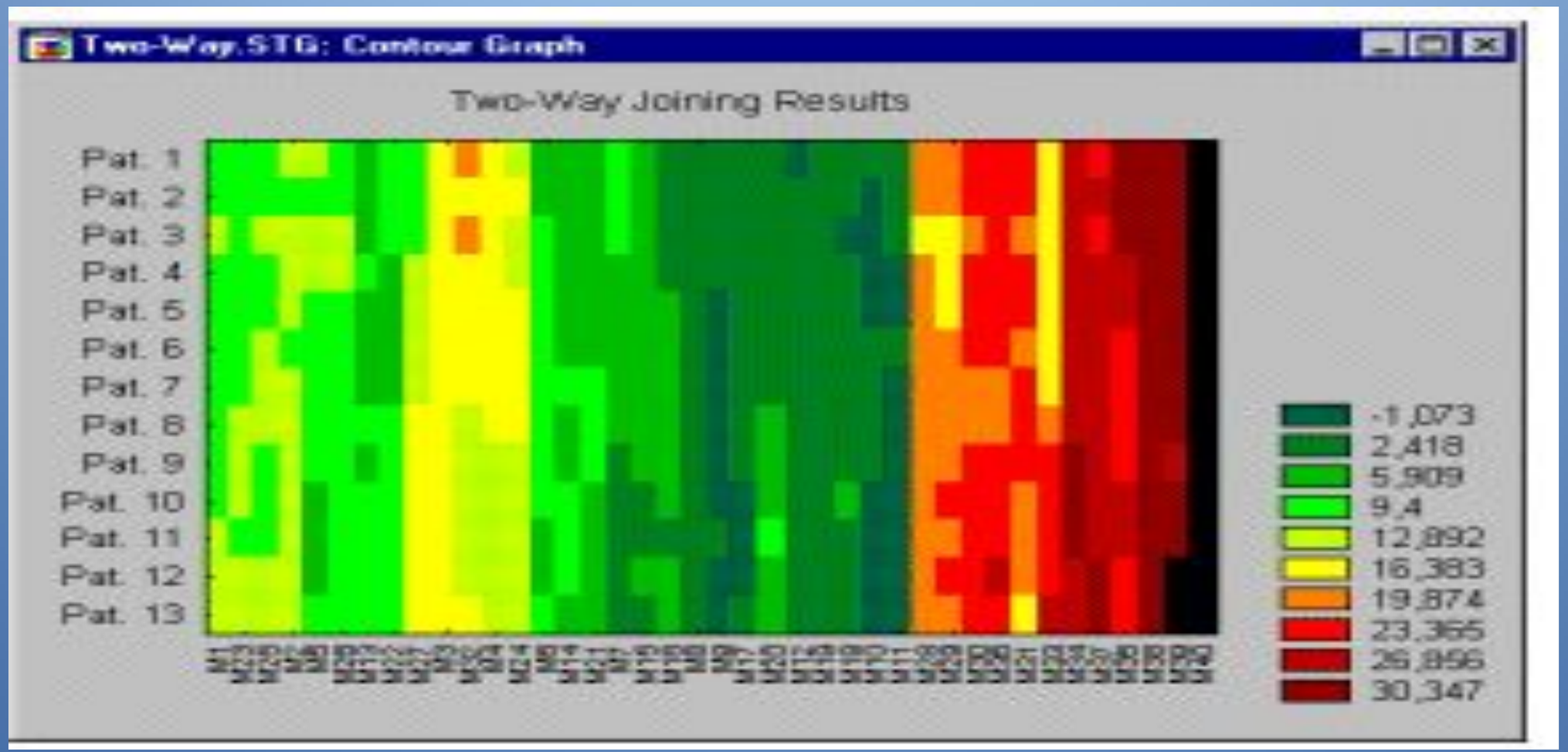
Использование меры сходства и расстояния между анализируемыми объектами

Типичный результат – **иерархическое дерево**

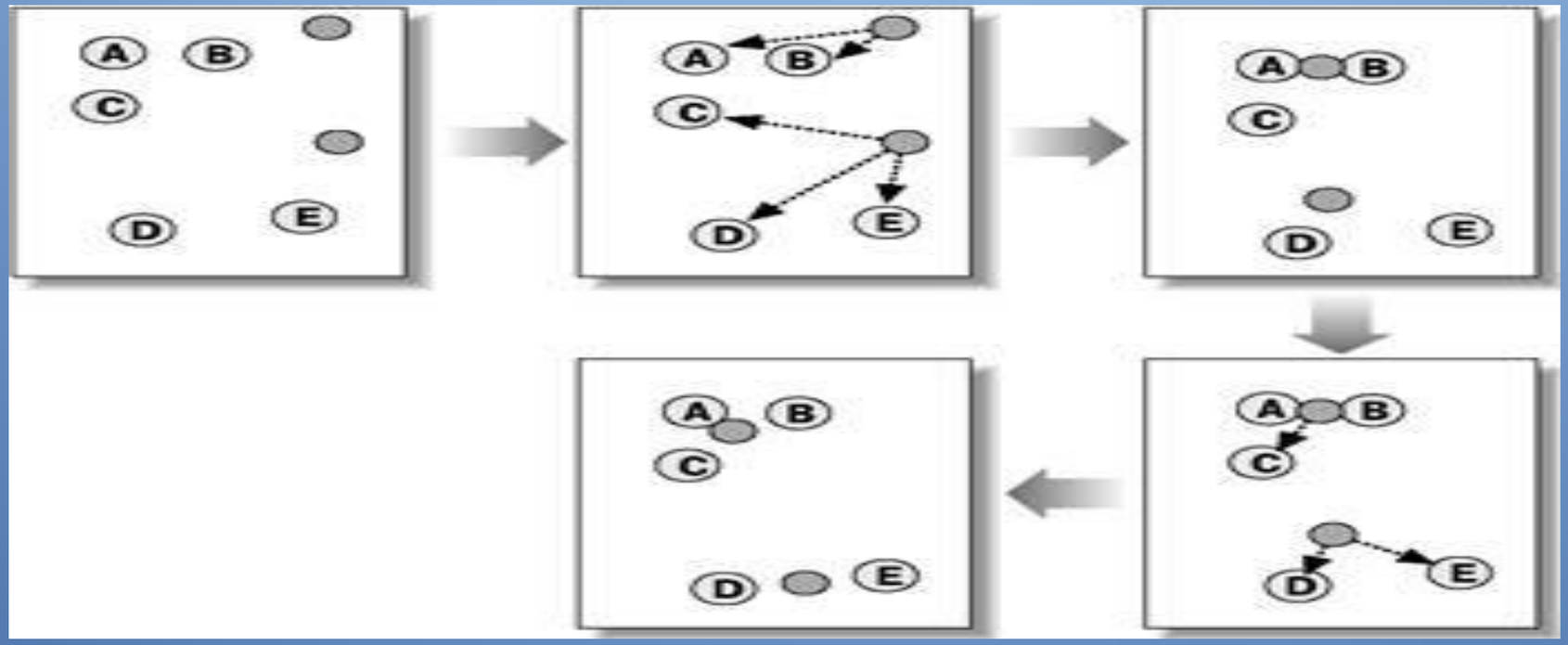


Двувходовое объединение

Наблюдения и **переменные** одновременно вносят вклад в обнаружение осмысленных кластеров



Метод K средних строит ровно K различных кластеров, расположенных на возможно больших расстояниях друг от друга



КЛАСТЕРНЫЙ ПОДХОД К АНАЛИЗУ УРОВНЯ УДОВЛЕТВОРЁННОСТИ ПЕРСОНАЛА МЕДИЦИНСКОГО УЧРЕЖДЕНИЯ

Сбор данных

- Метод анкетирования
- 165 респондентов
- Изучение удовлетворенности персонала текущими процессами деятельности медицинского центра

фото

Объект кластеризации –
персонал медицинского центра

Признаки кластеризации:

- доступность и качество информации
- фото
- корпоративная культура
- мотивация

Модель расчета

- Был рассчитан **индекс удовлетворенности** по каждому сотруднику Yi по каждой категории вопросов:

$$Yi = \frac{\sum_{i=1}^{n} Xi}{\sum_{i=1}^{n} maxXi}$$, где

i – номер респондента,

фото

$\sum_{i=1}^{n} \boldsymbol{Xi}$ - сумма ответов по каждой категории вопросов для каждого служащего,

$\sum_{i=1}^{n} \boldsymbol{maxXi}$ - максимально возможное значение суммы по категории вопросов

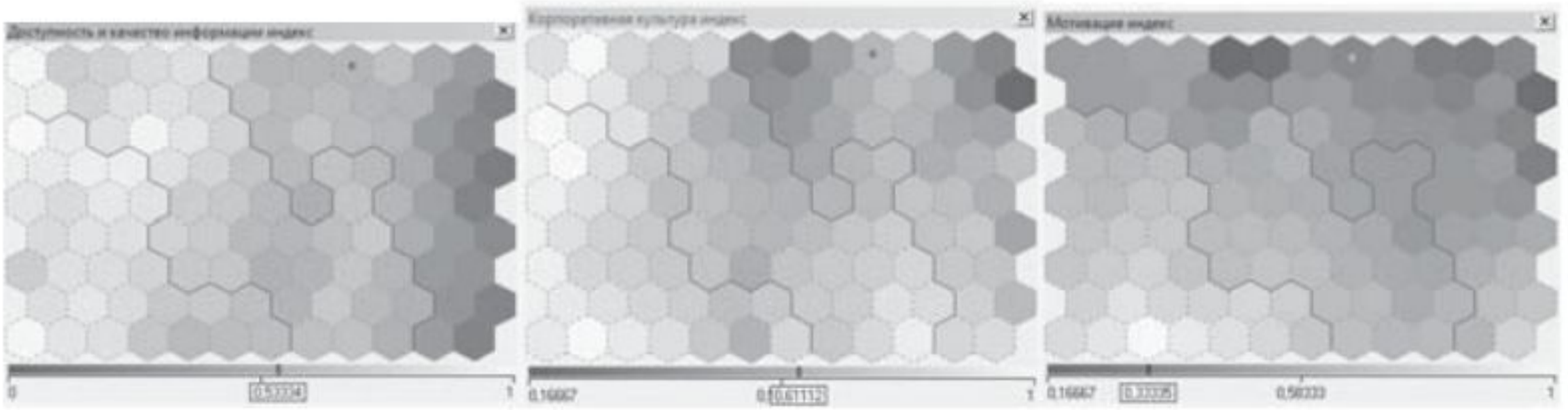
Кластерный анализ

Метод сетей Кохонена
Метод k-средних

Аналитическая платформа Deductor Academic 5.2

Сформировано 3 кластера:
фото
кластер 1 – **высокая** степень удовлетворенности,
кластер 2 – **средняя** степень удовлетворенности,
кластер 3 – **низкая** степень удовлетворенности.

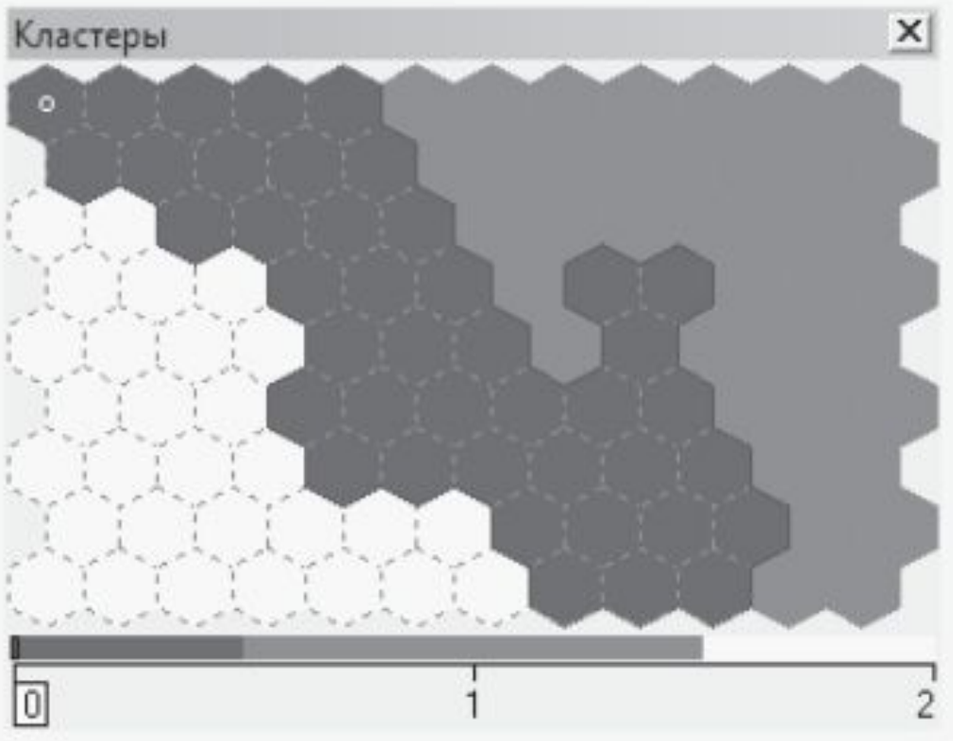
Кластерный анализ



а) б) в)

Рис. 1. Карта Кохонена для индексов удовлетворенности признаков:
а) доступность и качество информации
б) корпоративная культура
в) мотивация

Кластерный анализ



	Кластер 1	Кластер 2	Кластер 3
Количество объектов	44	67	54
Средний возраст	49 лет	43 года	44 года
Средний стаж работы в институте	14 лет	13 лет	13 лет
Средний стаж работы в отрасли	21 год	16 лет	18 лет
Распределение сотрудников по категориям, %:			
1) администрация;	1) 21%	1) 10%	1) 6%
2) врачи;	2) 34%	2) 46%	2) 39%
3) медицинский персонал;	3) 25%	3) 18%	3) 41%
4) сервисный персонал.	4) 21%	4) 25%	4) 15%
ИУ* по категории «Доступность и качество информации»	0,78	0,67	0,4
ИУ по категории «Корпоративная культура»	0,81	0,73	0,58
ИУ по категории «Мотивация»	0,76	0,56	0,46

Рис. 2. Карта Кохонена: разделение по кластерам

Табл. 1. Характеристика полученных кластеров

Применение методов кластерного анализа для обработки данных психологических исследований

.

Сбор данных и объект кластеризации

Экспертные оценки

9 респондентов

Исследование **структуры** команды (малой группы, ориентированной на решение деловой задачи и состоящей из молодых специалистов (инженеров-программистов), коллективно принимающих решение, выполняющих сложные работы в различном составе и качественном описании характеристик каждой подгруппы

фото

Сбор данных и объект кластеризации

№	Завис. от групп. стандарт.	Ответст.	Труд. актив.	Работо-способн.	Понимание цели	Мотивация
1	2.0	7.0	9.0	8.0	10.0	3.0
2	4.0	2.0	8.0	8.0	8.0	1.0
3	2.0	3.0	9.0	7.0	8.0	1.0
4	7.0	3.0	5.0	6.0	4.0	0.0
5	2.0	2.0	5.0	3.0	7.0	2.0
6	4.0	3.0	5.0	5.0	5.0	2.0
7	5.0	4.0	4.0	5.0	5.0	3.0
8	6.0	1.0	4.0	4.0	7.0	0.0
9	5.0	3.0	3.0	5.0	4.0	2.0

Матрица смешения для коллектива из 9 человек

Матрица расстояний, полученная с использованием метрики Евклида

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
C_1	0.00	6.16	5.00	10.3	8.72	8.43	8.77	10.5	10.3
C_2	6.20	0.00	2.65	6.30	6.32	5.39	6.56	6.20	7.30
C_3	5.00	2.65	0.00	7.70	5.92	5.83	7.21	7.50	8.10
C_4	10.3	6.32	7.68	0.00	6.93	3.87	4.12	4.40	3.60
C_5	8.70	6.32	5.92	6.90	0.00	3.61	4.80	4.80	5.20
C_6	8.40	5.39	5.83	3.90	3.61	0.00	2.00	4.20	2.40
C_7	8.80	6.56	7.21	4.10	4.80	2.00	0.00	4.90	2.00
C_8	10.5	6.24	7.48	4.40	4.80	4.24	4.90	0.00	4.50
C_9	10.3	7.28	8.12	3.60	5.20	2.45	2.00	4.50	0.00

Дерево классификации

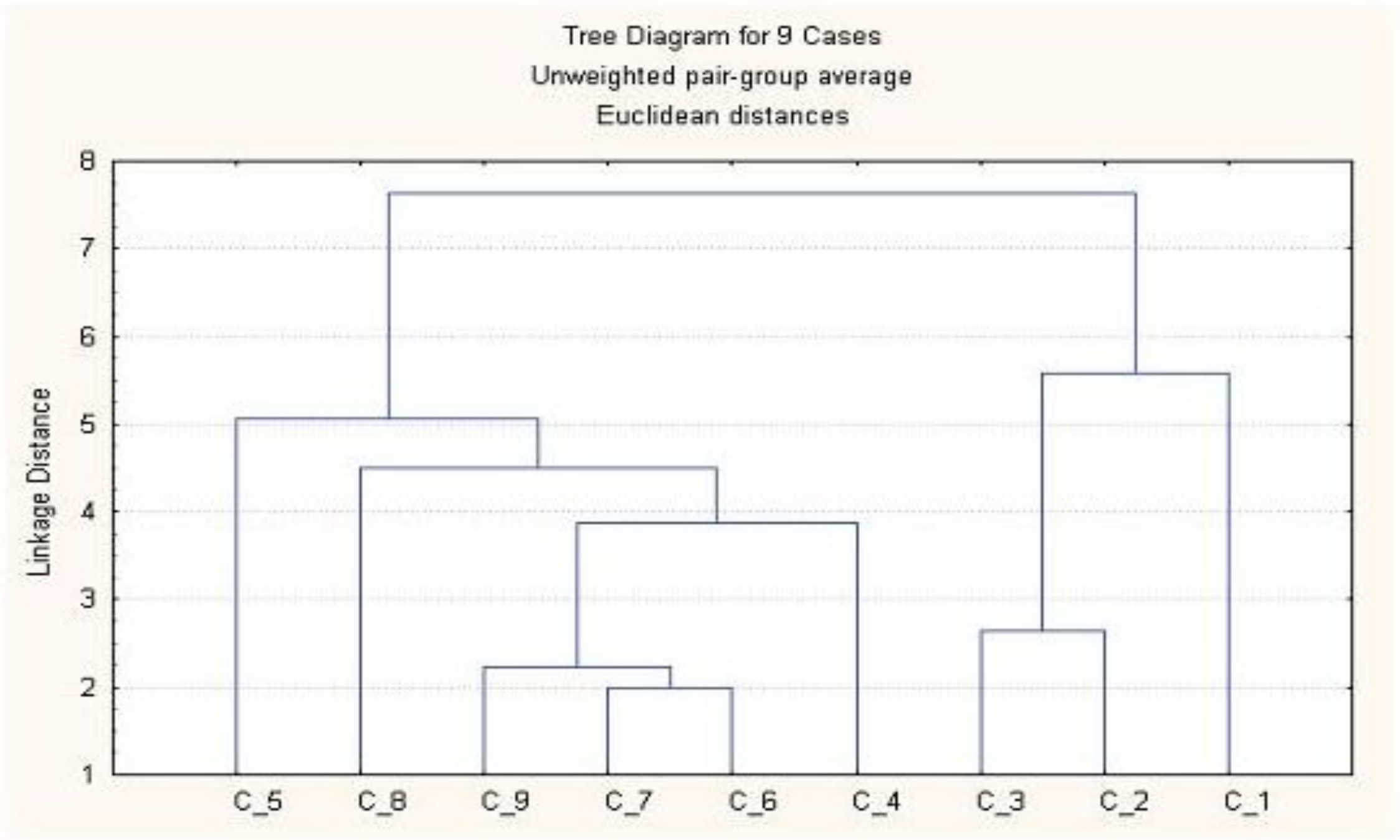
Для определения «естественного» числа кластеров, на которые может быть разбита совокупность объектов применялся следующий **критерий**: на каждом уровне иерархической кластеризации выполнялось разбиение множества на данное число классов. Для каждой пары кластеров оценивалась **отношение** среднего внутрикластерного расстояния к межкластерному:

$$\pi = \frac{a_i + a_j}{2b_{ij}}$$

Оценка «естественного» разбиения производится по формуле:

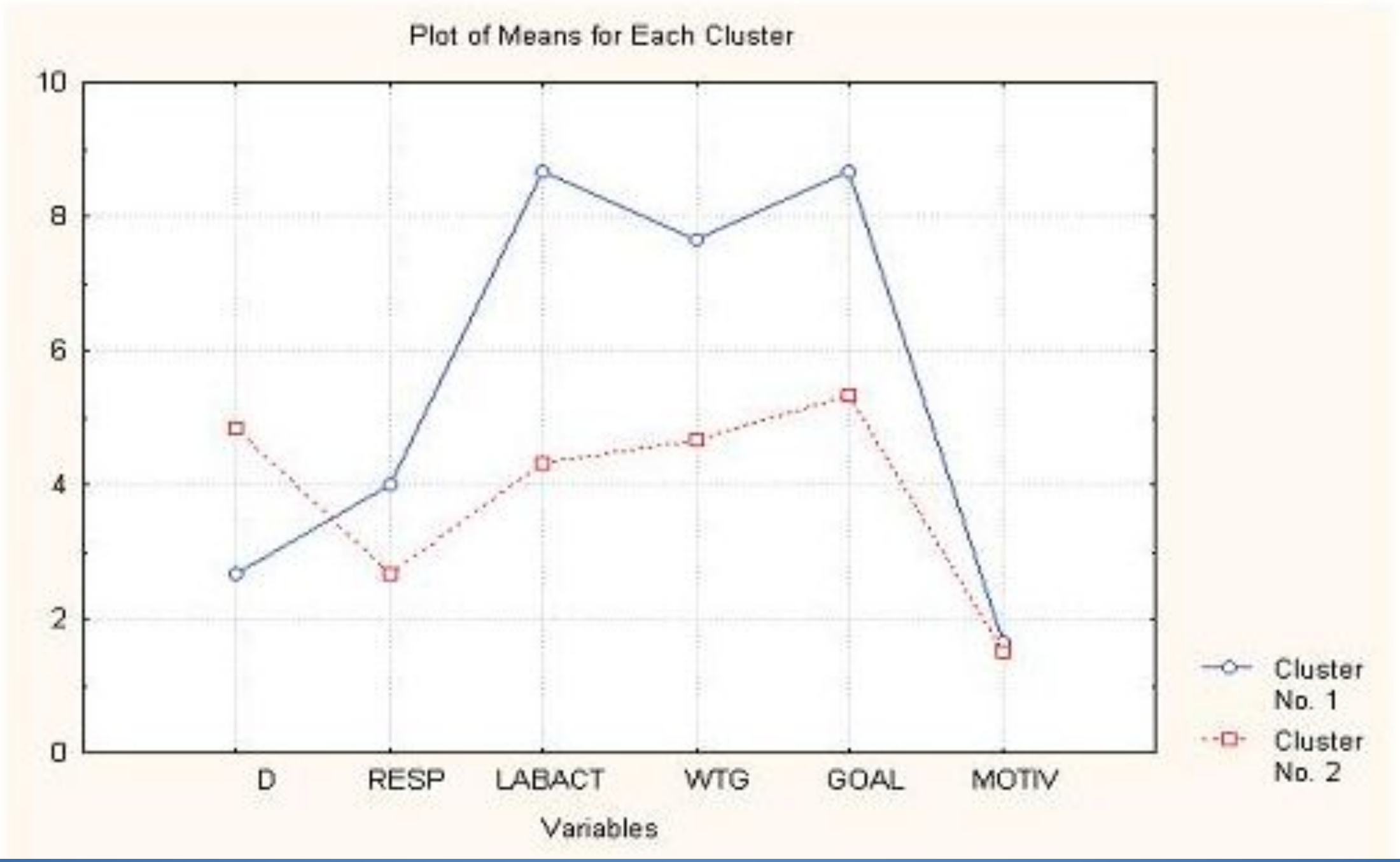
$$S = \frac{1}{k}\sum_{i=1}^{\tilde{g}} \max_j \pi_{ij}$$

Дерево классификации



При помощи **метода к-среднего** реализуется процедура построения **усредненных профилей** каждого класса, что дает возможность проводить качественный **анализ выраженности признаков** у представителей каждого класса.

Усредненные профили классов



Результаты сравнительного анализа, демонстрирующие значимые отличия классов по трем характеристикам:

фото **трудовая активность, работоспособность и понимание цели**

Результаты

Номера объектов первого класса

	1	3	2
Distance	1.484488	1.097134	0.693889

Номера объектов второго класса

	8	4	9	7	6	5
Distance	1.357421	1.566430	0.535758	0.855267	1.272938	0.822147

фото

	Межкл. расст.	Внутрикл. расст.	Степ.свободы	F	p
D	9.38	17.50	7	3.75	0.093821
RESP	3.55	19.33	7	1.28	0.293890
LABACT	**37.55**	**4.00**	**7**	**65.72**	**0.000084**
WTG	**18.00**	**6.00**	**7**	**21.00**	**0.002536**
GOAL	**22.22**	**12.00**	**7**	**12.96**	**0.008735**
MOTIV	0.055	10.16	7	0.038	0.850495

Кластерный анализ рынка модельной обуви города Красноярска

Цели и задачи

- **Сегментировать** целевую аудиторию
- **Составить портрет** потребителя: рациональность и эмоции
- **Выровнять** карту восприятия товара продавцом и потребителем

фото

- **Оптимизировать** рекламные сообщения

ЗАО «Ионесси»

- Российский производитель обуви
- **Проблема:** отрицательное отношение к обуви российского производства

- **Окружение:** высоко конкурентная среда

фото

- **Задача:** провести поведенческий анализ аудитории

Выбор признаков сегментирования

Поведенческие: отношение к продукции предприятия

фото

Социально-демографические: пол, возраст, уровень дохода

Анкета: поведенческие признаки

1.	**Насколько Вы знакомы с деятельностью компании «Ионесси»?**	
	1. Я никогда не слышал об этой компании.	3,3%
	2. Я знаю название этой компании, и ничего более.	7,4%
	3. Я знаю, что эта компания занимается производством обуви.	89,3%
2.	**Вы знакомы с продукцией, производимой компанией «Ионесси»?**	
	1. Нет, я никогда не видел продукцию, произведенную компанией.	37,3%
	2. Да, я видел продукцию, произведенную компанией, но никогда не приобретал.	38%.
	3. Да, я приобретал продукцию, произведенную компанией.	9,9%
	4. Да, я сейчас имею в своем гардеробе продукцию, произведенную компанией.	4,1%
3.	**Как одним словом Вы охарактеризуете свое отношение к продукции компании «Ионесси»?**	
	1. Отрицательное.	25,6%
	2. Безразличное.	45,5%
	3. Благожелательное.	18,2%

4. Насколько Вы оцениваете качество продукции компании «Ионесси»?

Шкала: -2 = очень низкое качество, +2 = очень высокое качество.

-2	-1	0	+1	+2
3,3%	17,4%	43%	19,8%	5,8%

5. Насколько Вы оцениваете продукцию компании «Ионесси», с точки зрения соотношения цена/качество?

Шкала: -2 = цена не соответствует качеству, +2 = цена соответствует качеству.

-2	-1	0	+1	+2
1,7%	14,9%	42,2%	22,3%	8,2%

6. Насколько Вы оцениваете ассортимент продукции компании «Ионесси»?

Шкала: -2 = слишком узкий ассортимент, +2 = достаточно широкий и разнообразный ассортимент.

-2	-1	0	+1	+2
8,3%	28,1%	34,7%	14,9%	3,3%

Анкета: соц-дем

9.	**Пол**	
	1. Женщина	52,9%
	2. Мужчина	47,1%
10.	**Возраст**	
	1. 18-35 лет	33,9%
	2. 36-55 лет	44,6%
	3. Более 56 лет	21,5%
11.	**Размер семьи**	
	1. Один человек	9,9%
	2. Два человека	15,7%
	3. 3-4 человека	65,3%
	4. 5 человек и более	9,1%
12.	**Ежемесячный доход на семью**	
	1. До 3000 рублей	11,5%
	2. 3000-5999 рублей	37,2%
	3. 6000-9999 рублей	38%
	4. 10000-20000 рублей	8,3%
	5. Более 20000 рублей	5%

Модель измерения отношения к продукции

$$A_o^j = \mathring{a} в_i^j \times e_i^j$$

гA_o^j – отношение респондента j к продукции ЗАО «Ионесси»

$в_i^j$ –сила мнения респондента j, что продукция ЗАО «Ионесси» имеет характеристику i

e_i^j – оценка значимости характеристики i для респондента

фото

i = 1, …, n, n – число значимых характеристик
j = 1, …, m, m – количество респондентов

Характеристики продукции

1. Высокое качество
2. Соответствие цены качеству
3. Достаточно широкий и разнообразный ассортимент
4. Соответствие тенденциям современной моды

Отношение	Интервал	Количество человек
Отрицательное	[-16; -5]	21 чел.
Безразличное	(-5; +5)	51 чел.
Благожелательное	[+5; +16]	36 чел.

Определение уровня дохода

$\mathbf{d^j = D^j / k}$

$\mathbf{d^j}$ – ежемесячный доход на одного человека семьи респондента j

$\mathbf{D^j}$ – ежемесячный доход на семью респондента j

k – размер семьи респондента j

j = 1, …, m, m – количество респондентов

Если d^j £ 1500 руб. Þ - **низким**

Если 1500 руб. < d^j £ 5000 руб. Þ – **средний**

Если d^j > 5000 руб. Þ – **высокий**

фото

Низкий	45 чел.
Средний	67 чел.
Высокий	9 чел.

Сегменты

Сегмент I — «отрицательно настроенные» потребители. 30% респондентов. 83% женщины. Средний возраст женщин – 32, мужчин – 34 года.

Сегмент II — «безразличные» потребители. 53% респондентов. Почти поровну мужчин и женщин (52% и 48%). 66% - от 36 до 55 лет.

Сегмент III — «благожелательные» потребители. 17% респондентов. 67% мужчин. Самый старший сегмент: 91% старше 35 лет.

СПАСИБО ЗА ВНИМАНИЕ