НАФИ

ТЕМА 9

КЛАСТЕРНЫЙ АНАЛИЗ

ОГЛАВЛЕНИЕ

1. КЛАСТЕРНЫЙ АНАЛИЗ: ПОНЯТИЕ И НАЗНАЧЕНИЕ

1. Кластерный анализ: понятие и назначение

ЧТО ТАКОЕ КЛАСТЕРНЫЙ АНАЛИЗ?

1. **Кластерный анализ** предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

2. **Кластерный анализ** – группа методов, используемых для классификации объектов или событий в относительно гомогенные (однородные) группы, которые называют кластерами (clusters).

1. Кластерный анализ: понятие и назначение

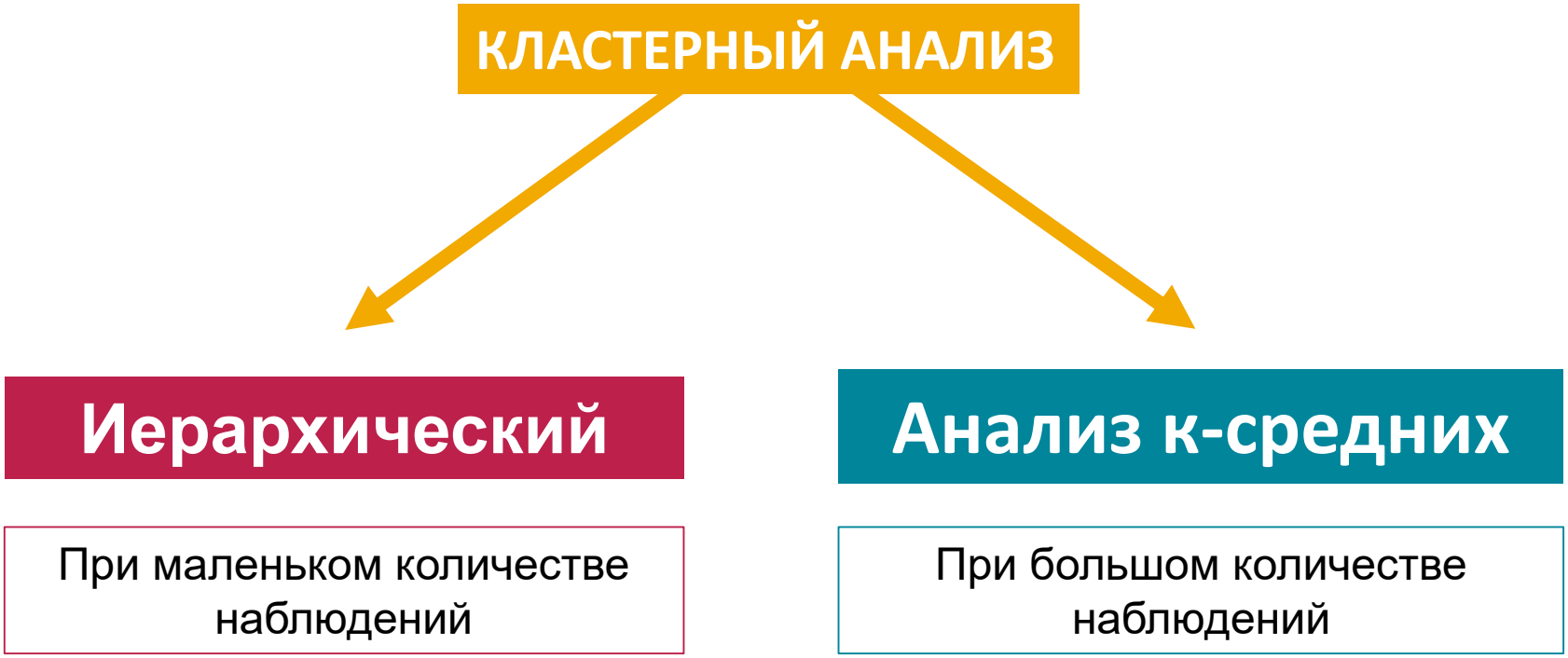
- В **факторном анализе** группируются столбцы, т.е. цель – анализ структуры множества признаков и выявление обобщенных факторов.
- В **кластерном анализе** – группируются строки, т.е. цель – анализ структуры множества объектов.
- Кластерный анализ выполняет **классификацию объектов.**
- Каждый объект (респондент) – точка в пространстве признаков.
- **Задача кластерного анализа** – выделение «сгущений» точек, разбиение совокупности на однородные подмножества объектов (сегментация).

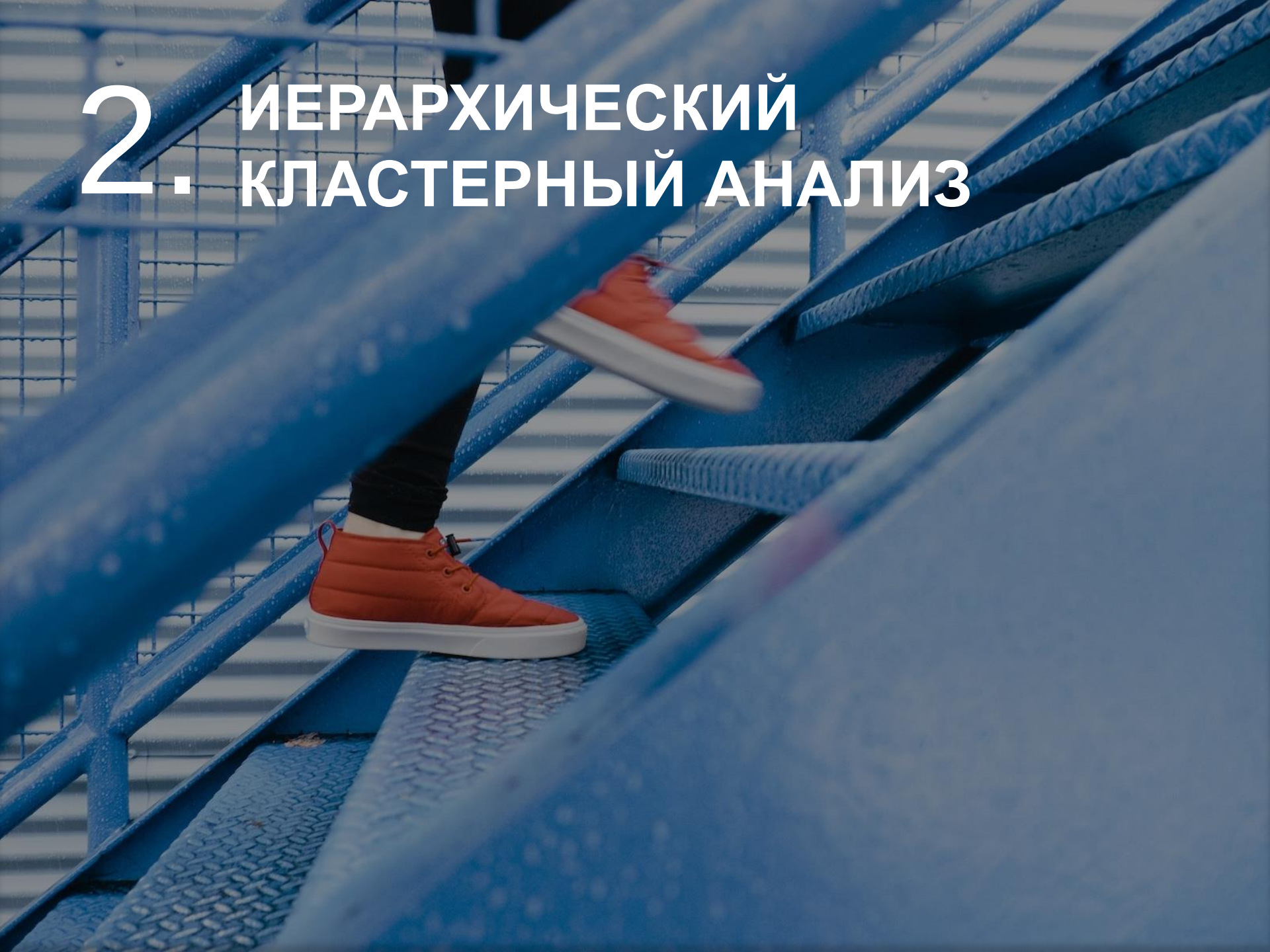
Кластерный анализ в теории

Кластерный анализ на практике

Как очертить границу кластеров? Сколько их следует выделить?

1. Кластерный анализ: понятие и назначение





2. ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ

2. Иерархический кластерный анализ

- Каждое наблюдение образовывает сначала свой **отдельный кластер.**
- На первом шаге анализа **два соседних кластера объединяются в один**.
- Этот процесс продолжается до тех пор, пока не останутся **только два кластера.**
- Расстояние между кластерами является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров (Between-groups linkage (Связь между группами))

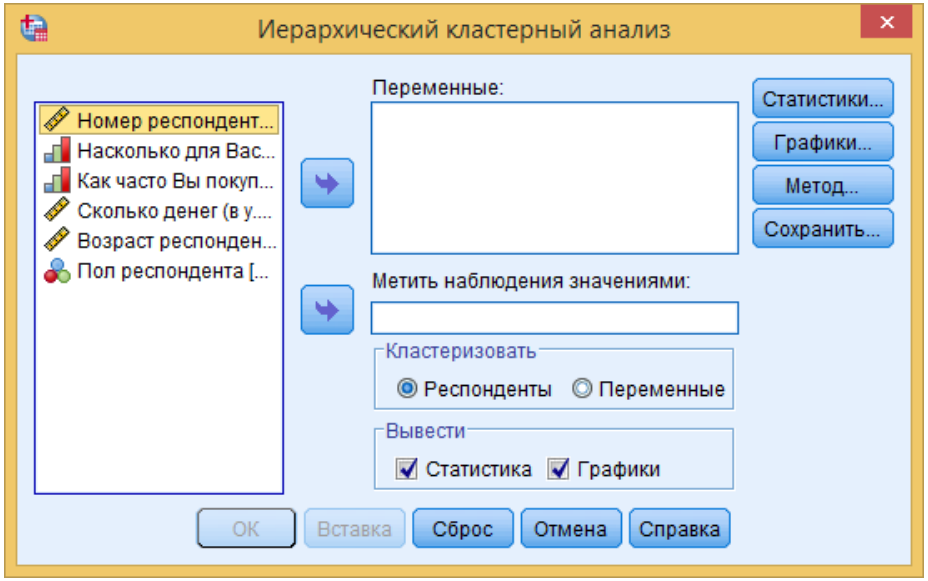
2.1 ЭТАПЫ КЛАСТЕРНОГО АНАЛИЗА

2.1. Этапы кластерного анализа

2.2 ВЫБОР СПОСОБА ИЗМЕРЕНИЯ РАССТОЯНИЯ

2.2. Выбор способа измерения расстояния

1. Команда **«Анализ»** → **«Классификация»** → **«Иерархический кластерный анализ»**



Выбор меры сходства объектов зависит от типа переменной и шкалы, к которой она относится.

Для каждого типа данных существует несколько способов измерения расстояния или определения меры сходства объектов. Наиболее используемыми для интервальных данных являются:

- Евклидово расстояние (Euclidian Distance)
- Квадрат Евклидова расстояния (Squared Euclidian distance)

2.2. Выбор способа измерения расстояния

Самой распространенной мерой для определения расстояния между двумя точками на плоскости, образованной координатными осями x и y, является **Евклидова мера**:

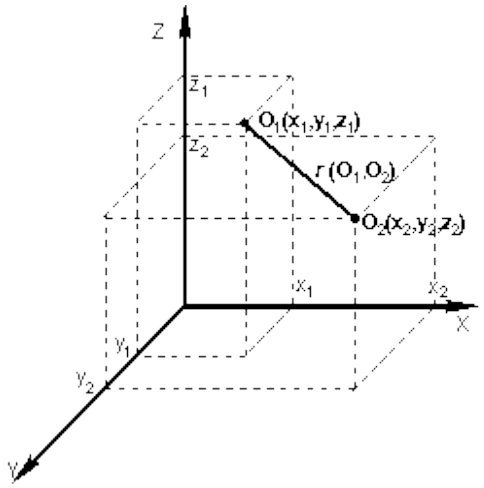
Квадрат евклидова расстояния (Squared Euclidian distance)

$$d(X,Y)=\sum_{i=1}^{m}(X_i-Y_i)^2$$

На расстояния могут сильно влиять различия между осями, по координатам которых вычисляются эти расстояния. Например, если одна из осей измерена в сантиметрах, а потом переведена в миллиметры, то окончательное евклидово расстояние (или квадрат евклидова расстояния), вычисляемое по координатам, сильно изменится, и, как следствие, результаты кластерного анализа могут сильно отличаться от предыдущих.

2.2. Выбор способа измерения расстояния

Благодаря возведению в квадрат лучше учитываются большие разности. Эта мера должна всегда использоваться при построении кластеров центройдным, медианным методом или методом Варда (Уорда).



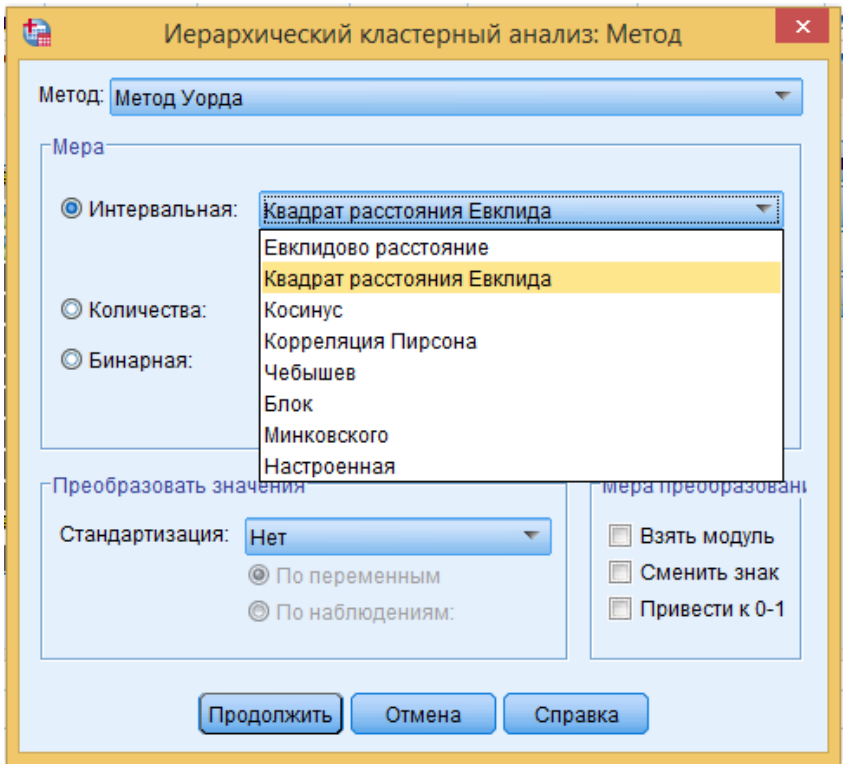
$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

2.2. Выбор способа измерения расстояния

Меры близости между объектами (меры подобия)

Для количественных (интервальных или метрических) переменных:

- **Евклидово расстояние** – это наименьшее расстояние между x и y. В двух- или трёхмерном случае – это прямая, соединяющая данные точки.
- **Квадрат евклидового расстояния** – устанавливается по умолчанию.
- **Корреляция Пирсона** – применима, если кластеризация наблюдений осуществляется только на основании двух переменных.
- **Блок (Block)** – это дистанционная мера, называемая также расстоянием Манхэттена или Хемминга, определяется суммой абсолютных разностей пар значений.



Меры близости между объектами (меры подобия)

Также для количественных (интервальных или метрических) переменных:

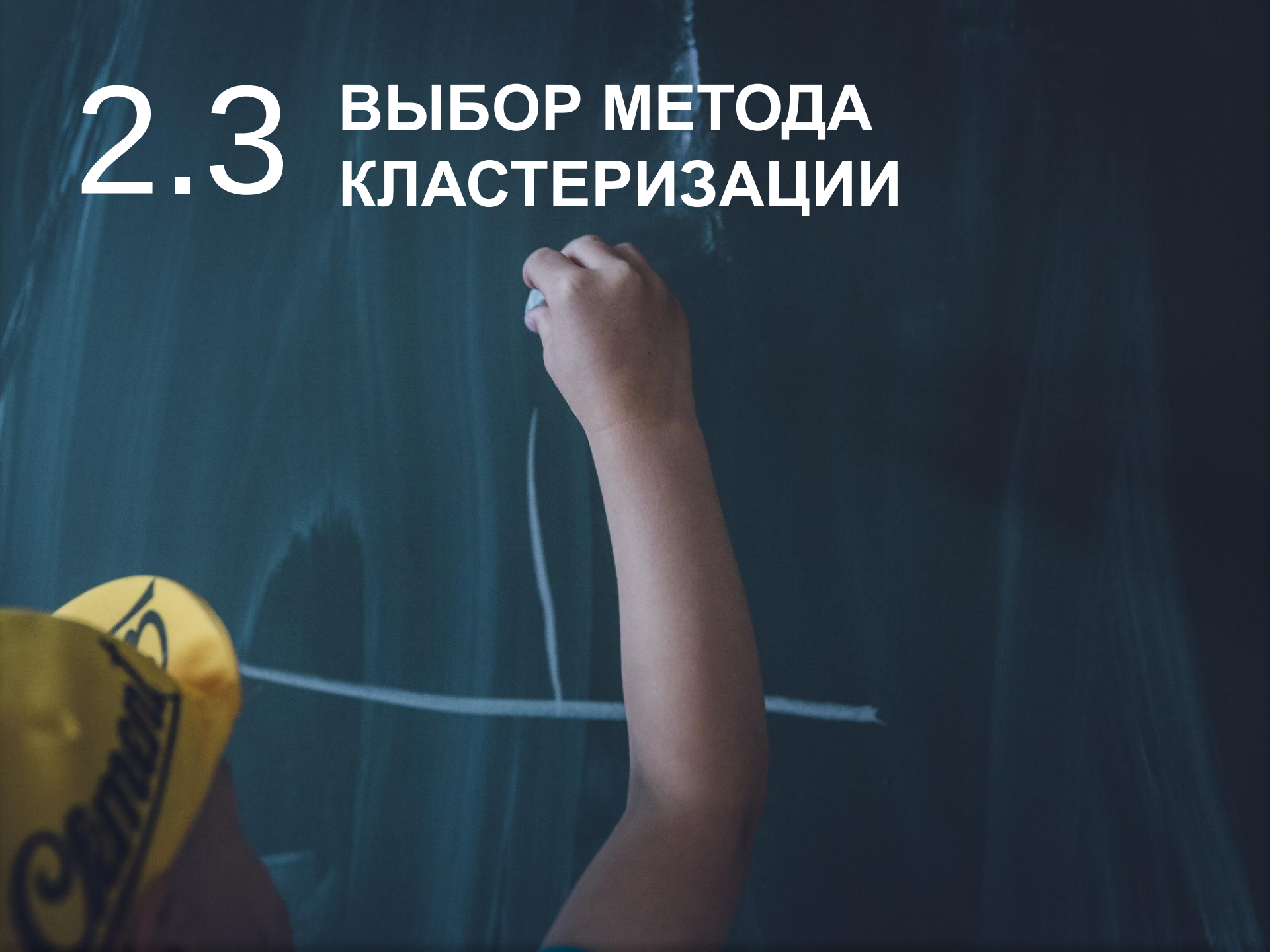
- **Расстояние Чебышева (Chebychev)** – вычисление расстояния как максимума абсолютного значения разности между элементами. Используется при определении двух объектов как "различные", если они отличаются по какому-то одному измерению.
- **Расстояние Минковского (Minkowski) –** равно корню r-ой степени из суммы абсолютных разностей пар значений взятых в r-ой степени. В SPSS при расчете этого расстояния допускается применение только квадратного корня, в то время как степень разности значений можно выбрать в пределах от 1 до 4. Если эту степень взять равной 2, то получим евклидово расстояние.

Отдельно выделяются меры близости для переменных, значения которых отображают частоты, и для бинарных переменных, которые указывают на факт осуществления события. В файле данных это должно быть закодировано при помощи двух численных значений (в SPSS – 0 и 1).

2.2. Выбор способа измерения расстояния

Меры близости между объектами (меры подобия)

Показатели	*Формулы*
Для количественных шкал	
Линейное расстояние	$d_{1ij} = \sum_{l=1}^{m} \lvert x_i^l - x_j^l \rvert$
Евклидово расстояние	$d_{Eij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^2 \right)^{1/2}$
Квадрат евклидово расстояние	$d_{Eij}^2 = \sum_{l=1}^{m} (x_i^l - x_j^l)^2$
Обобщенное степенное расстояние Минковского	$d_{Pij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^P \right)^{1/P}$
Расстояние Чебышева	$d_{ij} = \max_{1 \le i,j \le l} \lvert x_i - x_j \rvert$
Расстояние городских кварталов (Манхэттенское расстояние)	$d_M(x_i, x_j) = \sum_{l=1}^{k} \lvert x_i^l - x_j^l \rvert$

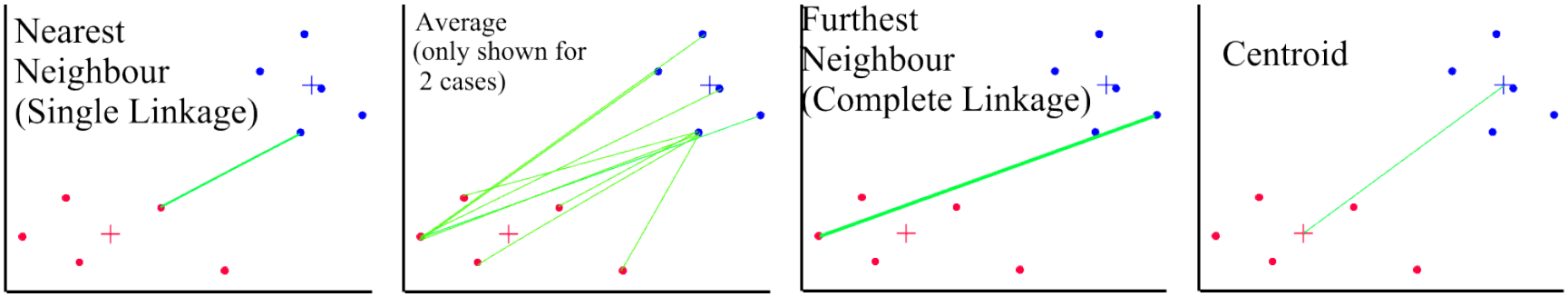


2.3 ВЫБОР МЕТОДА КЛАСТЕРИЗАЦИИ

2.3. Выбор метода кластеризации

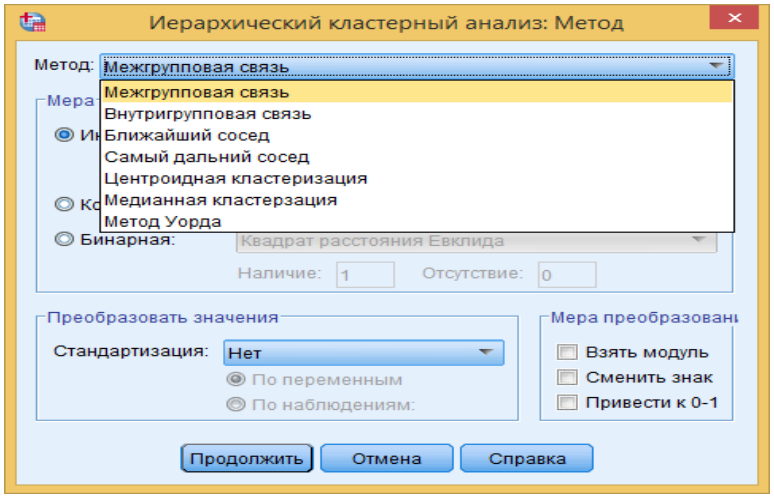
Метод кластеризации – это способ вычисления расстояний между кластерами. Существуют следующие основные методы кластеризации:

- Межгрупповая связь (Between-groups linkage)
- Внутригрупповая связь (Within-groups linkage)
- Ближайший сосед (Nearest neighbor)
- Самый дальний сосед (Furthest neighbor)
- Центроидная кластеризация (Centroid clustering)
- Медианная кластеризация (Median clustering)
- Метод Варда (Уорда)(Ward's method)



2.3. Выбор метода кластеризации

- **Межгрупповая связь (Between-groups linkage)** - дистанция между кластерами, которая равна среднему значению всех расстояний между всеми возможными парами точек из обоих кластеров. Информация, необходимая для расчёта дистанции, находится на основании всех теоретически возможных пар наблюдений. Данный метод устанавливается по умолчанию.
- **Внутригрупповая связь (Within-groups linkage)** - дистанция между двумя кластерами рассчитывается на основании всех возможных пар наблюдений, принадлежащих обоим кластерам, причём учитываются также и пары наблюдений, образующиеся внутри кластеров.
- **Ближайший сосед (Nearest neighbor)** - дистанция между двумя кластерами определяется как расстояние между парой наблюдений, расположенных друг к другу ближе всего, причём каждое наблюдение берётся из своего кластера.



2.3. Выбор метода кластеризации

- **Самый дальний сосед (Furthest neighbor)** - дистанция между двумя кластерами определяется как расстояние между самыми удалёнными друг от друга значениями наблюдений, причём каждое наблюдение берётся из своего кластера.

- **Центроидная кластеризация (Centroid clustering)** - в обоих кластерах рассчитываются средние значения переменных относящихся к ним наблюдений. Затем расстояние между двумя кластерами рассчитывается как дистанция между двумя осредненными наблюдениями.

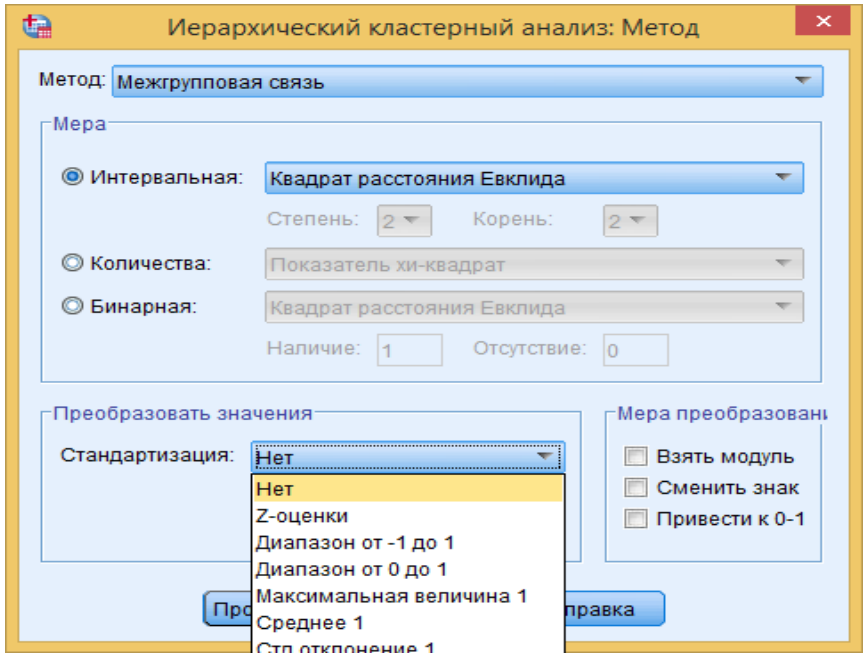
- **Медианная кластеризация (Median clustering)** - тот же центроидный метод, но центр объединенного кластера вычисляется как среднее всех объектов.

- **Метод Варда (Ward-Method)** - сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных. Затем вычисляются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до этого кластерного среднего значения. Эти дистанции суммируются. Потом в один новый кластер объединяются те кластеры, которые дают наименьший прирост общей суммы дистанций.

Стандартизация данных

- Согласно формуле евклидовой меры, переменная, имеющая большие значения, практически полностью доминирует над переменной с малыми значениями.

- Для решения этой проблемы используется **z-преобразование** (стандартизация) значений переменных, которая приводит значения всех преобразованных переменных к единому диапазону значений, а именно от -3 до +3.

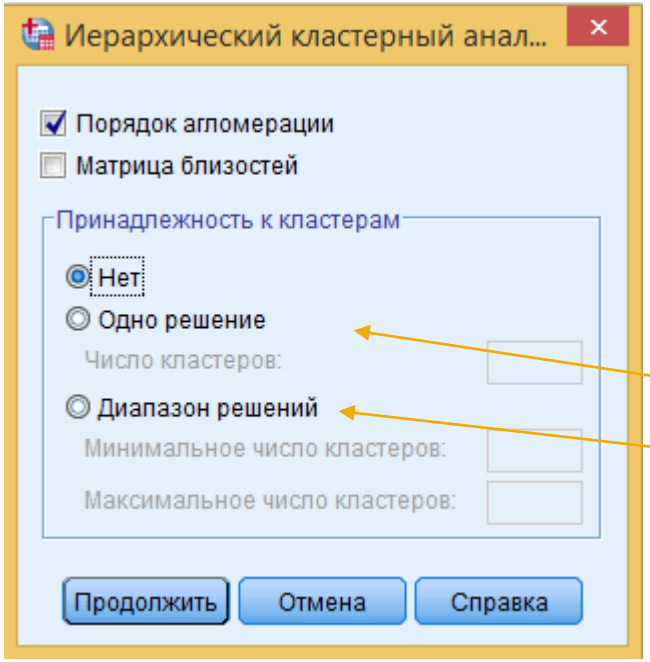


Стандартизация данных

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sigma_{X_j}}$$

Z-стандартизация	Из значений вычитается среднее и затем они делятся на стандартное отклонение.
Разброс от -1 до 1	Линейным преобразованием переменных добиваются разброса значений от -1 до 1.
Разброс от 0 до 1	Линейным преобразованием переменных добиваются разброса значений от 0 до 1.
Максимум 1	Значения переменных делятся на их максимум.
Среднее 1	Значения переменных делятся на их среднее.
Стандартное отклонение 1	Значения переменных делятся на стандартное отклонение.

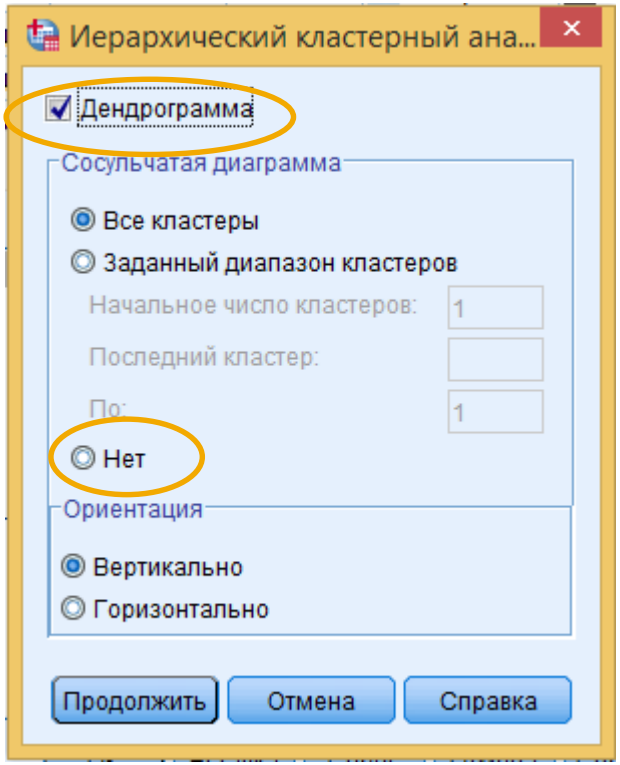
Количество кластеров



Во вкладке «Статистики» можно задать число кластеров:

- конкретное количество
- диапазон значений

Количество кластеров



Во вкладке «Графики» отметить «Дендрограмма» для вывода древовидной диаграммы и отменить вывод накопительной диаграммы.

2.3. Выбор метода кластеризации

Протокол объединения объектов

Первоначально имеется 499 кластеров. На первом шаге объединены 497 и 500 респонденты.

«Объединенный кластер» – содержит кластеры, объединяемые на данном этапе. В нашем случае, на первом этапе объединились кластеры 497 и 500.

«Коэффициент» - расстояние между кластерами.

Порядок агломерации (кластеров)

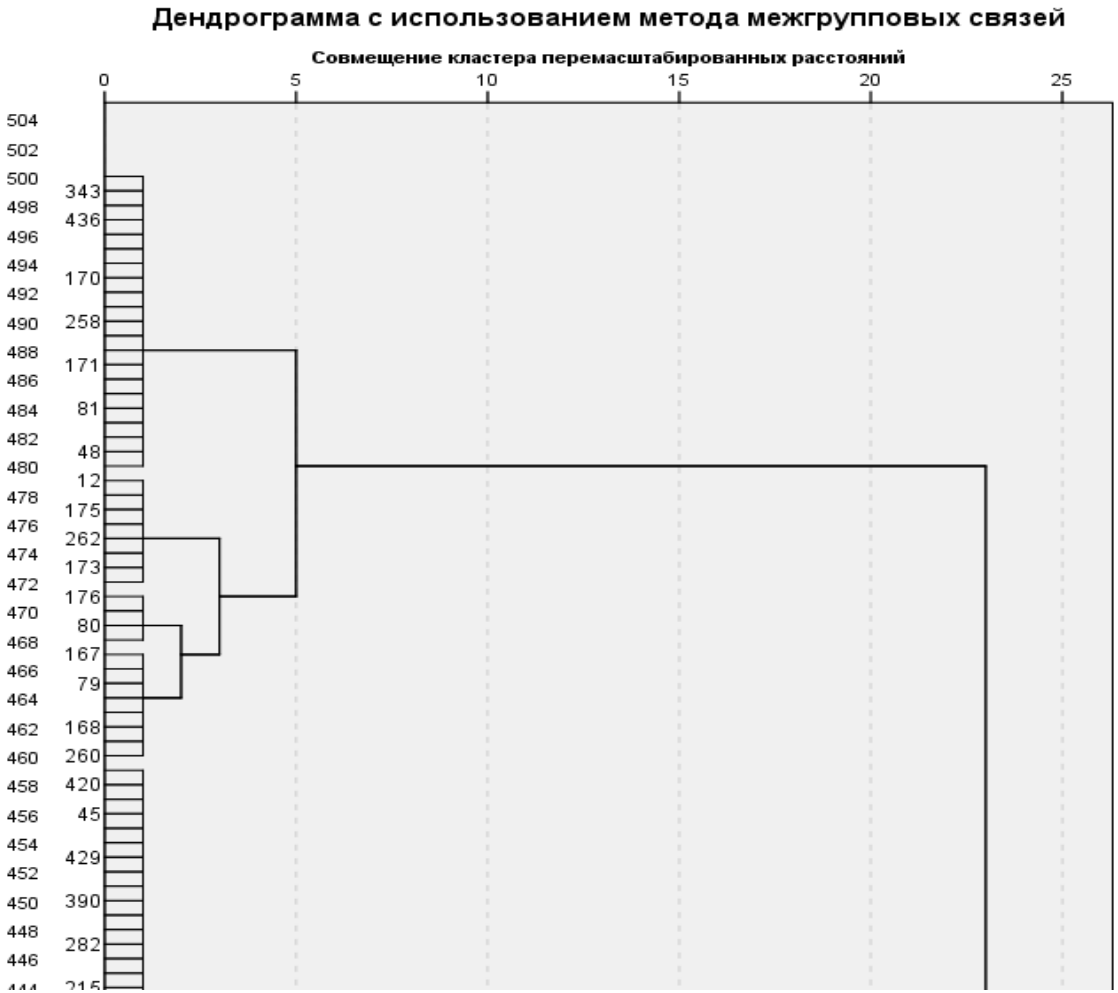
Этап	Объединенный кластер		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	497	500	,000	0	0	4
2	478	499	,000	0	0	23
3	491	498	,000	0	0	10
4	22	497	,000	0	1	7
5	483	496	,000	0	0	18
6	494	495	,000	0	0	7
7	22	494	,000	4	6	16
8	460	493	,000	0	0	41
9	392	492	,000	0	0	108
10	13	491	,000	0	3	25
11	476	490	,000	0	0	25
12	261	489	,000	0	0	238
13	439	488	,000	0	0	62

«Этап первого появления кластера» - показывает, на каком шаге до этого появлялись первый и второй объединяемые кластеры.

«Следующий этап» - показывает, на каком шаге появится кластер, объединенный на этом этапе.

2.3. Выбор метода кластеризации

Фрагмент дендрограммы



2.4 ПРИНЯТИЕ РЕШЕНИЯ О ЧИСЛЕ КЛАСТЕРОВ

2.4. Принятие решения о числе кластеров

1. Необходимо руководствоваться практическими и теоретическими соображениями. Исходя из цели исследования, например, может быть необходимо три кластера.

2. В иерархической кластеризации в качестве критерия используются расстояния. Необходимо смотреть на **коэффициент в протоколе объединения** (расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений).

   - Когда мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить. Иначе будут объединены кластеры, находящиеся на большом расстоянии друг от друга.

   - Оптимальным считается число кластеров равное разности количества наблюдений и количества шагов, после которого коэффициент увеличивается скачкообразно.

3. Размеры кластеров должны быть значимыми.

2.5 ИНТЕРПРЕТАЦИЯ И ПРОФИЛИРОВАНИЕ КЛАСТЕРОВ

2.5. Интерпретация и профилирование кластеров

- **Интерпретация и профилирование** кластеров включает проверку кластерных центроидов.

- **Центроиды** – средние значения объектов по каждой из переменных. Позволяют описывать кластеры.

2.6 ОЦЕНКА КАЧЕСТВА КЛАСТЕРИЗАЦИИ

2.6. Оценка качества кластеризации

- Необходимо выполнять кластерный анализ одних и тех же данных, но с использованием **различных способов измерения расстояния**.
- Сравнить результаты, полученные на основе различных способов расстояния, чтобы определить, насколько совпадают полученные результаты.
- Разбить данные на **две равные части** случайным образом. Выполнить кластерный анализ отдельно для каждой половины. Сравнить кластерные центроиды двух подвыборок.
- Случайным образом **удалить некоторые переменные**. Выполнить кластерный анализ по сокращенному набору переменных. Сравнить результаты с полученными на основе полного набора переменных.

3. КЛАСТЕРНЫЙ АНАЛИЗ МЕТОДОМ k - СРЕДНИХ

3. Кластерный анализ методом k-средних

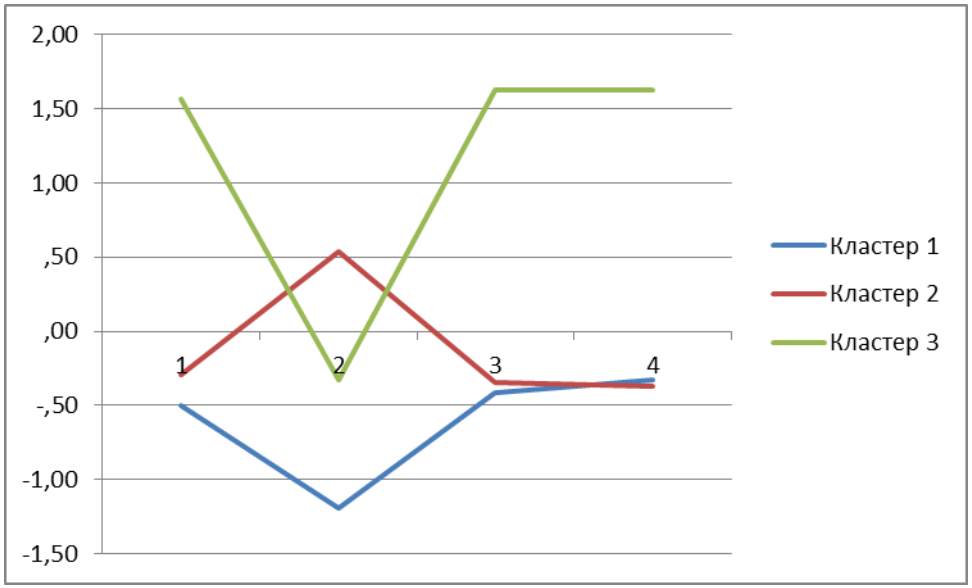
Сначала определяется **центр кластера**, а затем группируют все объекты в пределах заданного от центра порогового значения.

Недостатки:
- Чувствительность к выбросам
- Необходимо заранее задавать количество кластеров, а не как в иерархическом анализе, получать это в качестве результата

Проблему с выбором числа кластеров можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров.

Достоинства:
- Простота использования
- В качестве метрики используется Евклидово расстояние
- Возможность наглядной интерпретации кластеров с использованием графика «Средних значений в кластерах»

График «Средних значений в кластерах»

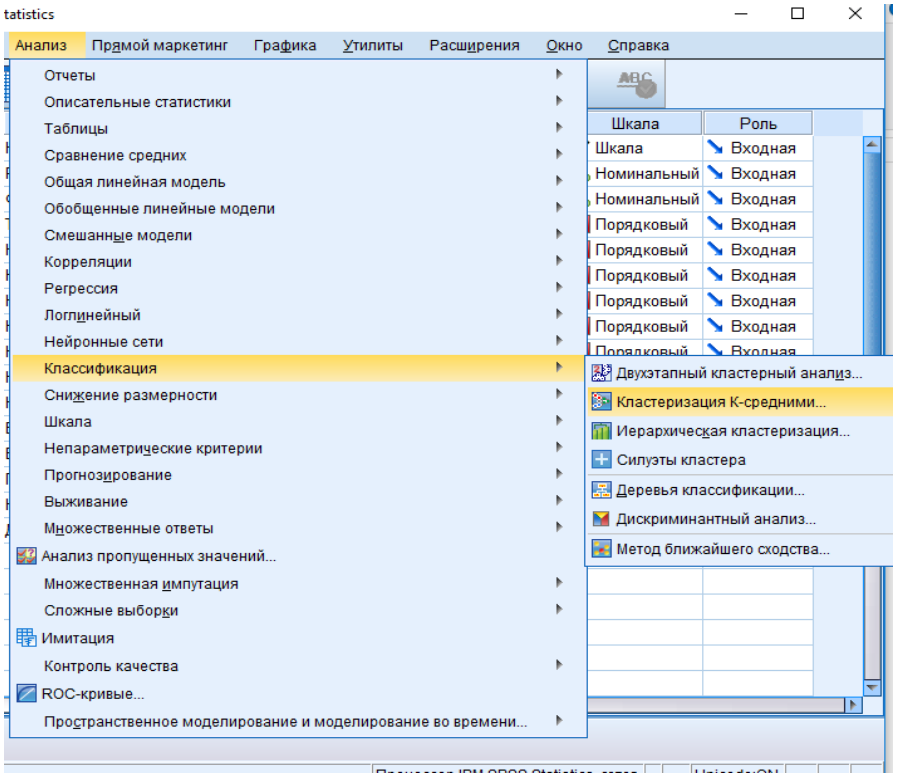


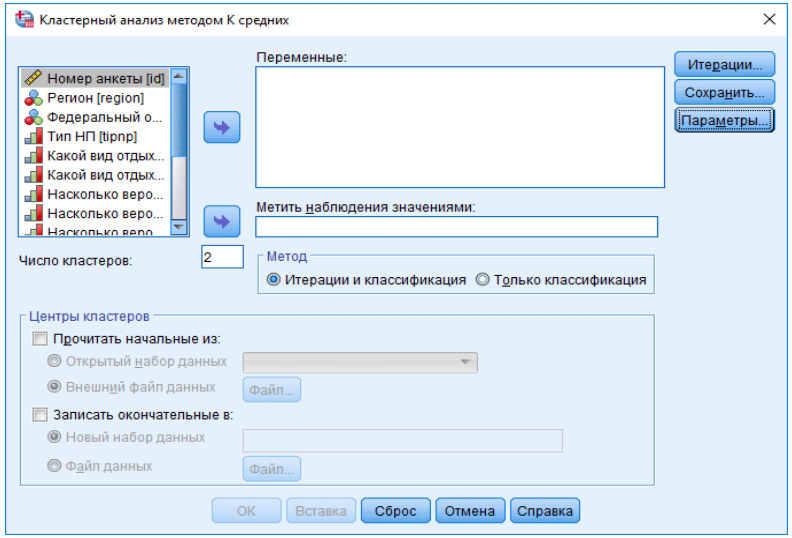
- По горизонтали отложены участвующие в классификации переменные
- По вертикали – средние значения переменных для каждого кластера
- Из графика видно, что кластеры достаточно разнообразны и не дублируют друг друга

3. Кластерный анализ методом k-средних

Методы

- Последовательный пороговый метод (Sequential threshold method)
- Параллельный пороговый метод (Parallel threshold method)
- Метод оптимизирующего распределения (Optimizing partitioning method)

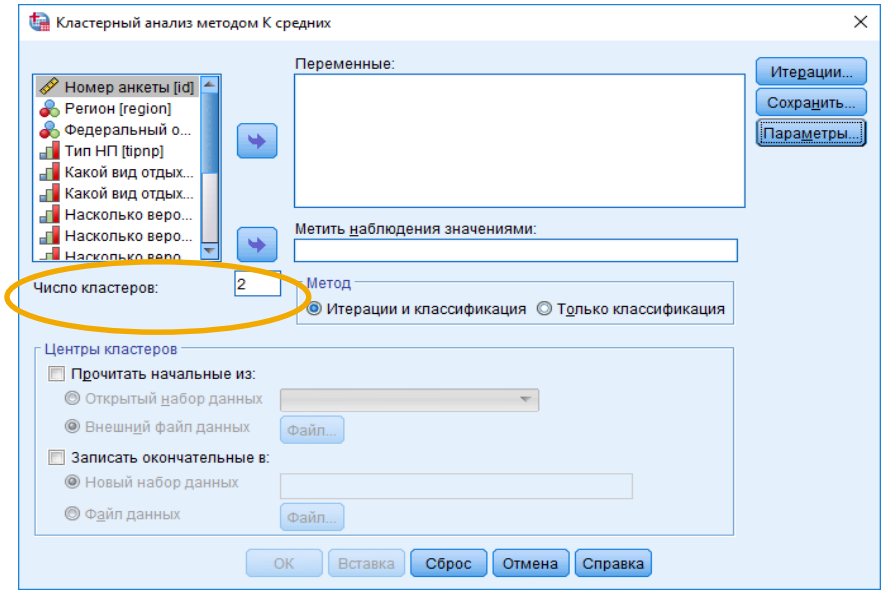




Исходные кластерные центры – это значения любых трех объектов.

3. Кластерный анализ методом k-средних

Кластеризация методом К-средних



- Необходимо задать число кластеров.
- Во вкладке «Итерации» задать количество итераций равное 100 (по умолчанию 10 итераций может оказаться недостаточным)
- Во вкладке «Сохранить» отметить «Принадлежность к кластеру», чтобы использовать для дальнейшего анализа.
- Нажать ОК.

Кластерная принадлежность объектов

Принадлежность к кластерам

Номер наблюдения	Кластеризовать	Расстояние
1	1	17,939
2	1	18,847
3	1	16,855
4	1	23,909
5	1	24,860
6	1	20,948
7	1	24,007
8	1	18,280
9	1	17,323
10	1	24,918
11	2	22,036
12	2	20,656
13	3	11,795
14	3	10,372

Число наблюдений в каждом кластере

Кластеризовать	1	245,000
	2	41,000
	3	214,000
Допустимо		500,000
Пропущенные		,000

3. Кластерный анализ методом k-средних

Конечные кластерные центры и расстояния

Конечные центры кластеров

	Кластеризовать		
	1	2	3
Какую сумму Вы планируете потратить на отдых, если соберетесь в ближайший отпуск поехать за границу?	5	6	97
Какую сумму Вы планируете потратить на отдых, если соберетесь ближайший отпуск провести в России?	4	97	4
Возраст	39	37	45
Доход на 1 члена семьи	7	8	7
Насколько вероятно, что в предстоящем отпуске Вы... Воспользуетесь услугами туроперагента	2	2	3

Расстояния между конечными центрами кластеров

Кластеризовать	1	2	3
1		93,129	92,256
2	93,129		130,558
3	92,256	130,558	

Расстояния между кластерными центрами указывают, насколько хорошо разделены кластеры.

3. Кластерный анализ методом k-средних

В кластеризации методом k-средних программа перемещает объекты (т.е. наблюдения) из одних групп (кластеров) в другие для того, чтобы получить наиболее значимый результат при проведении дисперсионного анализа (ANOVA).

ANOVA

	Кластеризовать		Ошибка			
	Средний квадрат	ст.св.	Средний квадрат	ст.св.	F	Знач.
Какую сумму Вы планируете потратить на отдых, если соберетесь в ближайший отпуск поехать за границу?	516677,323	2	1,294	497	399424,609	,000
Какую сумму Вы планируете потратить на отдых, если соберетесь ближайший отпуск провести в России?	163594,846	2	2,747	497	59554,073	,000
Возраст	2561,901	2	185,146	497	13,837	,000
Доход на 1 члена семьи	50,024	2	1,600	497	31,271	,000
Насколько вероятно, что в предстоящем отпуске Вы... Воспользуетесь услугами туроперагента	31,552	2	1,174	497	26,877	,000

F-критерий следует использовать только для целей описания, так как кластеры выбраны так, чтобы разница между наблюдениями в разных кластерах была максимальной. Наблюдаемые уровни значимости не скорректированы для этого, и поэтому их нельзя использовать для проверки гипотезы о равенстве средних кластеров.

Проверка результатов кластерного анализа

1. Выполнить иерархический кластерный анализ, выбрать число кластеров.
2. Сохранить все решения (отнесение к кластеру каждого респондента).
3. Каждое решение проверить методом Краскала-Уоллиса и выбрать, какое решение наилучшее (суммарная ошибка должна быть самой маленькой).

«Анализ» → «Непараметрические критерии» → «Для K независимых выборок»

- **«Список проверяемых переменных»** - вводим список переменных, ка которых построены кластеры.
- Вводим результат кластерного решения, **«Задать диапазон»** - заполнять каждый раз в зависимости от того количества кластеров, которые в данный момент тестируются.
- В итоге, по каждой зависимой переменной Тест покажет ошибку (Asymp.Sig.). Ее необходимо суммировать.
- Где суммарная ошибка будет меньше (логично предположить, что и по каждому критерию она будет минимальной), ту модель и следует выбрать.

Литература по Теме 9

1. **Бююль А., Цеффель П. SPSS: искусство обработки информации. – М., 2005**
   - Глава 20. Кластерный анализ
2. **Наследов А. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. – СПб., 2013**
   - Глава 21. Кластерный анализ

НАФИ