План лекции

5.1. Кластерный анализ: цели и задачи

Исходные данные для кластерного анализа

$$\begin{pmatrix} y_1 & x_{11} & x_{21} & \boxtimes & x_{p1} \\ y_2 & x_{12} & x_{22} & \boxtimes & x_{p2} \\ y_3 & x_{13} & x_{23} & \boxtimes & x_{p3} \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ y_n & x_{1n} & x_{21} & \boxtimes & x_{pn} \end{pmatrix}$$

Основные понятия:

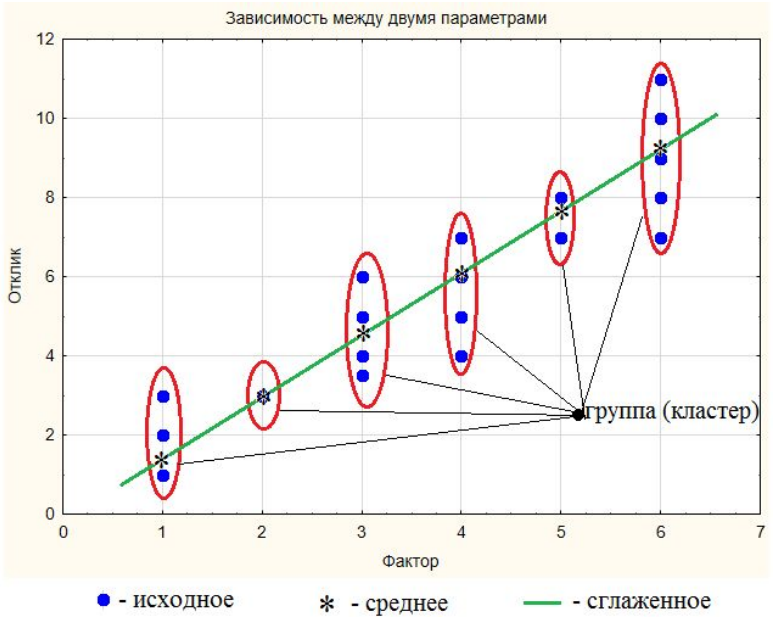
Класс, таксон, группа

Мера сходства

Процедура кластеризации

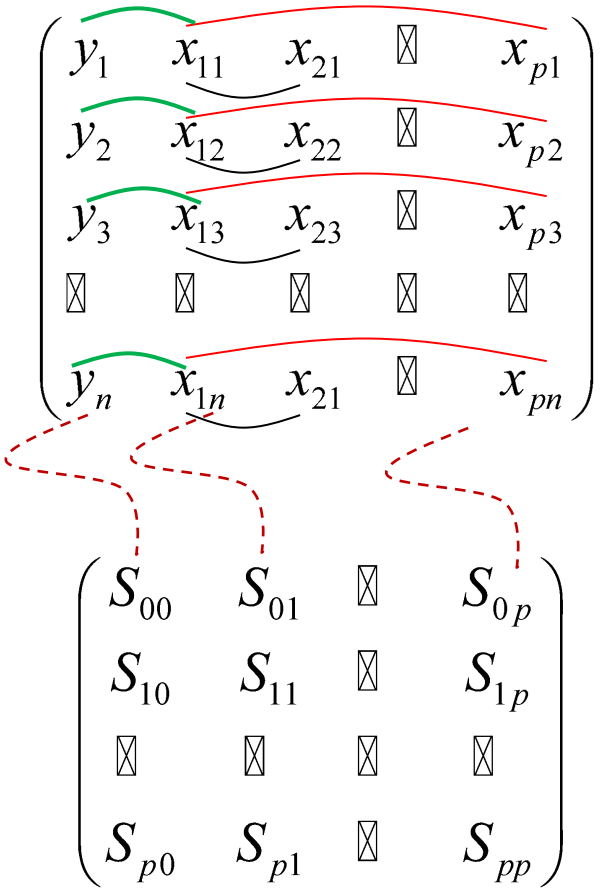
Цель кластеризации: разбиение всего множества наблюдений на однородные группы для их использования при исследовании взаимосвязей между признаками

Задачи кластерного анализа данных:
– определение мер сходства для наблюдений и признаков;
– реализация процедур кластерного анализа.



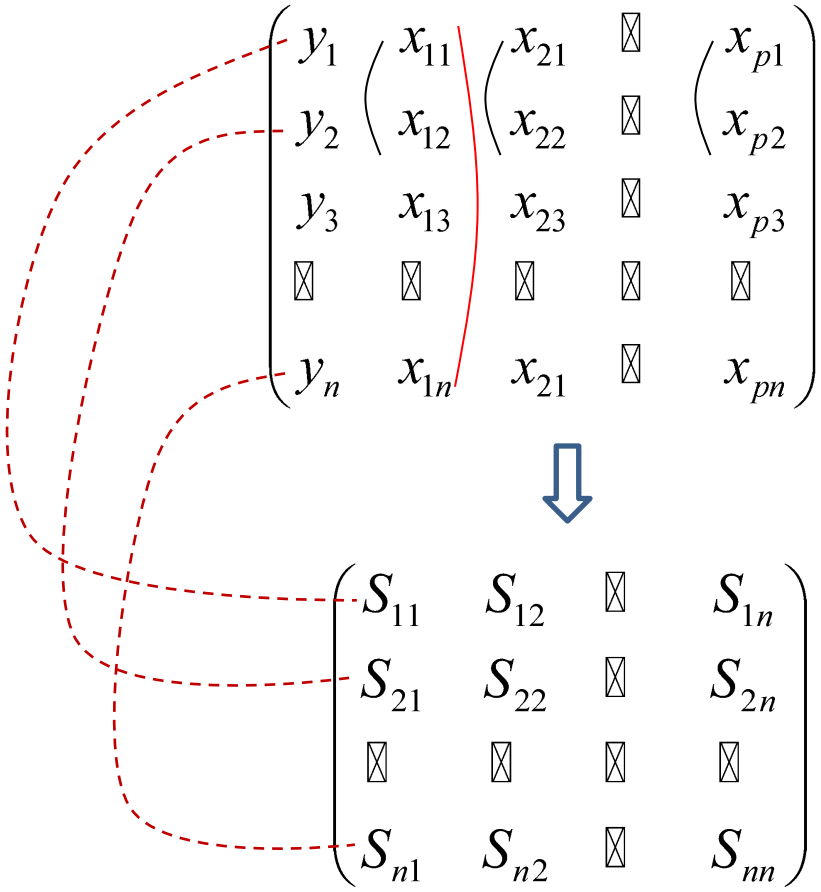
5.2. Меры сходства признаков в общем наборе данных

Сходство между факторами



Матрица сходства по факторам

Сходство между наблюдениями



Матрица сходства по наблюдениями

5.2. Меры сходства признаков в общем наборе данных

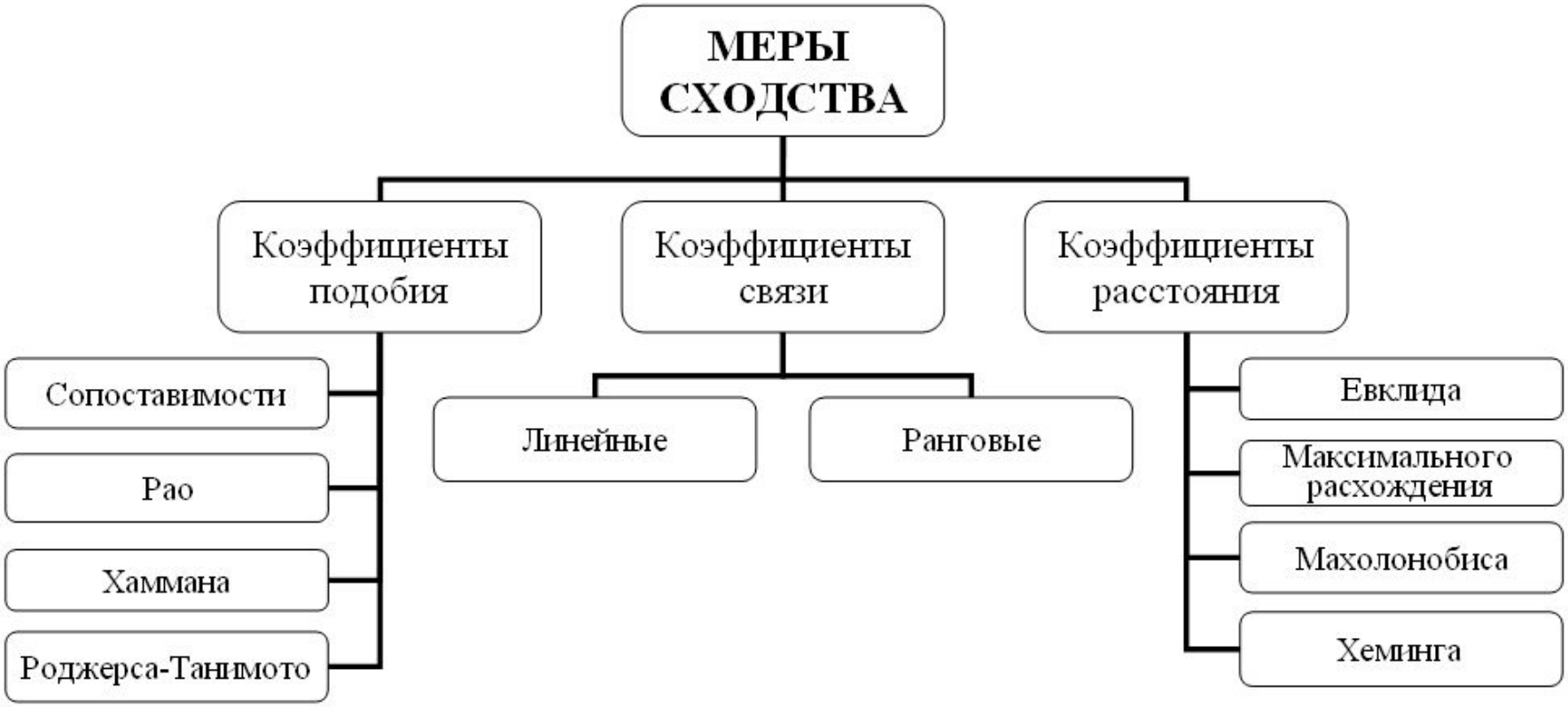


Рис. Схема классификации мер сходства для кластеризации признаков и наблюдений

5.2. Меры сходства признаков в общем наборе данных

Порядок вычисления:

1. Подготовить матрицу исходных данных.

2. Перевести значения наблюдаемых признаков в бинарный вид.

3. Выровнять количество бинарных признаков во всех исходных данных по длине максимального значения в каждом столбце.

4. Выполнить расчет коэффициентов подобия по соответствующей формуле для каждой пары признаков или наблюдений.

5. Записать вычисленные коэффициенты на соответствующие места в матрице.

Формулы для вычисления:

Коэффициент совстречаемости

$$S_{kl} = \frac{P_{kl}}{S}$$

Коэффициент Рао

$$S_{kl} = \frac{P_{kl}^{1,1}}{S}$$

Коэффициент Хаммана

$$S_{kl} = \frac{P_{kl} - Q_{kl}}{S}$$

Коэффициент Роджерса и Танимото

$$S_{kl} = \frac{P_{kl}^{1,1}}{P_{k}^{1,1} + P_{l}^{1,1} - P_{kl}^{1,1}}$$

Обозначение: S – количество сравниваемых бинарных признаков; k, l – номера строк (столбцов), выбранных для рассмотрения; P – количество совпадений; Q – количество несовпадений.

5.2. Меры сходства признаков в общем наборе данных

Исходные данные

	Y	X1	X2	X3	X4
k	1	2	4	1	2
l	2	3	4	2	2

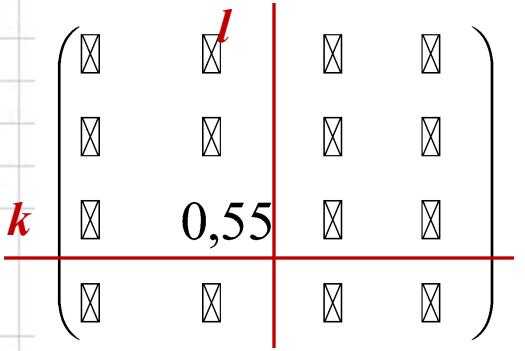
Бинарное представление

k	01	10	100	01	10
l	10	11	100	10	10

Количество разрядов		11
Количество совпадений		6
Количество совпадений 1		3
Количество несопавдений		5

Коэффициент совстречамости	Коэффициент Рао	Коэффициент Хаммана	Коэффициент Роджерса и Танимото
0,55	*0,27*	*0,09*	*0,38*

Положение коэффициента в матрице



$$S_{kl} = \frac{P_{kl}}{S} \qquad S_{kl} = \frac{P_{kl}^{1,1}}{S} \qquad S_{kl} = \frac{P_{kl} - Q_{kl}}{S} \qquad S_{kl} = \frac{P_{kl}^{1,1}}{P_k^{1,1} + P_l^{1,1} - P_{kl}^{1,1}}$$

Для каждого вида коэффициентов строится новая матрица!

5.2. Меры сходства признаков в общем наборе данных

Линейный коэффициент корреляции является количественной оценкой линейной взаимосвязи между двумя выбранными объектами, в частном случае – столбцами или строками данных.

$$r_{ij} = \frac{1}{n}\sum_{k=1}^{n} Z_{ki} \cdot Z_{kj} = \frac{Z_i^T \cdot Z_j}{n}$$

$$Z_i = \begin{pmatrix} Z_{1i} \\ Z_{2i} \\ \square \\ Z_{ni} \end{pmatrix} \qquad Z_{ki} = \frac{X_{ki} - \overline{X}_i}{\sigma_i}$$

$$i, j = \overline{1, p};\ k = \overline{1, n};$$

Матрица парной корреляции:

$$\begin{pmatrix} r_{yy} & r_{yx_1} & \dots & r_{yx_p} \\ r_{x_1y} & r_{x_1x_1} & \dots & r_{x_1x_p} \\ \dots & \dots & \dots & \dots \\ r_{x_py} & r_{x_px_1} & \dots & r_{x_px_p} \end{pmatrix}$$

Свойства коэффициентов:

а) $r_{kl} \in [-1;1]$;

б) если $r_{kl}=0$, то выбранные признаки не зависимы, при условии наличия нормального распределения;

в) если $|r_{kl}|=1$, то между выбранными величинами существует функциональная зависимость, при условии наличия нормального распределения;

г) если $r_{kl}<0$, то между выбранными зависимость убывающая, если $r_{kl}>0$, то между выбранными зависимость возрастающая;

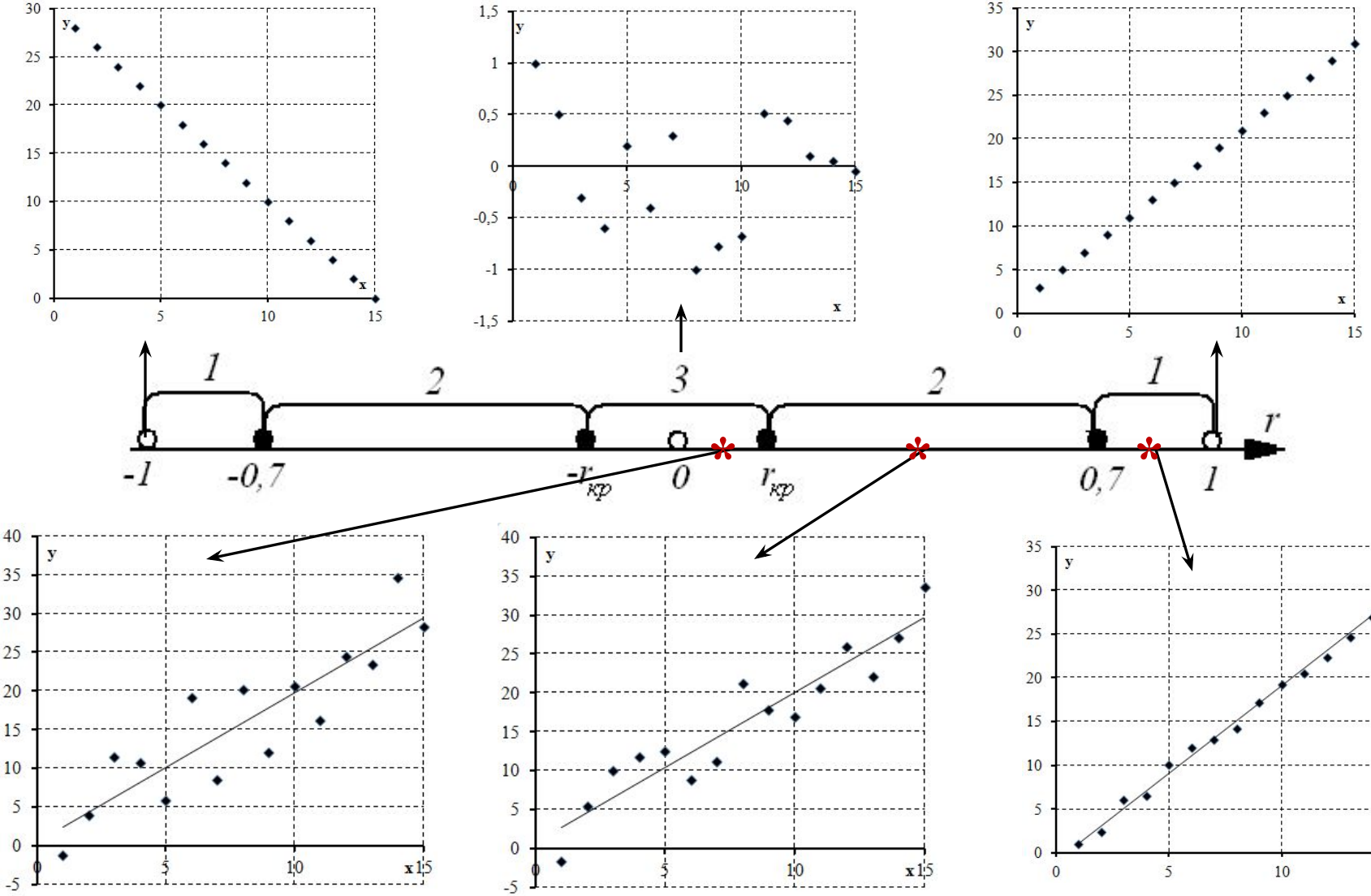
д) $r_{kk}=1$, k=0, 1, 2, …, p;

е) для остальных возможных значений коэффициента корреляции между признаками существует стохастическая (вероятностная зависимость).

Свойства матрицы:

1. Если из этой матрицы удалить строку и столбец соответствующие функции отклика, то будет получена матрица межфакторной корреляции.
2. Матрицы симметричная относительно главной диагонали.

5.2. Меры сходства признаков в общем наборе данных



Обозначения: 1 – область сильной линейной зависимости; 2 – область значимой линейной зависимости; 3 – область слабой линейной зависимости; $r_{кр}$ – критическое значение линейного коэффициента корреляции.

5.2. Меры сходства признаков в общем наборе данных

Алгоритм проверки:

1) выдвигается гипотеза H_0 о том, что линейный коэффициент корреляции попадает в область значимости;
2) рассчитывается величина *t*-статистики:

$$t_{факт} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

3) проверяется неравенство

$$t_{факт} \geq t_{табл}(\alpha, \nu)$$

4) если неравенство истинно, то нет оснований отвергать выдвинутую гипотезу.

Гипотеза проверяется для каждого коэффициента в матрице парной корреляции, за исключением главной диагонали.

Пример оценки значимости:

Дано: r=0,34; n=127; p=5; α=5%.

Определить: значимость r.

Решение:

Выдвинем гипотезу H_0 о том, что линейный коэффициент корреляции попадает в область значимости.

Рассчитаем величину *t*-статистики

$$t_{факт} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{|0{,}34|\sqrt{127-2}}{\sqrt{1-0{,}34^2}} = \frac{0{,}34 \cdot 15}{0{,}94} = 5{,}42$$

Находим табличное значение:

$$t_{табл}(\alpha, \nu) = t_{табл}(5\%, 125) = 1{,}97.$$

Проверяем неравенство:

$$5{,}42 \geq 1{,}97.$$

Вывод: неравенство истинно, нет оснований отвергать гипотезу на 5 %-ом уровне значимости.

5.2. Меры сходства признаков в общем наборе данных

Под *ранговой корреляцией* понимается статистическая связь между ранжировками.

Исходные данные представлены ранжировками m экспертов n альтернатив в виде матрицы

$$\left\| r_{ij} \right\|$$

$$r_l = (r_{l1}, \ldots, r_{ln}) \qquad r_k = (r_{k1}, \ldots, r_{kn})$$

где $i = 1, \ldots, m$, $j = 1, \ldots, n$, где r_{ij} – ранговая оценка i-го эксперта для j-й альтернативы.

Коэффициент ранговой корреляцией Спирмена

$$\rho_{lk} = \frac{\frac{1}{6}\left(n^3 - n\right) - S_{lk}^2 - T_l - T_k}{\sqrt{\left(\frac{1}{6}\left(n^3 - n\right) - 2T_l\right)\left(\frac{1}{6}\left(n^3 - n\right) - 2T_k\right)}}$$

$$S_{lk}^2 = \sum_{j=1}^{m} \left(r_{lj} - r_{kj}\right)^2 \qquad T_i = \frac{1}{2}\sum_{d=1}^{H} \left(h_d^3 - h_d\right)$$

где T_i – показатель связанных рангов в i-и ранжировке; H_i – число групп равных рангов в i-и ранжировке; h_d – число равных рангов в d-й группе связанных рангов в i-и ранжировке.

Проверка статистически значимого отличия от нуля рангового коэффициента корреляции проводится при «не слишком малых» n ($n > 10$) и заданном уровне значимости критерия с помощью неравенства

$$\left|\rho_{lk}\right| > t\left(\frac{\alpha}{2}, n-2\right)\sqrt{\frac{1-\rho_{lk}^2}{n-2}}$$

где $t(Q,v)$ – 100 Q%-ная точка распределения Стьюдента с ν степенями свободы, Q = α/2.

Выполнение неравенства приводит к необходимости отвергнуть гипотезу об отсутствии статистически значимой ранговой корреляционной связи, то есть мнения двух экспертов *признаются согласованными*.

5.2. Меры сходства признаков в общем наборе данных

Расстояние Евклида между объектами обычно оценивается метрикой:

$$d_{kl} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(Z_{kj} - Z_{lj}\right)^2}$$

$$d_{kl} = \sqrt{\sum_{j=1}^{m} w_j \left(Z_{kj} - Z_{lj}\right)^2}$$

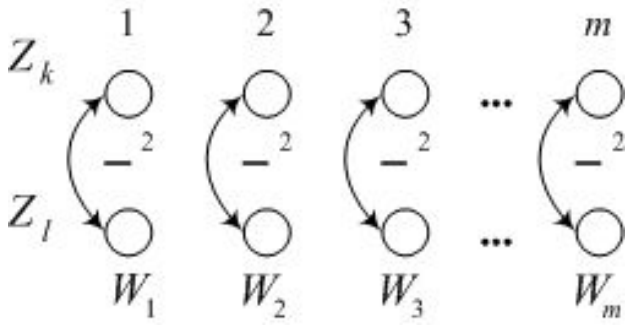


Рис. Схема расчета расстояния между объектами

Максимальное расхождение (расстояние Чебышева)

$$d_{kl} = \underset{1 \le j \le m}{\text{макс}} \left| Z_{kj} - Z_{lj} \right|$$

Расстояние Махалонобиса

$$d_{kl}^2 = \left(Z_k - Z_l\right) R^{-1} \left(Z_k - Z_l\right)^T$$

Расстояние Хемминга (расстояние городских кварталов или Манхэттенское расстояние)

$$d_{kl} = \frac{1}{m}\sum_{j=1}^{m}\left| Z_{kj} - Z_{lj} \right|$$

5.2. Меры сходства признаков в общем наборе данных

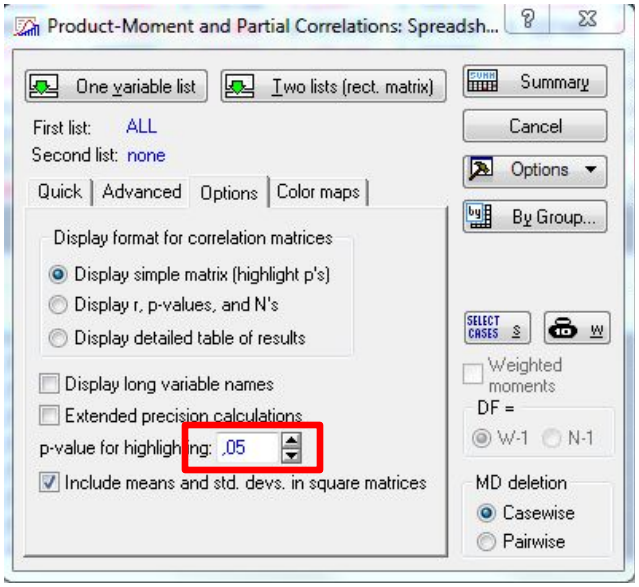


Рис. 1. Настройка уровня значимости

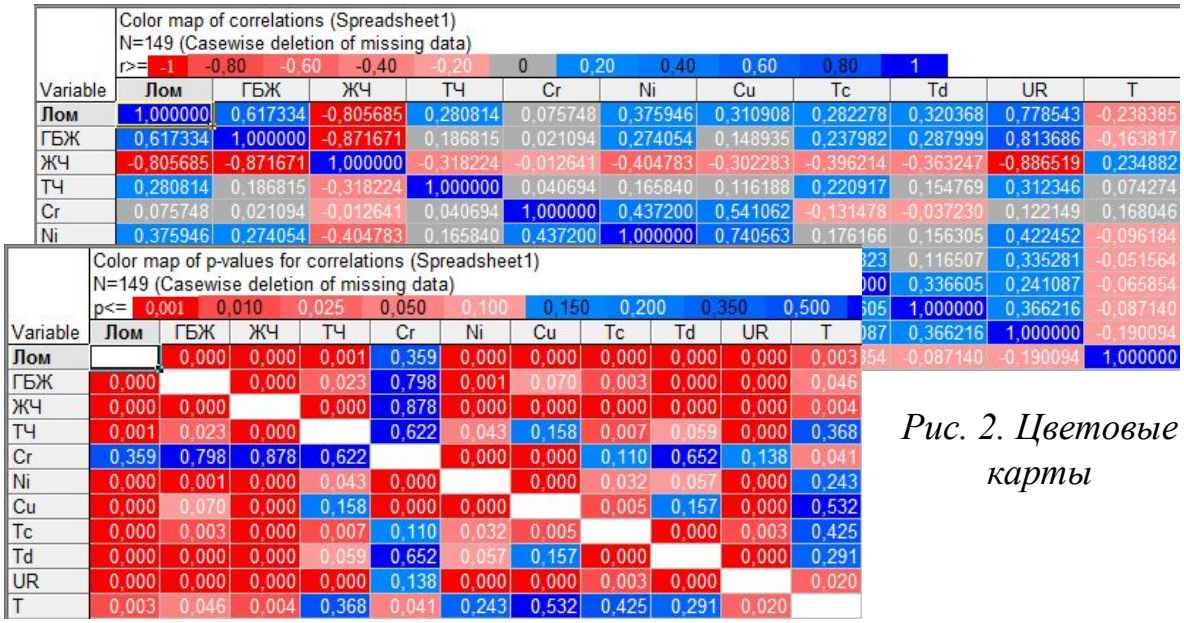


Рис. 2. Цветовые карты

Correlations (Spreadsheet1)
Marked correlations are significant at p < ,05000
N=149 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Лом	ГБЖ	ЖЧ	ТЧ	Cr	Ni	Cu	Tc	Td	UR	T
Лом	132,908	33,85624	1,000000	0,617334	-0,805685	0,280814	0,075748	0,375946	0,310908	0,282278	0,320368	0,778543	-0,238385
ГБЖ	8,435	10,37537	0,617334	1,000000	-0,871671	0,186815	0,021094	0,274054	0,148935	0,237982	0,287999	0,813686	-0,163817
ЖЧ	74,121	38,65840	-0,805685	-0,871671	1,000000	-0,318224	-0,012641	-0,404783	-0,302283	-0,396214	-0,363247	-0,886519	0,234882
ТЧ	0,709	4,30305	0,280814	0,186815	-0,318224	1,000000	0,040694	0,165840	0,116188	0,220917	0,154769	0,312346	0,074274
Cr	0,037	0,01718	0,075748	0,021094	-0,012641	0,040694	1,000000	0,437200	0,541062	-0,131478	-0,037230	0,122149	0,168046
Ni	0,076	0,01872	0,375946	0,274054	-0,404783	0,165840	0,437200	1,000000	0,740563	0,176166	0,156305	0,422452	-0,096184
Cu	0,146	0,04619	0,310908	0,148935	-0,302283	0,116188	0,541062	0,740563	1,000000	0,228823	0,116507	0,335281	-0,051564
Tc	56,173	6,54340	0,282278	0,237982	-0,396214	0,220917	-0,131478	0,176166	0,228823	1,000000	0,336605	0,241087	-0,065854
Td	35,121	14,66883	0,320368	0,287999	-0,363247	0,154769	-0,037230	0,156305	0,116507	0,336605	1,000000	0,366216	-0,087140
UR	242,138	73,91272	0,778543	0,813686	-0,886519	0,312346	0,122149	0,422452	0,335281	0,241087	0,366216	1,000000	-0,190094
T	1629,966	16,01326	-0,238385	-0,163817	0,234882	0,074274	0,168046	-0,096184	-0,051564	-0,065854	-0,087140	-0,190094	1,000000

Рис. 3. Матрица парной корреляции

5.2. Меры сходства признаков в общем наборе данных

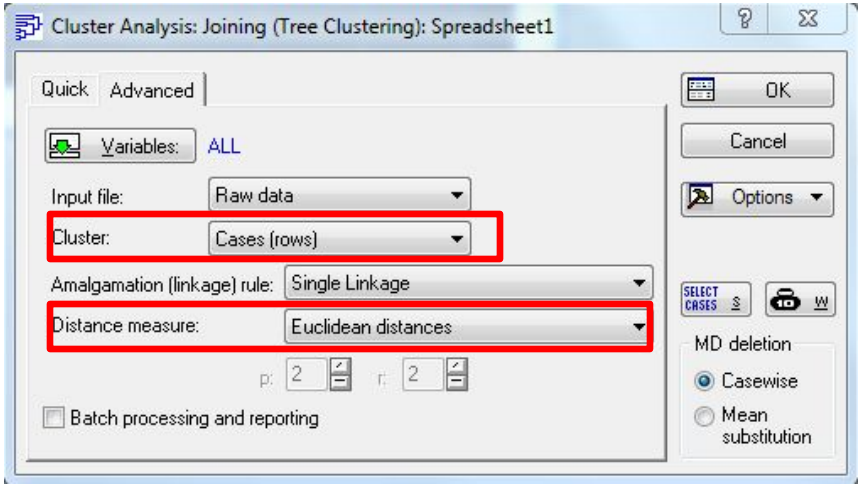


Рис. 1. Настройка объектов и метода для расчета расстояния

Euclidean distances (Spreadsheet1)

Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13
C_1	0	38	18	104	37	34	27	64	33	46	34	33	11
C_2	38	0	44	127	57	44	29	93	43	67	60	25	41
C_3	18	44	0	108	42	28	30	66	32	36	33	36	12
C_4	104	127	108	0	92	134	114	54	133	133	110	127	103
C_5	37	57	42	92	0	56	58	67	54	56	29	65	42
C_6	34	44	28	134	56	0	42	92	12	26	37	38	34
C_7	27	29	30	114	58	42	0	73	44	61	57	15	24
C_8	64	93	66	54	67	92	73	0	93	91	76	86	61
C_9	33	43	32	133	54	12	44	93	0	31	35	38	35
C_10	46	67	36	133	56	26	61	91	31	0	31	61	45
C_11	34	60	33	110	29	37	57	76	35	31	0	60	37
C_12	33	25	36	127	65	38	15	86	38	61	60	0	32
C_13	11	41	12	103	42	34	24	61	35	45	37	32	0
C_14	34	66	27	107	40	37	55	67	38	26	20	60	33
C_15	24	37	20	124	48	13	33	83	14	32	33	31	24
C_16	30	27	29	128	61	27	20	87	29	50	52	12	28
C_17	29	12	35	118	50	41	22	83	40	61	52	22	31
C_18	25	44	25	104	56	45	17	59	48	58	54	29	18

Squared Euclidean distances (Spreadsheet1)

Variable	Лом	ГБЖ	ЖЧ	ТЧ	Cr	Ni	Cu	Tc	Td	UR	T
Лом	0	2429930	1217894	2764272	2800183	2798581	2795739	1034819	1579185	2179266	334182200
ГБЖ	2429930	0	983483	25096	26439	26327	26148	357049	140912	8777753	391837200
ЖЧ	1217894	983483	0	1042587	1038957	1038180	1036710	305181	540629	5985775	360893800
ТЧ	2764272	25096	1042587	0	2807	2796	2781	465594	208132	9466783	395556500
Cr	2800183	26439	1038957	2807	0	0	2	475880	215249	9541844	395881800
Ni	2798581	26327	1038180	2796	0	0	1	475216	214825	9538901	395862800
Cu	2795739	26148	1036710	2781	2	1	0	474034	214085	9533689	395828800
Tc	1034819	357049	305181	465594	475880	475216	474034	0	94656	5933242	369093300
Td	1579185	140912	540629	208132	215249	214825	214085	94656	0	7108445	379062200
UR	2179266	8777753	5985775	9466783	9541844	9538901	9533689	5933242	7108445	0	287897000
T	334182200	391837200	360893800	395556500	395881800	395862800	395828800	369093300	379062200	287897000	0

Рис. 2. Матрицы расстояний Евклида для строк и столбцов

5.3. Процедуры кластерного анализа данных

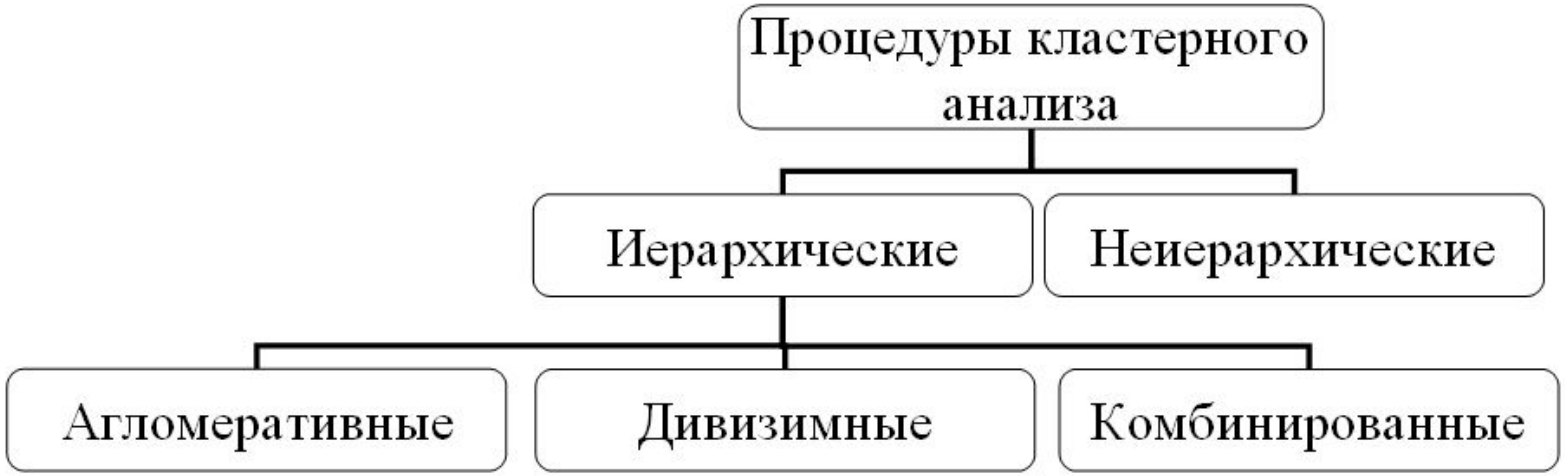


Рис. 1. Схема классификации процедур кластеризации

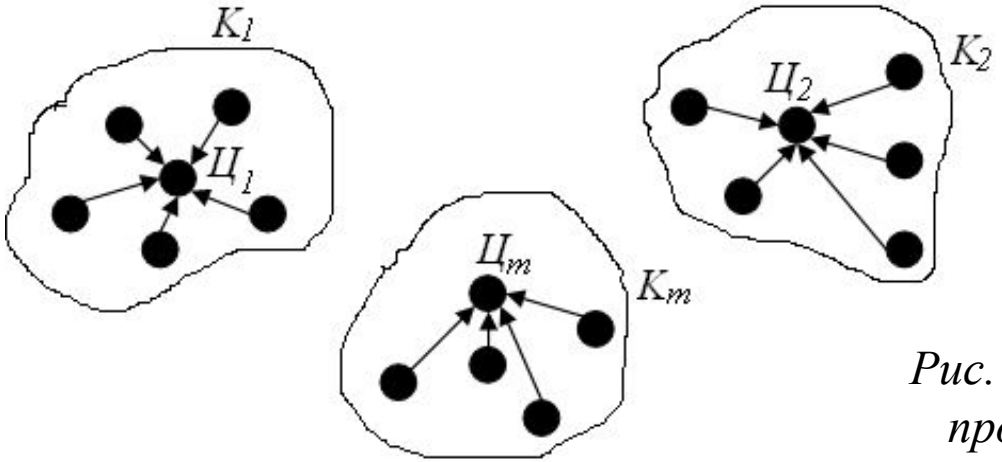


Рис. 2. Схема неиерархической процедуры кластеризации

5.3. Процедуры кластерного анализа данных

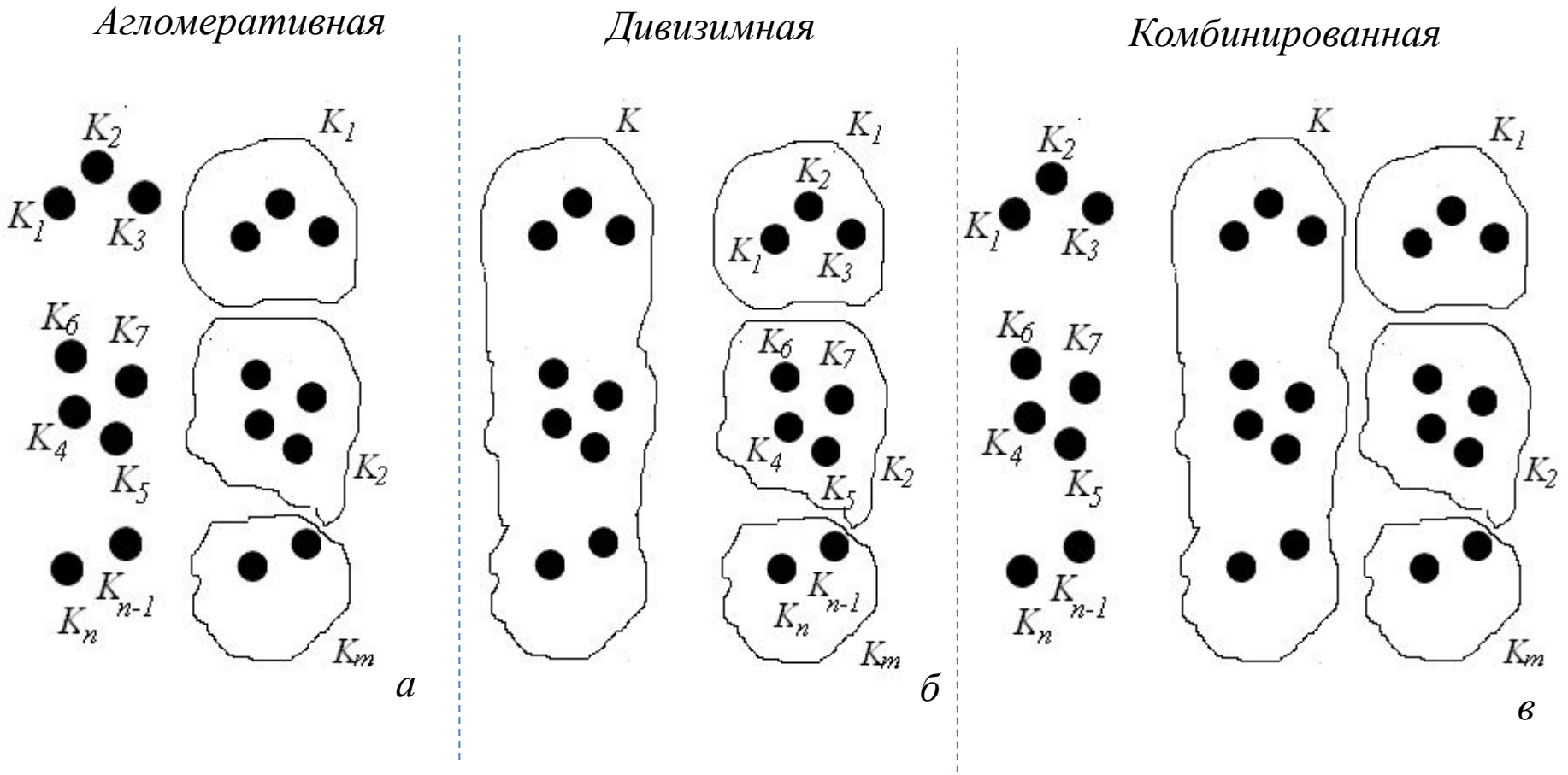


Рис. Схемы иерархических процедур кластеризации

5.3.2. Агломеративная процедура кластеризации по расстоянию

Алгоритм:

а) в исходной матрице сходства (расстояния) находят два различных, но наиболее подобных (ближайших) объекта (кластера) *p* и *q*;

б) кластеры *p* и *q* объединяют в один общий кластер *r*;

в) составляется новая матрица расстояний, в которой сохраняются прежние, кроме *p* и *q*, кластеры, но вводится новый кластер *r*, причем расстояние от любого сохранившегося кластера *s* до кластера *r* определяется как

$$d_{sr} = \alpha_p d_{ps} + \alpha_q d_{qs} + \beta d_{pq} + \gamma \left| d_{ps} - d_{qs} \right|$$

где d_{ps}, d_{qs}, d_{pq} – расстояния между кластерами по предыдущей матрице, α_p, α_q, β, γ – параметры, определяемые методом расчета.

Методы расчета расстояния между кластерами:

1) медианный:

$\alpha_p = \alpha_q = ½$, $\beta = -¼$, $\gamma = 0$;

2) простого среднего:

$\alpha_p = \alpha_q = ½$, $\beta = \gamma = 0$;

3) группового среднего:

$$\alpha_p = \frac{n_p}{n_p + n_q} \qquad \alpha_q = \frac{n_q}{n_p + n_q}$$

$\beta = \gamma = 0$, где n – число объектов в соответствующей группе;

4) центроидный:

$$\alpha_p = \frac{n_p}{n_p + n_q} \qquad \alpha_q = \frac{n_q}{n_p + n_q}$$

$$\beta = -\frac{n_q}{(n_p + n_q)^2} \qquad \gamma = 0$$

5.3.2. Агломеративная процедура кластеризации по расстоянию

Пусть по результатам наблюдений построена матрица расстояний

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 3 & 4 & 2 \\ 2 & 3 & 0 & 0{,}5 & 3 \\ 3 & 4 & 0{,}5 & 0 & 4 \\ 2 & 2 & 3 & 4 & 0 \end{pmatrix}$$

Требуется выполнить кластеризацию методом простого среднего.

Решение:

Определим минимальное расстояние в матрице: $d_{43}=0{,}5$.
Следовательно, $p=4$ и $q=3$.
Вводим новый кластер с номером $s=6$.
Выполним расчет всех остальных расстояний:

$$d_{16}=\frac{1}{2}d_{41}+\frac{1}{2}d_{31}+0\cdot d_{43}+0\cdot\left|d_{41}-d_{31}\right|=$$

$$=\frac{1}{2}\cdot 3+\frac{1}{2}\cdot 2=\frac{5}{2}=2{,}5;$$

$$d_{26}=\frac{1}{2}d_{42}+\frac{1}{2}d_{32}+0\cdot d_{43}+0\cdot\left|d_{42}-d_{32}\right|=$$

$$=\frac{1}{2}\cdot 4+\frac{1}{2}\cdot 3=\frac{7}{2}=3{,}5;$$

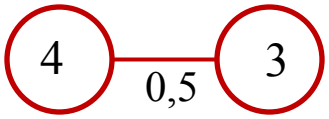
$$d_{56}=\frac{1}{2}d_{45}+\frac{1}{2}d_{35}+0\cdot d_{43}+0\cdot\left|d_{45}-d_{35}\right|=$$

$$=\frac{1}{2}\cdot 4+\frac{1}{2}\cdot 3=\frac{7}{2}=3{,}5.$$

Новая матрица:

$$\begin{pmatrix} 0 & 1 & 2 & 2{,}5 \\ 1 & 0 & 2 & 3{,}5 \\ 2 & 2 & 0 & 3{,}5 \\ 2{,}5 & 3{,}5 & 3{,}5 & 0 \end{pmatrix}$$

Первый кластер:



Процесс итерационный…

5.3.3. Метод вроцлавской таксономии

Дендрит *– это такая ломаная, которая может разветвляться, но не может содержать замкнутых ломаных, и которой соединены две любые точки множества признаков.*

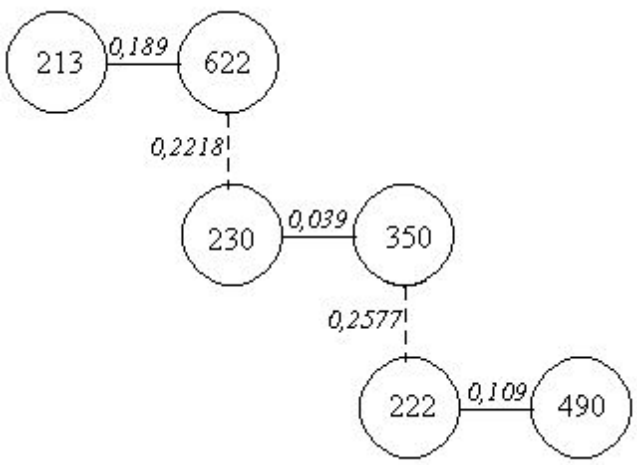


Рис. Вид дендрита

Преимущества:

Использует матрицу расстояний, но не требует их пересчета.

Алгоритм метода вроцлавской таксономии:

1. Из матрицы расстояний выбираются элементы с близкими расстояниями. Поиск проводится путем *нахождения наименьших чисел в каждом столбце* (или строке) матрицы расстояний.

2. Выполнить построение дендритов первого порядка.

3. Выполнить объединение скоплений дендритов первого порядка в дендриты второго порядка. Объединение выполняется до тех пор пока не будет получен единый дендрит.

4. Упорядочить связи дендрита по убыванию длины рассчитать отношение:

$$i_2 = \frac{d_1}{d_2},\ i_3 = \frac{d_2}{d_3},\quad \ldots \quad i_{n-1} = \frac{d_{n-2}}{d_{n-1}},$$

5. Найти все *k,* для которого выполняется соотношение $i_{k-1} < i_k$ (для $k=2, 3, \ldots, n-2$) и выбрать из них минимальное.

6. Разорвать *k-1* связь.

5.3.3. Метод вроцлавской таксономии

Пусть по результатам наблюдений построена матрица расстояний

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 3 & 4 & 2 \\ 2 & 3 & 0 & 0{,}5 & 3 \\ 3 & 4 & 0{,}5 & 0 & 4 \\ 2 & 2 & 3 & 4 & 0 \end{pmatrix}$$

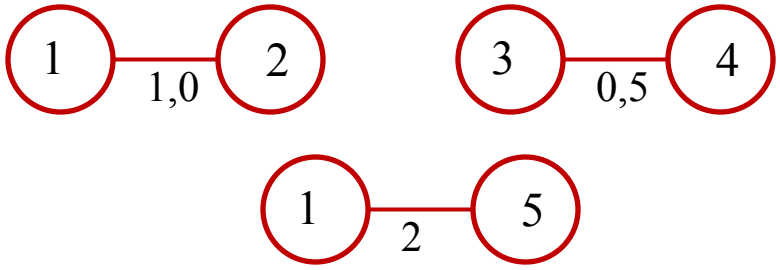
Требуется выполнить кластеризацию методом вроцлавской таксономии.

Решение:

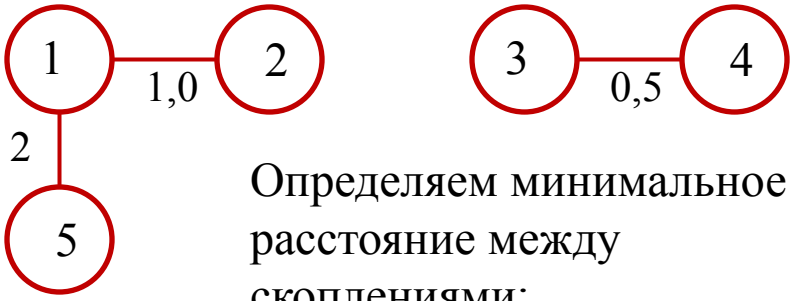
Определим минимальное расстояние в каждом столбце матрицы:

$d_{21}=1$, $d_{12}=1$, $d_{43}=0{,}5$, $d_{34}=0{,}5$, $d_{15}=2$.

Получаем дендриты первого порядка с учетом повторений:



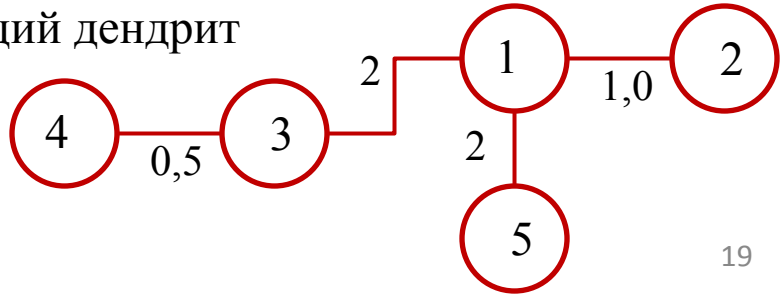
Дендриты второго порядка:



Определяем минимальное расстояние между скоплениями:

$Min\{d_{13}=2, d_{23}=3, d_{53}=3, d_{14}=3, d_{42}=4, d_{45}=4\}=d_{13}=2.$

Объединяем 1 и 3 группу. Получаем общий дендрит



5.3.3. Метод вроцлавской таксономии

Упорядочивание связей:

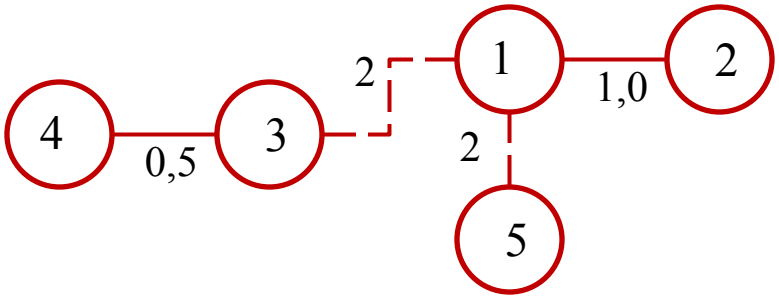
d_i	S_{ij}	Значение	i_k	Значение
d_1	S_{15}	2		
d_2	S_{13}	2	i_2	1
d_3	S_{12}	1	i_3	2
d_4	S_{34}	0,5	i_4	2

$i_2 < i_3$

Количество кластеров: 3.

Количество разрываемых связей: 2.

Новые кластеры:



Состав групп:

Номер кластера	Состав кластера
1	1, 2
2	3, 4
3	5

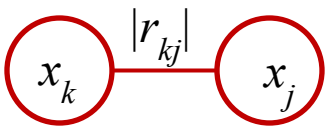
Результат: новая матрица наблюдений и состав каждой группы

5.3.4. Метода корреляционных плеяд

Алгоритм метода корреляционных плеяд:

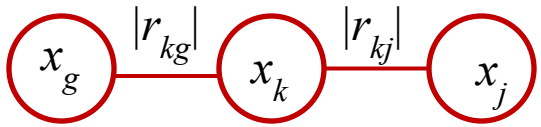
1. В матрице коэффициентов межфакторной корреляции находится наибольший по абсолютной величине коэффициент корреляции (не считая диагональных) – r_{kj}.

2. Строится дендрит первого уровня между факторами с номерами k и j с указанием над связью абсолютного значения $|r_{kj}|$.



3. Находим наибольшие по абсолютному значению коэффициенты корреляции в столбцах k и j, исключая r_{kj} и из выбранных находим наибольший по абсолютному значению – $|r_{km}|$.

4. Строится дендрит второго уровня между факторами с номерами k и g с указанием над связью абсолютного значения $|r_{kg}|$.



5. Находим признаки, наиболее тесно связанные с двумя последними рассмотренными, и, повторяя процедуру выбора, выбираем из двух соответствующих коэффициентов корреляции наибольший по абсолютной величине.

6. Продолжая построение, на каждом шаге находим признак, наиболее тесно связанный с одним из двух признаков, отобранных на предыдущем этапе. Построение чертежа завершим, когда в нем окажется m кружков (m – число признаков).

7. Выбираем пороговую величину h и исключаем из схемы связи, соответствующие *меньшим,* чем h коэффициентам парной корреляции, например по значимости коэффициента парной корреляции.

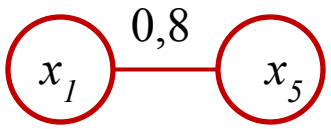
8. Разрываем все связи с коэффициентом корреляции ниже критического при заданном уровне значимости.

9. Для факторов внутри группы достаточно определить линейные взаимосвязи.

5.3.4. Метода корреляционных плеяд

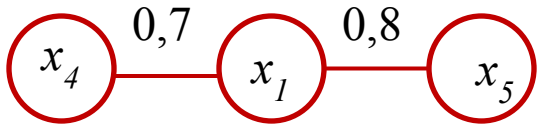
Матрица межфакторной корреляции – Итерация 1

	A	B	C	D	E	F	G
1	Матрица межфакторной корреляции						
2		1	2	3	4	5	6
3	1	1	0,1	0,45	0,7	-0,8	0,54
4	2	0,1	1	0,35	0,24	0,7	0,6
5	3	0,45	0,35	1	0,15	0,45	0,75
6	4	0,7	0,24	0,15	1	0,32	0,1
7	5	**-0,8**	0,7	0,45	0,32	1	-0,56
8	6	0,54	0,6	0,75	0,1	-0,56	1



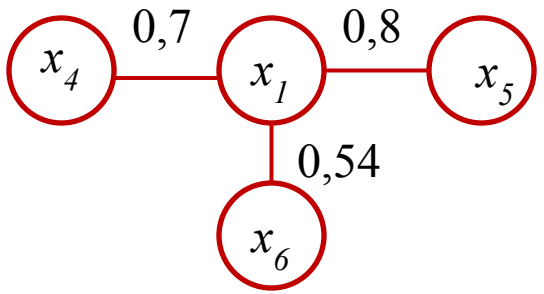
Матрица межфакторной корреляции – Итерация 2

	A	B	C	D	E	F	G
1	Матрица межфакторной корреляции						
2		1	2	3	4	5	6
3	1	1	0,1	0,45	0,7	-0,8	0,54
4	2	0,1	1	0,35	0,24	**0,65**	0,6
5	3	0,45	0,35	1	0,15	0,45	0,75
6	4	**0,7**	0,24	0,15	1	0,32	0,1
7	5	-0,8	0,65	0,45	0,32	1	-0,56
8	6	0,54	0,6	0,75	0,1	-0,56	1



Матрица межфакторной корреляции – Итерация 3

	A	B	C	D	E	F	G
1	Матрица межфакторной корреляции						
2		1	2	3	4	5	6
3	1	1	0,1	0,45	0,7	-0,8	0,54
4	2	0,1	1	0,35	0,24	0,65	0,6
5	3	0,45	0,35	1	0,15	0,45	0,75
6	4	0,7	0,24	0,15	1	0,32	0,1
7	5	-0,8	0,65	0,45	**0,32**	1	-0,56
8	6	**0,54**	0,6	0,75	0,1	-0,56	1



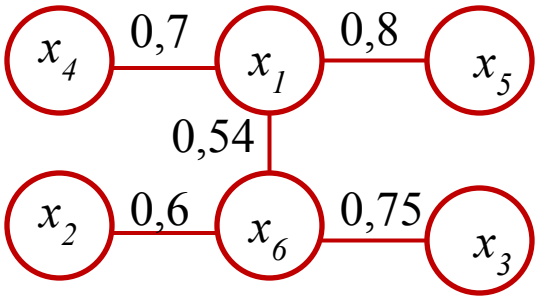
Матрица межфакторной корреляции – Итерация 4

	A	B	C	D	E	F	G
1	Матрица межфакторной корреляции						
2		1	2	3	4	5	6
3	1	1	0,1	0,45	0,7	-0,8	0,54
4	2	0,1	1	0,35	0,24	0,65	0,6
5	3	**0,45**	0,35	1	0,15	0,45	**0,75**
6	4	0,7	0,24	0,15	1	0,32	0,1
7	5	-0,8	0,65	0,45	0,32	1	-0,56
8	6	0,54	0,6	0,75	0,1	-0,56	1

5.3.4. Метода корреляционных плеяд

Матрица межфакторной корреляции – Итерация 4

	A	B	C	D	E	F	G
1	Матрица межфакторной корреляции						
2		1	2	3	4	5	6
3	1	1	0,1	0,45	0,7	-0,8	0,54
4	2	0,1	1	0,35	0,24	0,65	**0,6**
5	3	0,45	0,35	1	0,15	0,45	0,75
6	4	0,7	0,24	0,15	1	0,32	0,1
7	5	-0,8	0,65	0,45	0,32	1	-0,56
8	6	0,54	**0,6**	0,75	0,1	-0,56	1



Примем количестве наблюдений равным 30.

Критическое значение t-статистики:

t(5%, 28)=1,17.

Используем выражение критерия Стьюдента:

$$t_{факт} = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$

Выразим значение коэффициента корреляции:

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

Вычислим критическое значение коэффициента:

$$r = \frac{1{,}17}{\sqrt{30-2+1{,}17^2}} \approx 0{,}21$$

Все факторы образуют один кластер.

Между факторами можно установить линейные зависимости:

$$\begin{cases} x_4 = f(x_1) + \varepsilon_1; \\ x_5 = f(x_1) + \varepsilon_2; \\ x_2 = f(x_6) + \varepsilon_3; \\ x_3 = f(x_6) + \varepsilon_4; \\ x_1 = f(x_6) + \varepsilon_5. \end{cases}$$

5.3.5. Метода k-средних или алгоритм Лойда

Алгоритм метода k-средних:

1. Из исходного множества данных ***случайным*** образом выбираются *k* записей, которые будут служить начальными центрами кластеров (центроидами или эталонами).

Количество классов *k* ***назначается*** исследователем.

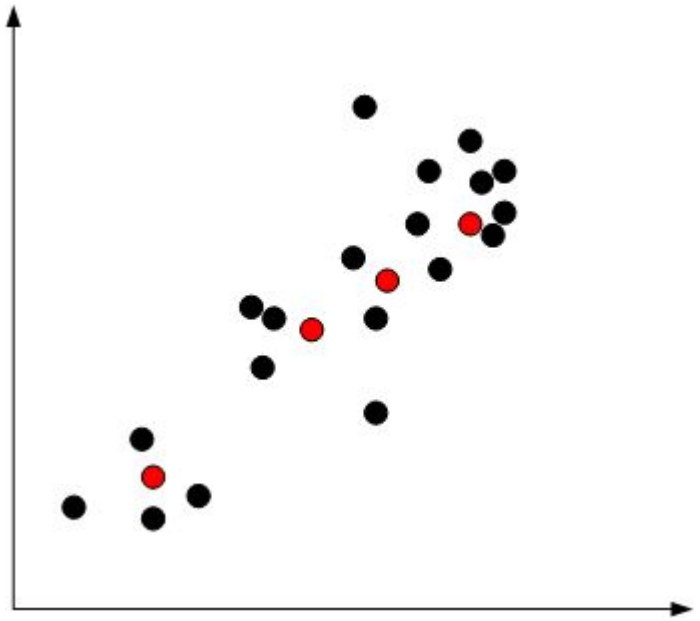
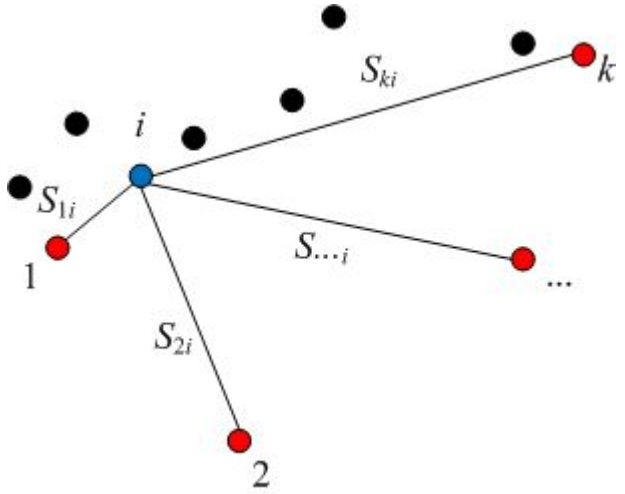


Рис. Исходные данные и выбранные центроиды

2. Для каждой точки определяется расстояние до центроида и выбирается принадлежность к классу.

В качестве метрики ***чаще*** всего используется расстояние ***Евклида***.



Номер класса – это номер центроида с минимальным расстоянием до выбранной точки *i*:

5.3.5. Метода k-средних или алгоритм Лойда

3. Вычисляются внутригрупповая дисперсия в каждом кластере:

$$D_j^l = \frac{1}{k_j}\sum_{i=1}^{k_j}\left(x_i - \mu_j\right)^2.$$

l – номер итерации, μ_j – центроид класса j.

4. Вычисляются центры тяжести новых кластеров, т.е. значение центроида в новом кластере:

$$\mu_j = \frac{1}{k_j}\sum_{i=1}^{k_j} x_i.$$

5. Шаги 2, 3, 4 повторяются, пока не будет найдена стабильная конфигурация (то есть кластеры перестанут изменяться) или число итераций не превысит заданное пользователем.

Особенности метода: результат зависим от начального выбора центроидов.

Результаты кластеризации

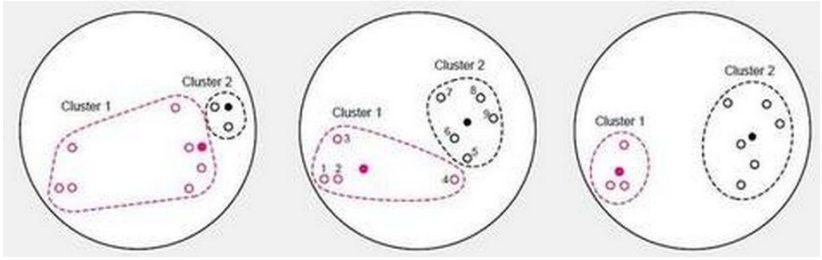


Рис. Круги Эйлера

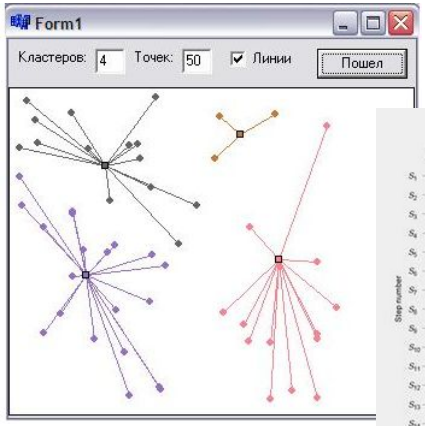


Рис. Лучевая диаграмма

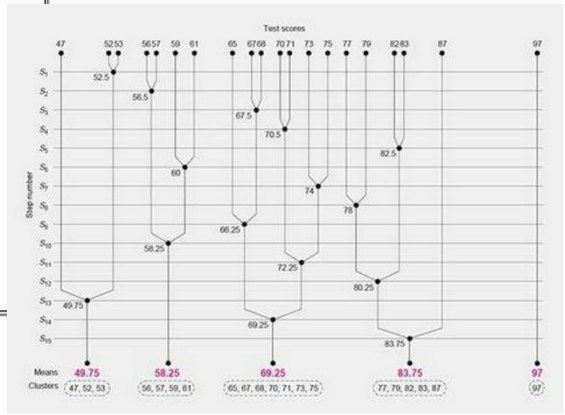
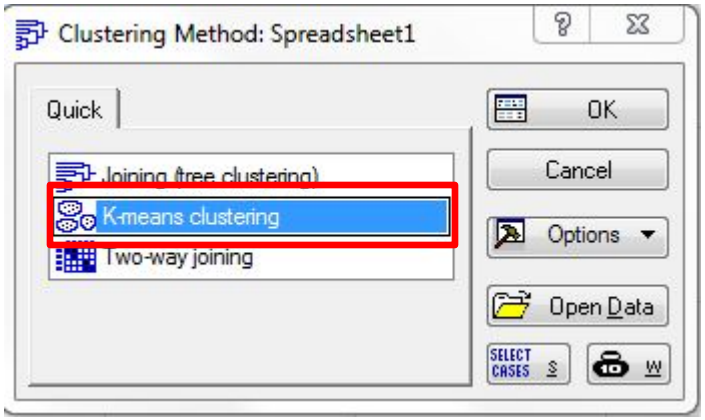
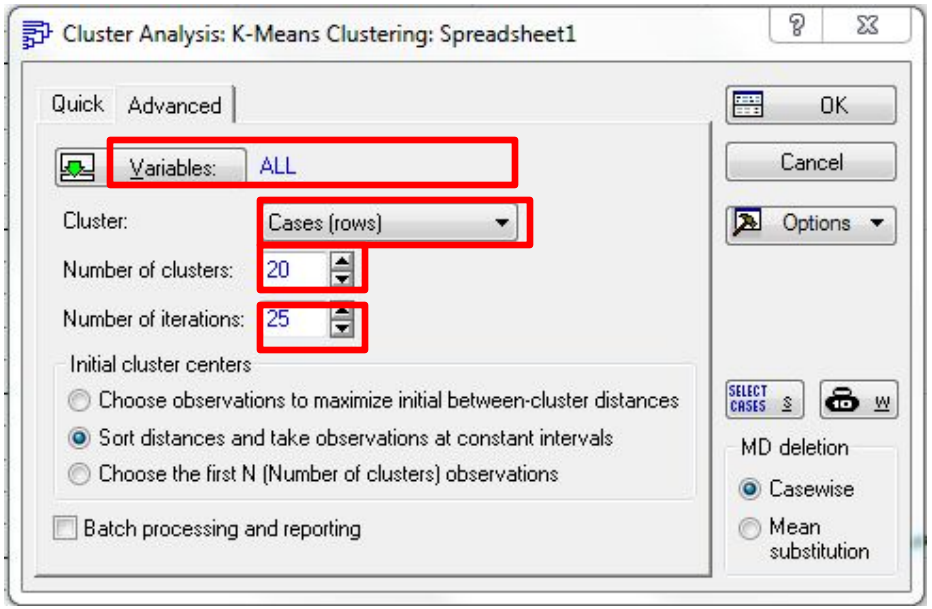


Рис. Дендрограмма

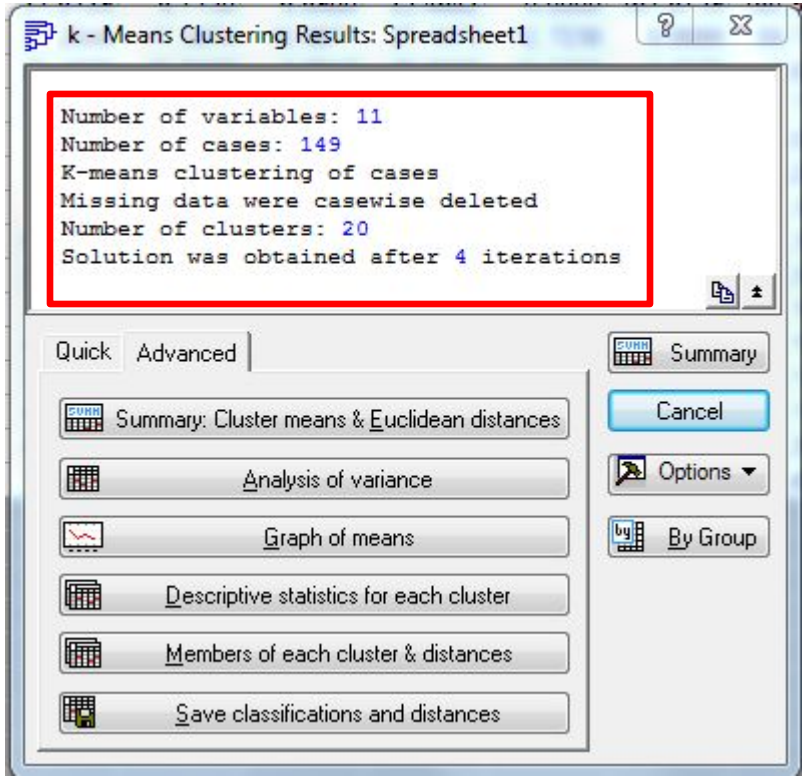
5.3.5. Метода k-средних или алгоритм Лойда



а



б



в

Рис. Диалоговые окна для настройки параметров кластеризации: а – выбор процедура кластеризации; б – определение исходных данных и параметров; в – меню результатов

5.3.5. Метода k-средних или алгоритм Лойда

Euclidean Distances between Clusters (Spreadsheet1)
Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12	No. 13	No. 14	No. 15	No. 16	No. 17	No. 18	No. 19	No. 20
No. 1	0,00000	291,1926	209,8734	137,2547	577,2233	280,2576	180,3748	147,3947	81,2225	96,4073	317,569	556,888	2974,527	205,496	486,125	728,933	1641,633	1020,197	3188,033	2778,567
No. 2	17,06437	0,0000	343,8545	461,7862	131,7401	475,8032	585,8610	648,5784	616,5494	359,6868	1156,301	72,134	4020,625	692,191	1321,920	1803,661	677,755	2131,969	1661,812	1369,014
No. 3	14,48701	18,5433	0,0000	66,3731	484,8667	75,9168	309,3147	282,7715	250,9991	229,3018	725,838	405,133	3545,649	655,456	1189,541	1269,212	1180,421	1407,033	2842,435	2591,742
No. 4	11,71557	21,4892	8,1470	0,0000	669,4567	47,4720	120,7156	79,0628	98,3636	198,6941	403,214	618,162	3077,472	425,408	797,356	788,840	1700,909	912,504	3538,371	3173,762
No. 5	24,02547	11,4778	22,0197	25,8739	0,0000	550,5674	737,5182	949,6027	934,6619	468,1992	1658,853	144,378	4489,133	907,184	1791,409	2315,182	573,542	2850,321	1339,098	1012,680
No. 6	16,74089	21,8129	8,7130	6,8900	23,4642	0,0000	161,8739	185,9285	257,8140	278,1129	674,496	556,662	3432,464	580,749	1103,104	1072,779	1537,206	1220,633	3295,196	2934,316
No. 7	13,43037	24,2046	17,5874	10,9871	27,1573	12,7230	0,0000	59,7209	122,1225	184,2908	317,066	866,258	2941,487	177,661	509,700	521,828	2227,600	846,290	3979,120	3421,238
No. 8	12,14062	25,4672	16,8158	8,8917	30,8156	13,6356	7,7279	0,0000	60,4720	272,5157	174,464	920,785	2745,719	267,328	449,005	407,836	2317,397	537,249	4264,668	3777,528
No. 9	9,01235	24,8304	15,8430	9,9178	30,5722	16,0566	11,0509	7,7764	0,0000	148,6013	135,919	933,645	2755,972	200,636	385,366	424,188	2234,719	674,287	4132,103	3677,362
No. 10	9,81872	18,9654	15,1427	14,0959	21,6379	16,6767	13,5754	16,5081	12,1902	0,0000	494,205	641,305	3215,246	175,442	628,403	882,760	1645,792	1423,182	3093,145	2636,514
No. 11	17,82045	34,0044	26,9414	20,0802	40,7290	25,9711	17,8063	13,2085	11,6584	22,2307	0,000	1622,727	2500,539	281,089	159,203	111,904	3323,458	298,612	5497,797	4932,155
No. 12	23,59847	8,4932	20,1279	24,8629	12,0157	23,5937	29,4323	30,3444	30,5556	25,3240	40,283	0,000	4573,939	1151,070	1926,848	2391,630	372,507	2586,759	1298,761	1117,695
No. 13	54,53923	63,4084	59,5453	55,4750	67,0010	58,5872	54,2355	52,3996	52,4974	56,7031	50,005	67,631	0,000	2802,938	2580,214	2577,826	6500,100	2725,592	8710,994	8011,436
No. 14	14,33512	26,3095	25,6019	20,6254	30,1195	24,0987	13,3289	16,3502	14,1646	13,2455	16,766	33,927	52,943	0,000	202,811	439,678	2606,518	1046,235	4152,651	3496,323
No. 15	22,04624	35,3582	34,4897	28,2375	42,3250	33,2130	22,5766	21,1897	19,6307	25,0680	12,618	43,896	50,796	14,241	0,000	173,092	3745,283	639,558	5640,681	4927,060
No. 16	26,99875	42,4695	35,6260	28,0863	48,1163	32,7533	22,8436	20,1950	20,5958	29,7113	10,578	48,904	50,772	20,968	13,156	0,000	4461,171	257,427	6847,709	6117,194
No. 17	40,51707	26,0337	34,3573	41,2421	23,9487	39,2072	47,1975	48,1394	47,2728	40,5684	57,649	19,300	80,623	51,054	61,199	66,792	0,000	4689,788	469,070	552,055
No. 18	31,94053	46,1733	37,5104	30,2077	53,3884	34,9376	29,0911	23,1786	25,9670	37,7251	17,280	50,860	52,207	32,346	25,289	16,045	68,482	0,000	7433,865	6853,767
No. 19	56,46267	40,7653	53,3145	59,4842	36,5937	57,4038	63,0803	65,3044	64,2814	55,6161	74,147	36,038	93,333	64,441	75,104	82,751	21,658	86,220	0,000	81,040
No. 20	52,71211	37,0002	50,9092	56,3362	31,8226	54,1693	58,4913	61,4616	60,6412	51,3470	70,229	33,432	89,507	59,130	70,193	78,212	23,496	82,787	9,002	0,000

а

Descriptive Statistics for Cluster 1 (
Cluster contains 3 cases

Variable	Mean	Standard Deviation	Variance
Лом	127,500	0,70000	0,4900
ГБЖ	0,000	0,00000	0,0000
ЖЧ	80,000	0,00000	0,0000
ТЧ	0,000	0,00000	0,0000
Cr	0,034	0,01300	0,0002
Ni	0,091	0,02343	0,0005
Cu	0,165	0,01100	0,0001
Tc	54,500	10,33199	106,7500
Td	33,667	1,15470	1,3333
UR	221,430	7,55584	57,0907
T	1619,333	5,85947	34,3333

б

Descriptive Statistics for Cluster 2
Cluster contains 16 cases

Variable	Mean	Standard Deviation	Variance
Лом	138,600	11,51700	132,6413
ГБЖ	20,450	2,04026	4,1627
ЖЧ	54,375	5,43906	29,5833
ТЧ	0,000	0,00000	0,0000
Cr	0,035	0,01705	0,0003
Ni	0,071	0,01604	0,0003
Cu	0,113	0,04523	0,0020
Tc	55,125	3,28380	10,7833
Td	34,875	1,99583	3,9833
UR	265,413	7,86392	61,8412
T	1627,625	17,62905	310,7833

в

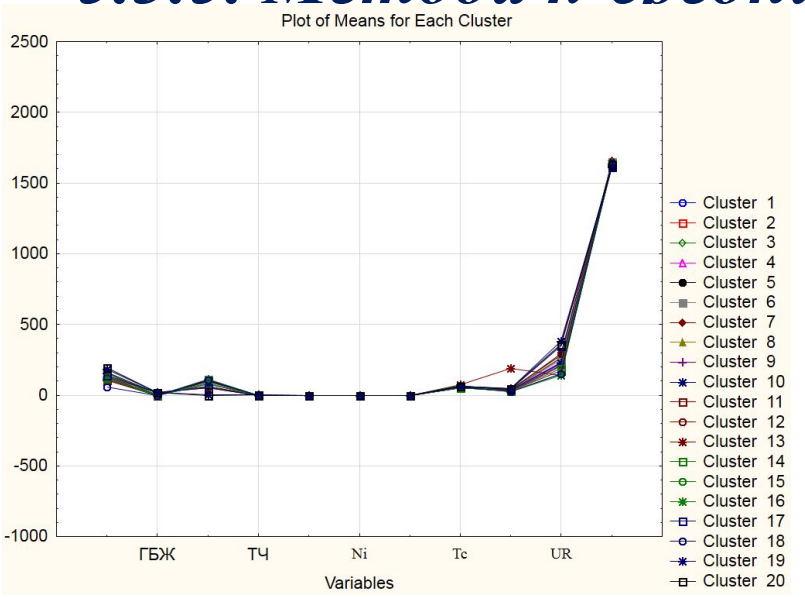
Descriptive Statistics for Cluster 20
Cluster contains 11 cases

Variable	Mean	Standard Deviation	Variance
Лом	196,386	15,75544	248,2340
ГБЖ	21,250	1,39696	1,9515
ЖЧ	0,000	0,00000	0,0000
ТЧ	4,745	10,56138	111,5427
Cr	0,037	0,01697	0,0003
Ni	0,092	0,01945	0,0004
Cu	0,175	0,04287	0,0018
Tc	60,545	4,00908	16,0727
Td	46,364	5,42720	29,4546
UR	358,027	12,48143	155,7862
T	1628,727	17,64705	311,4182

г

Рис. Результаты кластеризации: а – расстояние Евклида между кластерами; б, в, г – выборочные характеристики 1, 2 и 20 классов

5.3.5. Метода k-средних или алгоритм Лойда



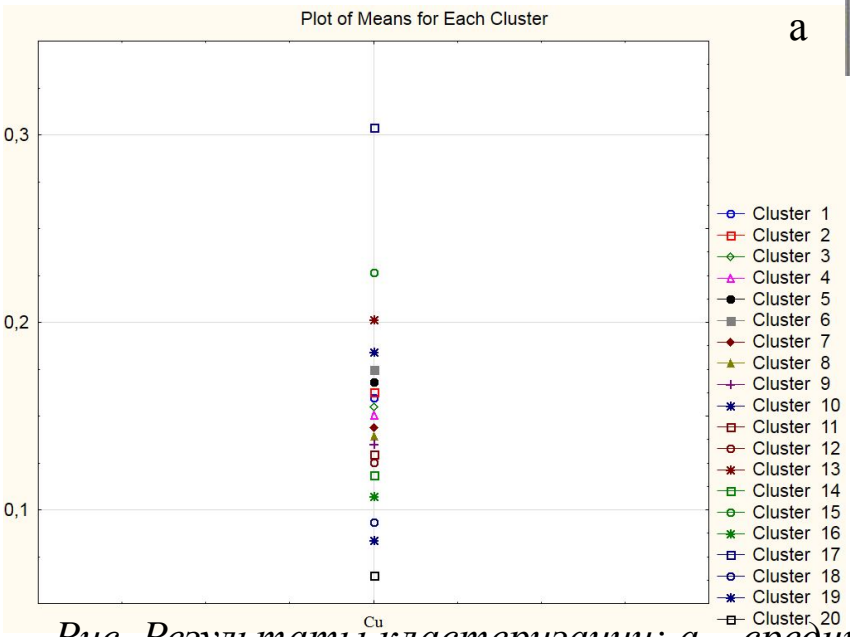
а

Case No.	Members of Cluster Number 1 and Distances from Respective Cluster contains 6 cases Distance
C_145	10,78673
C_147	6,78272
C_148	5,33395
C_151	7,03791
C_153	6,80468
C_154	5,19078

Case No.	Members of Cluster Number 3 (and Distances from Respective Cluster contains 4 cases Distance
C_179	2,623671
C_191	3,945241
C_225	4,337188
C_226	3,865659

Case No.	Members of Cluster Number 2 (and Distances from Respective Cluster contains 21 cases Distance
C_2	9,73546
C_6	4,48620
C_7	9,24443
C_9	5,29147
C_10	10,69679
C_12	8,37786
C_15	2,51544
C_16	5,50681
C_17	8,06678
C_20	6,19040
C_21	5,93368
C_23	7,57999
C_24	9,79409
C_26	9,87817
C_29	3,32884
C_30	7,32021
C_31	8,33931
C_33	10,02140
C_34	4,45520
C_35	9,17369
C_36	6,24298

в



б

Variable	Analysis of Variance (Spreadsheet1) Between SS	df	Within SS	df	F	signif. p
Лом	155716,8	19	13927,45	129	75,9102	0,000000
ГБЖ	15022,4	19	909,58	129	112,1334	0,000000
ЖЧ	213677,7	19	7504,17	129	193,3270	0,000000
ТЧ	392,9	19	2347,50	129	1,1364	0,323427
Cr	0,0	19	0,03	129	1,9987	0,012481
Ni	0,0	19	0,04	129	2,9993	0,000124
Cu	0,1	19	0,22	129	2,8134	0,000298
Tc	2581,4	19	3755,37	129	4,6670	0,000000
Td	30953,2	19	892,65	129	235,4297	0,000000
UR	795658,4	19	12879,02	129	419,4498	0,000000
T	15465,7	19	22485,18	129	4,6699	0,000000

г

Рис. Результаты кластеризации: а – средние по классам; б – значение среднего в каждом классе для одной переменной; в – состав групп; и – дисперсионный анализ групп

Задания к практическому занятию

Задание 1

Для исходных данных выполнить расчет матрицы коэффициентов сопоставимости по факторам и наблюдениям, матрицы парной корреляции, матрицы расстояний (способ расчета расстояния согласовать с ведущим преподавателем).

Задание 2

Выполнить кластеризацию факторов по методу корреляционных плеяд.

Задание 3

Выполнить кластеризацию наблюдений. Количество классов не должно быть менее 30. Выбор процедуры кластеризации согласовать с ведущим преподавателем.

Задание 4

Оформить результат предварительной обработки данных в виде письменного отчета. В отчете отобразить: исходные данные, матрицы мер сходства и их анализ, дендрит кластеризации, состав групп, новую матрицу исходных данных.

Контрольные вопросы

1. Определите цели и задачи кластеризации.
2. Мера сходства, принципы расчета и построения матрицы коэффициентов.
3. Приведите классификацию мер сходства.
4. Коэффициенты подобия и порядок расчета.
5. Коэффициент корреляции: назначение, способы расчета и оценки значимости.
6. Показатели расстояния : назначение, способы расчета и оценки значимости.
7. Классификация процедур кластеризации и принцип их проведения.
8. Алгоритм процедура кластеризации по расстоянию.
9. Метод вроцлавской таксономии.
10. Метод корреляционных плеяд.
11. Метод *k*-средних.
12. Особенности реализации процедур кластеризации в пакете Statistica.
13. Дендрит и его назначение в процедурах кластеризации.
14. Основные результаты кластерного анализа.