МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ГЕОГРАФИИ

Карпиченко Александр Александрович

доцент кафедры почвоведения и геоинформационных систем

3. КЛАСТЕРНЫЙ АНАЛИЗ

При проведении географических исследований, как правило, возникает проблема *объединения по сходству (кластеризация)* объектов, которые характеризуются *множеством признаков, выраженных в разных единицах измерения*. Для этой цели используется *кластерный анализ*. Наилучшие результаты кластерный анализ дает в сочетании с факторным, поскольку первый удобен для проведения классификации объектов, а второй – для ее обоснования, поскольку исследует связи между объектами.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Методологические особенности кластерного анализа сводятся к выявлению единой меры, охватывающей ряд исследуемых признаков. Эти признаки объединяются с помощью метрики (расстояния) в один кластер сходства группируемых объектов.

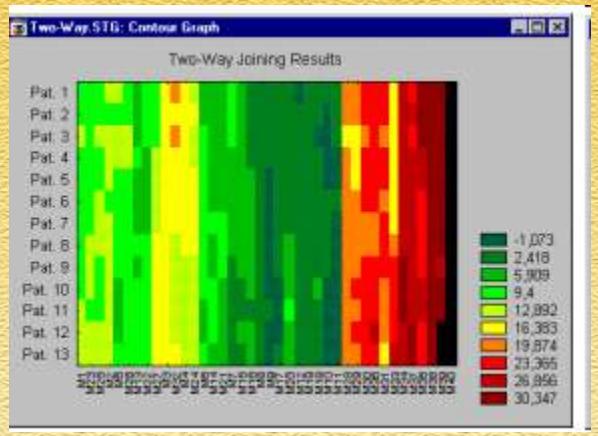
Кластерный анализ удобно использовать для классификации или типологии, при проведении районирования и т.д.

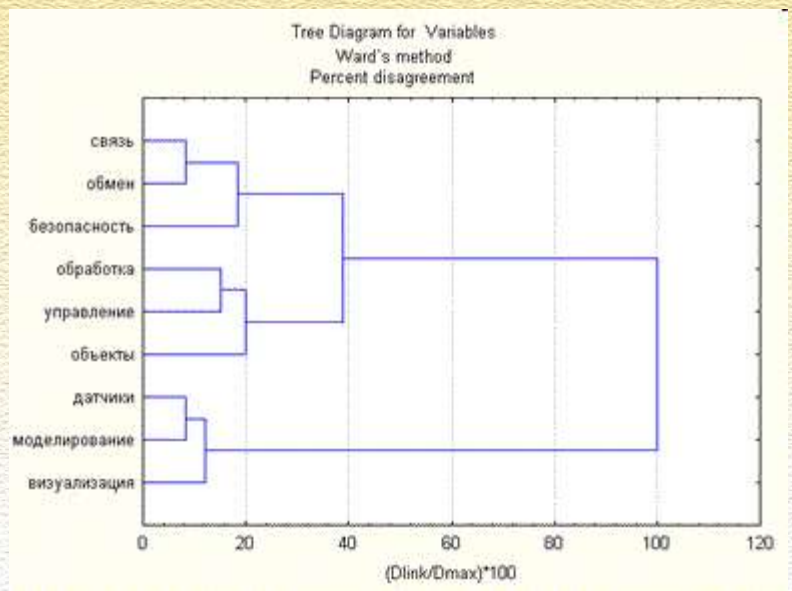
3. КЛАСТЕРНЫЙ АНАЛИЗ

Кластерный анализ – это общее название множества вычислительных процедур, используемых при создании классификации. *Главная цель кластерного анализа – нахождение групп схожих объектов в выборке данных*. Эти группы удобно называть кластерами. Не существует общепринятого определения термина *«кластер» (от cluster (англ.) — гроздь, скопление)*, однако считается, что кластеры обладают некоторыми свойствами, наиболее важными из которых являются плотность, дисперсия, размеры, форма и отделимость. Исходя из их свойств кластеры – это непрерывные области некоторого пространства с относительно высокой плотностью точек, отделенные от других таких же областей областями с относительно низкой плотностью точек.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Примеры результатов кластерного анализа





3. КЛАСТЕРНЫЙ АНАЛИЗ

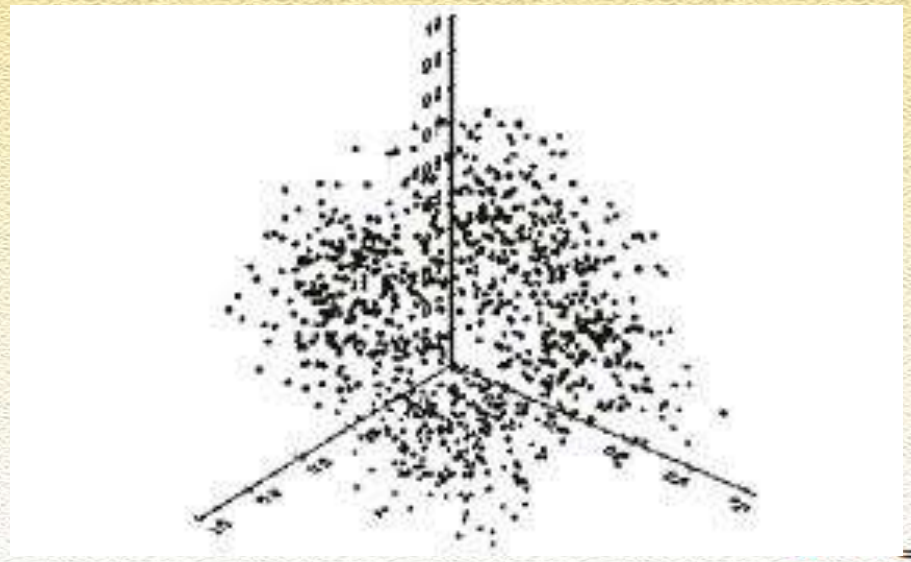
Применяя процедуры кластерного анализа, **всегда следует помнить**, что:

1. Многие методы кластерного анализа – довольно простые процедуры, которые, как правило, не имеют достаточного статистического обоснования (то есть большинство методов являются эвристическими).
2. Методы кластерного анализа разрабатывались для многих дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.
3. Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.
4. Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные, и эта структура может не совпадать с искомой «реальной».

3. КЛАСТЕРНЫЙ АНАЛИЗ

Суть применения кластерного анализа в том, что состояние любого объекта может быть описано с использованием *многомерного признака*, или *многомерной случайной величины* (x_1, x_2,…, x_n).

Исследование нескольких аналогичных объектов обязывает проводить разбиение совокупности объектов на однородные группы, т.е. провести их классификацию по сходству признаков (x_1, x_2 …), при этом объекты из одного класса должны быть сходными по характеризующим их признакам.



3. КЛАСТЕРНЫЙ АНАЛИЗ

В зависимости от специальности и природы используемых методов исследователи называют классификацию многомерных наблюдений как *распознавание образов с обучением /учителем/* (численной таксономией), *кластер-анализом без обучения, дискриминантным* анализом.

Подобные способы реализованы в ряде ГИС-программ, где могут применяться для картографирования и дешифрирования аэро- и космоснимков.

3. КЛАСТЕРНЫЙ АНАЛИЗ

При относительной формализации методов кластерного анализа они носят эвристический (теоретический) характер, реализуют принцип здравого смысла. Для оценки сходства объектов по ряду признаков используют три типа мер:

- *коэффициент подобия* – для группировки объектов и признаков, если уровни показателей являются действительно целыми числами;
- *коэффициенты связи* – чаще применяются для группировки признаков с использованием коэффициента корреляции;
- *показатели расстояния* – характеризуют степень взаимной удаленности признаков и применяются в основном для кластеризации объектов; признаки объектов должны быть независимыми, что предварительно можно уточнить с помощью корреляционного анализа.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Многомерное наблюдение может быть интерпретировано геометрически в виде точки в многомерном пространстве. *Геометрическая близость точек в пространстве означает близость физических состояний объектов, их однородность.* Решающим в интерпретации остается выбор масштаба метрики, т.е. задание расстояния между объектами, которые объединяют или разъединяют объекты. В результате разбиения объектов на группы по сходству признаков образуются *кластеры (таксоны, образы)*. Необходимость разбиений совокупности объектов на однородные группы возникает при проведении социально-экономических, землеустроительных, географических исследований и т. д.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Выбор метрики (меры близости) является важнейшим моментом исследования, который определяет окончательный вариант разбиения объектов на группы. Это зависит от цели исследования, физической и статистической природы вектора наблюдений, полноты априорных сведений о характере вероятностного распределения.

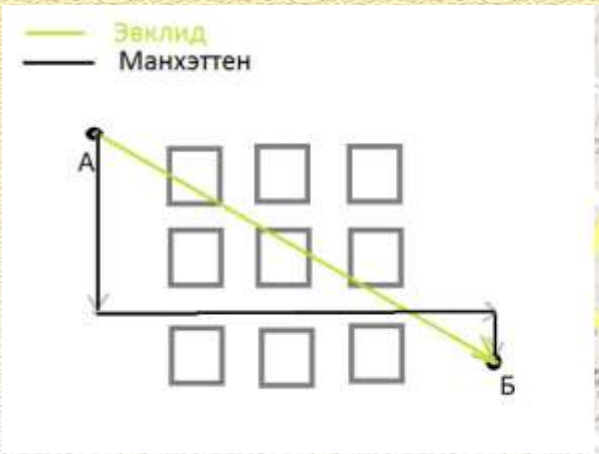
В задачах кластер-анализа широко используются следующие метрики: Эвклида, манхэттенское расстояние, Чебышева, Минковского, Хемминга, меры близости, задаваемые потенциальной функцией. Эвклидова метрика наиболее употребительна.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Обычно среднее Эвклидовое расстояние рассчитывается по формуле:

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

где m – число признаков x; Zk_j, Zl_j – стандартизированные значения признака j для k и l объектов соответственно.



3. КЛАСТЕРНЫЙ АНАЛИЗ

Расчет упрощается, если в качестве метрики использовать l_1-*норму*:

$$d_{kl} = \sqrt{\sum_{j=1}^{m} (Z_{k_j} - Z_{l_j})^2} \qquad (3.3)$$

Эти метрики применяются в следующих случаях:

- наблюдения x извлекаются из генеральных совокупностей, описываемых многомерным нормальным законом с ковариационной матрицей (совместное изменение двух признаков), где компоненты x взаимно независимы и имеют одинаковую дисперсию;
- компоненты $x_1, x_2, \ldots, x_p$ вектора наблюдений x однородны по своему физическому смыслу и все важны;
- факторное пространство совпадает с геометрическим; понятие близости объектов соответственно совпадает с понятием геометрической близости в этом пространстве.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Метод дендритов. Исследуемые объекты, разделенные на кластеры, можно изобразить в виде дендрограммы, которая представляет собой графическое изображение матрицы расстояний или сходства. Такой анализ объектов исследования носит название метода дендритов. Имея *n* объектов, можно построить большое количество дендрограмм, которые соответствуют избранной процедуре кластеризации. Для конкретной матрицы расстояний или сходства существует только одна дендрограмма.

3. КЛАСТЕРНЫЙ АНАЛИЗ

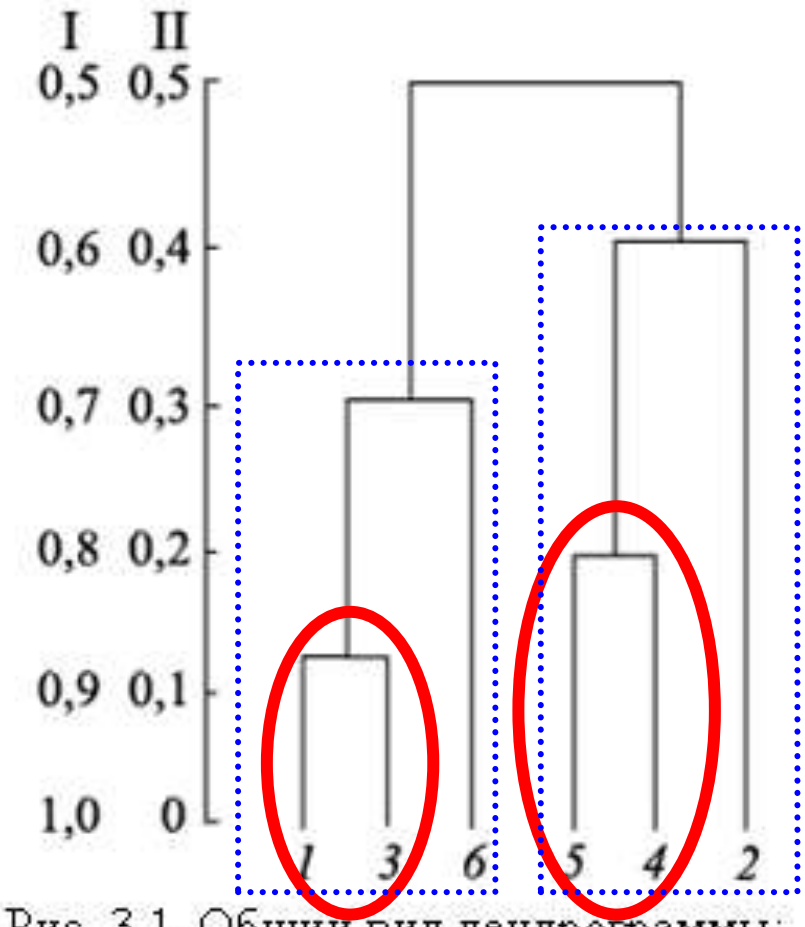
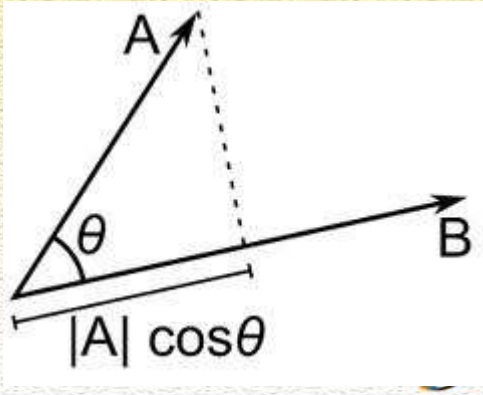


Рис. 3.1. Общий вид дендрограммы: I – сходство, II – расстояние

Представим дендрограмму с шестью объектами (n = 6) (рис. 3.1). Объекты 1 и 3 наиболее близки, т. е. наименее удалены друг от друга, поэтому объединяются в один кластер на уровне сходства, равном 0,9 (образуют 1-й шаг). Объекты 4 и 5 объединяются при уровне сходства 0,8 (2-й шаг). На 3-м и 4-м шагах процесса образуются кластеры 1, 3, 6 и 5, 4, 2, соответствующие уровню сходства соответственно 0,7 и 0,6. Окончательно все объекты группируются в один кластер при уровне сходства 0,5.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Вид дендрограммы зависит от выбора меры сходства или расстояния и метода кластеризации. Например, разработаны алгоритмы кластерного анализа, позволяющие проводить классификацию (группировку) многомерных наблюдений (строк и столбцов матрицы *x*) с помощью следующих мер сходства: выборочного коэффициента корреляции, модуля выборочного коэффициента корреляции, косинуса угла между векторами, модуля косинуса угла между векторами, эвклидова расстояния и т. д.



3. КЛАСТЕРНЫЙ АНАЛИЗ

Решение задач классификации объектов с использованием кластерного анализа проводится в определенной последовательности. Многомерный анализ делится на три этапа:

- составляется таблица исходной информации с указанием объектов и их признаков;
- проводится нормализация исходной информации с использованием среднего квадратического отклонения;
- по нормализованным данным рассчитывается метрика, строится дендрограмма и проводится содержательная интерпретация полученных результатов.

3. КЛАСТЕРНЫЙ АНАЛИЗ

На первом этапе при формировании таблицы выбор объекта зависит от места и масштаба исследования. Каждый объект должен быть пространственно локализован и одного ранга (уровня). Показатели должны отражать существенные черты или свойства исследуемых объектов и характеризовать их всесторонне.

На втором этапе нормализация значений исходных показателей по объектам проводится потому, что исходные данные выражены обычно в разных единицах измерения и проводить между ними арифметические действия невозможно без перевода их в безразмерные единицы.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Наиболее распространенный способ нормализации показателей проводится с использованием среднего квадратического отклонения (σ) по формуле:

$$\hat{Z}_{ij} = (Z_{ij} - \bar{Z}_{ij}) / \sigma_j \qquad (3.7);$$

$$\sigma_j = \sqrt{\frac{\sum (Z_{ij} - \bar{Z}_{ij})^2}{N_j}}, \qquad (3.8)$$

где $\hat{Z}_{ij}$ – нормализованная безразмерная величина; Z_{ij} – индивидуальные значения по столбцам матрицы; $\bar{Z}_{ij}$ – среднее значение по столбцам матрицы; σ_j – среднее квадратическое отклонение по столбцам; N_j – объем выборки по столбцам.

Составляется матрица нормализованных показателей.

3. КЛАСТЕРНЫЙ АНАЛИЗ

На третьем этапе по нормализованным показателям рассчитывается метрика по одному из способов, учитывая условия задачи. Классификацию объектов производят приемами таксономического или факторного анализа.

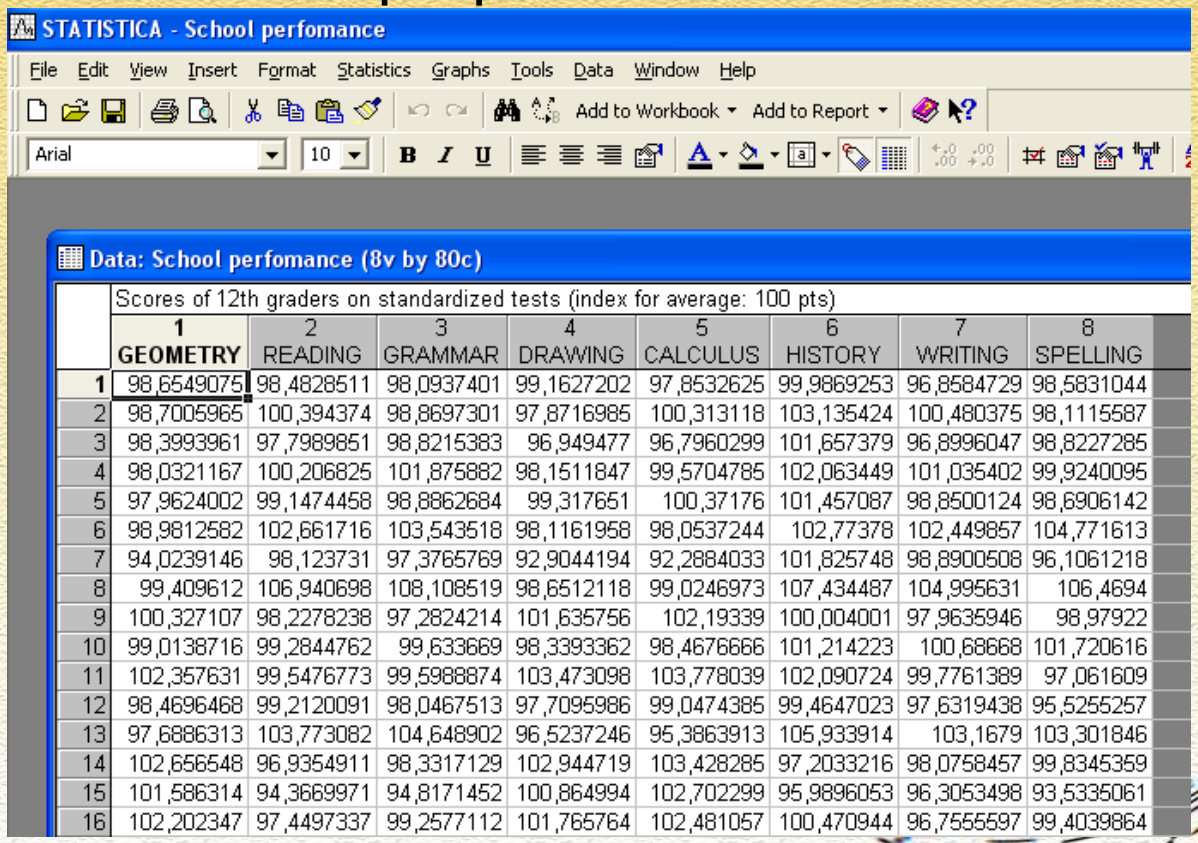
При количестве координат (показателей) в многомерном пространстве более трех графически интерпретировать таксономические расстояния невозможно. Поэтому таксономические расстояния определяют на основе функции расстояний. Чаще всего используется эвклидова метрика.

На основе матрицы таксономических расстояний производится группировка объектов с использованием разных приемов, из них наиболее распространенные – вроцлавская таксономия, дендрограмма Берри, метод дендритов.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Выполнение кластерного анализа в программе StatSoft STATISTICA 6.0.

Рассмотрим на примере одного из учебных файлов, поставляемых вместе с программой.



Scores of 12th graders on standardized tests (index for average: 100 pts)

	1 GEOMETRY	2 READING	3 GRAMMAR	4 DRAWING	5 CALCULUS	6 HISTORY	7 WRITING	8 SPELLING
1	98,6549075	98,4828511	98,0937401	99,1627202	97,8532625	99,9869253	96,8584729	98,5831044
2	98,7005965	100,394374	98,8697301	97,8716985	100,313118	103,135424	100,480375	98,1115587
3	98,3993961	97,7989851	98,8215383	96,949477	96,7960299	101,657379	96,8996047	98,8227285
4	98,0321167	100,206825	101,875882	98,1511847	99,5704785	102,063449	101,035402	99,9240095
5	97,9624002	99,1474458	98,8862684	99,317651	100,37176	101,457087	98,8500124	98,6906142
6	98,9812582	102,661716	103,543518	98,1161958	98,0537244	102,77378	102,449857	104,771613
7	94,0239146	98,123731	97,3765769	92,9044194	92,2884033	101,825748	98,8900508	96,1061218
8	99,409612	106,940698	108,108519	98,6512118	99,0246973	107,434487	104,995631	106,4694
9	100,327107	98,2278238	97,2824214	101,635756	102,19339	100,004001	97,9635946	98,97922
10	99,0138716	99,2844762	99,633669	98,3393362	98,4676666	101,214223	100,68668	101,720616
11	102,357631	99,5476773	99,5988874	103,473098	103,778039	102,090724	99,7761389	97,061609
12	98,4696468	99,2120091	98,0467513	97,7095986	99,0474385	99,4647023	97,6319438	95,5255257
13	97,6886313	103,773082	104,648902	96,5237246	95,3863913	105,933914	103,1679	103,301846
14	102,656548	96,9354911	98,3317129	102,944719	103,428285	97,2033216	98,0758457	99,8345359
15	101,586314	94,3669971	94,8171452	100,864994	102,702299	95,9896053	96,3053498	93,5335061
16	102,202347	97,4497337	99,2577112	101,765764	102,481057	100,470944	96,7555597	99,4039864

3. КЛАСТЕРНЫЙ АНАЛИЗ

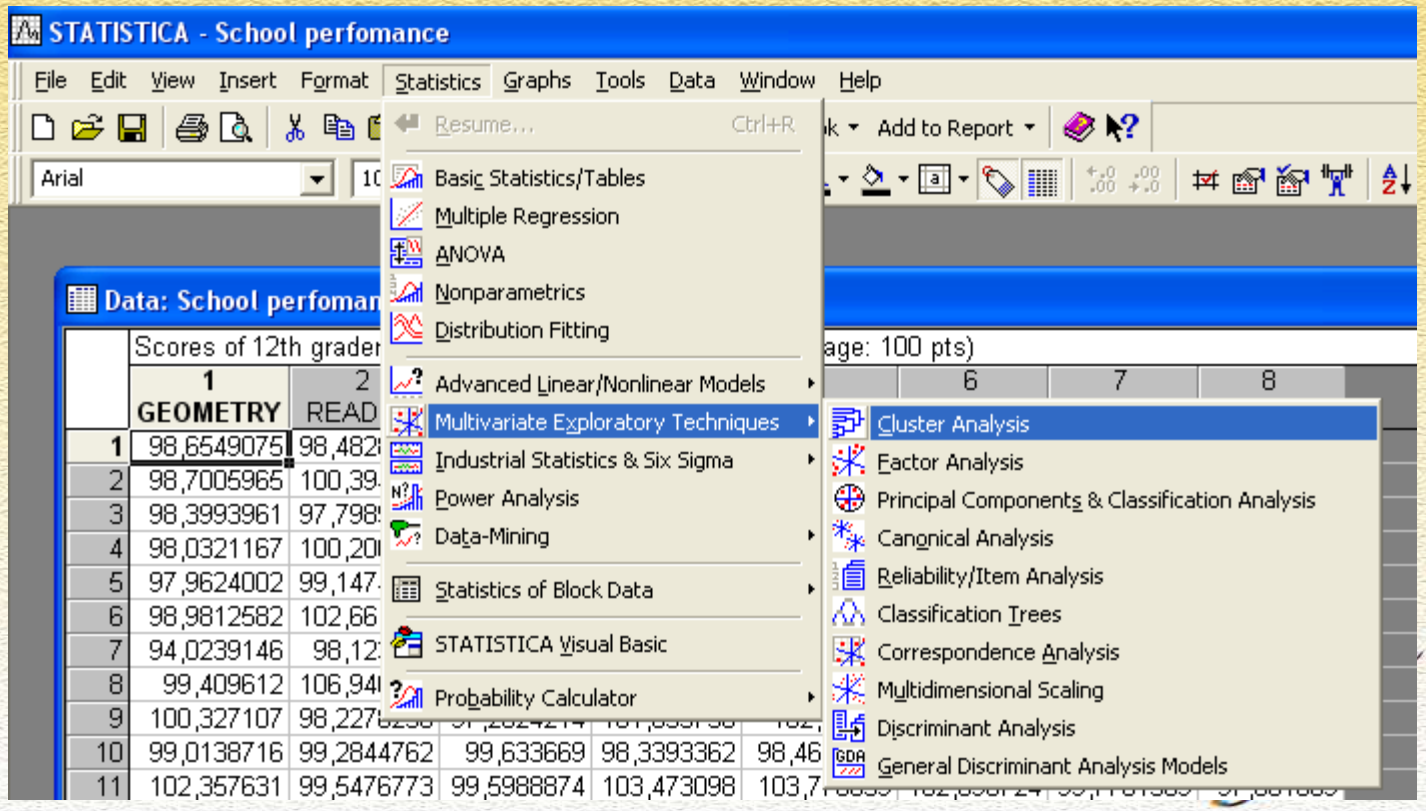
В файле приведены результаты школьников по стандартизированным тестам по 8 предметам (среднее значение 100): GEOMETRY (Геометрия), READING (Чтение), GRAMMAR (Грамматика), DRAWING (Рисование или Черчение), CALCULUS (Вычисление), HISTORY (История), WRITING (Письмо), SPELLING (Орфография).

Data: School perfomance (8v by 80c)

	Scores of 12th graders on standardized tests (index for average: 100 pts)							
	1 GEOMETRY	2 READING	3 GRAMMAR	4 DRAWING	5 CALCULUS	6 HISTORY	7 WRITING	8 SPELLING
1	98,6549075	98,4828511	98,0937401	99,1627202	97,8532625	99,9869253	96,8584729	98,5831044
2	98,7005965	100,394374	98,8697301	97,8716985	100,313118	103,135424	100,480375	98,1115587
3	98,3993961	97,7989851	98,8215383	96,949477	96,7960299	101,657379	96,8996047	98,8227285
4	98,0321167	100,206825	101,875882	98,1511847	99,5704785	102,063449	101,035402	99,9240095
5	97,9624002	99,1474458	98,8862684	99,317651	100,37176	101,457087	98,8500124	98,6906142
6	98,9812582	102,661716	103,543518	98,1161958	98,0537244	102,77378	102,449857	104,771613
7	94,0239146	98,123731	97,3765769	92,9044194	92,2884033	101,825748	98,8900508	96,1061218
8	99,409612	106,940698	108,108519	98,6512118	99,0246973	107,434487	104,995631	106,4694
9	100,327107	98,2278238	97,2824214	101,635756	102,19339	100,004001	97,9635946	98,97922
10	99,0138716	99,2844762	99,633669	98,3393362	98,4676666	101,214223	100,68668	101,720616
11	102,357631	99,5476773	99,5988874	103,473098	103,778039	102,090724	99,7761389	97,061609

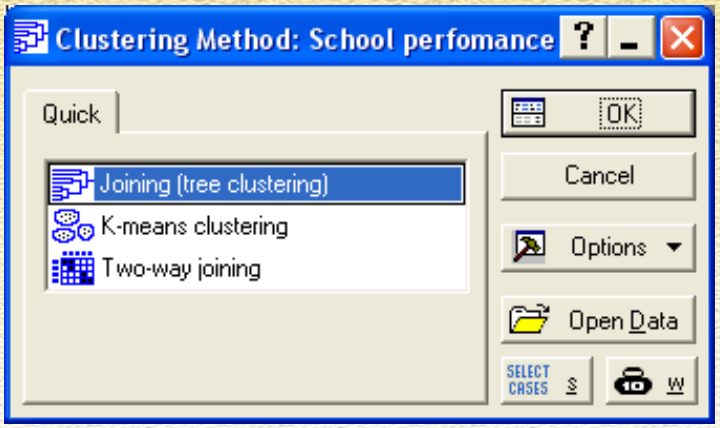
3. КЛАСТЕРНЫЙ АНАЛИЗ

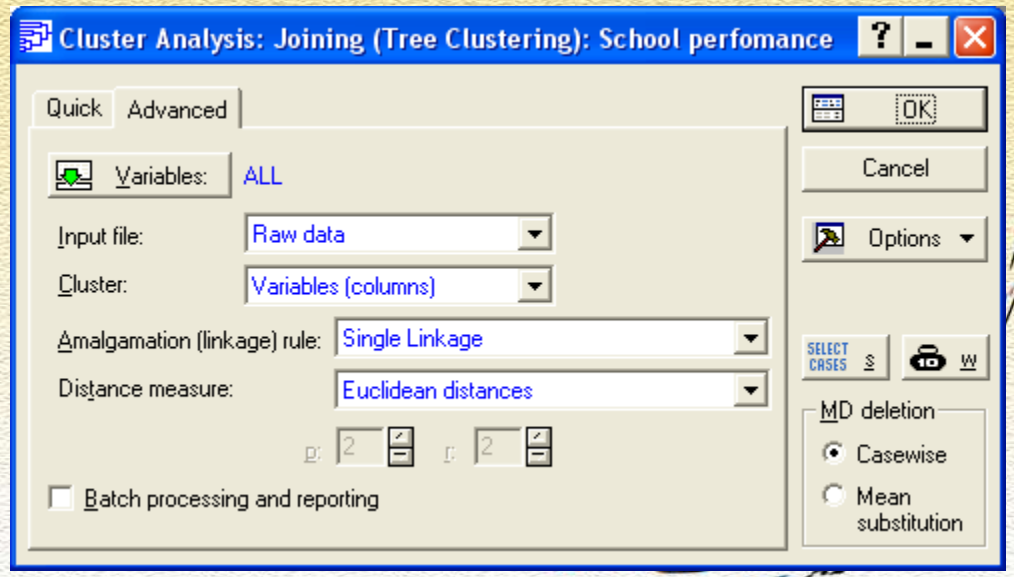
Проводим кластерный анализ, для чего выполняем следующие действия: *Statistics – Multivariate Exploratory Techniques – Cluster analysis – Joining tree clustering* (оно выбрано по умолчанию).



3. КЛАСТЕРНЫЙ АНАЛИЗ

. В следующем диалоговом окне выбираем закладку *Advanced* – жмем на кнопку *Variables*, там отмечаем все переменные (выделяем левой клавишей мыши при нажатой клавише *Shift* или просто кликаем на кнопке *Select All*) – *OK*. В полях *Input file* ставим *Raw data*, *Cluster – Cases (rows)*, *Amalgamation (linkage) rule – Single Linkage*, *Distance Measure – Euclidean distances.*





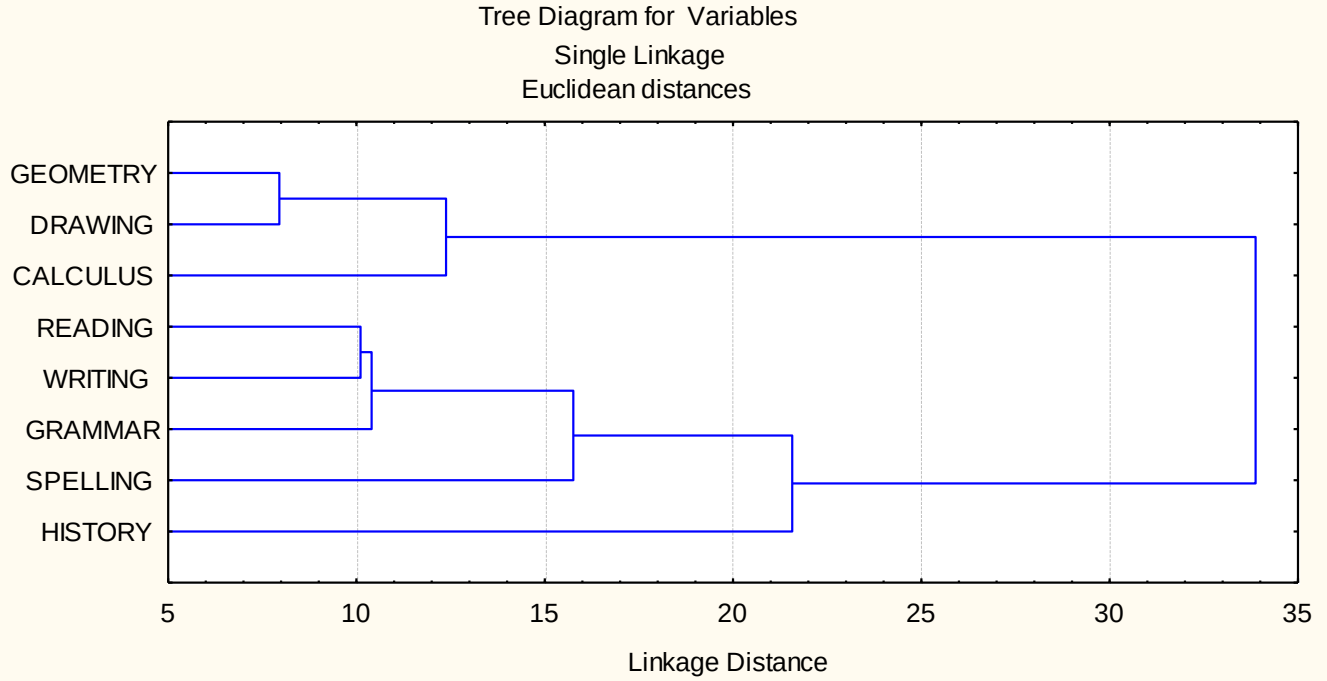
3. КЛАСТЕРНЫЙ АНАЛИЗ

В программе STATISTICA реализованы следующие методы кластеризации: иерархический агломеративный (объединительный) метод – joining (tree clustering), итеративный метод k-средних (k-means clustering) и двухвходовое объединение (two-way joining).

joining (tree clustering)

В агломеративных методах происходит последовательное объединение наиболее близких объектов в один кластер. Процесс такого последовательного объединения можно показать на графике в виде дендрограммы, или дерева объединения.

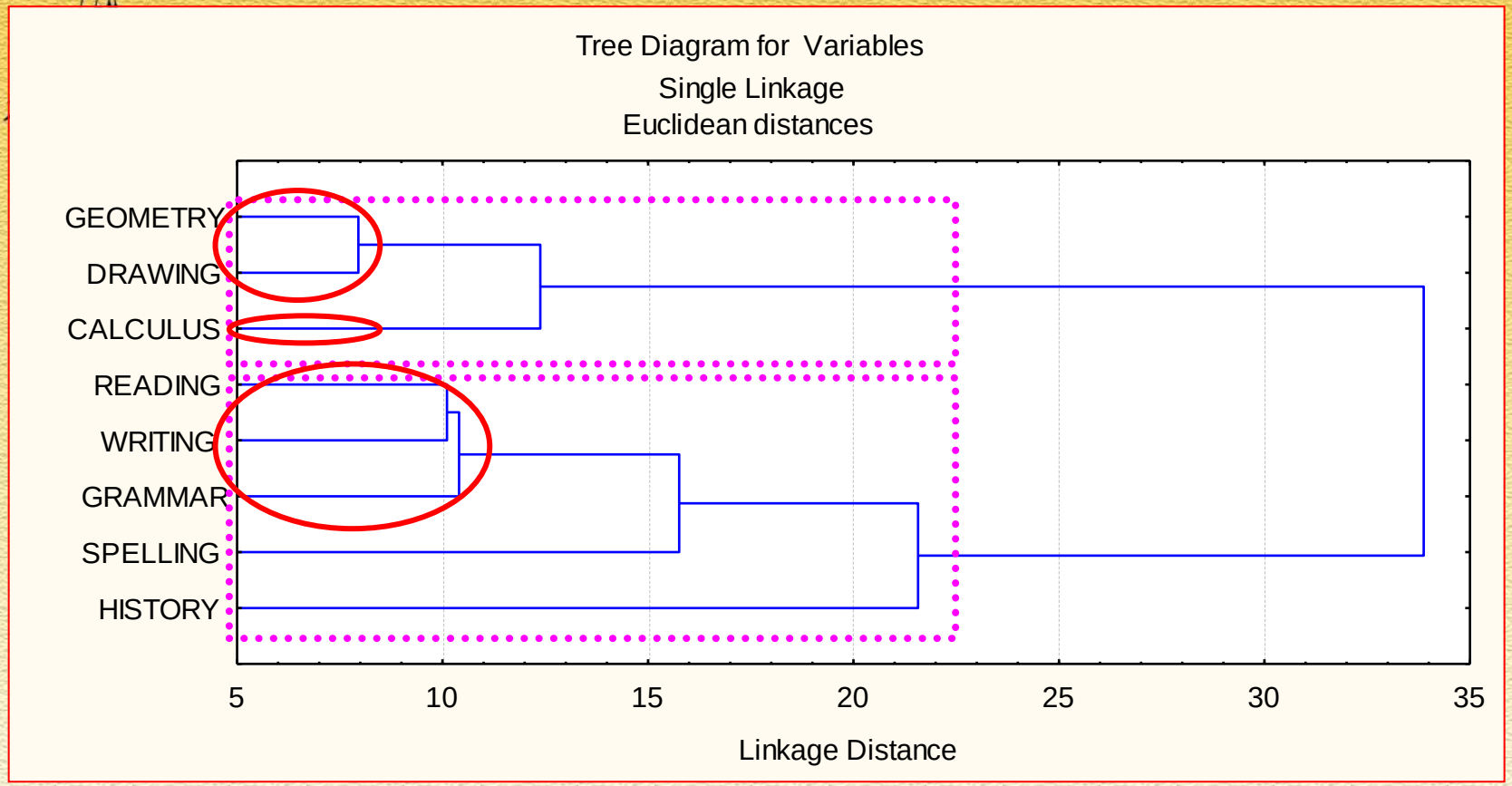
3. КЛАСТЕРНЫЙ АНАЛИЗ



Variable	Means and Standard Deviations	
	Mean	Std.Dev.
GEOMETRY	99,8523	1,798619
READING	99,7669	3,295847
GRAMMAR	99,6856	3,512469
DRAWING	99,7891	2,072439
CALCULUS	100,0655	2,337676
HISTORY	101,9269	3,490726
WRITING	99,8200	3,377652
SPELLING	99,9867	3,757352

Variable	Euclidean distances (School perfomance)							
	GEOMETRY	READING	GRAMMAR	DRAWING	CALCULUS	HISTORY	WRITING	SPELLING
GEOMETRY	0,0	33,9	35,3	8,0	12,4	39,8	34,4	36,8
READING	33,9	0,0	10,4	35,1	36,2	21,6	10,1	17,2
GRAMMAR	35,3	10,4	0,0	36,7	37,6	23,4	13,7	18,7
DRAWING	8,0	35,1	36,7	0,0	12,4	41,3	35,8	37,9
CALCULUS	12,4	36,2	37,6	12,4	0,0	41,2	36,2	38,6
HISTORY	39,8	21,6	23,4	41,3	41,2	0,0	22,9	24,9
WRITING	34,4	10,1	13,7	35,8	36,2	22,9	0,0	15,8
SPELLING	36,8	17,2	18,7	37,9	38,6	24,9	15,8	0,0

3. КЛАСТЕРНЫЙ АНАЛИЗ



Результаты показывают, что предметы разделены на два больших кластера (группы), которые условно можно обозначить как «гуманитарные» и «естественные». Можно выделить иерархию по евклидовым расстояниям. Например, геометрия и черчение ближе друг к другу, чем к сходному с ними вычислению. Подобное наблюдается и в блоке гуманитарных дисциплин.

Контрольные вопросы по лекции

Контрольные вопросы:

- Какие условия применения кластерного анализа?
- Какие метрики используются при кластерном анализе?
- Зачем необходимо нормализовывать данные перед проведением кластерного анализа?
- Сколько этапов включает в себя кластерный анализ?
- Почему кластерный анализ удобен для классификации по множеству признаков?