ЛАБОРАТОРНАЯ РАБОТА №4

Кластерный анализ с использованием возможностей ПО Statistica 12.0

Цель работы: провести иерархическую классификацию данных кадастровой оценки сельскохозяйственных земель методами одиночной связи и Варда, используя Евклидово расстояние; провести классификацию переменных этими же методами; выполнить два варианта классификации объектов методом k-средних, задав в первом случае 3 класса, во втором - 5 классов.

Исходные данные; материалы кадастровой оценки сельскохозяйственного предприятия.

Выполнение работы

Войдите в STATISTICA и создайте новый файл данных для своего варианта . Импортируйте данные кадастровой оценки из файла excel с результатами оценки агрохимические показатели рабочих участков (лист Т2 книги excel. Сохраните данные.

Щелкнув на кнопке **Анализ (Statistics),** откройте меню и затем выберите раздел **Многомерный разведочный анализ** (**Multivariate Exploratory Technique**), затем перейдите в раздел **Кластерный анализ (Cluster Analysis) (рис 4.1).**

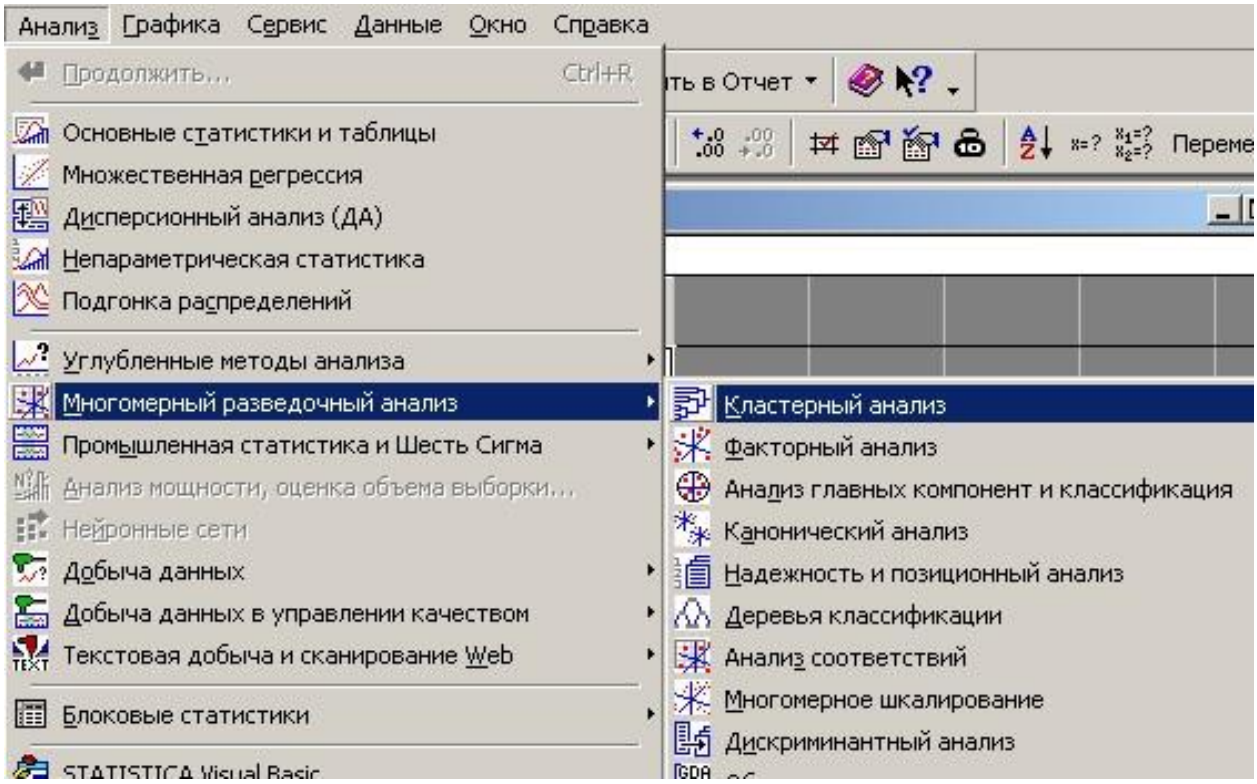

Рис 4.1.Меню кластерного анализа

Кластерный анализ – это группа методов, используемых для классификации объектов в относительно однородные группы (кластеры). Эти методы не являются строгими со статистической точки зрения. Кластерный анализ используется обычно на начальной стадии исследования, когда не существует еще гипотез относительно классов, в которые объединяются объекты. Выделяют аггломеративные и итеративные дивизивные методы кластерного анализа. Аггломеративные методы кластеризации – это иерархические методы, при которых на начальном этапе каждый объект находится в отдельном кластере.

На следующих этапах происходит объединение объектов в более крупные кластеры на основании понижения некоторого порога, например, увеличения расстояния между объектами. Иными словами, чем выше уровень агрегации, тем меньше сходства между членами в соответствующем классе. Итеративные дивизивные методы кластеризации состоят в том, что выполняется разбиение объектов, объединенных в один или несколько крупных кластеров, на фиксированное число кластеров, как правило, более мелких. При этом образуются новые кластеры так, чтобы они были настолько различны, насколько это возможно.

Выберите пункт **Иерархическая классификация** (**Joining –tree clustering)** дендрограммы. Нажмите **ОК.** Для выполнения второй части задания нужно будет в этом же меню выбрать пункт **Кластеризация методом - kсредних** (**K-means clustering**)) (рис 4.2).

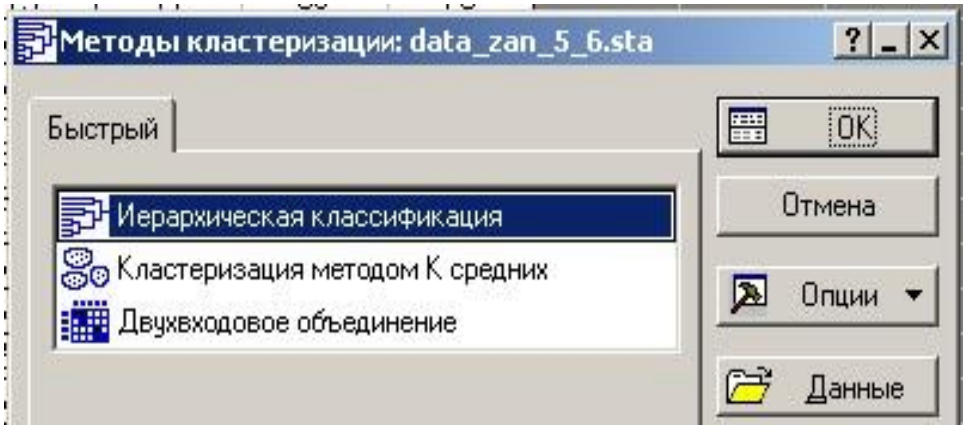

Рис 4.2. Выбор метода кластерного анализа.

ИЕРАРХИЧЕСКАЯ КЛАССИФИКАЦИЯ. Выберите закладку **Дополнительно (Advanced).** Выберите переменные (**Variables**), по которым будет проводиться анализ (C, PHS, IL, G, V). Обратите внимание, что **Файл данных** (**Input file**) может содержать данные как в исходном виде, так и в виде

матрицы расстояний (distance matrix). В поле **Объекты (Cluster)** выберите **Наблюдения-строки (Cases -rows)** (рис 4.3).

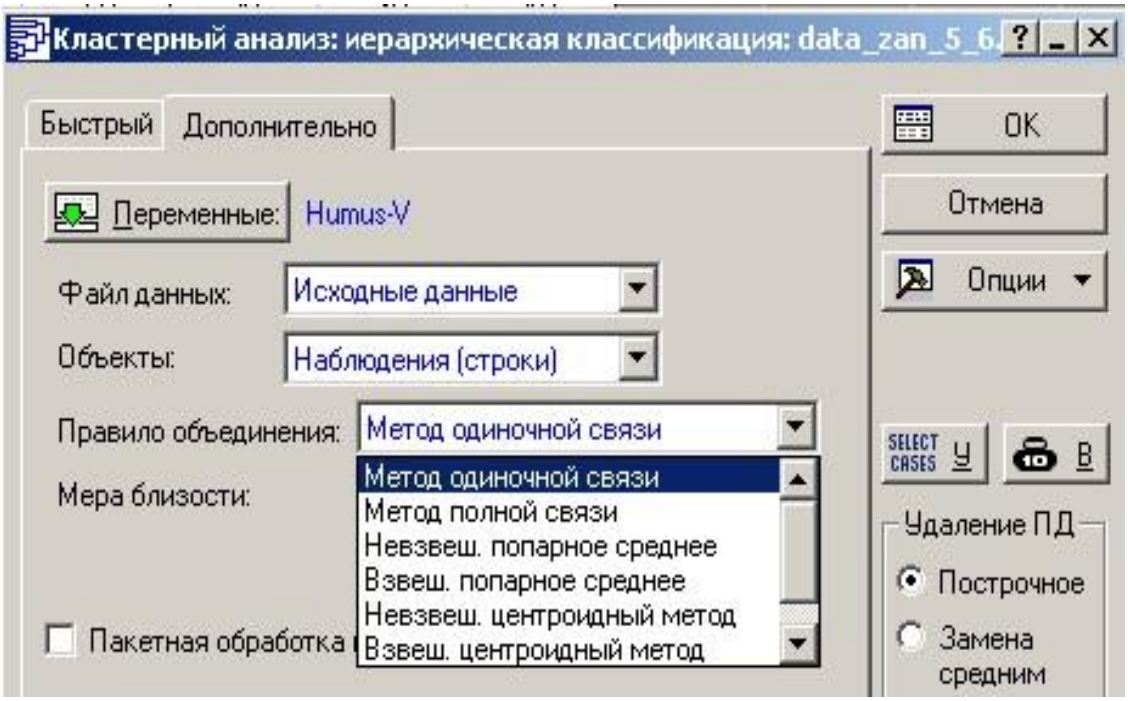

Рис 4.3. Настройка параметров кластерного анализа.

Выберите правило объединения (**Amalgamation –linkage rule**) и подходящую **Меру близости** между объектами (**Distance measure**) (рис 4.4).

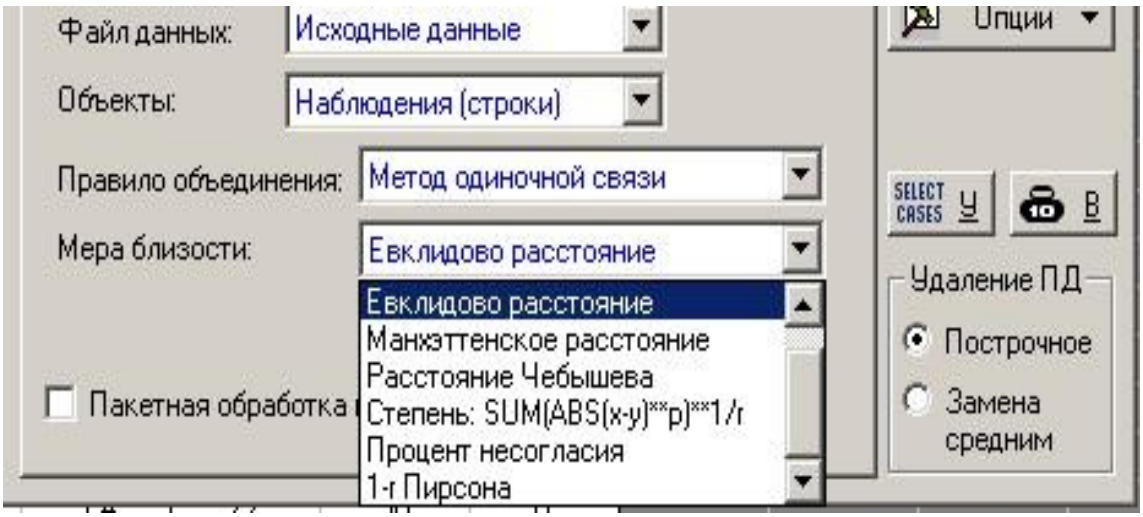

Рис 4.4. Настройка параметров кластерного анализа.

В таблице 4.1. приведены возможные варианты перевода названий методов объединения и мер расстояния.

Таблица 4.1. Перевод терминов с англоязычной версии программы

Joining rule- Методы объединения		**Distance measure – Меры расстояния**	
Single linkage	Метод одиночной связи (ближайшего соседа)	Squared Euclidean distances	Квадрат Евклидова расстояния
Complite linkage	Метод полной связи (дальнего соседа)	Euclidean distances	Евклидово расстояние
Unweighted pair group average	Невзвешенный метод “средней связи”, невзвешенное попарное среднее	City (Manchattan)-block	Манхэттенское расстояние
		Chebyshev distance metric	Расстояние Чебышева
Weighted pair group average	Взвешенный метод средней связи	Power	Степенное
Weighted centroid pair group (median)	Взвешенный центроидный метод	Percent disagreement	Процент несовпадений (используется для качественных признаков)
Ward method	Метод Уорда (Варда)	Pearson r	Коэффициент корреляции (1-r Пирсона)

Проведите иерархический кластерный анализ **Методом одиночной связи** (**Single Linkage**) с использованием **Евклидового расстояния** (**Euclidean distances**). Задав начальные установки, нажмите **ОК.**

Евклидово расстояние – это геометрическое расстояние в многомерном пространстве, то есть аналог физического расстояния. Метод одиночной связи (ближайшего соседа) предполагает, что расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в сравниваемых кластерах. В результате формируются кластеры, представленные длинными "цепочками" объектов.

Следующая панель (рис 4.5) дает информацию о выбранных ранее условиях (число случаев, число переменных, число пропусков, способ присоединения и мера близости).

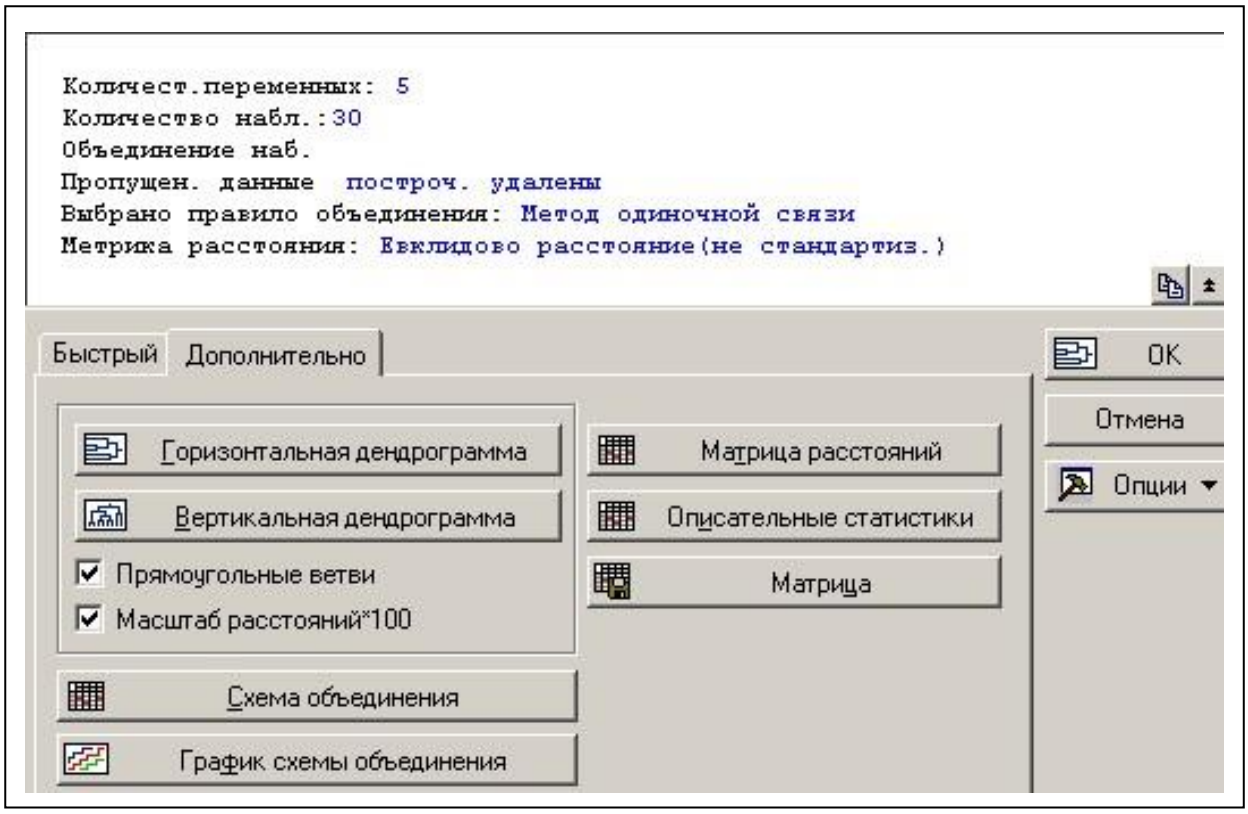

Рис 4.5. Выбор параметров представления результатов анализа.

Появляется возможность построить горизонтально (**Horizontal hierarchical tree plot**) или вертикально (**Vertical icicle plot**) расположенную дендрограмму. Нажмите соответствующую кнопку, чтобы построить каждую из дендрограмм. Посмотрите рисунки.

Для продолжения анализа в нижнем левом углу нажмите на свернутую панель кластерного анализа (**Joining results**). По умолчанию дендрограмма строится с ветвями, соединяющимися под прямыми углами **Прямоугольные ветви** (**Rectangular branches**). Посмотрите, что получится, если значок выбора снять (дерево получится с острыми углами). Вторая галочка позволяет масштабировать ось расстояния на рисунке дендрограммы, то есть перейти к процентам от максимального расстояния (**Scale tree to dlink/dmax *100%**).

Постройте вертикально расположенную дендрограмму с прямоугольными ветвями и с масштабированным расстоянием (рис 4.6).

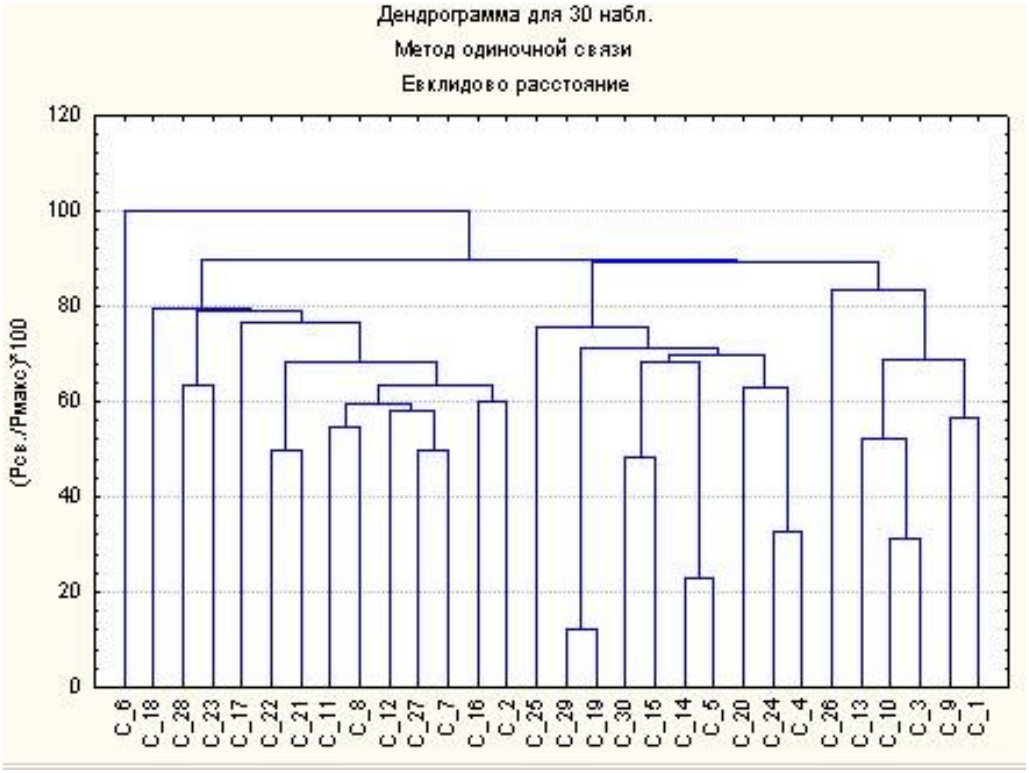

Рис 4.6. Вертикальная дендрограмма.

На графике по оси абсцисс отложены объекты (наблюдения). В данном случае – это 30 горизонтов, соответствующие 5 разрезам дерново-подзолистой почвы. По оси ординат отложено Евклидово расстояние между объектами и группами объектов, рассчитанное по свойствам объектов (наблюдений). В группы объединяются объекты (и/или их группы), находящиеся на самом близком расстоянии.

Дважды щелкнув по графику можно перейти в режим оформления, где можно заменить номера объектов (наблюдений) на их имена. Для этого в появившемся меню выберите вкладку Единицы, заданные пользователем (**Custom Units**) (рис 4.7). Для сохранения имени горизонта в строке используйте клавишу **Enter**. Замените порядковые номера наблюдений названиями. Нажмите **OK**. Сохраните график в файле результатов Excel.

Проведите иерархический кластерный анализ методом Варда с использованием Евклидового расстояния. Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод Варда минимизирует сумму квадратов для любых двух кластеров, которые могут быть сформированы на каждом шаге. При использовании данного метода получаются кластеры малого размера. Результаты сохраните в файле Excel.

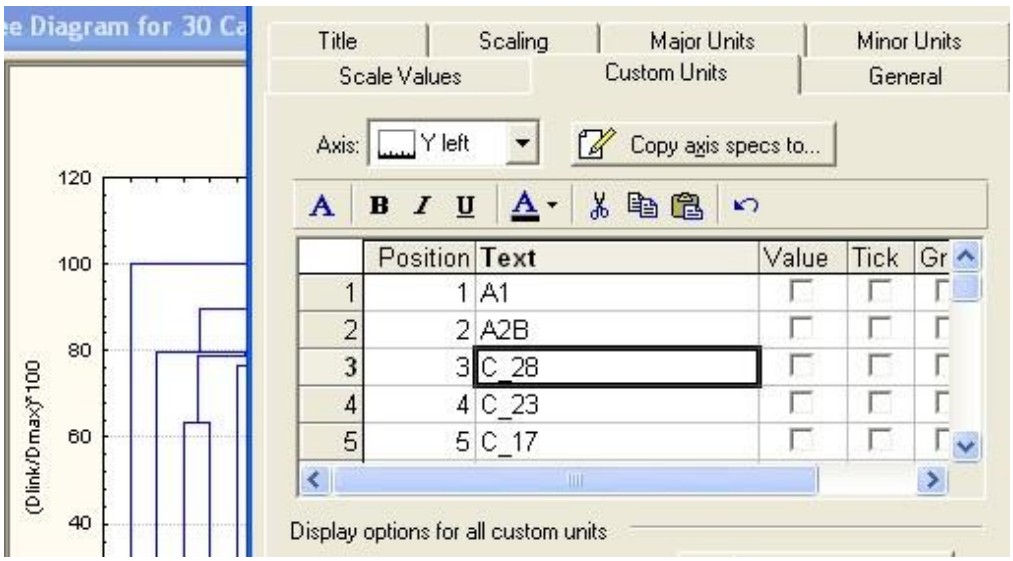

Рис 4.7. Настройка подписей данных

На этой же панели меню, где строятся дендрограммы, можно сохранить в виде таблицы порядок объединения объектов - схема объединения (**Amalgamation schedule**), график схемы объединения (**Graph of Amalgamation schedule**), матрицу расстояний между объектами (**Distance matrix**), а также среднее и стандартное отклонение для полученных классов – Описательные статистики (**Descriptive statistics**).

СРАВНЕНИЕ ПЕРЕМЕННЫХ. Кластерный анализ позволяет также оценивать близость переменных между собой. Для этого на первой панели в поле **Объекты** (**Cluster)** выберите **Variables (Columns)** (рис 4.8).

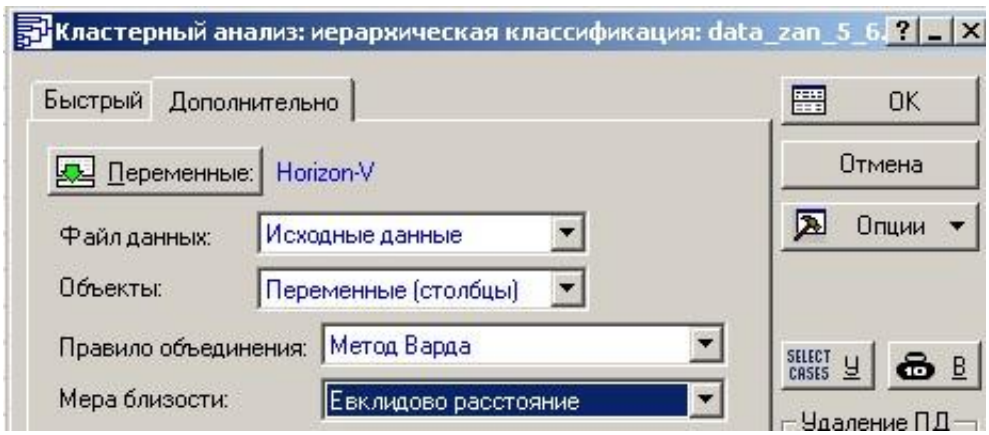

Рис 4.8. Настройка классификации методом Варда

Для 5 переменных проведите иерархический кластерный анализ методом одиночной связи и методом Варда с использованием Евклидового расстояния. Полученные два графика (рис 4.9.) сохраните в файле Excel.

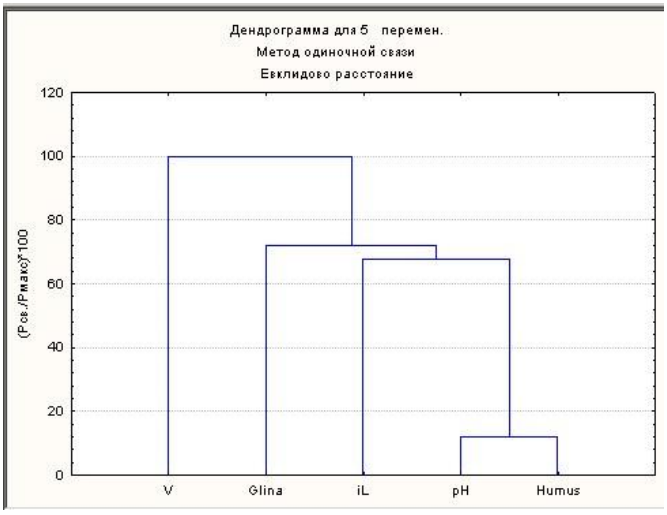

Рис 4.8. График близости переменных

МЕТОД К-СРЕДНИХ**.** Вернитесь в самое начало анализа и выберите **Кластеризацию методом к-средних** (**K-means clustering**). По методу К средних будет построено К кластеров, расположенных на возможно больших расстояниях друг от друга. Расчеты начинаются К кластеров, в которые объекты объединены случайным образом. Процедура состоит в изменении принадлежности объектов к кластерам так, чтобы: изменчивость внутри кластеров сделать минимальной, изменчивость между кластерами - максимальной. Эта оценка производиться с помощью дисперсионного анализа. Выберите закладку **Дополнительно (Advanced)** (рис 4.10).

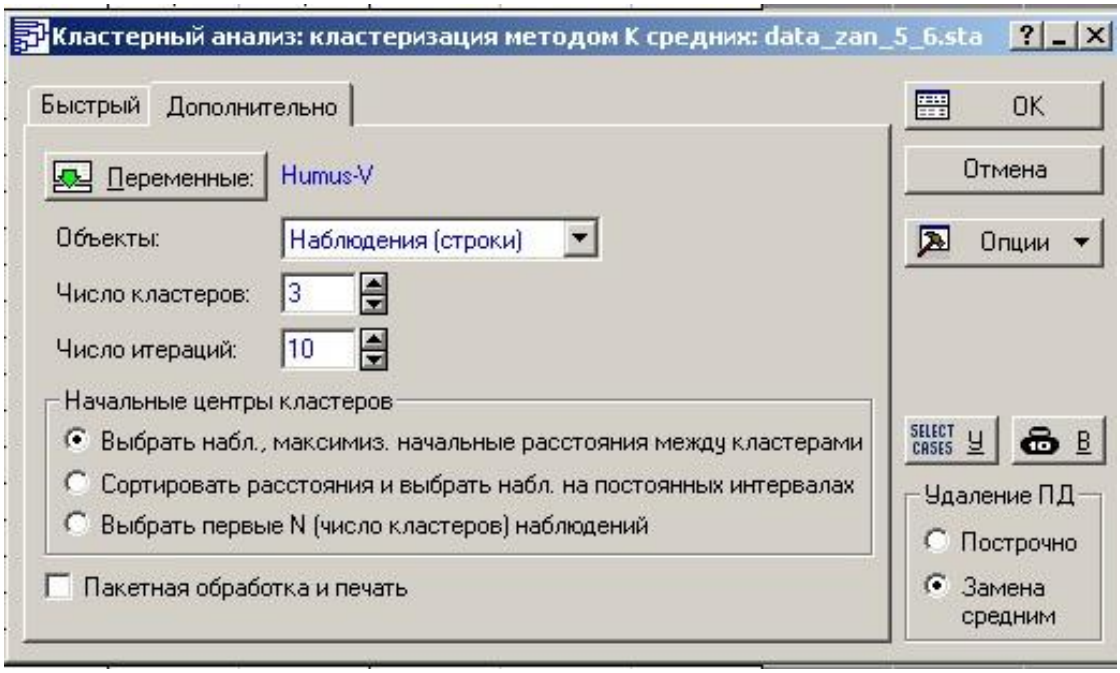

Рис 4.10. Настройка параметров кластеризации методом к-средних

Необходимо произвести выбор переменных (**Variables**), по которым будет проводиться анализ (C, PHS, IL, G, V) и выбор типа анализа (для объектов или для самих переменных) в окошке **Объекты** (**Cluster),** - точно такой, как и при иерархической классификации.

Укажите переменные: **C, PHS, IL, G, V,** и выберите анализ объектов-наблюдений (**Cases (row)**). Затем нужно задать **Число кластеров** (**Number of clusters**) и число итераций для расчетов (**Number of iterations**). Кроме этого, можно разным способом задать **Начальные центры кластеров** (**Initial cluster centers**).

Для ваших данных проведите кластеризацию методом k-средних, задав 3 кластера. Число итераций возьмите по умолчанию, равное 10. Начальные центры классов задайте через одинаковые интервалы в ранжированном ряду расстояний **Сортировать расстояния и выбрать наблюдения на постоянных интервалах** (**Sort distances and take observations at constant intervals**). Нажмите **ОК**.

Результирующая панель содержит информацию о заданных ранее условиях кластерного анализа (рис 4.11).

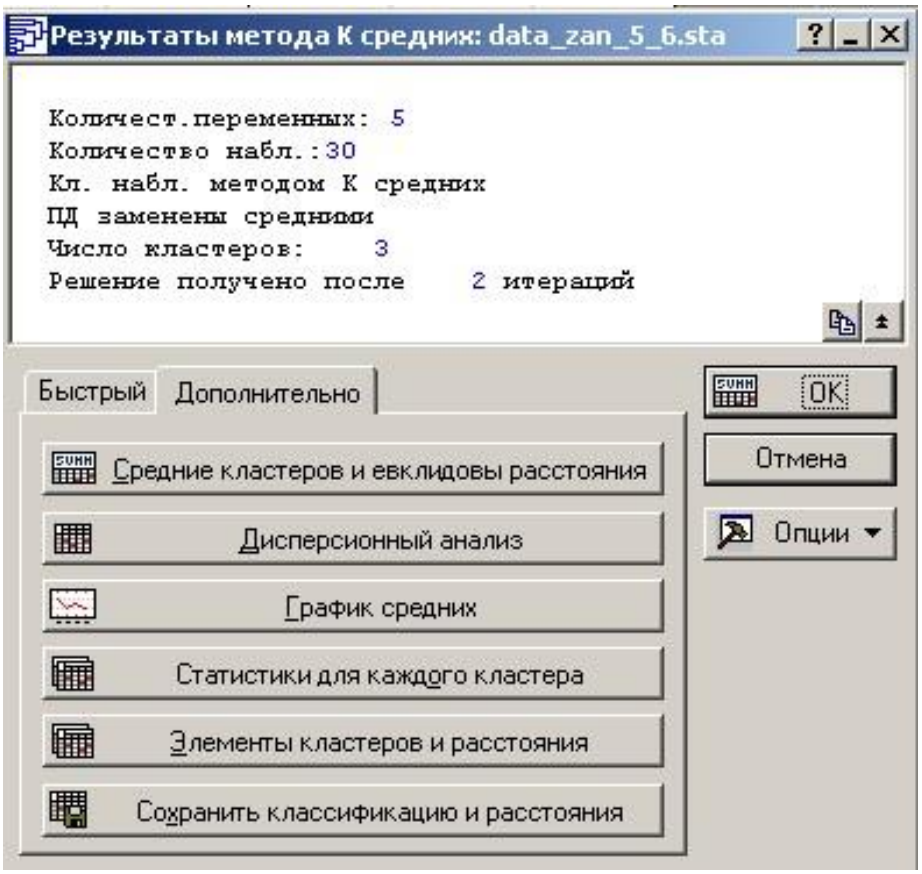

Рис 4.11. Выбор параметров представления результатов кластеризации методом К средних.

Панель позволяет оценить качество классификации с помощью таблицы **Дисперсионного анализа** (**Analysis of variance)**, получить таблицу средних значений признаков для кластеров и таблицу расстояний между кластерами – **Средние кластеров и Евклидовы расстояния** (**Cluster means & Euclidean distances**), построить графики средних значений для кластеров – **График средних** (**Graph of means**), получить описательные статистики для каждого класса (**Descriptive statistics for each cluster**), получить таблицу принадлежности объектов к каждому классу Э**лементы кластеров и расстояния** (**Members of each cluster & distances**).

Проанализируйте результаты, оценив качество классификации при помощи таблицы дисперсионного анализа (**Analysis of variance)** (таблица 4.2).

Таблица 4.2. Оценка качества параметров кластеризации.

Признаки	Between SS	df	Within SS	df	F	signif. p
	Сумма кв. между классами	Число ст. свободы	Общая сумма кв. внутри классов	Число ст. свободы		Уровень значимости
C	41,253422	2	89,541245	27	6,219717	0,0060027
PHS	0,4869745	2	1,8676891	27	3,519941	0,0438099
IL	2881,6445	2	291,72192	27	133,3537	1,015E-14
G	2422,0554	2	256,64453	27	127,4048	1,774E-14
V	0,5615084	2	0,5753129	27	13,17607	0,0001016

Например, из данной таблицы видно, что для всех почвенных свойств уровень значимости меньше 0,05 и, следовательно, нулевая гипотеза о равенстве средних по выделенным кластерам отвергается. Варьирование между выделенными кластерами превышает внутриклассовое варьирование. Значения F-статистики, полученные для каждого признака, являются индикатором того, насколько хорошо соответствующий признак разделяет кластеры.

Постройте график средних (рис 4.12) и таблицу принадлежности объектов к каждому классу. Результаты сохраните в файле Excel.

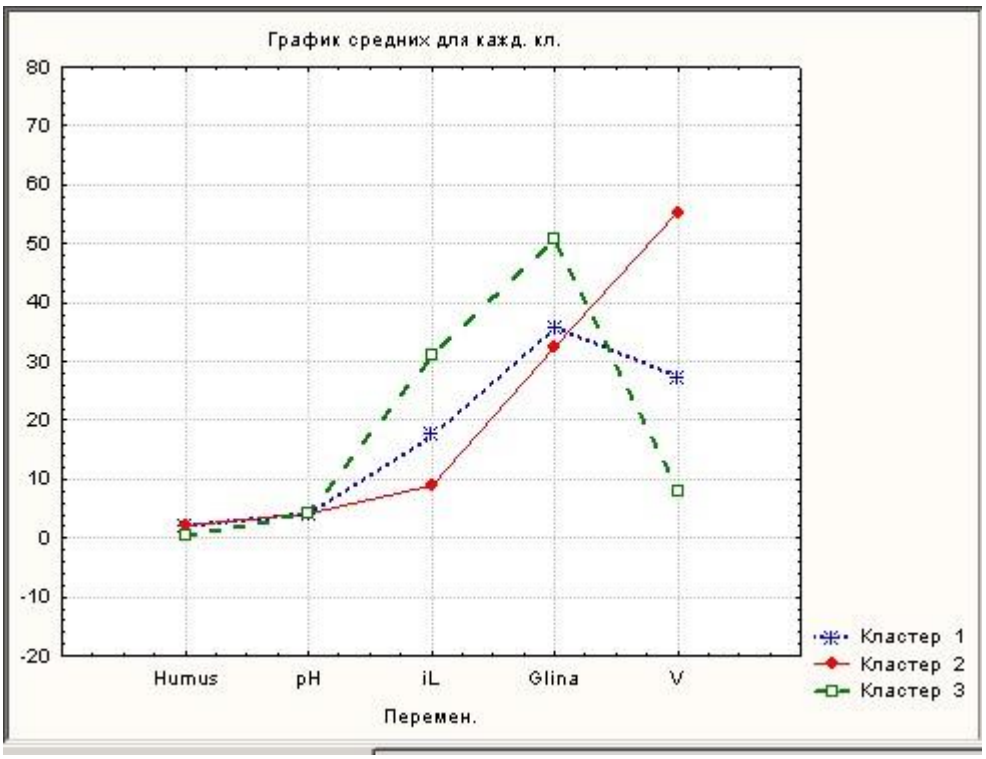

Рис 4.12 График средних

При копировании в отчет таблиц принадлежности объектов к кластерам их необходимо транспонировать и заменить порядковые номера объектов на номера участков

Повторите анализ, задав 5 классов. Результаты сохраните в файле Excel. Распечатайте отчет.