Министерство образования и науки Российской Федерации

Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

О. И. Бантикова, Е.Н. Седова, О.С. Чудинова

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА

КЛАССИФИКАЦИЯ БЕЗ ОБУЧЕНИЯ (НЕПАРАМЕТРИЧЕСКИЙ СЛУЧАЙ)

Методические указания
к лабораторному практикуму, курсовой работе, дипломному проектированию и самостоятельной работе студентов специальности 080116.65,
направлений подготовки 231300.62, 080500.62

Рекомендовано к изданию Редакционно-издательским советом Государственного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет»

Оренбург
ИПК ГОУ ОГУ
2011

УДК 519.22 (076.5)
ББК 22.172 я7
Б23

Рецензент - кандидат экономических наук, доцент С.В. Дьяконова

Бантикова, О.И.

Б23 Методы кластерного анализа. Классификация без обучения (непараметрический случай): методические указания к лабораторному практикуму, курсовой работе, дипломному проектированию и самостоятельной работе студентов / О.И. Бантикова, Е.Н. Седова, О.С. Чудинова; под ред. А.Г. Реннера; Оренбургский гос. ун-т. – Оренбург: ГОУ ОГУ, 2011.– 93 с.

Методические указания к семинарским занятиям, лабораторному практикуму, самостоятельной работе студентов, в том числе для выполнения расчетно-графических заданий, курсовых и дипломных работ, связанных с анализом многомерных статистических данных. Предназначены для студентов специальности 080116.65 – Математические методы в экономике, направлений подготовки 231300.62 – Прикладная математика, 080500.62 - Бизнес-информатика и других специальностей и направлений, изучающих дисциплины, связанные с математическим анализом многомерных статистических данных.

УДК 519.22 (076.5)
ББК 22.172 я7

Содержание

Введение

Для большинства социально-экономических явлений и процессов типична ситуация, связанная с разбросом значений показателей, их характеризующих и, таким образом, с неоднородностью объектов (стран, муниципальных образований, предприятий, семей и т.д.) по уровню (состоянию) развития.

Выявление особенностей, внутренних связей между объектами позволит выработать эффективные рекомендации по исправлению диспропорции в уровне развития, но требует предварительного разбиения (классификации) всей совокупности наблюдений на однородные, в определенном смысле, группы объектов, схожих между собой по набору показателей, их характеризующих.

Решение подобной задачи при небольшом наборе показателей традиционно осуществлялось методами комбинационной группировки, в противном случае (при наличии большого набора показателей) требуется использование специальных методов кластерного, дискриминантного анализа и статистических пакетов, их реализующих.

Целью изучения данного раздела является выработка практических навыков проведения многомерной классификации методами кластерного анализа в пакетах Statistica 7.0, Stata и последующего анализа результатов.

1 Теоретические аспекты кластерного анализа

1.1 Постановка задач многомерной классификации

В общем случае под **классификацией** понимается разделение рассматриваемой совокупности объектов или явлений на однородные, в определенном смысле, группы (классы), либо отнесение каждого из заданного множества объектов к одному из заранее заданных классов.

Исходная информация о классифицируемых объектах $O_1, O_2, ..., O_n$, каждый из которых характеризуется k признаками, может быть представлена в виде матрицы X типа «объект-свойство»:

$$\underset{n*k}{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

где x_{ij} - значение *j*-го признака на *i*-м объекте наблюдения;

или в виде матрицы парных сравнений объектов γ:

$$\underset{n*n}{\gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2n} \\ \dots & \dots & \dots & \dots \\ \gamma_{n1} & \gamma_{n2} & \dots & \gamma_{nn} \end{pmatrix},$$

где γ_{ij} - соотношение между i-ым и j-ым объектами.

Также могут быть даны обучающие выборки $X^{(1)}, X^{(2)}, ..., X^{(p)}$, где

$$\underset{n_j*k}{X^{(j)}} = \begin{pmatrix} x_{11}^{(j)} & x_{12}^{(j)} & \dots & x_{1k}^{(j)} \\ x_{21}^{(j)} & x_{22}^{(j)} & \dots & x_{21}^{(j)} \\ \dots & \dots & \dots & \dots \\ x_{n_j1}^{(j)} & x_{n_j2}^{(j)} & \dots & x_{n_jk}^{(j)} \end{pmatrix}, \; j = \overline{1, p}$$

Обучающая выборка - матрица типа «объект-свойство», содержащая значения k признаков для n_j объектов, для которых заранее известно, что они принад-

лежат к j - ому классу. p - число обучающих выборок, а соответственно и количество классов. Именно по этим k классам и необходимо распределить n объектов, подлежащих классификации.

Если известна матрица X и обучающие выборки, то говорят, что решается задача классификации «с обучением», если обучающих выборок нет, а известна только матрица X, то решается задача «без обучения».

Выбор методов классификации обусловлен априорной информацией, на основе которой она осуществляется. При этом априорная информация состоит из двух частей: информация о законе распределения классов; о наличии (отсутствии) обучающих выборок. Классификация методов разбиения объектов на однородные, в определенном смысле, группы, в зависимости от наличия априорной и предварительной выборочной информации, представлена в таблице 1.

Таблица 1 – Методы многомерной классификации

Априорные сведения о классах (генеральных совокупностях)	Априорная выборочная информация	
	Отсутствие обучающих выборок	Наличие обучающих выборок
Не известен вид закона распределения генеральной совокупности	Классификация без обучения, непараметрический случай: непараметрический кластерный анализ	Классификация с обучением, непараметрический случай: непараметрический дискриминантный анализ
Известен вид закона распределения генеральной совокупности (не известны параметры распределения)	Классификация без обучения, параметрический случай: параметрический кластерный анализ (расщепление смесей вероятностных распределений)	Классификация с обучением, параметрический случай: параметрический дискриминантный анализ

1.2 Постановка задачи непараметрического кластерного анализа

Необходимо разбить анализируемую совокупность объектов $O = \{O_1, O_2, ...O_n\}$, которые описаны с помощью матрицы X_{n*k} или γ_{n*n} на сравнительно небольшое число однородных, в определенном смысле, групп или классов.

Для осуществления процедуры разбиения вводится величина $\rho(O_i, O_j)$, характеризующая либо различия между любой парой исследуемых объектов с помощью расстояния $d(O_i, O_j)$[1], либо сходство с помощью меры близости $r(O_i, O_j)$[2].

Если задана функция $d(O_i, O_j)$, то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу.

Требования к функциям, определяющим расстояние и меру близости объектов:

1. симметричности: $d(O_i, O_j) = d(O_j, O_i)$, $r(O_i, O_j) = r(O_j, O_i)$
2. максимального сходства объекта с самим собой:
   $d(O_i, O_i) = 0$, $r(O_i, O_i) = \max\limits_{1 \le j \le n} r(O_i, O_j)$
3. монотонного убывания меры близости по расстоянию:
   $d(O_k, O_j) \ge d(O_i, O_j) \Rightarrow r(O_k, O_j) \le r(O_i, O_j)$

1.3 Расстояния между объектами и классами объектов

Наиболее часто используемые расстояния между объектами:

1. Обобщенное расстояние Махаланобиса:
   $$d_O(O_i, O_j) = \sqrt{(O_i - O_j)^T \Delta^T \Sigma^{-1} \Delta (O_i - O_j)},$$

где Σ - ковариационная матрица генеральной совокупности, из которой извлечена выборка;

[1] Применяется, как правило, при решении задачи классификации объектов.
[2] Применяется, как правило, при решении задачи классификации признаков.

Δ - некоторая симметричная неотрицательно-определенная матрица весовых коэффициентов признаков.

2. Обычное евклидово расстояние: $d_E(O_i, O_j) = \sqrt{\sum_{l=1}^{k}(x_{il} - x_{jl})^2}$.

Данная мера различия объектов используется в трех случаях:

- наблюдения извлекаются из нормально распределенной генеральной совокупности с ковариационной матрицей вида $\Sigma = \sigma^2 E_k$ (компоненты вектора X взаимно некоррелированные и имеют одинаковую дисперсию);
- компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;
- признаковое пространство совпадает с геометрическим пространством.

3. Взвешенное евклидово расстояние: $d_{BE}(O_i, O_j) = \sqrt{\sum_{l=1}^{k}\omega_l(x_{il} - x_{jl})^2}$.

Данная мера различия объектов применяется в случаях, когда каждой компоненте вектора наблюдений X удается приписать некоторый «вес» ω_l, пропорциональный степени важности признака в задачи классификации.

4. Расстояние Минковского: $d_M(O_i, O_j) = \left(\sum_{l=1}^{k}\left|x_{il} - x_{jl}\right|^p\right)^{\frac{1}{p}}$.

5. Хеммингово расстояние (манхеттеновское расстояние, расстояние city-block) часто применяется как мера различия объектов, задаваемых дихотомическими признаками (частный случай расстояния Минковского при $p = 1$): $d_H(O_i, O_j) = \sum_{l=1}^{k}\left|x_{il} - x_{jl}\right|$.

6. Расстояние Чебышева (частный случай расстояния Минковского при $p = \infty$.):

$$d_{CH}(O_i, O_j) = \max_{1 \le l \le k}\left|x_{il} - x_{jl}\right|.$$

7. «Корреляционное» расстояние:

$$d_{corr}(O_i,O_j)=1-\frac{\sum_{l=1}^{k}\left(x_{il}-\bar{x}^{i}\right)\cdot\left(x_{jl}-\bar{x}^{j}\right)}{\sqrt{\sum_{l=1}^{k}\left(x_{il}-\bar{x}^{i}\right)^2\cdot\sum_{l=1}^{k}\left(x_{jl}-\bar{x}^{j}\right)^2}},\ \text{где}\ \bar{x}^{i}=\frac{1}{k}\sum_{l=1}^{k}x_{il},\bar{x}^{j}=\frac{1}{k}\sum_{l=1}^{k}x_{jl}.$$

8. Расстояние Канберра: $d_{CANB}(O_i,O_j)=\sum_{l=1}^{k}\frac{\left|x_{il}-x_{jl}\right|}{\left|x_{il}\right|+\left|x_{jl}\right|}$

9. Угловое расстояние, при котором объекты рассматриваются как векторы в многомерном пространстве: $d_{angular}(O_i,O_j)=1-\frac{\sum_{l=1}^{k}x_{il}x_{jl}}{\sqrt{\sum_{l=1}^{k}x_{il}^2\cdot\sum_{l=1}^{k}x_{jl}^2}}$

Внимание!

Если кластерный анализ применяется для решения задачи классификации признаков, то для измерения их сходства используется мера близости $r\left(X_i,X_j\right)$, в качестве которой могут выступать различные коэффициенты связи: парный коэффициент корреляции, корреляционное отношение, коэффициенты ранговой корреляции и т.д.

При реализации процедур кластерного анализа приходится рассчитывать расстояние не только между объектами, но и между классами объектов.

Пусть S_i - i-ый класс, состоящий из n_i объектов;

$\rho(S_l,S_m)$ - расстояние между классами S_l и S_m.

Наиболее часто используемые расстояния между классами:

1. расстояние, измеряемое по принципу «ближнего соседа»:

$$\rho_{\min}(S_l,S_m)=\min_{O_i\in S_l,O_j\in S_m}d(O_i,O_j);$$

2. расстояние, измеряемое по принципу «дальнего соседа»:

$$\rho_{\max}(S_l,S_m)=\max_{O_i\in S_l,O_j\in S_m}d(O_i,O_j);$$

3. расстояние, измеряемое по «центрам тяжести» групп:

$$\rho_{ЦТ}(S_l, S_m) = d(\overline{X}(l), \overline{X}(m)),$$

где $\overline{X}(l)$, $\overline{X}(m)$ - вектора средних арифметических значений признаков, характеризующих соответственно l-ый и m-ый классы;

4. расстояние, измеряемое по принципу «средней связи»:

$$\rho_{ср}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{O_i \in S_l} \sum_{O_j \in S_m} d(O_i, O_j);$$

5. обобщенное расстояние Колмогорова:

$$\rho_{\tau}^{(K)}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{O_i \in S_l} \sum_{O_j \in S_m} d^{\tau}(O_i, O_j) \right]^{1/\tau}$$

Если $S(m,q) = S_m \cup S_q$ - группа элементов, полученная путем объединения кластеров S_m и S_q, то обобщенное расстояние Колмогорова имеет вид:

$$\rho_{\tau}^{(K)}(S_l, S(m,q)) = \left[\frac{n_m \left(\rho_{\tau}^{(K)}(S_l, S_m)\right)^{\tau} + n_q \left(\rho_{\tau}^{(K)}(S_l, S_q)\right)^{\tau}}{n_m + n_q} \right]^{1/\tau}$$

6. Обобщенная формула расчета расстояния между классами объектов S_l и $S(m,q)$:

$$\rho(S_l, S(m,q)) = \alpha\rho(S_l, S_m) + \beta\rho(S_l, S_q) + \gamma\rho(S_m, S_q) + \delta\left|\rho(S_l, S_m) - \rho(S_l, S_q)\right|,$$

где $\alpha, \beta, \gamma, \delta$ - числовые коэффициенты, значение которых определяет специфику процедуры, ее алгоритм.

1.4 Иерархические методы кластерного анализа

Агломеративные кластер-процедуры

Основной принцип работы иерархических агломеративных процедур состоит в последовательном объединении групп элементов сначала самых близких, а затем все более отдаленных друг от друга.

На первом шаге каждый объект рассматривается как отдельный класс. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и, с учетом принятого расстояния между классами, пересчитывается матрица расстояний, размер которой снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Алгоритм иерархической классификации предусматривает геометрическое представление в виде дендрограммы (рисунок 1).

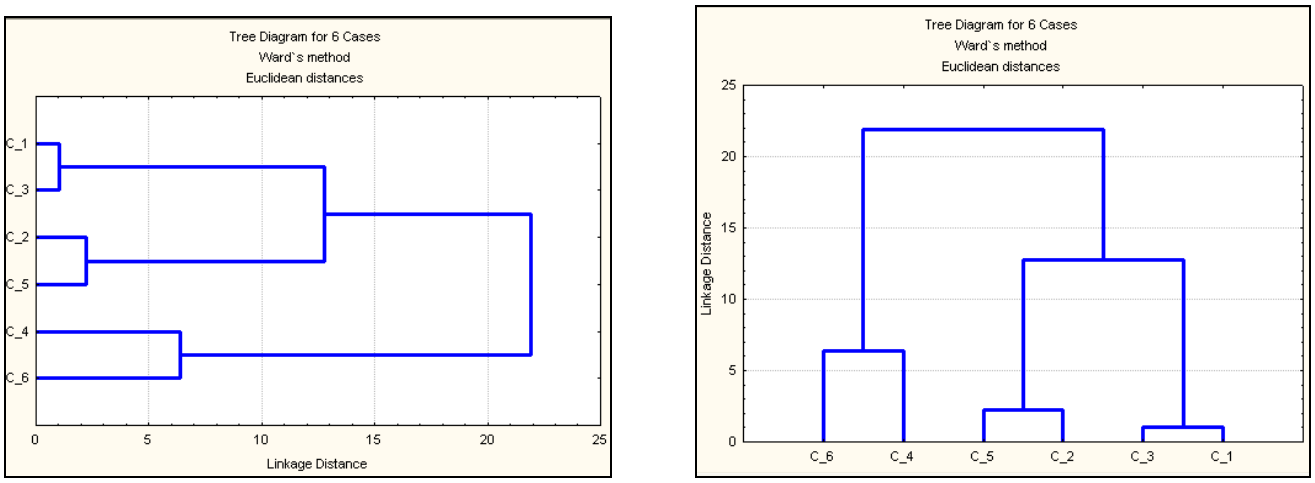


Рисунок 1 – Горизонтальная и вертикальная дендрограммы объединения классов иерархическими агломеративными методами

Если ставится задача разбиения объектов на несколько групп, то при реализации агломеративных кластер-процедур устанавливается пороговое значение расстояния ρ. Если расстояние между классами превосходит ρ, то дальнейшего объединения классов не происходит.

К агломеративным методам кластерного анализа относят:

- метод одиночной связи;
- метод полной связи;
- метод средней связи;
- метод Уорда.

Дивизимные кластер-процедуры

Основной принцип работы иерархических дивизимных процедур состоит в последовательном разделении групп элементов сначала самых далеких, а затем все более приближенных друг к другу.

Первоначально считается, что все n объектов объединены и составляют один кластер. Среди множества объектов на основе матрицы расстояний определяются наиболее удаленные друг от друга и берут их за основу двух новых кластеров. Оставшиеся $(n-2)$ объектов распределяются по образованным двум классам по принципу: объект следует отнести к тому классу, расстояние до которого наименьшее. Затем в этих двух классах находят наиболее удаленные друг от друга объекты, которые следует отнести к разным классам и т.д. Преимущество дивизимных кластер-процедур состоит в том, что все расчеты осуществляются на основе исходной матрицы расстояний. В отличие от агломеративных кластер-процедур ее не нужно пересчитывать на каждом шаге.

Общая схема работы агломеративных и дивизимных кластер-процедур приведена на рисунке 2:

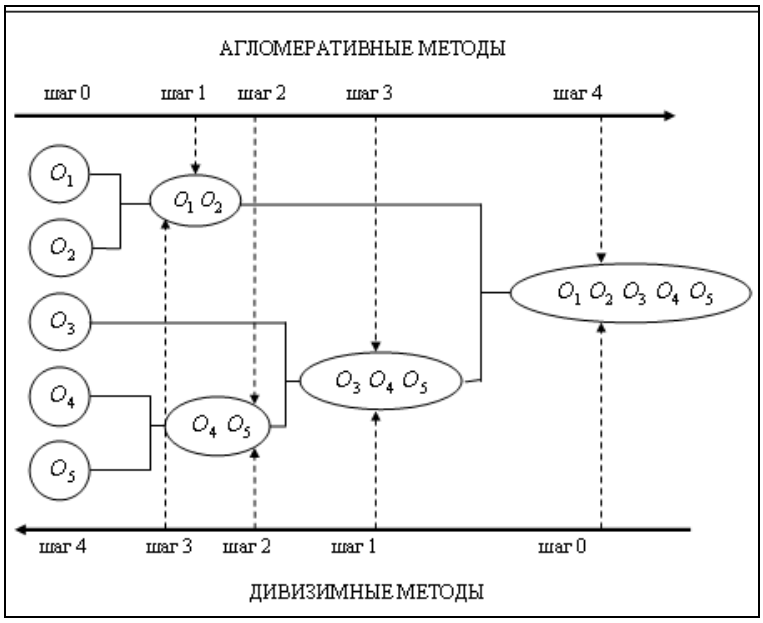


Рисунок 2 – Процесс последовательного объединения (разделения) классов иерархическими методами кластерного анализа

1.5 Итерационные методы кластерного анализа

Сущность этих методов заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.).

Метод k - средних

Для реализации данного метода изначально задается число классов, на которые необходимо разбить имеющуюся совокупность из n объектов. Для того чтобы задать начальные условия необходимо иметь либо дополнительную информацию о количестве кластеров, либо предварительно оценить число кластеров с помощью иерархических кластер-процедур.

Для начала процедуры классификации задаются p случайно выбранных объектов – эталоны (ε). Каждому эталону приписывается порядковый номер, который, одновременно, является номером класса. Из оставшихся n-p объектов извлекается объект и проверяется, к какому из эталонов он находится ближе. Данный объект присоединяется к тому эталону, для которого наблюдается минимальное расстояние, то есть $\min \rho_{il}$, где $1 \le l \le p$. Веса и эталоны пересчитываются по правилу:

$$\varepsilon_i^{\nu} = \begin{cases} \dfrac{\omega_i^{\nu-1} \cdot \varepsilon_i^{\nu-1} + O_{p+\nu}}{\omega_i^{\nu-1} + 1} & \text{если } \rho(O_{p+\nu}, \varepsilon_i^{\nu-1}) = \min\limits_{1 \le j \le p} \rho(O_{p+\nu}, \varepsilon_j^{\nu-1}) \\ \varepsilon_i^{\nu-1} & \text{в другом случае} \end{cases}$$

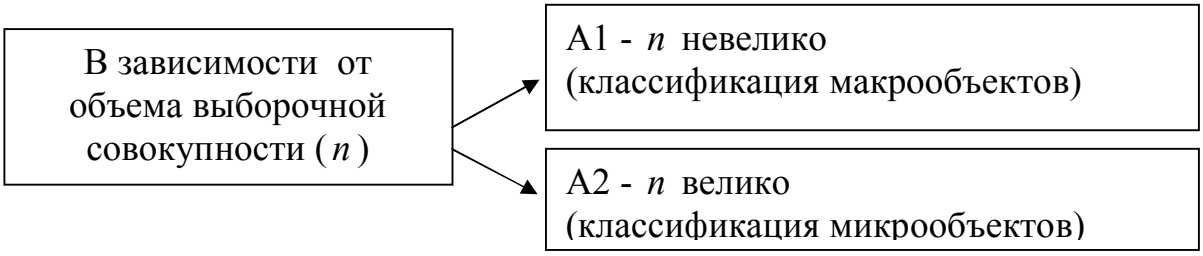
$$\omega_i^{\nu} = \begin{cases} \omega_i^{\nu-1} + 1 & \text{если } \rho(O_{p+\nu}, \varepsilon_i^{\nu-1}) = \min\limits_{1 \le j \le p} \rho(O_{p+\nu}, \varepsilon_j^{\nu-1}) \\ \omega_i^{\nu-1} & \text{в другом случае} \end{cases}$$

где ω - «вес» класса, ν - номер итерации.

При этом нулевое приближение строится с помощью случайно выбранных p точек исследуемой совокупности: $\varepsilon_i^0 = O_i$, $\omega_i^0 = 1$, $i = 1,2,...,p$.

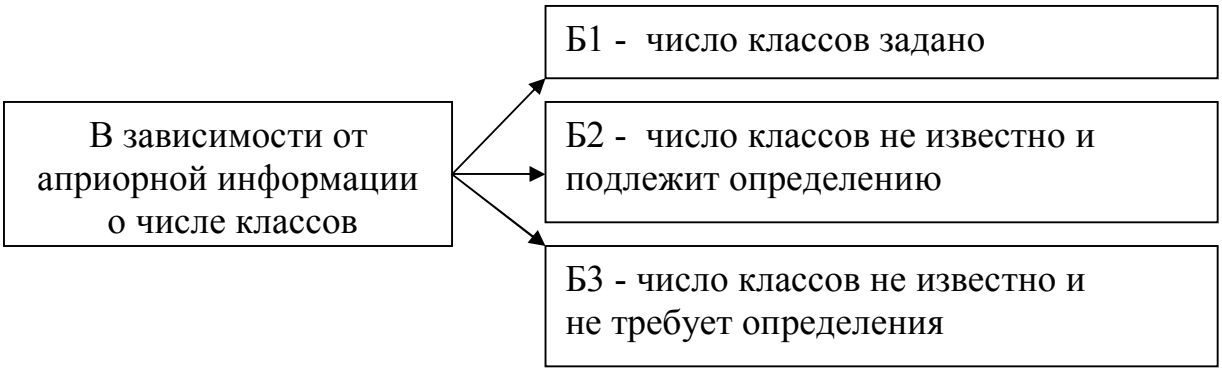
На следующем шаге выбирается (i+1)-ый объект и для него повторяется вся процедура. Через n-p шагов все объекты будут отнесены к одному из p кластеров. Для достижения устойчивого разбиения, все объекты опять присоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается, в противном случае алгоритм повторяется.

Все задачи кластерного анализа можно разделить по следующим критериям:



✓ В **случае А1** ведется речь о классификации сравнительно небольших по объему совокупностей наблюдений, состоящих как правило из нескольких десятков наблюдений, сюда могут быть отнесены задачи классификации макрообъектов, таких как страны, города, фирмы, предприятия.

✓ В **случае А2** речь идет о классификации достаточно больших массивов многомерных наблюдений (n – порядка нескольких сотен и тысяч) - классификация индивидуумов, семей, изделий.



Такое разделение задач классификации хотя и условно, но весьма необходимо с точки зрения принципиального различия идей и методов, на основе которых конструируются кластер-процедуры.

Так, иерархические кластер-процедуры предназначены в основном для решения задач типа **А1Б1, А1Б2, А1Б3,** итерационные кластер-процедуры **- А2Б2, А2Б1.**

1.6 Функционалы качества разбиения

При использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают, во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Возникает задача выбора «лучшего» разбиения. С этой целью вводится понятие так называемого функционала качества разбиения $Q(S)$, где $S = \{S_1, S_2, ..., S_p\}$ - результаты разбиения объектов на классы.

Под наилучшим разбиением S^* понимается то разбиение, на котором достигается экстремум выбранного функционала качества. Выбор того или иного функционала качества, как правило, осуществляется весьма произвольно и опирается скорее на эмпирические и профессионально – интуитивные соображения, чем на какую-либо строгую формализованную систему.

Наиболее часто используемые функционалы качества:

1. Сумма внутриклассовых дисперсий:

$$Q_1(S) = \sum_{l=1}^{p} \sum_{O_i \in S_l} d^2(O_i, \overline{X}(l)) \to \min,$$

где p – число классов;

S_l – l-ый класс в классификации S;

$\overline{X}(l)$ – центр класса S_l.

2. Сумма попарных внутриклассовых расстояний между объектами:

$$Q_2(S)=\sum_{l=1}^{p}\sum_{O_j O_j \in S_{il}} d^2(O_i,O_j)\to \min$$

3. Обобщенная внутриклассовая дисперсия:

$$Q_3(S)=\sum_{l=1}^{p}\sum_{j=1}^{k} S_j^2(l)\to \min,$$

где $S_j^2(l)$ - оценка дисперсии j-ого признака l-ого класса.

1.7 Критерии определения оптимального числа классов

При использовании методов кластерного анализа возникает задача определения оптимального количества классов. Частично это позволяет сделать уже визуальный анализ дендрограммы: например, довольно большой разрыв между уровнями, соответствующими разбиению на p_0 и разбиению на $p_1 > p_0$ классов говорит о том, что оптимальное количество классов равно p0. Можно использовать и более формальные критерии, которых в литературе известно более тридцати. Исследования показали, что одними из наиболее эффективных являются индекс Калински и Харабаза и индекс Дуды и Харта.

Индекс Калински и Харабаза (1) сравнивает степень «разброса» данных внутри кластеров и между кластерами и рассчитывается как скорректированное на количество классов p и объем выборки n отношение следа матрицы межгруппового рассеяния B к следу матрицы внутригруппового рассеяния W:

$$G1(p)=\frac{trace(B)/(p-1)}{trace(W)/(n-p)} \qquad (1)$$

То значение p, при котором индекс принимает максимальное значение, и есть оптимальное количество классов.

Для расчета $G1(p)$ можно также использовать формулу (2):

$$G1(p) = \frac{RR^2/(p-1)}{(1-RR^2)/(n-p)}, \quad (2)$$

где $RR^2 = 1 - \frac{SSE}{SST}$,

$SSE = \sum_{g=1}^{p}\sum_{i=1}^{n_g}\sum_{j=1}^{k}\left(x_{ij}^{g} - \bar{x}_j^{g}\right)^2$ - сумма квадратов расстояний от объектов до центров их классов;

n_g - количество объектов в классе g, $g = 1,...,p$;

$\bar{x}_j^{g}$ - среднее значение j-го признака в классе g, $j = 1,...,k$;

$SSE = \sum_{g=1}^{p}\sum_{i=1}^{n_g}\sum_{j=1}^{k}\left(x_{ij}^{g} - \bar{x}_j\right)^2$ - сумма квадратов расстояний от объектов до общего среднего;

$\bar{x}_j$ - среднее значение j-го признака, $j = 1,...,k$;.

Чем больше значение данного индекса, тем лучше разделены классы.

1.8 Интерпретация результатов классификации

Для содержательной интерпретации результатов наилучшей, с точки зрения функционала качества, классификации определяются средние значения показателей в каждом кластере. График средних значений, благодаря своей наглядности, позволяет охарактеризовать каждый класс и провести сравнительный анализ классов. Очень желательно, чтобы в результате сравнительного анализа каждому классу было дано название.

1.9 Вопросы для практическо-семинарских занятий

Группа А – базовые вопросы по лекционному материалу

1) В чем состоит принципиальное отличие методов многомерной классификации от комбинационных группировок?

2) Что понимается под классификацией?

3) Что понимается под термином «классификация без обучения»?

4) Что понимается под термином «непараметрический случай»?

5) В чем заключается постановка задачи непараметрического кластерного анализа?

6) Что понимается под однородностью объектов в кластерном анализе?

7) Каким требованиям должны удовлетворять расстояние и мера близости?

8) Привести расстояния между объектами и дать рекомендации по их применению.

9) Из каких соображений выбираются весовые коэффициенты для взвешенного евклидова расстояния, какими свойствами они должны обладать?

10) Какие характеристики могут выступать в качестве меры близости объектов или признаков?

11) Привести расстояния между классами объектов.

12) В чем состоит основной принцип работы иерархических кластер-процедур?

13) В чем отличие агломеративных и дивизимных кластер-процедур?

14) Какие методы относятся к итерационным кластер-процедурам?

15) Охарактеризовать принцип работы метода к-средних?

16) Представить графически процесс последовательного объединения объектов в классы.

17) Представить графически процесс последовательного разделения объектов в классы.

18) Как оценивается качество полученного разбиения совокупности на классы?

19) Из каких соображений дается содержательная интерпретация результатов классификации.

Группа В – вопросы, требующие самостоятельной подготовки

1) Как проводится классификация объектов, если характеризующие их признаки имеют разные единицы измерения [3, с.246]?

2) С помощью каких метрик можно измерить различие (сходство) между объектами, если характеризующие их признаки измерены в порядковой шкале [3, с.83]?

3) С помощью каких метрик можно измерить различие (сходство) между объектами, если характеризующие их признаки измерены в номинальной шкале [6, с.85]?

4) Проиллюстрировать метод медианной связи для измерения расстояния между классами S_l и $S(m,q)$ [4, с.480].

5) К какому расстоянию сводится обобщенное расстояние Колмогорова, если $\tau \to +\infty$, $\tau \to -\infty$, $\tau = 1$ [3, с.249]?

6) О каком расстоянии между классами объектов идет речь, если заданы следующие числовые коэффициенты обобщенной формулы $\alpha = \beta = -\delta = 1/2 \quad \gamma = 0$ [3, с.249]?

7) О каком расстоянии между классами объектов идет речь, если заданы следующие числовые коэффициенты обобщенной формулы $\alpha = \beta = \delta = 1/2 \quad \gamma = 0$ [3, с.249]?

8) О каком расстоянии между классами объектов идет речь, если заданы следующие числовые коэффициенты обобщенной формулы $\alpha = \dfrac{n_m}{n_m + n_q} \quad \beta = \dfrac{n_q}{n_m + n_q} \quad \gamma = \delta = 0$ [3, с.249]?

9) О каком расстоянии между классами объектов идет речь, если заданы следующие числовые коэффициенты обобщенной формулы $\alpha = \beta = 1/2 \;\; \gamma = -0{,}25 \;\; \delta = 0$ [3, с.249]?

10) В чем суть методов полной, одиночной и средней связи [4, с.476]?

11) В чем особенность и преимущество метода Уорда [3, с.247]?

12) Привести алгоритм иерархических агломеративных кластер-процедур [4, с.479].

13) Охарактеризовать итерационный метод поиска сгущений [4, с.493].

14) Охарактеризовать итерационный метод взаимного поглощения [4, с.496].

15) Привести алгоритм итерационных кластер-процедур (на примере метода к-средних) [4, с.493].

16) Обосновать выбор метрики и рассчитать расстояние между объектами O_1 и O_2, характеризующимися показателями X_1- рентабельность ($x_{11} = 23{,}4; x_{21} = 17{,}5$) и X_2- производительность труда ($x_{12} = 9{,}1; x_{22} = 5{,}2$).

17) Рассчитать расстояние между объектами O_1 и O_2, характеризующимися показателями X_1- расходы на питание ($x_{11} = 2; x_{21} = 12$) и X_2- расходы на развлечения ($x_{12} = 10; x_{22} = 9$) по взвешенной евклидовой метрике, выбрав весовые коэффициенты пропорционально степени важности признака в задачи классификации.

18) Обосновать выбор метрики для расчета расстояния между объектами, характеризующимися показателями X_1 - наличие квартиры и X_2- наличие автомобиля.

19) Обосновать выбор метрики для расчета расстояния между объектами, характеризующимися показателями X_1 - успеваемость по дисциплине I; X_2- успеваемость по дисциплине II.

20) На основе матрицы расстояний $D = \begin{pmatrix} 0 & & & \\ 4{,}49 & 0 & & \\ 2{,}16 & 3{,}26 & 0 & \\ 3{,}53 & 1{,}92 & 2{,}68 & 0 \end{pmatrix}$ проиллюстрировать работу дивизимного метода классификации объектов.

21) В кластер S_1 входят четыре объекта (O_1, O_2, O_3, O_4), расстояние от которых до объекта O_5 составляет соотвественно: 2,5,6,7. Определить расстояние от объекта O_5 до кластера S_1 используя принципы «ближнего соседа», «дальнего соседа», «средней связи».

2 Содержание лабораторной работы

Выполнение лабораторной работы включает в себя следующие этапы:

– ознакомление с формулировкой задания к лабораторной работе и порядком её выполнения в пакетах прикладных программ;

– выполнение расчетов на компьютере;

– анализ полученных результатов;

– подготовку письменного отчета по лабораторной работе;

– защиту лабораторной работы.

3 Задание к лабораторной работе

1) Выбрать предмет исследования, а также набор показателей, характеризующих данное явление или процесс[3].

2) По данным Приложения А (таблица А.2) с помощью методов кластерного анализа:

- провести классификацию муниципальных образований с помощью иерархических агломеративных методов кластерного анализа;

[3] Полный перечень показателей, характеризующих муниципальные образования Оренбургской области, приведен в приложении А (таблица А.1)

- провести классификацию муниципальных образований с помощью метода К-средних.

3) Сравнить классификации, полученные с помощью агломеративных кластер-процедур и метода К-средних, обосновать выбор окончательного варианта классификации;

4) Дать экономическую интерпретацию результатов классификации.

4 Порядок выполнения работы

4.1 Порядок выполнения работы в пакете Statistica

Порядок выполнения лабораторной работы рассмотрен на примере, где **целью** исследования является проведение многомерной классификации муниципальных образований Оренбургской области по показателям, характеризующим демографическое состояние региона.

Объектом исследования выступают города и районы Оренбургской области.

Предмет исследования - демографическое состояние региона, характеризующееся следующими показателями:

x_1 - общий коэффициент рождаемости (‰);

x_2 - общий коэффициент смертности (‰);

x_3 - удельный вес населения в трудоспособном возрасте (%);

x_4 - удельный вес населения старше трудоспособного возраста (%);

x_5 - коэффициент миграционного прироста, снижения (‰).

Исходные данные для анализа представлены в виде матрицы X. Фрагмент таблицы с исходными данными в пакете Statistica 7.0 представлен на рисунке 3.

STATISTICA - [Data: data* (5v by 47c)]

File Edit View Insert Format Statistics Graphs Tools Dat

Add to

Arial 10

Города и районы	Демографическое состояние региона				
	1 x1	2 x2	3 x3	4 x4	5 x5
Абдулинский	8,5	20,8	53,69	30,93	-22
Адамовский	14,1	12,4	62,31	16,29	-9,6
Акбулакский	15,5	11,9	62,71	16,74	-3,4
Александровский	12,3	13,6	61,69	19,82	-2,3
Асекеевский	12,5	17,2	59,44	23,32	-11
Беляевский	14	14,7	61,76	19,21	1,7
Бугурусланский	13,5	19,4	60,42	22,54	-8,1
Бузулукский	15,3	18,6	59,12	23,24	2,8
Гайский	15,8	17,7	59,43	20,59	-16
Грачевский	13,1	15,2	60,03	22,17	-15,6
Домбаровский	15,7	11,7	62,36	15,47	-8,2
Илекский	15,1	15,9	59,89	21,06	1,3
Кваркенский	14,5	16,7	60,41	19,2	-15,4
Красногвардейский	15,2	13	60,37	19,17	-6,9
Кувандыкский	14,7	15,9	58,77	21,23	-18,3
Курманаевский	13	18,2	60,63	22,93	-5,6
Матвеевский	10,4	17,5	58,95	23,67	-8,6
Новоорский	14,6	16	61,11	19,27	-7
Новосергиевский	14,3	19	59,47	21,7	1,8
Октябрьский	13,7	15,3	61,05	21,49	8,6

Рисунок 3 – Исходные данные для анализа

Внимание!

Если исходные признаки, по которым производится классификация объектов, имеют разные единицы измерения, то необходимо перейти к стандартизованным переменным одним из следующих способов:

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{S_j};\quad x_{ij}^{*}=\frac{x_{ij}}{x_{\max j}};\quad x_{ij}^{*}=\frac{x_{ij}}{\bar{x}_j};\quad x_{ij}^{*}=\frac{x_{ij}}{x_{\min j}};\ x_{ij}^{*}=\frac{x_{ij}}{100};\ x_{ij}^{*}=x_{ij}\cdot 100,$$

где x_{ij} - исходное значение j-го признака на i-ом объекте наблюдения;

x_{ij}^{*} - нормированное значение исходного j-го признака на i-ом объекте наблюдения;

$\bar{x}_j$ - среднее значение j-го признака;

S_j - выборочное среднеквадратическое отклонение j-го признака;

$x_{\max j}$ - максимальное значение j-го признака;

$x_{\min j}$ - минимальное значение j-го признака.

Для приведения исходных переменных к стандартизованному виду можно воспользоваться операцией центрирования и нормирования данных. Для этого в пакете Statistica 7.0 необходимо выбрать пункты меню **Data/Standardize.** Вид экрана представлен на рисунке 4.

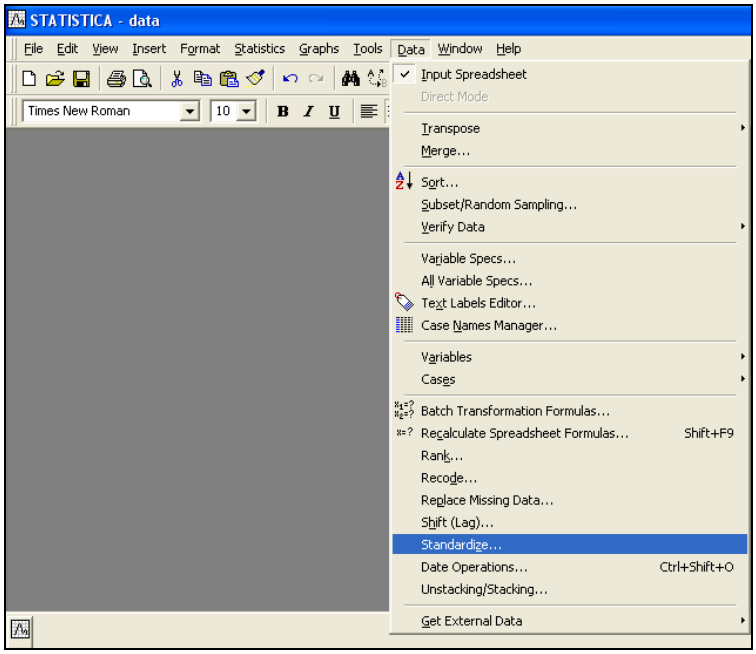


Рисунок 4 – Осуществление операции стандартизации данных

Результаты преобразования данных представлены на рисунке 5.

STATISTICA - [Data: data* (5v by 47c)]

File Edit View Insert Format Statistics Graphs Tools Data Window Help

Add to Workbook

Arial 10

Города и районы	Демографическое состояние региона				
	1 x1	2 x2	3 x3	4 x4	5 x5
Абдулинский	-2,422649	2,1344385	-2,57926	3,1862766	-2,147446
Адамовский	0,3246204	-1,115374	0,24794	-1,17882	-0,633558
Акбулакский	1,0114378	-1,308815	0,379133	-1,044647	0,1233866
Александровский	-0,558431	-0,651115	0,044592	-0,126307	0,2576831
Асекеевский	-0,460314	0,7416618	-0,69337	0,9172614	-0,804481
Беляевский	0,275562	-0,225544	0,06755	-0,308186	0,7460342
Бугурусланский	0,0302701	1,5928031	-0,37194	0,6846948	-0,450426
Бузулукский	0,9133211	1,2832972	-0,79832	0,8934085	0,8803308
Гайский	1,158613	0,935103	-0,69665	0,1032783	-1,414919
Грачевский	-0,165963	-0,032103	-0,49986	0,5743748	-1,366084
Домбаровский	1,1095546	-1,386191	0,264339	-1,423313	-0,462635
Илекский	0,8152043	0,2387147	-0,54578	0,2434146	0,6971991
Кваркенский	0,520854	0,5482206	-0,37522	-0,311167	-1,341667
Красногвардейский	0,8642627	-0,883244	-0,38834	-0,320112	-0,303921
Кувандыкский	0,6189707	0,2387147	-0,91311	0,2941022	-1,695721
Курманаевский	-0,215022	1,1285442	-0,30307	0,8009781	-0,145207
Матвеевский	-1,49054	0,8577266	-0,85408	1,0216183	-0,51147
Новоорский	0,5699123	0,2774029	-0,14564	-0,290296	-0,316129
Новосергиевский	0,4227372	1,4380502	-0,68353	0,4342385	0,758243
Октябрьский	0,1283868	0,0065852	-0,16532	0,3716244	1,5884399

Рисунок 5 – Центрировано-нормированные значения признаков

Для реализации кластерного анализа с помощью пакета Statistica 7.0 после запуска программы и ввода исходных данных необходимо выбрать пункт меню **Statistics – Критерии,** подпункты **Multivariate Exploratory Techniques/ Cluster Analysis – Кластерный анализ.** Вид экрана представлен на рисунке 6.

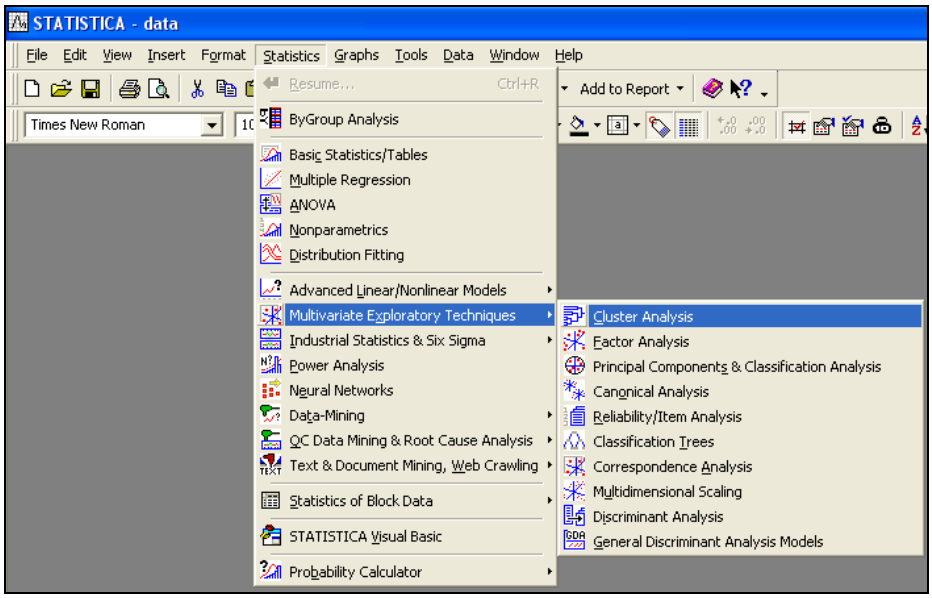


Рисунок 6 – Вызов диалога кластерного анализа

На экране появится окно, изображенное на рисунке 7, в котором содержатся основные процедуры кластерного анализа:

Joining (tree clustering) – иерархические агломеративные методы;

K-mean clustering – метод к-средних;

Two-way joining – метод двухстороннего присоединения, в котором классифицируются и объекты, и признаки одновременно.

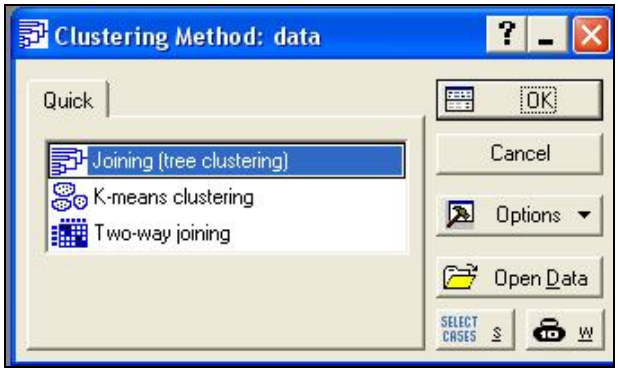


Рисунок 7 – Основные процедуры кластерного анализа

Классификация муниципальных образований иерархическими агломеративными методами кластерного анализа

Выбор процедуры Joining (tree clustering) и нажатие на кнопку OK позволяют перейти к окну функциональных возможностей модуля «Иерархические агломеративные методы», в котором необходимо выбрать переменные для анализа и задать основные параметры классификации.

Выбор переменных для анализа осуществляется нажатием на кнопку Variables: – **Переменные** на форме **Cluster analysis: Joining**. Вид формы отбора признаков для анализа представлен на рисунке 8.

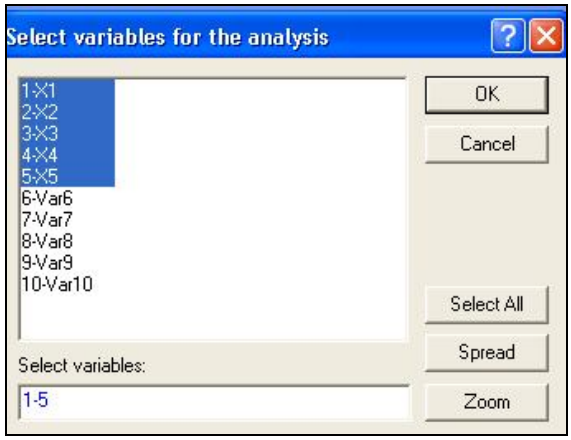


Рисунок 8 – Выбор переменных для анализа

В поле **Input file** следует задать вид входной информации:

Raw data – матрица типа «объект-свойство»;

Distance matrix – матрица расстояний.

Так как исходные данные представлены в виде матрицы X типа «объект-свойство», то в поле **Input file** следует установить **Raw data.** Вид формы задания типа входной информации представлен на рисунке 9.

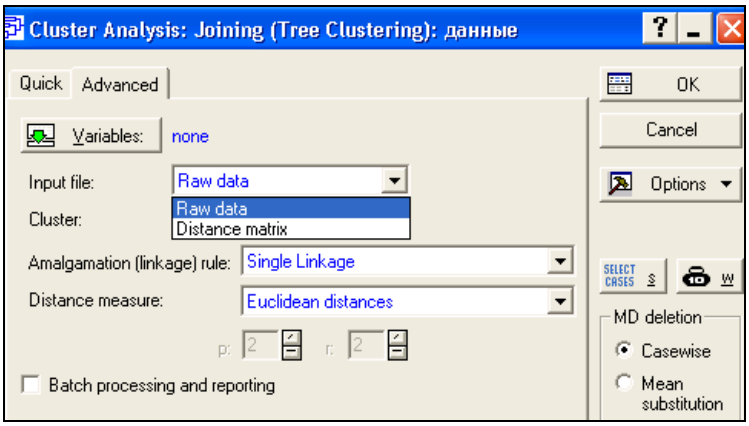


Рисунок 9 – Форма задания типа входной информации

В поле **Claster** устанавливают объект классификации:

Cases (rows)/строки – классификация объектов наблюдения;

Variables (columns)/столбцы – классификация признаков.

Так как необходимо провести классификацию объектов – муниципальных образований, то в поле **Claster** необходимо установить режим **Cases (rows).** Форма задания режима классификации представлена на рисунке 10.

Рисунок 10 – Форма задания режима классификации

На следующем этапе необходимо определить правило объединения кластеров. При нажатии на кнопку **Amalgamation (linkage) rule,** появляется окно, в котором предложены различные методы объединения кластеров. Вид экрана представлен на рисунке 11.

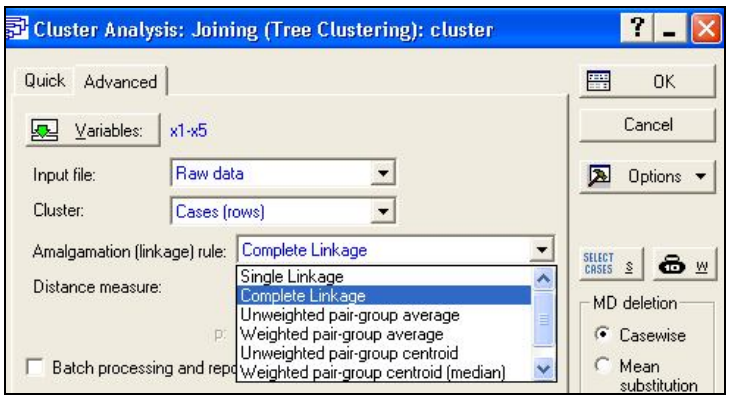


Рисунок 11 – Методы объединения кластеров

В пакете Statistica реализованы следующие агломеративные методы классификации:

Single linkage – метод «одиночной связи»;

Complete linkage – метод «полных связей»;

Unweighted pair group average – метод «средней связи»;

Weighted pair group average – взвешенный метод средней связи**;**

Unweighted pair group centroid – центроидный метод (невзвешенный);

Weighted pair group centroid – взвешенный центроидный метод;

Ward′s method – метод Уорда.

Поскольку метод «одиночной связи» не позволяет определить наиболее подходящее число классов в исследуемой совокупности объектов, воспользуемся для классификации, например, методом «полных связей».

Далее необходимо задать метрику расстояний. При нажатии на кнопку **Distance matrix,** появляется окно, представленное на рисунке 12, в котором предложены следующие метрики для расчета расстояний:

Squared euclidean distance – квадратичное евклидово расстояние;

Euclidean distance – обычное евклидово расстояние;

City-block (Manhattan) distances – манхеттенское расстояние;

Chebychev distance metric – расстояние Чебышева;

Power distance – специальный класс метрических функций (расстояние Минковского).

В качестве метрики расстояния между объектами выберем обычное евклидово расстояние.

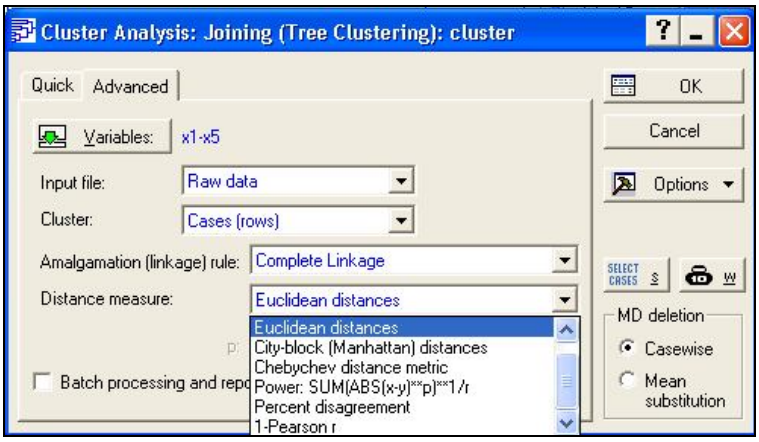


Рисунок 12 – Метрики расстояний между объектами

После задания всех необходимых параметров и нажатия кнопки 

, будут произведены вычисления, и на экране появится форма **Joining Results**, содержащая результаты кластерного анализа. Вид формы представлен на рисунке 13.

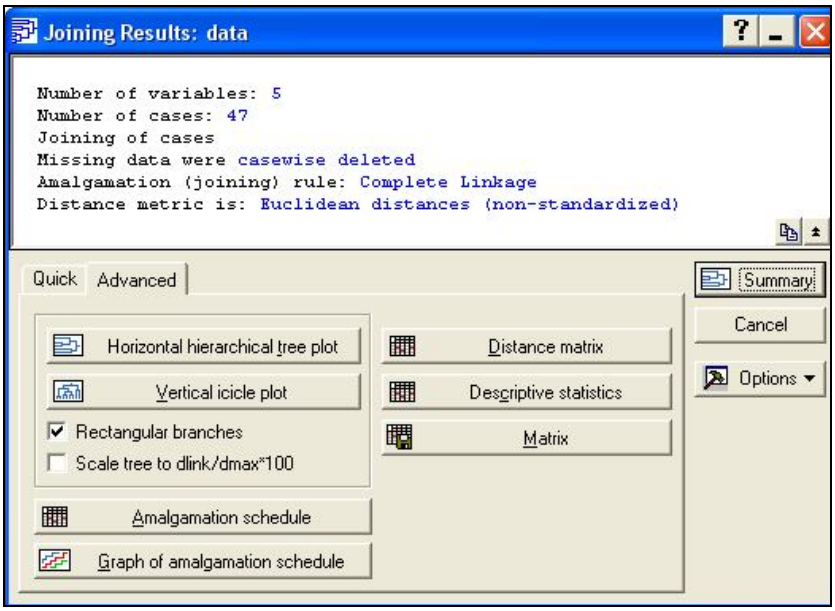


Рисунок 13 – Вид окна для вывода результата расчетов кластерного анализа

Для построения вертикальной дендрограммы необходимо нажать кнопку 

. График объединения классов представлен на рисунке 14.

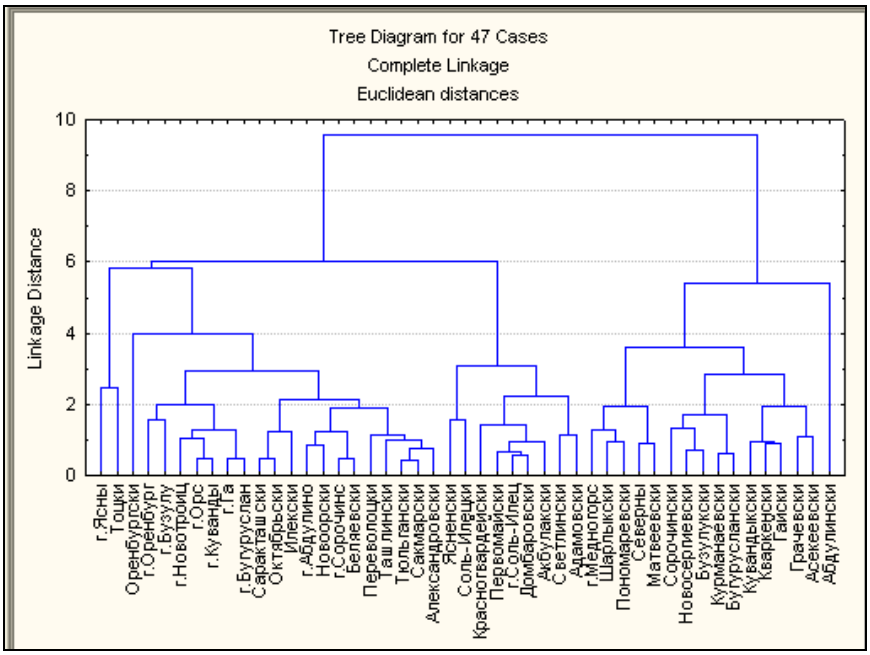


Рисунок 14 – Дендрограмма объединения классов методом «полных связей»

Методом «полных связей» при пороговом значении расстояния $\rho_{пор} = 5{,}9$ все города и районы Оренбургской области разбиваются на три класса $S_1 = \{S_{11}, S_{12}, S_{13}\}$, состав которых приведен в таблице 2.

Таблица 2 – Результаты классификация муниципальных образований Оренбургской области методом «полных связей»

Номер кластера	Количество объектов в кластере	Состав класса
1	2	3
$\{S_{11}\}$	22	**Города:** Ясный, Оренбург, Бузулук, Новотроицк, Орск, Кувандык, Гай, Бугуруслан, Абдулино, Сорочинск. **Районы:** Оренбургский, Тоцкий, Саракташ ский, Октябрьский, Илекский, Новоорский, Беляевский, Переволоцкий, Ташлинский, Тюльганский, Сакмарский, Александровский.

Продолжение таблицы 2

1	2	3
$\{S_{12}\}$	9	**Город:** Соль-Илецк. **Районы:** Ясненский, Соль-Илецкий, Красногвардейский, Первомайский, Домбаровский, Акбулакский, Светлинский, Адамовский.
$\{S_{13}\}$	16	**Город:** Медногорск. **Районы:** Шарлыкский, Пономаревский, Северный, Матвеевский, Сорочинский, Новосергиевский, Бузулукский, Курманаевский, Бугурусланский, Кувандыкский, Кваркенский, Гайский, Грачевский, Асекеевский, Абдулинский.

В данном случае уровень порогового значения выбирается из тех соображений, чтобы получить небольшое количество кластеров.

Далее рассчитываются средние значения показателей в каждом классе (приложение Б, таблица Б.1). Графическое изображение информации о средних значениях признаков в классах представлено на рисунке 15.

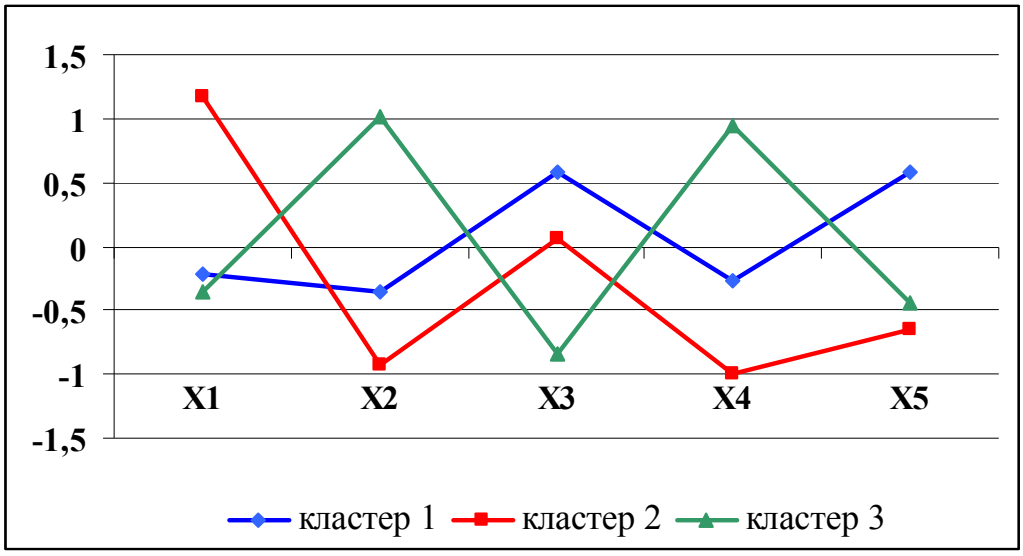


Рисунок 15 – График средних значений признаков в каждом кластере

Анализируя график средних значений в классах, можно сделать следующие выводы:

- Первый класс, куда вошло большинство городов Оренбургской области, характеризуется по сравнению с остальными классами наибольшими средними значениями таких показателей, как удельный вес населения в трудоспособном возрасте (X_3) и миграционный прирост населения (X_5), при этом на достаточно низком уровне зафиксировано среднее значение общего коэффициента рождаемости (X_1).

- Объекты второго класса характеризуются самым высоким по сравнению с первым и третьим классами средним значением общего коэффициента рождаемости (X_1) и самым низким средним значением общего коэффициента смертности (X_2), что свидетельствует о значительном естественном приросте населения. В то же время объекты данного класса характеризуются низким уровнем механического движения населения, о чем свидетельствует коэффициент миграционного прироста(X_5), который для объектов второго класса ниже, чем для объектов других классов.

- Третий класс схож со вторым только по показателю миграционного прироста населения (X_5), который находится на достаточно низком уровне. По всем остальным показателям объекты третьего класса являются полной противоположностью объектам второго класса: на фоне низкого уровня рождаемости (X_1) зафиксирован самый высокий уровень смертности (X_2) в среднем по классу. Муниципальные образования данного класса характеризуются наименьшим удельным весом населения в трудоспособном возрасте (X_3) и наибольшим удельным весом населения старше трудоспособного возраста (X_4).

С помощью метода «полных связей» получено достаточно неравномерное распределение объектов по классам, так во второй класс вошло 9 объектов, в то время как в первый класс -22 объекта. Данный недостаток можно устранить методом Уорда.

После задания в поле **Amalgamation (linkage) rule Ward′s method** (метод Уорда) и нажатия кнопки OK (форма окна представлена на рисунке 16)**,** будут

произведены вычисления, и на экране появится форма **Joining Results**, содержащая результаты кластерного анализа указанным методом.

Рисунок 16 – Методы объединения кластеров

Вертикальная дендрограмма объединения классов методом Уорда представлена на рисунке 17.

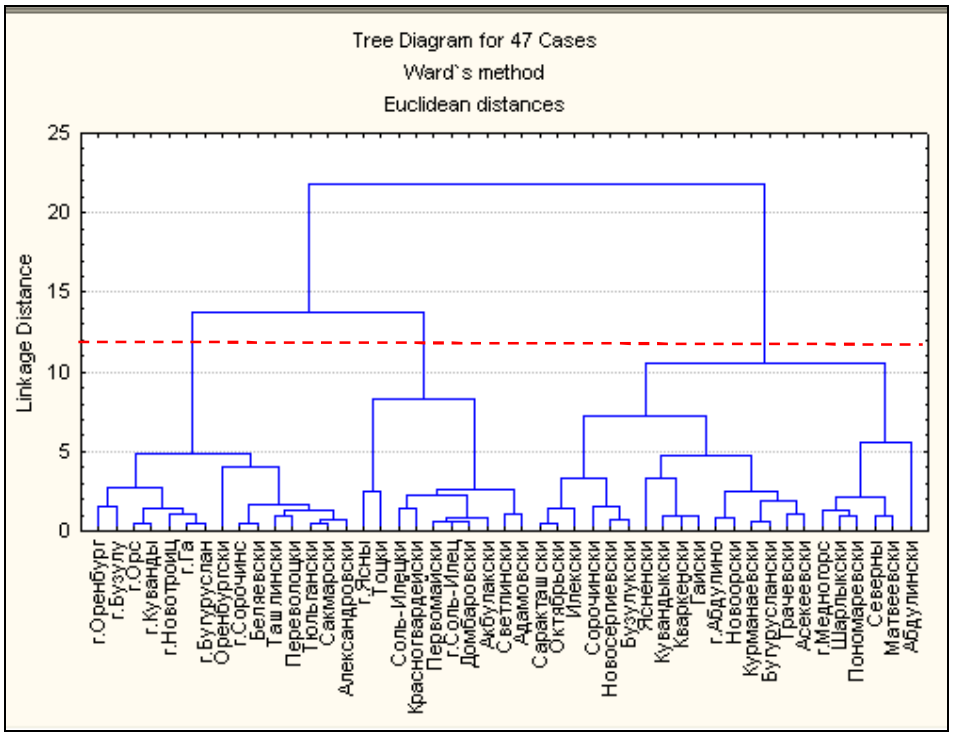


Рисунок 17 – Дендрограмма объединения классов методом Уорда

Методом Уорда при пороговом значении расстояния $\rho_{пор} = 12$ все города и районы Оренбургской области разбиваются на три класса $S_2 = \{S_{21}, S_{22}, S_{23}\}$, состав которых приведен в таблице 3.

Таблица 3 – Результаты классификация муниципальных образований Оренбургской области методом Уорда

Номер кластера	Количество объектов в кластере	Состав класса
$\{S_{21}\}$	15	**Города:** Оренбург, Бузулук, Орск, Кувандык, Новотроицк, Гай, Бугуруслан, Сорочинск. **Районы:** Оренбургский, Беляевский, Ташлинский, Переволоцкий, Тюльганский, Сакмарский, Александровский.
$\{S_{22}\}$	10	**Города:** Ясный, Соль-Илецк. **Районы:** Тоцкий, Соль-Илецкий, Красногвардейский, Первомайский, Домбаровский, Акбулакский, Светлинский, Адамовский.
$\{S_{23}\}$	22	**Города:** Абдулино, Медногорск. **Районы:** Саракташский, Октябрьский, Илекский, Сорочинский, Новосергиевский, Бузулукский, Ясненский, Кувандыкский, Кваркенский, Гайский, Новоорский, Курманаевский, Бугурусланский, Грачевский, Асекеевский, Шарлыкский, Пономаревский, Северный, Матвеевский, Абдулинский.

Средние значения в каждом классе, представленные в приложении Б (таблица Б.2) и на рисунке 18, позволяют сделать следующие выводы:

- Первый класс муниципальных образований Оренбургской области, преимущественно города и примыкающие к ним районы, характеризуется наибольшим средним значением миграционного прироста населения (X_5) и наименьшим средним значением общего коэффициента рождаемости (X_1).

- Объекты второго класса, напротив, характеризуется наибольшим средним значением рождаемости (X_2), но достаточно низким средним значением миграционного прироста населения (X_5). Самое низкое среднее значение зафиксировано для таких показателей, как общий коэффициент смертности (X_2), удельный вес населения старше трудоспособного возраста (X_4).

- Третий класс объектов характеризуется наибольшим средним значением таких показателей, как общий коэффициент смертности (X_2) и удельный вес населения старше трудоспособного возраста (X_4). Что касается среднего значения удельного веса населения в трудоспособном возрасте (X_3), то для городов и районов третьего класса оно значительно ниже, чем для объектов первого и второго класса.

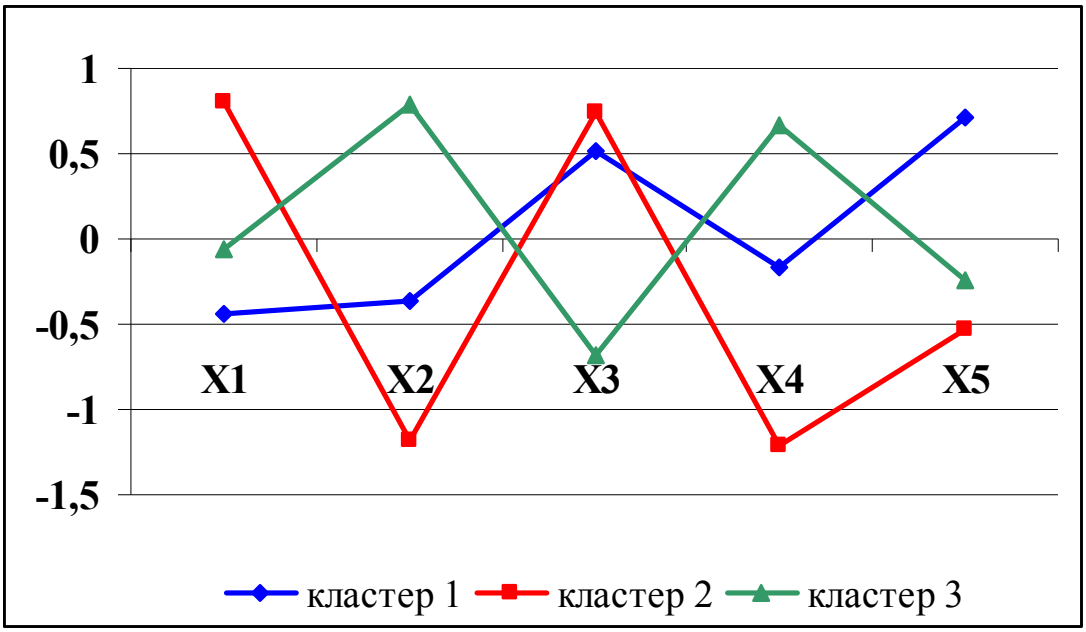


Рисунок 18 - График средних значений признаков в каждом кластере

Кнопка Amalgamation schedule на форме результатов **Joining Results** предназначена для вывода на экран протокола объединения классов. Протокол объединения классов методом Уорда представлен на рисунке 19.

График изменения расстояния между объединенными классами может быть получен нажатием на кнопку Graph of amalgamation schedule. График изменения расстояния при объединении кластеров методом Уорда представлен в приложении Б (рисунок Б.1).

Amalgamation Schedule (data)
Ward`s method
Euclidean distances

linkage distance	**Obj. No. 1**	Obj. No. 2	Obj. No. 3	Obj. No. 4
,4267506	Сакмарский	Тюльганский		
,4623814	г.Кувандык	г.Орск		
,4867061	г.Бугуруслан	г.Гай		
,4896898	Беляевский	г.Сорочинск		
,4958209	Октябрьский	Саракташский		
,5756568	Домбаровский	г.Соль-Илецк		
,6221970	Бугурусланский	Курманаевский		
,6317909	Домбаровский	г.Соль-Илецк	Первомайский	
,7028964	Александровский	Сакмарский	Тюльганский	
,7096063	Бузулукский	Новосергиевский		
,8725668	Новоорский	г.Абдулино		
,8826820	Акбулакский	Домбаровский	г.Соль-Илецк	Первомайский
,9147993	Гайский	Кваркенский		
,9228597	Матвеевский	Северный		
,9338040	Пономаревский	Шарлыкский		
,9685140	Гайский	Кваркенский	Кувандыкский	

Рисунок 19 – Протокол объединения кластеров

Для просмотра матрицы расстояний необходимо нажать на кнопку Distance matrix. Матрица обычных евклидовых расстояний между объектами представлена на рисунке 20.

Euclidean distances (data)

Case No.	**Абдулинский**	Адамовский	Акбулакский	Александровский	Асекеевский
Абдулинский	0,00	6,89	7,45	5,91	4,04
Адамовский	6,89	0,00	1,06	1,71	3,06
Акбулакский	7,45	1,06	0,00	1,97	3,50
Александровский	5,91	1,71	1,97	0,00	2,17
Асекеевский	4,04	3,06	3,50	2,17	0,00
Беляевский	6,36	1,87	1,66	1,07	2,44
Бугурусланский	4,51	3,36	3,64	2,59	1,12
Бузулукский	5,43	3,71	3,53	2,84	2,24
Гайский	5,28	2,84	3,15	2,98	1,92
Грачевский	4,64	2,36	2,94	1,99	1,08
Домбаровский	7,55	0,88	0,72	2,36	3,67
Илекский	5,91	2,55	2,30	1,83	2,15
Кваркенский	5,38	2,11	2,63	2,32	1,70
Красногвардейский	6,36	1,26	1,22	1,62	2,50
Кувандыкский	4,92	2,56	3,05	2,66	1,63
Курманаевский	4,56	3,13	3,37	2,10	0,90
Матвеевский	3,58	3,64	4,14	2,42	1,09
Новоорский	5,81	1,75	1,94	1,59	1,81
Новосергиевский	5,31	3,45	3,41	2,53	1,99

Рисунок 20 – Матрица расстояний

Кнопка Descriptive statistics на форме результатов предназначена для вывода на экран описательных статистик для каждого объекта: среднего арифметического и стандартного отклонения.

Кнопка Matrix на форме результатов позволяет сохранить матрицу расстояний в файле.

Классификация муниципальных образований методом К-средних

Использование различных методов иерархического агломеративного кластерного анализа приводит к различным результатам классификации. Метод *К*-средних позволяет получить более устойчивое разбиение, но требует задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.).

Выбор процедуры **K-mean clustering,** представленной на рисунке 7, и нажатие на кнопку OK позволяют перейти к окну функциональных возможностей метода К-средних, которое содержит следующие параметры:

Cluster – классификация признаков или объектов;

Number of clasters – число кластеров;

Number of iteration – число итераций (установленных по умолчанию 10 итераций, как правило, вполне достаточно для получения устойчивого разбиения).

Радио-кнопки в группе **Initial cluster centers** задают способ определения начальных эталонов классов.

Вид формы задания параметров классификации методом *К*-средних представлен на рисунке 21.

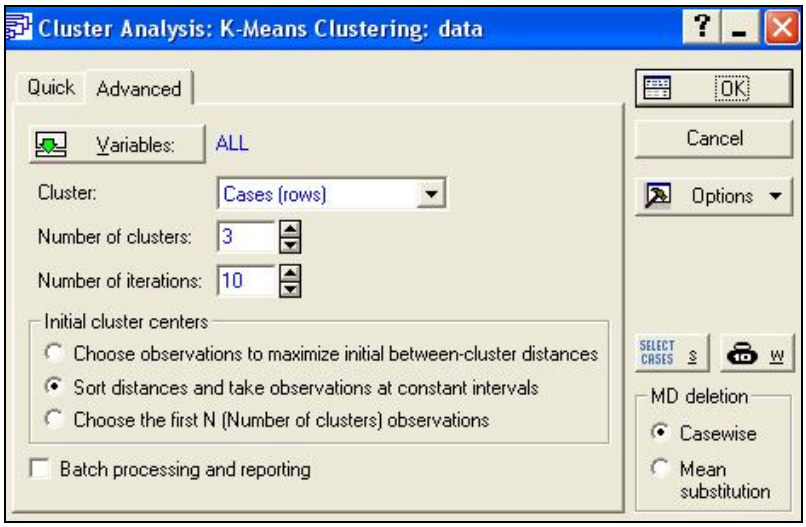


Рисунок 21 – Форма задания параметров классификации методом *K* -средних

С помощью иерархических агломеративных методов кластерного анализа было выявлено, что 47 муниципальных образований Оренбургской области целесообразно разбить на три класса (наглядной является дендрограмма объединения методом Уорда).

После нажатия кнопки OK будут произведены вычисления, и на экране появится форма результатов классификации **k-Means Clustering Results**, представленная на рисунке 22.

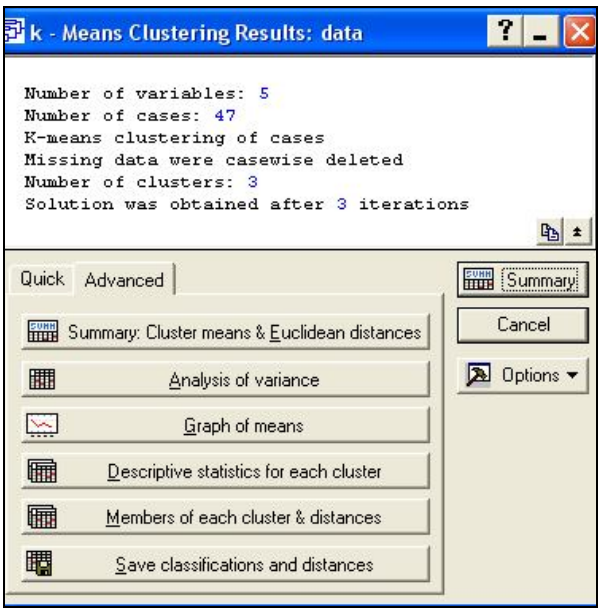


Рисунок 22 – Вид окна для вывода результатов классификации методом *K* -средних

В верхней части формы указаны параметры классификации, в нижней части расположены кнопки для вывода различной информации по кластерам.

Кнопка Members of each cluster & distances на форме результатов позволяет вывести на экран информацию о количестве объектов в кластерах, состав кластеров, а также евклидово расстояние от объектов до центра соответствующего класса. Результаты представлены на рисунках 23,24,25.

	Members of Cluster Number 1 and Distances from Respectiv Cluster contains 20 cases		
linkage	**Distance**		
Александровский	0,340506		
Беляевский	0,281132		
Илекский	0,661918		
Новоорский	0,667315		
Октябрьский	0,528780		
Оренбургский	1,228103		
Переволоцкий	0,463234		
Сакмарский	0,246217		
Саракташский	0,670688		
Ташлинский	0,504604		
Тюльганский	0,156300		
г.Абдулино	0,489344		
г.Бугуруслан	0,406516		
г.Бузулук	0,722880		
г.Гай	0,449031		
г.Кувандык	0,448579		
г.Новотроицк	0,680501		

Рисунок 23 – Состав кластера 1

	Members of Cluster Number 2 and Distances from Respectiv Cluster contains 11 cases		
linkage	**Distance**		
Адамовский	0,331865		
Акбулакский	0,385424		
Домбаровский	0,246801		
Красногвардейский	0,637220		
Первомайский	0,378690		
Светлинский	0,736587		
Соль-Илецкий	0,762347		
Тоцкий	1,724758		
Ясненский	1,094173		
г.Соль-Илецк	0,210520		
г.Ясный	1,105544		

Рисунок 24 – Состав кластера 2

linkage	Members of Cluster Number 3 (and Distances from Respective Cluster contains 16 cases **Distance**
Абдулинский	1,818957
Асекеевский	0,219679
Бугурусланский	0,388565
Бузулукский	0,826171
Гайский	0,889613
Грачевский	0,669316
Кваркенский	0,845459
Кувандыкский	0,842611
Курманаевский	0,289158
Матвеевский	0,517347
Новосергиевский	0,705733
Пономаревский	0,666051
Северный	0,716263
Сорочинский	0,567721
Шарлыкский	0,614279
г.Медногорск	0,800440

Рисунок 25 – Состав кластера 3

Классификация муниципальных образований на три класса методом К-средних $S_3 = \{S_{31}, S_{32}, S_{33}\}$ представлена в таблице 4.

Таблица 4 – Результаты классификация муниципальных образований Оренбургской области методом K -средних

Номер кластера	Количество объектов в кластере	Состав класса
1	2	3
$\{S_{31}\}$	20	**Города:** Абдулино, Бугуруслан, Бузулук, Гай, Кувандык, Новотроицк, Оренбург, Орск, Сорочинск. **Районы:** Александровский, Беляевский, Илекский, Новоорский, Октябрьский, Оренбургский, Переволоцкий, Сакмарский, Саракташский, Ташлинский, Тюльганский.

Продолжение таблицы 4

1	2	3
$\{S_{32}\}$	11	**Города:** Соль-Илецк, Ясный. **Районы:** Адамовский, Акбулакский, Домбаровский, Красногвардейский, Первомайский, Светлинский, Соль-Илецкий, Тоцкий, Ясненский.
$\{S_{33}\}$	16	**Город:** Медногорск. **Районы:** Абдулинский, Асекеевский, Бугурусланский, Бузулукский, Гайский, Грачевский, Кваркенский, Кувандыкский, Курманаевский, Матвеевский, Новосергиевский, Пономаревский, Северны, Сорочинский, Шарлыкский.

При нажатии на кнопку Summary: Cluster means & Euclidean distances появится окно, содержащее две таблицы. В первой таблице, представленной на рисунке 26, указаны средние значения признаков в каждом классе. Во второй таблице, представленной на рисунке 27, приведены расстояния между классами. Причем, ниже главной диагонали указаны расстояния между классами, рассчитанные по метрике обычного евклидового расстояния, а выше главной диагонали – расстояния между классами, рассчитанные по метрике квадратичного евклидового расстояния.

	Cluster Means (data)		
Variable	**Cluster No. 1**	Cluster No. 2	Cluster No. 3
x1	-0,249363	0,96238	-0,349932
x2	-0,206200	-1,10482	1,017316
x3	0,323376	0,62333	-0,832759
x4	-0,078750	-1,22842	0,942978
x5	0,711850	-0,65465	-0,439743

Рисунок 26 – Средние значения признаков в классах

Cluster Number	Euclidean Distances between C Distances below diagonal Squared distances above diago		
	No. 1	No. 2	No. 3
No. 1	0,000000	1,110974	1,042769
No. 2	1,054028	0,000000	2,621398
No. 3	1,021161	1,619073	0,000000

Рисунок 27 – Расстояния между классами

Как видно из рисунка 27 наименьшее расстояние наблюдается между первым и третьим классами (1,021161).

Кнопка Analysis of variance (анализ дисперсий) на форме результатов позволяет вывести на экран информацию о значениях сумм квадратов при расчете межгрупповой дисперсии (Between) и внутригрупповой дисперсии (Within) по каждому признаку, а также соответствующие им степени свободы. Результаты представлены на рисунке 28.

Variable	Analysis of Variance (data)					
	Between SS	df	Within SS	df	F	signif. p
x1	13,39080	2	32,60920	44	9,03418	0,000516
x2	30,83622	2	15,16378	44	44,73799	0,000000
x3	17,46118	2	28,53882	44	13,46047	0,000027
x4	30,95060	2	15,04940	44	45,24520	0,000000
x5	17,94275	2	28,05725	44	14,06911	0,000019

Рисунок 28 – Анализ дисперсий

Чтобы получить значения межгрупповых и внутригрупповых дисперсий, необходимо сумму квадратов поделить на соответствующее число степеней свободы. Рассчитанные таким образом межгрупповые и внутригрупповые дисперсии представлены в таблице 5.

Таблица 5 – Значения межгрупповых и внутригрупповых дисперсий

Признаки	Межгрупповая дисперсия	Внутригрупповая дисперсия
X_1	6,695	0,741
X_2	15,418	0,345
X_3	8,731	0,649
X_4	15,475	0,342
X_5	8,971	0,638

Таблица, представленная на рисунке 28, содержит также наблюденное значение F-критерия, а также значимость нулевой гипотезы о равенстве межгрупповой и внутригрупповой дисперсий. На уровне значимости 0,05 по всем признакам нулевая гипотеза отвергается. Это означает, что каждый из признаков вносит существенный вклад в разделение объектов на классы.

Кнопка Graph of means на форме результатов предназначена для вывода графического изображения информации, содержащейся в таблице, представленной на рисунке 26. График средних значений признаков в классах представлен на рисунке 29.

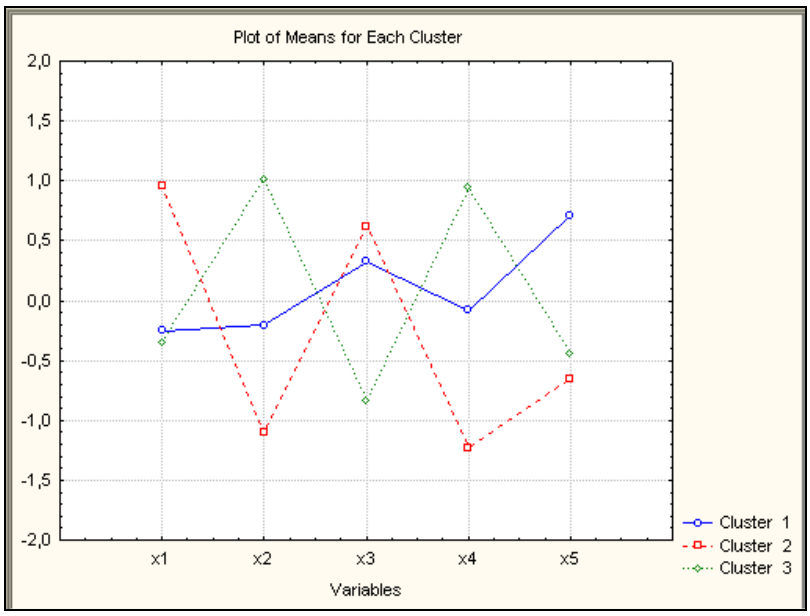


Рисунок 29 – График средних значений признаков в каждом кластере

Данный график, благодаря своей наглядности, оказывается полезным при интерпретации результатов классификации, которая приведена ниже.

Кнопка Descriptive statistics for each cluster позволяет вывести на экран результаты расчетов описательных статистик для каждого кластера: среднего арифметического, оценку среднего квадратичного отклонения, несмещенную оценку дисперсии по каждому признаку. Результаты расчетов представлены на рисунке 30.

Descriptive Statistics for Cluster 1
Cluster contains 20 cases

Variable	Mean	Standard Deviation	Variance
x1	-0,249363	0,586351	0,343807
x2	-0,206200	0,606607	0,367972
x3	0,323376	0,555507	0,308587
x4	-0,078750	0,343022	0,117664
x5	0,711850	0,748142	0,559716

Descriptive Statistics for Cluster 2
Cluster contains 11 cases

Variable	Mean	Standard Deviation	Variance
x1	0,96238	0,943912	0,890970
x2	-1,10482	0,567832	0,322434
x3	0,62333	1,344829	1,808566
x4	-1,22842	0,596331	0,355611
x5	-0,65465	0,581453	0,338088

Descriptive Statistics for Cluster 3
Cluster contains 16 cases

Variable	Mean	Standard Deviation	Variance
x1	-0,349932	1,069803	1,144478
x2	1,017316	0,574339	0,329865
x3	-0,832759	0,553173	0,306000
x4	0,942978	0,785607	0,617179
x5	-0,439743	0,967532	0,936118

Рисунок 30 – Результаты расчета описательных статистик для каждого кластера

Сравнение классификаций

С помощью метода «полных связей», метода Уорда и метода К-средних были получены различные классификации. Сводная таблица результатов классификаций муниципальных образований Оренбургской области, полученных различными методами кластерного анализа, приведена в приложении Б (таблица Б.3)

Для выбора лучшей классификации необходимо воспользоваться функционалами качества разбиения.

Наиболее удобным, с точки зрения реализации на ЭВМ, функционалом качества является сумма квадратов расстояний от каждого объекта до центра кластера (3):

$$Q_1(S) = \sum_{l=1}^{p} \sum_{O_i \in S_l} d^2(O_i, \overline{X}(l)) \to \min \qquad (3)$$

Для нахождения функционала качества разбиений, полученных с помощью иерархических агломеративных кластер-процедур необходимо:

1) в таблице с исходными данными оставить только те объекты, которые были отнесены к первому классу;

2) вычислить средние значения для каждого признака и добавить их в качестве последней строки в исходные данные;

3) рассчитать матрицу расстояний между объектами с помощью кнопки Distance matrix;

4) в последней строке (столбце) матрицы будут стоять расстояния от объектов, относящихся к первому классу, до центра первого класса;

5) с помощью пакета Excel рассчитать сумму квадратов расстояний;

6) проделать шаги 1-5 для каждого кластера;

7) просуммировать полученные значения квадратов расстояний для каждого кластера.

Рассчитаем функционал качества классификации, полученной методом Уорда.

На рисунке 31 представлены значения признаков для 15 объектов, отнесенных к первому классу; 10 объектов, отнесенных ко второму классу и 22 объектов, отнесенных к третьему классу. В последней строке введены средние значения каждого признака.

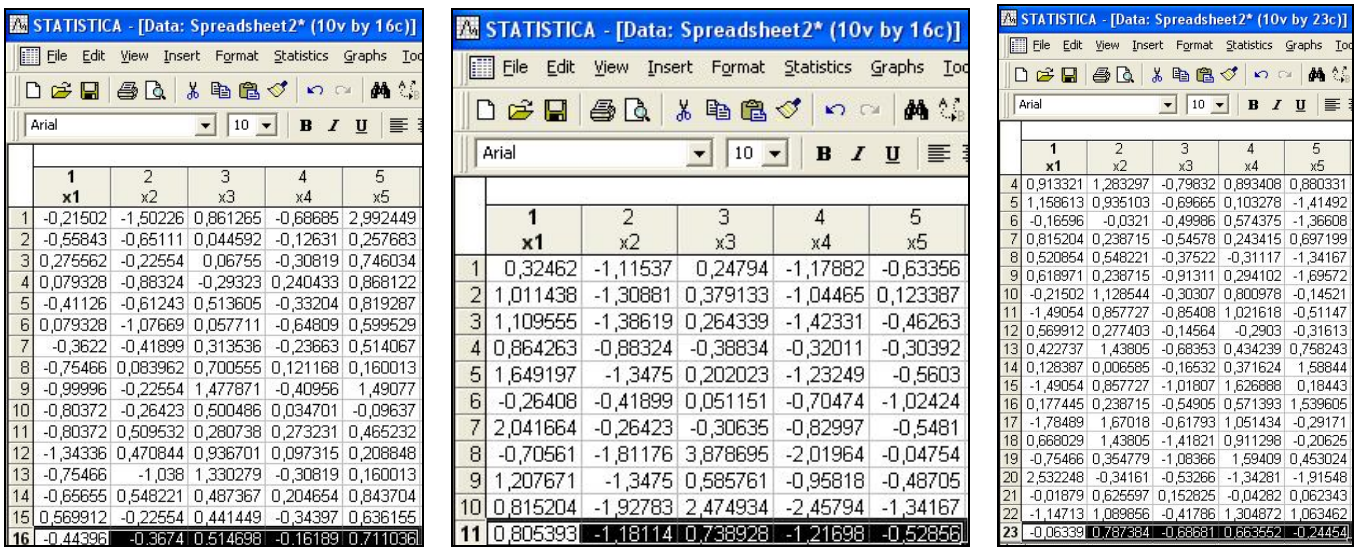
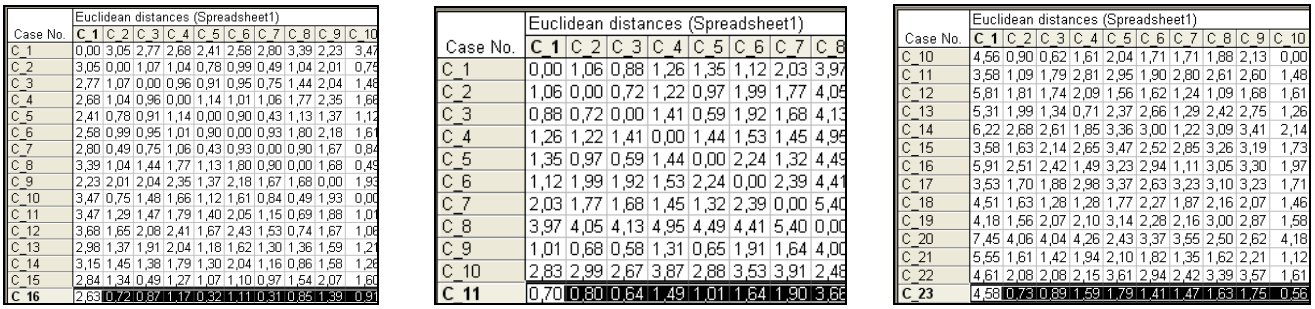


Рисунок 31 – Объекты, отнесенные к первому, второму, третьему классу соответственно

По имеющимся данным рассчитаны матрицы расстояний, представленные на рисунке 32.



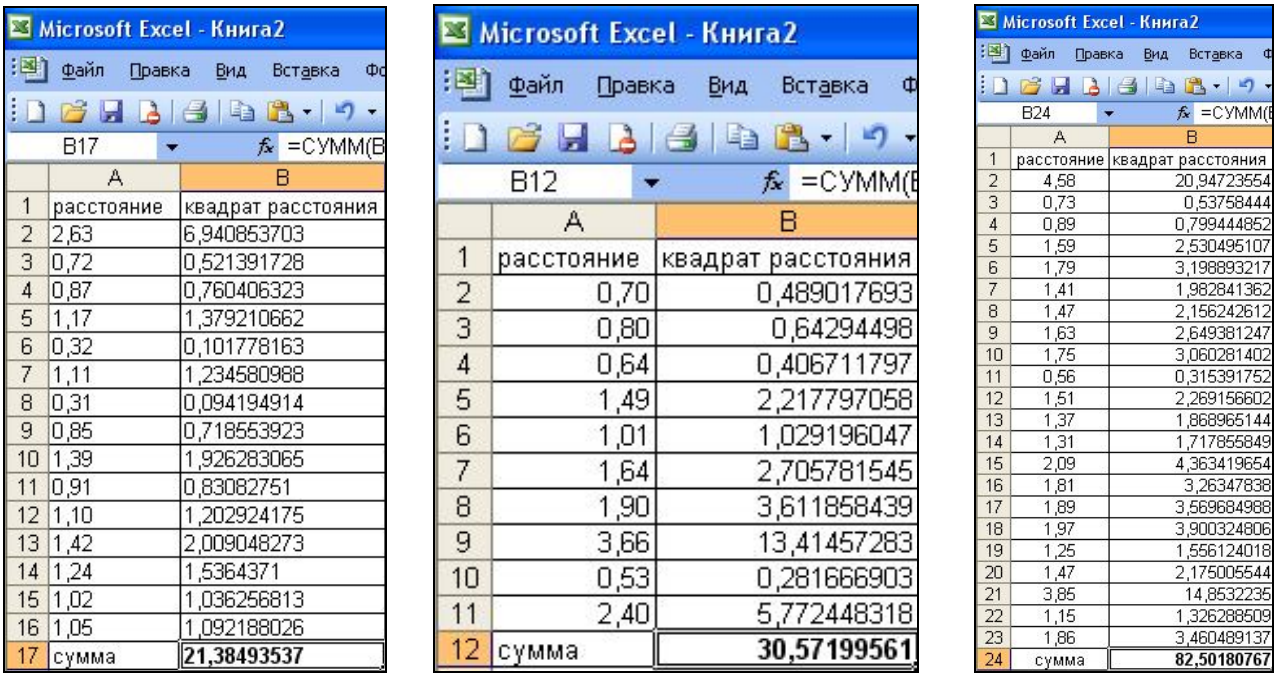
Case No.	Euclidean distances (Spreadsheet1)									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
C_1	0,00	3,05	2,77	2,68	2,41	2,58	2,80	3,39	2,23	3,47
C_2	3,05	0,00	1,07	1,04	0,78	0,99	0,49	1,04	2,01	0,75
C_3	2,77	1,07	0,00	0,96	0,91	0,95	0,75	1,44	2,04	1,46
C_4	2,68	1,04	0,96	0,00	1,14	1,01	1,06	1,77	2,35	1,66
C_5	2,41	0,78	0,91	1,14	0,00	0,90	0,43	1,13	1,37	1,12
C_6	2,58	0,99	0,95	1,01	0,90	0,00	0,93	1,80	2,18	1,61
C_7	2,80	0,49	0,75	1,06	0,43	0,93	0,00	0,90	1,67	0,84
C_8	3,39	1,04	1,44	1,77	1,13	1,80	0,90	0,00	1,68	0,49
C_9	2,23	2,01	2,04	2,35	1,37	2,18	1,67	1,68	0,00	1,93
C_10	3,47	0,75	1,48	1,66	1,12	1,61	0,84	0,49	1,93	0,00
C_11	3,47	1,29	1,47	1,79	1,40	2,05	1,15	0,69	1,88	1,01
C_12	3,68	1,65	2,08	2,41	1,67	2,43	1,53	0,74	1,67	1,06
C_13	2,98	1,37	1,91	2,04	1,18	1,62	1,30	1,36	1,59	1,21
C_14	3,15	1,45	1,38	1,79	1,30	2,04	1,16	0,86	1,58	1,26
C_15	2,84	1,34	0,49	1,27	1,07	1,10	0,97	1,54	2,07	1,60
C_16	2,63	0,72	0,87	1,17	0,32	1,11	0,31	0,85	1,39	0,91

Case No.	Euclidean distances (Spreadsheet1)							
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
C_1	0,00	1,06	0,88	1,26	1,35	1,12	2,03	3,97
C_2	1,06	0,00	0,72	1,22	0,97	1,99	1,77	4,05
C_3	0,88	0,72	0,00	1,41	0,59	1,92	1,68	4,13
C_4	1,26	1,22	1,41	0,00	1,44	1,53	1,45	4,95
C_5	1,35	0,97	0,59	1,44	0,00	2,24	1,32	4,49
C_6	1,12	1,99	1,92	1,53	2,24	0,00	2,39	4,41
C_7	2,03	1,77	1,68	1,45	1,32	2,39	0,00	5,40
C_8	3,97	4,05	4,13	4,95	4,49	4,41	5,40	0,00
C_9	1,01	0,68	0,58	1,31	0,65	1,91	1,64	4,00
C_10	2,83	2,99	2,67	3,87	2,88	3,53	3,91	2,48
C_11	0,70	0,80	0,64	1,49	1,01	1,64	1,90	3,66

Case No.	Euclidean distances (Spreadsheet1)									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
C_10	4,56	0,90	0,62	1,61	2,04	1,71	1,71	1,88	2,13	0,00
C_11	3,58	1,09	1,79	2,81	2,95	1,90	2,80	2,61	2,60	1,48
C_12	5,81	1,81	1,74	2,09	1,56	1,62	1,24	1,09	1,68	1,61
C_13	5,31	1,99	1,34	0,71	2,37	2,66	1,29	2,42	2,75	1,26
C_14	6,22	2,68	2,61	1,85	3,36	3,00	1,22	3,09	3,41	2,14
C_15	3,58	1,63	2,14	2,65	3,47	2,52	2,85	3,26	3,19	1,73
C_16	5,91	2,51	2,42	1,49	3,23	2,94	1,11	3,05	3,30	1,97
C_17	3,53	1,70	1,88	2,98	3,37	2,63	3,23	3,10	3,23	1,71
C_18	4,51	1,63	1,28	1,28	1,77	2,27	1,87	2,16	2,07	1,46
C_19	4,18	1,56	2,07	2,10	3,14	2,28	2,16	3,00	2,87	1,58
C_20	7,45	4,06	4,04	4,26	2,43	3,37	3,55	2,50	2,62	4,18
C_21	5,55	1,61	1,42	1,94	2,10	1,82	1,35	1,62	2,21	1,12
C_22	4,61	2,08	2,08	2,15	3,61	2,94	2,42	3,39	3,57	1,61
C_23	4,58	0,73	0,89	1,59	1,79	1,41	1,47	1,63	1,75	0,56

Рисунок 32 – Матрицы расстояний

В последней строке (столбце) данных матриц стоят расстояния от объектов до центра соответствующего класса. Результаты расчетов суммы квадратов расстояний представлены на рисунке 33.



Microsoft Excel - Книга2

Файл Правка Вид Вставка Фо

B17 =СУММ(B

	A	B
1	расстояние	квадрат расстояния
2	2,63	6,940853703
3	0,72	0,521391728
4	0,87	0,760406323
5	1,17	1,379210662
6	0,32	0,101778163
7	1,11	1,234580988
8	0,31	0,094194914
9	0,85	0,718553923
10	1,39	1,926283065
11	0,91	0,83082751
12	1,10	1,202924175
13	1,42	2,009048273
14	1,24	1,5364371
15	1,02	1,036256813
16	1,05	1,092188026
17	сумма	**21,38493537**

Microsoft Excel - Книга2

Файл Правка Вид Вставка Ф

B12 =СУММ(B

	A	B
1	расстояние	квадрат расстояния
2	0,70	0,489017693
3	0,80	0,64294498
4	0,64	0,406711797
5	1,49	2,217797058
6	1,01	1,029196047
7	1,64	2,705781545
8	1,90	3,611858439
9	3,66	13,41457283
10	0,53	0,281666903
11	2,40	5,772448318
12	сумма	**30,57199561**

Microsoft Excel - Книга2

Файл Правка Вид Вставка Ф

B24 =СУММ(B

	A	B
1	расстояние	квадрат расстояния
2	4,58	20,94723554
3	0,73	0,53758444
4	0,89	0,799444852
5	1,59	2,530495107
6	1,79	3,198893217
7	1,41	1,982841362
8	1,47	2,156242612
9	1,63	2,649381247
10	1,75	3,060281402
11	0,56	0,315391752
12	1,51	2,269156602
13	1,37	1,868965144
14	1,31	1,717855849
15	2,09	4,363419654
16	1,81	3,26347838
17	1,89	3,569684988
18	1,97	3,900324806
19	1,25	1,556124018
20	1,47	2,175005544
21	3,85	14,8532235
22	1,15	1,326288509
23	1,86	3,460489137
24	сумма	**82,50180767**

Рисунок 33 – Результаты расчетов суммы квадратов расстояний

Тогда значение функционала качества для классификации, полученной методом Уорда, рассчитывается следующим образом:

$$Q(S_2) = 21{,}38 + 30{,}57 + 82{,}50 = 134{,}45$$

Аналогичным образом можно рассчитать функционал качества для классификации, полученной методом «полных связей»:

$$Q(S_1) = 67{,}42 + 12{,}19 + 50 = 129{,}61$$

Значение функционала качества $Q(S_3)$ для классификации, полученной методом К-средних, рассчитывается на основе таблиц, представленных на рисунках 23-25.

$$Q(S_3) = 6{,}45 + 7{,}43 + 10 = 23{,}88$$

По выбранному функционалу качества наилучшей является классификация $S_3 = \{S_{31}, S_{32}, S_{33}\}$, полученная методом К-средних.

Содержательная интерпретация результатов классификации

Для того чтобы дать экономическую интерпретацию наилучшей с точки зрения функционала качества классификации, полученной методом К-средних, воспользуемся рисунком 29.

Первый класс муниципальных образований характеризуется более высокими по сравнению с другими кластерами средними значениями таких показателей, как удельный вес населения в трудоспособном возрасте (X_3) и миграционный прирост (X_5). Однако на достаточно низком уровне зафиксированы средние значения общего коэффициента рождаемости (X_1), смертности (X_2), а также удельного веса населения старше трудоспособного возраста (X_4). Прирост населения в муниципальных образованиях первого кластера происходит главным образом за счет механического движения населения, это объясняется тем, что в состав данного класса вошли практически все города Оренбургской области, где сосредоточены предприятия, предоставляющие торговые, культурные, медицинские, образовательные услуги, что весьма привлекательно для мигрантов.

Объекты второго класса с одной стороны характеризуются наибольшим средним значением общего коэффициента рождаемости (X_1), с другой стороны наименьшим средним значением общего коэффициента смертности (X_2), что, ско-

рее всего, связано с низким удельным весом населения старше трудоспособного возраста (X_4). Очевидно, что прирост населения в городе Соль-Илецк, а также в районах, вошедших во второй класс, происходит за счет естественного движения населения. Высокий уровень рождаемости во втором классе объясняется тем, что сельские жители более привержены традициям и ценностям, которых придерживались предыдущие поколения. Заметное воздействие на рождаемость оказывает и национальный состав этих районов. Некоторые народы (например, казахи) сохранили традиции многодетности, и там, где доля этих народов в населении выше, выше и показатели рождаемости.

Третий класс лидирует по значениям таких показателей как общий коэффициент смертности (X_2) и удельный вес населения старше трудоспособного возраста (X_4). Объекты, вошедшие в третий класс, характеризуются старением населения, что и обуславливает существенную естественную убыль населения. Прирост числа жителей в данных районах происходит только за счет незначительного миграционного притока. Переселенцами являются в основном либо сельские жители других регионов области, либо иммигранты из Казахстана и государств Центральной Азии, где уровень жизни в среднем ниже, чем в регионах России. Они, как правило, не обладают достаточными средствами для приобретения жилья и адаптации в городах Оренбургской области, поэтому вынуждены расселяться в сельской местности.

4.2 Порядок выполнения работы в пакете Stata

Одним из существенных достоинств пакета Stata является возможность работы с ним не только через кнопочный интерфейс (это удобно для первоначального знакомства с методами статистического анализа и с самим пакетом), но и через интерфейс командный, путем создания do-файлов, куда последовательно записываются все операции, которые нужно провести над анализируемыми данными. Второй вариант, безусловно, позволяет существенного повысить скорость и эффективность работы исследователя. Кроме того, каждый пользователь Stata может, запрограммировав нужный ему метод и создав соответствующий ado-файл, добавить к стан-

дартным реализованным в пакете методам новую команду. В настоящее время существуют целые базы таких ado-файлов, покрывающих самые современные методы анализа и оценивания.

Сначала опишем реализацию иерархических и итерационного методов кластерного анализа через кнопочный интерфейс, а затем обратимся к вопросам создания do-файла.

4.2.1 Порядок выполнения работы через кнопочный интерфейс Stata

После запуска Stata на экране появится основное окно программы (рисунок 41).

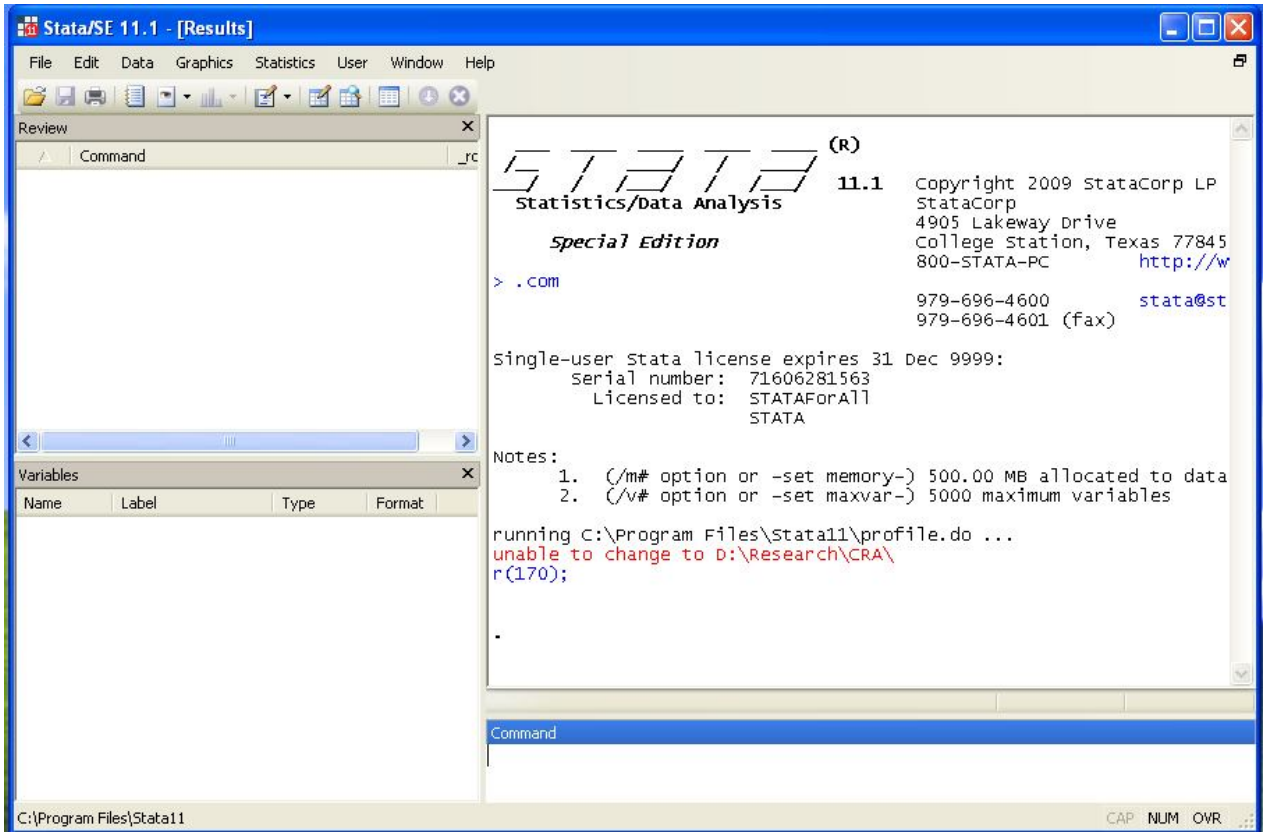
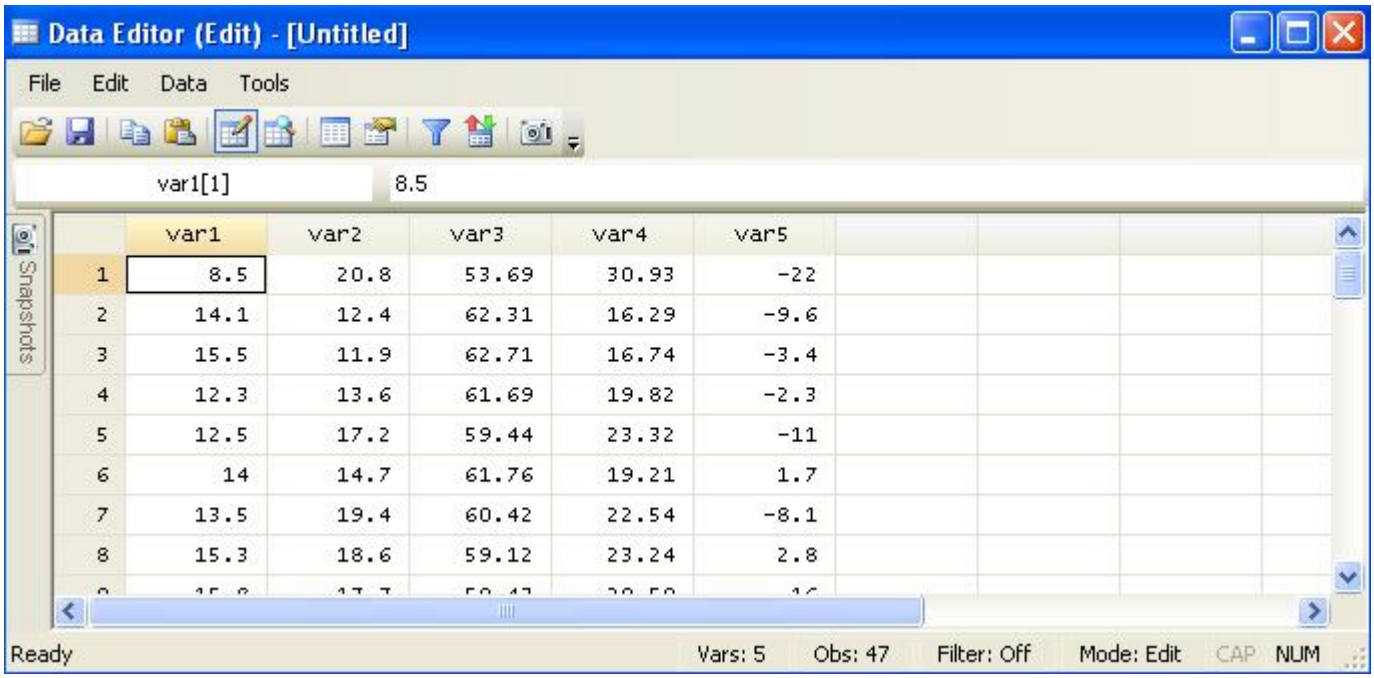


Рисунок 41 – Вид основных окон Stata после запуска

Для ввода данных используется редактор данных Data Editor, который можно вызвать кнопкой  или через главное меню программы **Window \ Data Editor**. Используя пункт меню **Paste** в окне Data Editor, вставим данные, предварительно ско-

пированные в буфер обмена (рисунок 42). Важно помнить, что разделителем целой и дробной части числе является точка.



	var1	var2	var3	var4	var5
1	8.5	20.8	53.69	30.93	-22
2	14.1	12.4	62.31	16.29	-9.6
3	15.5	11.9	62.71	16.74	-3.4
4	12.3	13.6	61.69	19.82	-2.3
5	12.5	17.2	59.44	23.32	-11
6	14	14.7	61.76	19.21	1.7
7	13.5	19.4	60.42	22.54	-8.1
8	15.3	18.6	59.12	23.24	2.8

Рисунок 42 – Вид окна Data Editor после вставки данных

Переименуем переменные, по умолчанию названные как var1, …, var5, в X1,…, X5. Для этого сделаем двойной щелчок левой клавишей мыши по заголовку переменной и в появившемся окне в поле **Name** внесем имя x1. В поле **Label** можно внести метку, или пояснение к переменной (рисунок 43). Нажатие кнопки **Apply** фиксирует внесенные изменения. Для перехода к редактированию имени следующей переменной удобно использовать расположенную на этой же форме кнопку  . Повторим описанные операции для каждой из оставшихся четырех переменных.

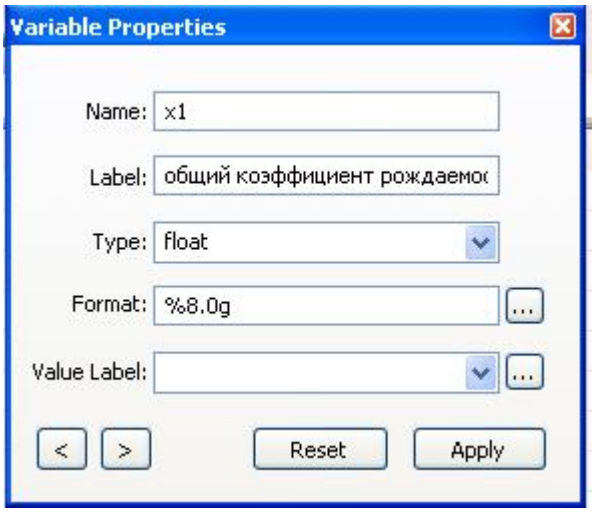


Рисунок 43 – Вид формы редактирования имени, формата и метки переменной

Поскольку рассматриваемые переменные имеют разные единицы измерения, разный масштаб, то имеет смысл перейти к стандартизированным данным одним из описанных в пункте 4.1 способом. Используем переход к $x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{S_j}$, $i = 1,...n;\ j = 1,...,k$. Для этого выберем пункт главного меню **Data** \ **Create or change data** \ **Create new variable (extended)** (рисунок 44).

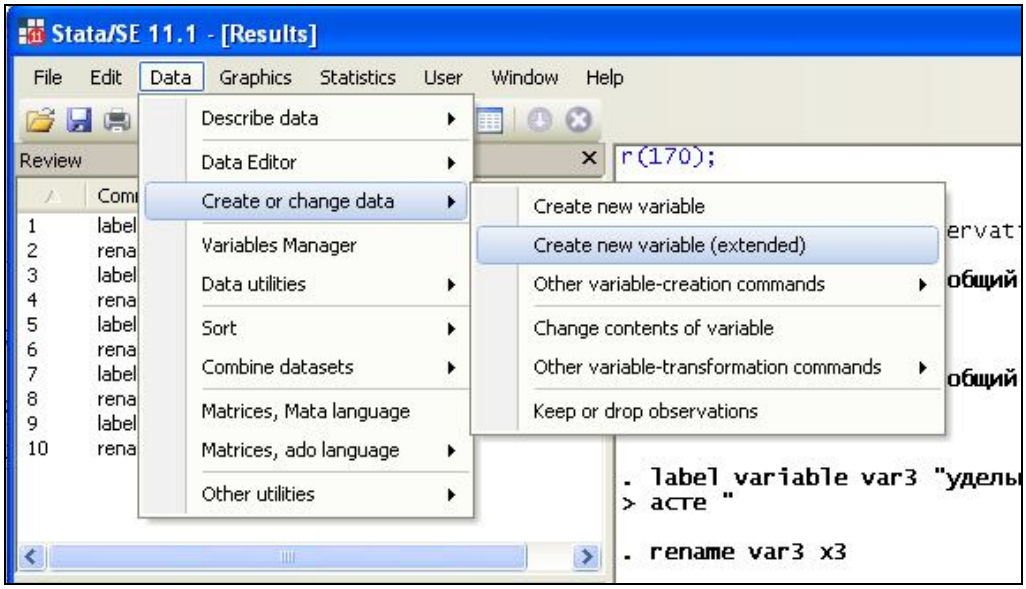


Рисунок 44 – Выбор пунктов меню при стандартизации переменных

В появившейся форме (рисунок 45) нужно в поле **Generate variable** ввести имя новой переменной (в нашем случае nx1), в поле **Expression** ввести имя преобразуемой переменной (в нашем случае x1). В списке функций **Egen function** выбрать группу **Standardized values**, Options-поля Mean и Standard deviation оставить по умолчанию равными 0 и 1 – это означает, что среднее значение новой переменной будет равно 0, а стандартное отклонение 1. Нажмем кнопку OK.

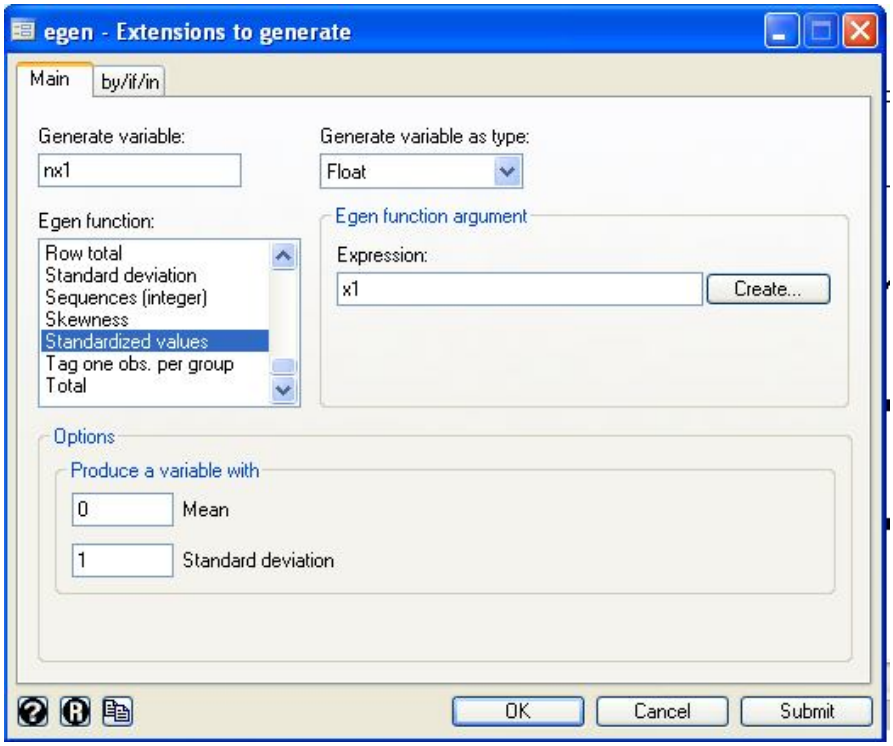


Рисунок 45 – Вид формы для преобразования переменных

Проделав аналогичные операции для оставшихся четырех переменных, получим следующий список переменных Variables (рисунок 46).

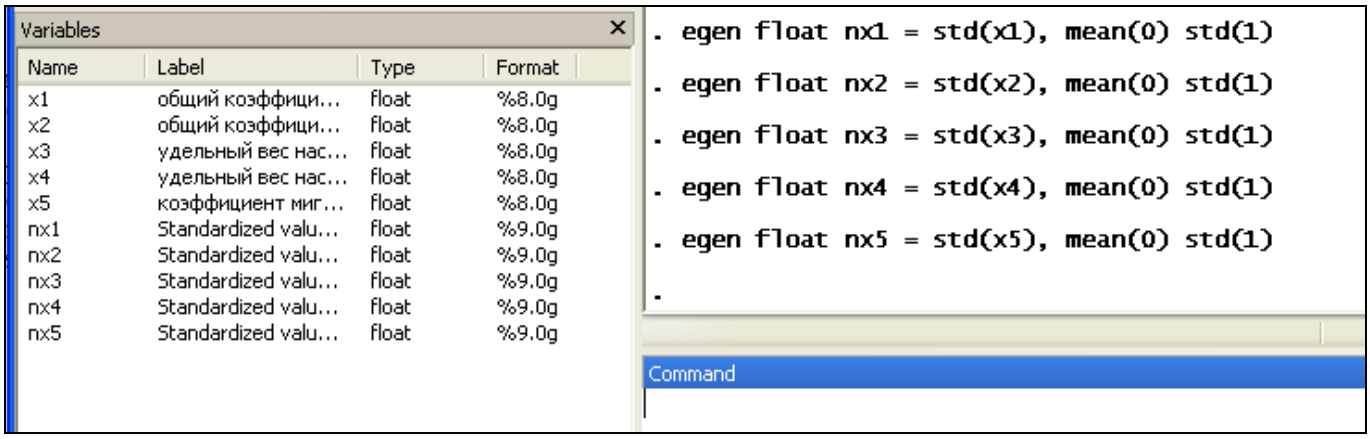


Рисунок 46 – Вид списка «Variables» после добавления стандартизированных переменных

В пакете Stata реализованы следующие агломеративные методы классификации:

Single linkage – метод «одиночной связи»;

Average linkage – метод «средней связи»;

Complete linkage – метод «полных связей»;

Weighted-average linkage – взвешенный метод средней связи**;**

Median linkage – метод медианной связи;

Centroid linkage – центроидный метод;

Ward′s linkage – метод Уорда.

Воспользуемся для классификации, например, методом «полных связей». Для запуска процедуры иерахической классификации в пакете Stata выберем пункт меню **Statistics \ Multivariate analysis \ Cluster analysis \ Cluster data \ Complete linkage**. В появившемся окне (рисунок 47) в поле **Variables** нужно задать переменные, которые будут учитываться при классификации. В части формы (Dis)similarity measure нужно указать тип анализируемых переменных: Continuous (непрерывные), Binary (бинарные) или Mixed (смешанные). Далее выбирается метрика, по которой будет рассчитываться расстояние между классифицируемыми переменными. В State реализованы такие метрики, как

L2 (Euclidean distance) **–** евклидово расстояние;

L2squared (squared Euclidean distance) - квадратичное евклидово расстояние;

L1 (absolute-value distance) – хеммингово расстояние или city-block;

Linfinity (maximum-value distance) - расстояние Чебышева

L(#) и – расстояние Минковского с аргументом #;

Lpower(#) - расстояние Минковского с аргументом # , возведенным в степень #;

Canberra – расстояние Канберра;

correlation – корреляционное расстояние;

angular - угловое расстояние.

Выберем обычное евклидово расстояние и нажмем кнопку ОК.

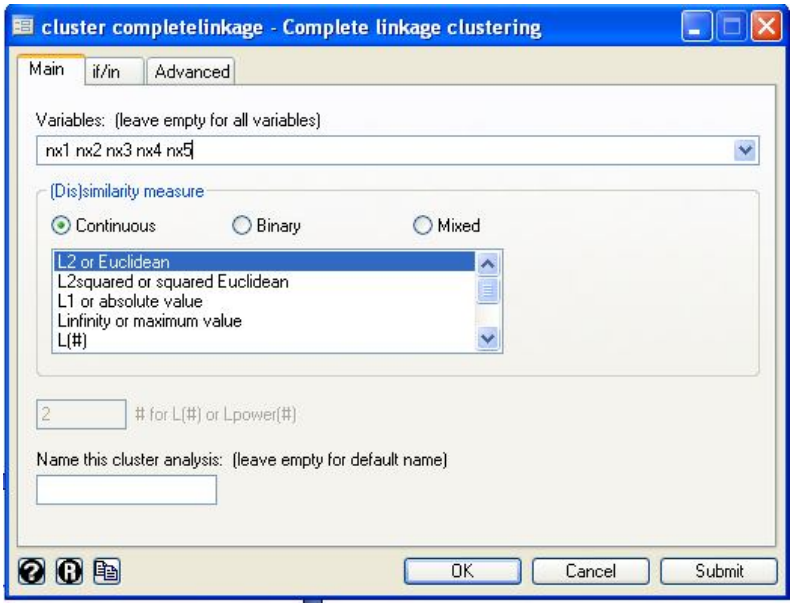


Рисунок 47 – Вид окна задания параметров иерархического кластерного анализа

Для построения дендрограммы воспользуемся пунктом меню **Statistics \ Multivariate analysis \ Cluster analysis \ Postclustering\ Dendrograms** (рисунок 48).

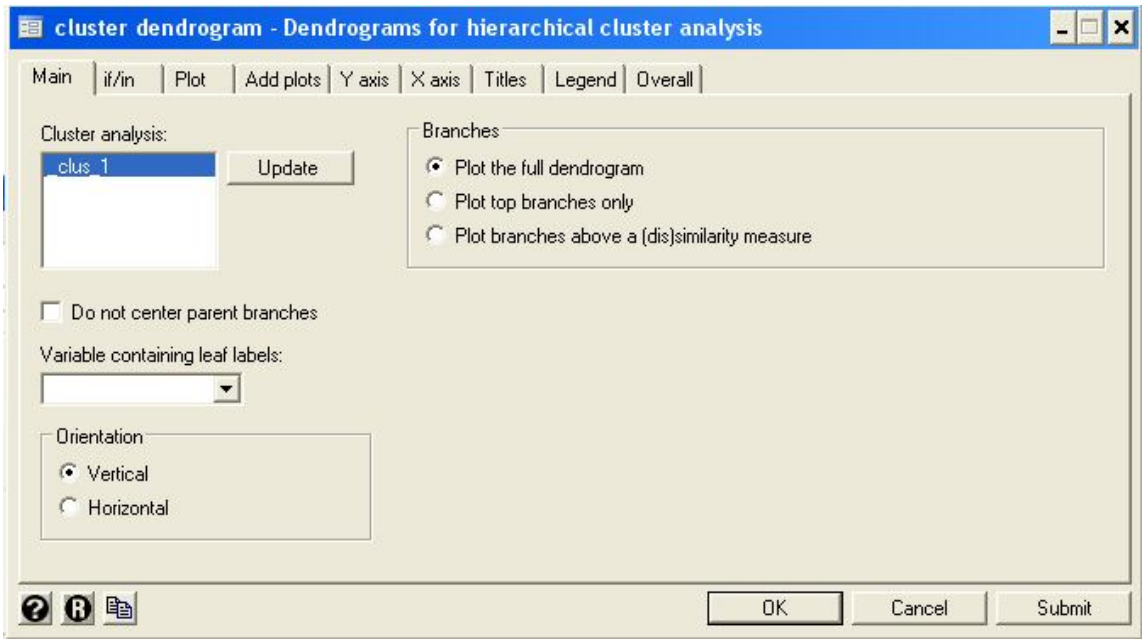


Рисунок 48 – Вид окна задания параметров построения дендрограммы

В поле **Cluster analysis** выбирается тот кластерный анализ, результаты которого нужно представить в виде дендрограммы. В нашем случае пока это только один вид анализа – метод полных связей, получивший по умолчанию имя _clus_1. В списке **Variable containing leaf labels** можно выбрать переменную, содержащую названия классифицируемых объектов (в нашем случае это районы Оренбургской области). Оставляем экспериментирование с данной возможностью на самостоятельную проработку читателя. В поле **Orientation** выберем ориентацию дендрограммы: Vertical (вертикальная, когда подписи объектов расположены по оси абсцисс) и Horizontal (горизонтальная, когда подписи объектов расположены по оси ординат). В поле **Branches** можно задать построение всей дендрограммы (Plot the full dendrogram), построение только заданного количества верхних ветвей дендрограммы (Plot top branches only) или построение только тех ветвей дендрограммы, которые находятся выше задаваемого порога (Plot branches above a (dis)similarity measure).

С помощью остальных вкладок этой формы можно настроить вид выводимой дендрограммы. При нажатии OK откроется окно редактора графиков Stata Graph, в котором будет представлена построенная дендрограмма. После изменения заголовка диаграммы, подбора размера шрифтов для каждой оси получаем дендрограмму следующего вида (рисунок 49).

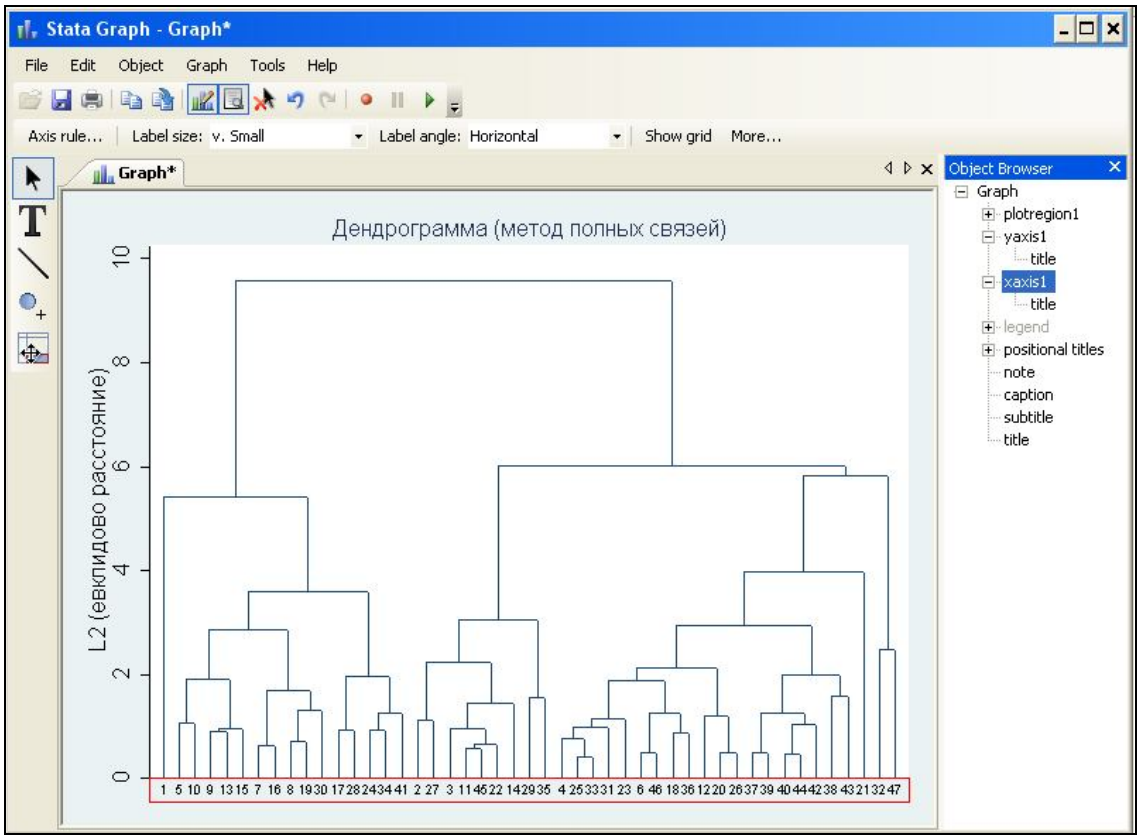


Рисунок 49 – Вид окна Stata Graph после редактирования дендрограммы (метод полных связей)

Анализируя дендрограмму, можно выдвинуть предположение, что в исследуемой совокупности объектов естественно выделяются 2 кластера (пороговое расстояние 8). Кроме визуального анализа дендрограммы для определения оптимального числа классов можно использовать так называемые stopping rules (критерии останова). В пакете Stata реализованы два наиболее эффективных критерия: индекс Калински и Харабаза и индекс Дуды и Харта. Для расчета этих индексов выберем пункт меню **Statistics \ Multivariate analysis \ Cluster analysis \ Postclustering \ Cluster analysis stopping rules**. В появившемся окне (рисунок 50) в поле Options укажем, что индекс нужно рассчитать только для разбиений на 2, 3,…, 9 классов. После нажатия кнопки ОК в окне появится таблица с результатами (отметим, что ее можно скопировать в отчет с сохранением табуляции). Рассчитаем также значения индекса Дуды и Харта (рисунок 51). По индексу Калински и Харабаза оптимальным следует признать количество классов, равное 2; анализируя индекс Дуды и Харта, видим, что его максимальные значения (0,7539 и 0,7315) достигаются для количест-

ва классов, равных 5 и 2 соответственно. Поскольку при выделении 5 классов один из классов содержит всего один объект, такую классификацию нельзя признать хорошей. Примем количество классов равным 2.

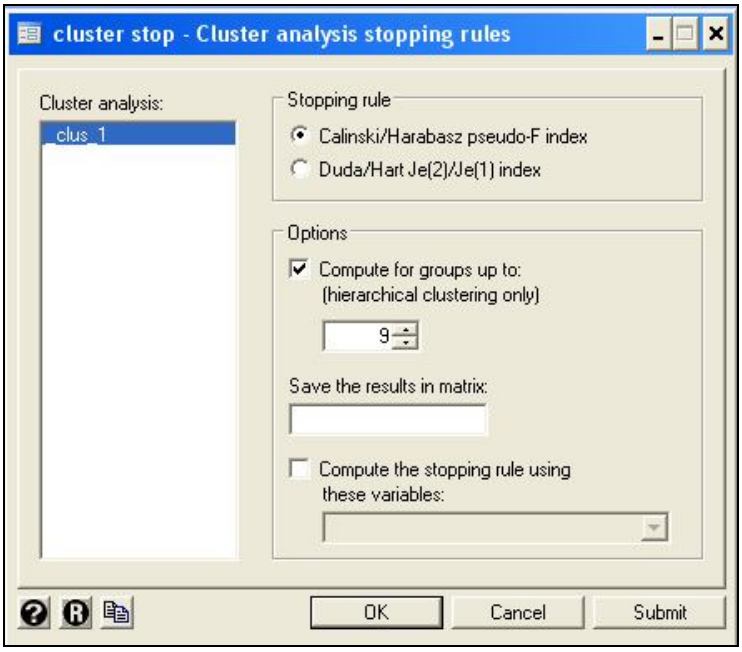


Рисунок 50 – Вид окна выбора индекса определения оптимального числа классов

Number of clusters	Calinski/ Harabasz pseudo-F
2	20.16
3	17.04
4	19.47
5	19.73
6	18.01
7	19.56
8	18.74
9	19.56

Number of clusters	Duda/Hart Je(2)/Je(1)	Duda/Hart pseudo T-squared
2	0.7315	10.64
3	0.5241	18.16
4	0.6471	7.63
5	0.7539	5.88
6	0.5832	9.29
7	0.5284	6.25
8	0.6804	7.98
9	0.4702	9.01

Рисунок 51 – Вид таблиц с результатами расчета индексов Калински и Харабаза, Дуды и Харта (метод полных связей)

Создадим переменную, которая каждому объекту поставит в соответствие номер класса, в который он был отнесен. Используем пункт меню **Statistics \ Multivariate analysis \ Cluster analysis \ Postclustering \ Summary variables from cluster analysis.** В появившемся окне (рисунок 52) в поле **Generate variable(s)** запишем имя переменной, в которую будет занесены номера классов. В поле **Function** можно вы-

брать Groups, если нужно сохранить результаты разделения на заданное в **Number of groups to form** количество групп, или Cut at value, если нужно сохранить результаты разделения при заданном пороговом расстоянии.

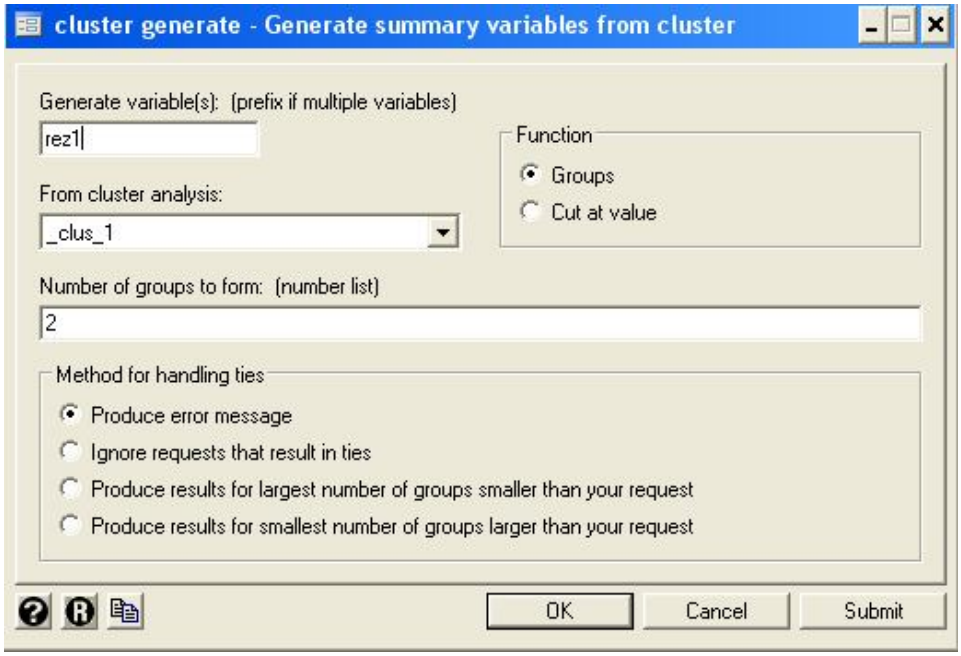
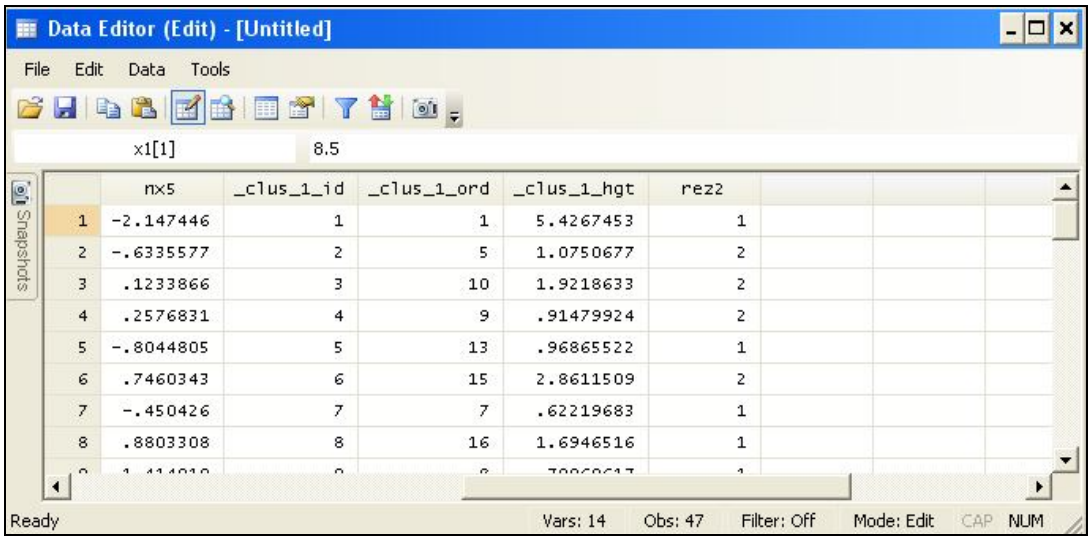


Рисунок 52 – Вид окна для создания переменной с результатами классификации

После нажатия кнопки ОК в файле с данными появится новая переменная rez1 (рисунок 53).



	mx5	_clus_1_id	_clus_1_ord	_clus_1_hgt	rez2
1	-2.147446	1	1	5.4267453	1
2	-.6335577	2	5	1.0750677	2
3	.1233866	3	10	1.9218633	2
4	.2576831	4	9	.91479924	2
5	-.8044805	5	13	.96865522	1
6	.7460343	6	15	2.8611509	2
7	-.450426	7	7	.62219683	1
8	.8803308	8	16	1.6946516	1

Рисунок 53 – Вид окна Data Editor после создания переменной с результатами классификации

Для подсчета количества элементов в каждом классе, используем описательную статистику **Statistics \ Summary, tables, and tests \ Tables \ Tables of summary statistics**. В появившемся окне в поле **Row variable** введем имя переменной с результатами классификации rez2 (рисунок 54).

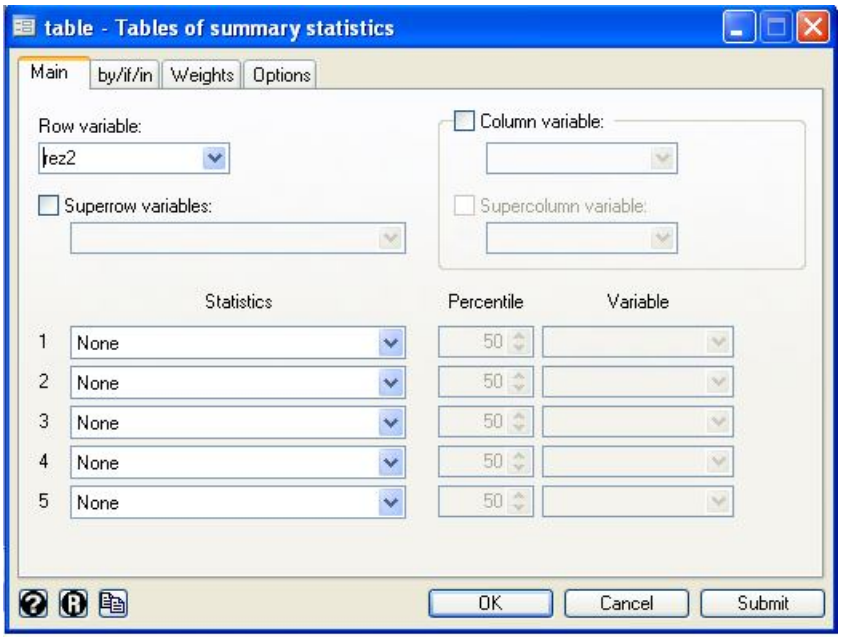


Рисунок 54 – Вид окна для подсчета количества объектов в классах

После нажатия ОК в основном окне программы появится таблица (рисунок 55).

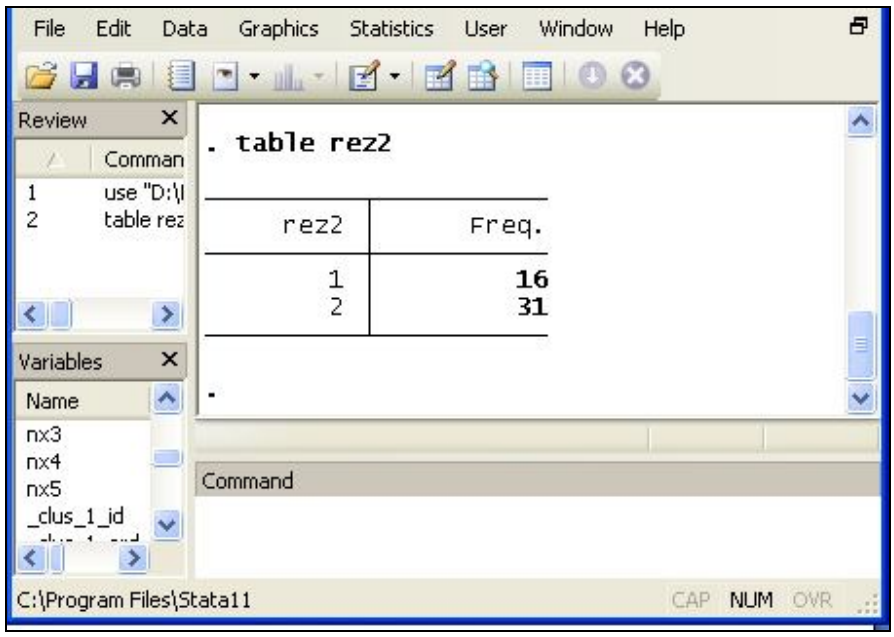


Рисунок 55 – Вид основного окна Stata после подсчета количества объектов в классах

Таким образом, первый класс содержит 16 объектов, а второй - 31.

Состав классов, выделенных методом полных связей, представлен в таблице 6.

Таблица 6 – Результаты классификации муниципальных образований Оренбургской области методом «полных связей» (пакет Stata)

Номер кластера	Количество объектов в кластере	Состав класса
$\{S_{11}\}$	16	**Города:** г. Медногорск **Районы:** Абдулинский, Асекеевский, Бугурусланский, Бузулукский, Гайский, Грачевский, Кваркенский, Кувандыкский, Курманаевский, Матвеевский, Новосергиевский, Пономаревский, Северный, Сорочинский, Шарлыкский.
$\{S_{12}\}$	31	**Города:** Абдулино, Бугуруслан, Бузулук, Гай, Кувандык, Новотроицк, Оренбург, Орск, Соль-Илецк, Сорочинск, Ясный. **Районы:** Адамовский, Акбулакский, Александровский, Беляевский**,** Домбаровский, Илекский, Красногвардейский, Новоорский, Октябрьский, Оренбургский, Первомайский, Переволоцкий, Сакмарский, Саракташский, Светлинский, Соль-Илецкий, Ташлинский, Тоцкий, Тюльганский, Ясненский.

Для интерпретации полученных результатов построим график средних значений всех признаков в каждом из выделенных классов. В Stata нет команды, которая бы выполняла построение такого графика, поэтому используем команду profileplot. Параметрами команды являются переменные, средние значения по которым

нужно рассчитать (в нашем случае nx1 nx2 nx3 nx4 nx5), и группирующая переменная (в нашем случае rez2). Наберем в командной строке **profileplot nx1 nx2 nx3 nx4 nx5, by(rez2)** и нажмем Enter. Результатом выполнения команды будет график следующего вида (рисунок 56).

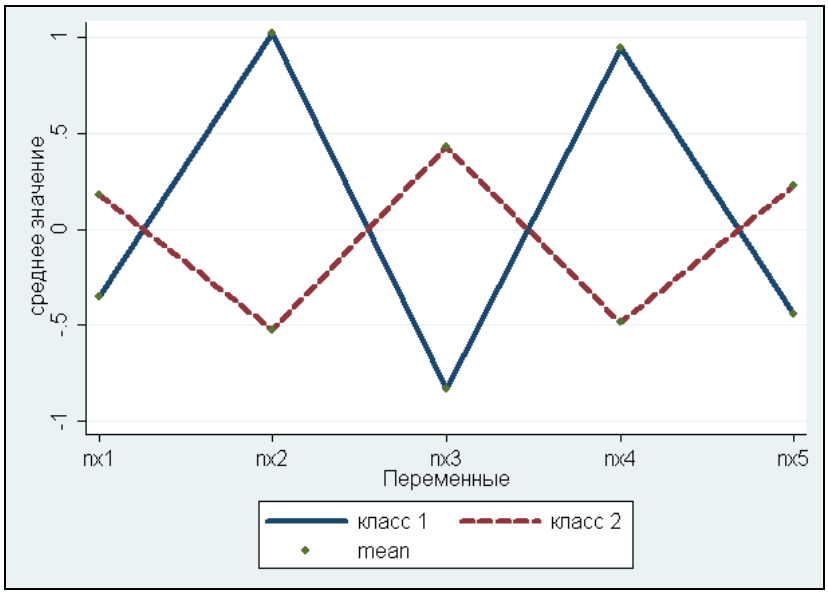


Рисунок 56 – График средних значений признаков в классах (метод полных связей)

Таким образом, можно сделать следующие выводы.

Большая часть объектов, вошедших в первый класс, - это районы, относящиеся к Западной зоне Оренбургской области. Объекты этого класса характеризуются сравнительно высокой смертностью и низкой рождаемостью, высоким удельным весом населения старше трудоспособного возраста и низким удельным весом населения в трудоспособном возрасте. В среднем эти районы непривлекательны и с миграционной точки зрения – для них характерен отрицательный миграционной прирост.

Во второй класс вошли практически все города Оренбургской области, большая часть сельских районов из Центральной зоны. Для них характерна более высокая рождаемость, низкая смертность, высокий удельный вес населения в трудоспособном и соответственно низкий удельный вес населения старше трудоспособного возраста. Это миграционно привлекательные города и районы – положительный миграционный прирост.

Очевидно, что первый класс можно назвать классом со сравнительно неблагоприятной демографической ситуацией, второй класс – со сравнительно благоприятной.

Отметим, что для наглядности различий между классами для построения этого графика использовались стандартизированные переменные. Табличное представление средних значений исходных признаков в классах можно получить, используя пункт меню **Statistics \ Summary, tables, and tests \ Summary and descriptive statistics \ Means** (рисунок 57).

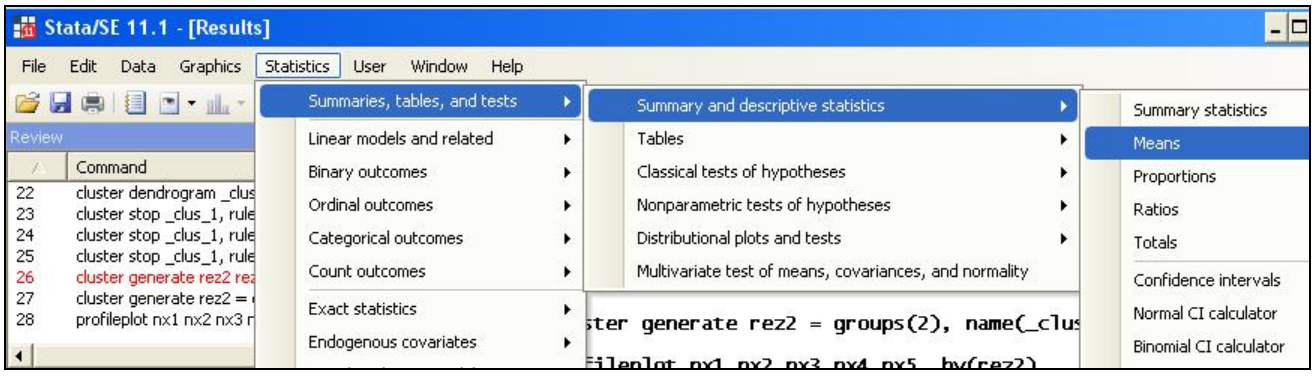


Рисунок 57 – Выбор пунктов меню при расчете средних значений признаков в классах

Использование этой команды позволит также получить доверительные интервалы для средних значений признаков. В появившемся окне (рисунок 58) в поле **Variables** выберем переменные X1, X2, X3, X4, X5. На вкладке **if/in/over** поставим галочку в **Group over subpopulations** и выберем в ставшем активным **списке Group variables** группирующую переменную – это наша переменная с номерами классов rez2 (рисунок 59). Нажмем ОК. Результаты выполнения команды представлены на рисунке 60.

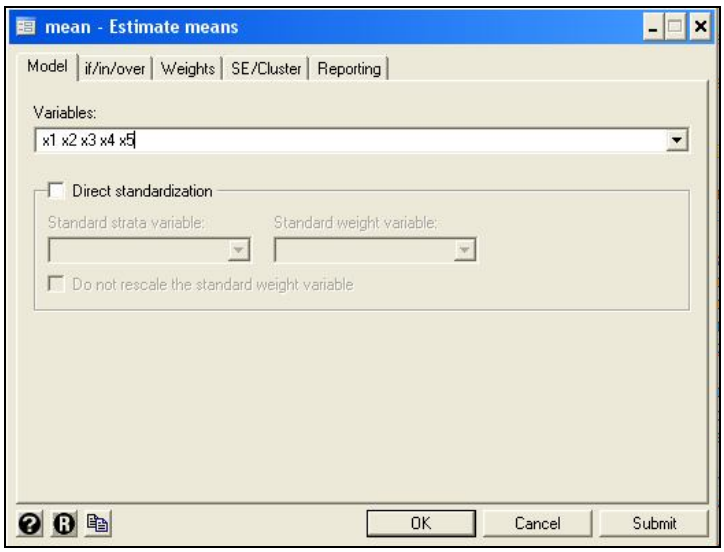


Рисунок 58 – Вид окна выбора переменных для расчета средних значений признаков в классах

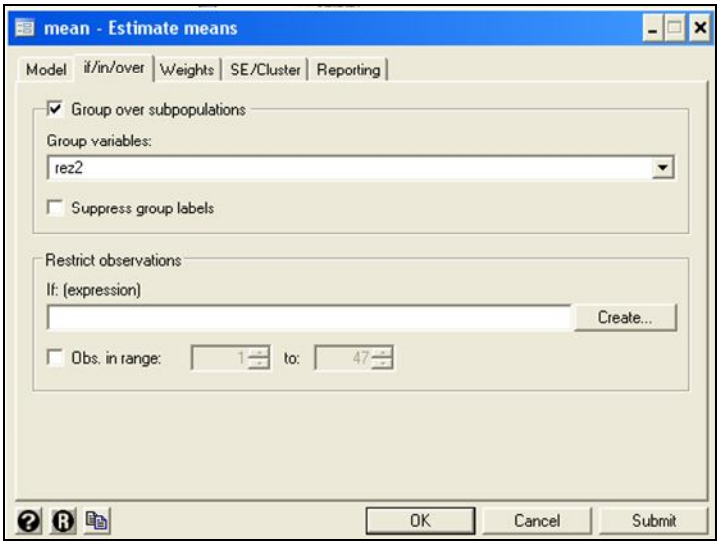


Рисунок 59 – Вид окна выбора группирующей переменной при расчете средних значений признаков в классах

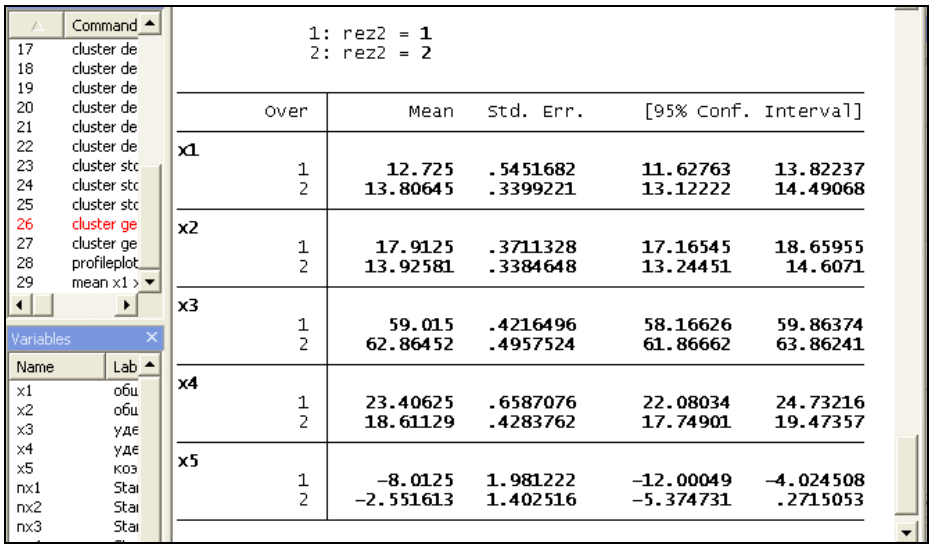


Рисунок 60 – Вид окна Stata после расчета средних значений признаков в классах (метод полных связей)

Выполнив аналогичные действия для метода Уорда, получаем дендрограмму (рисунок 61), значения индексов Калински и Харабаза, Дуды и Харта (рисунок 62), график средних значений (рисунок 63), таблицу результатов классификации (таблица 7). Из класса 2 в класс 1 перешли четыре района (Илекский, Октябрьский, Саракташский, Ясненский) и один город – Абдулино. Интерпретация классов совпадает с интерпретацией классов, выделенных методом полных связей, за исключением нивелирования различий в общем коэффициенте рождаемости.

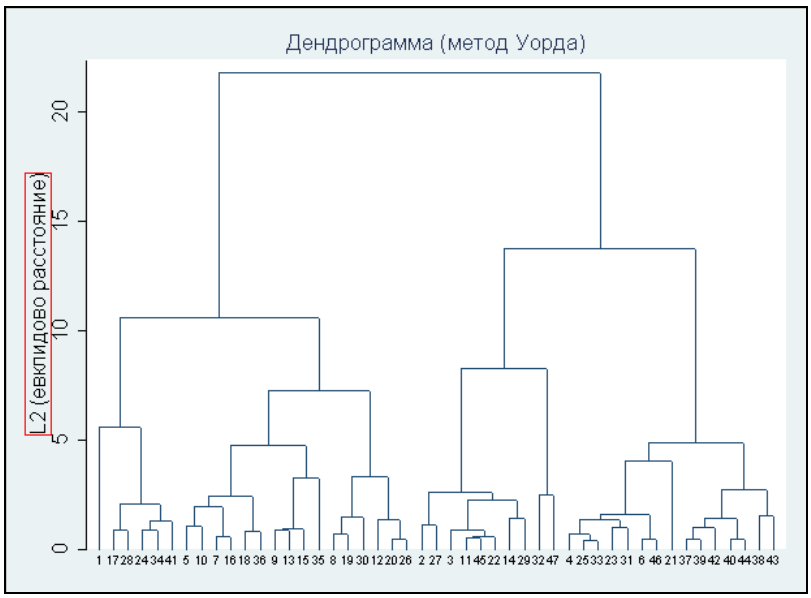


Рисунок 61 - Дендрограмма (метод Уорда)

Number of clusters	Calinski/ Harabasz pseudo-F
2	18.11
3	15.63
4	16.53
5	17.36
6	17.56
7	18.46
8	18.00
9	19.10

Number of clusters	Duda/Hart	
	Je(2)/Je(1)	pseudo T-squared
2	0.6375	13.08
3	0.6650	10.08
4	0.3415	15.43
5	0.6542	7.40
6	0.2343	13.08
7	0.6878	5.90
8	0.5550	6.41
9	0.3097	13.38

Рисунок 62 – Вид таблиц с результатами расчета индексов Калински и Харабаза, Дуды и Харта (метод Уорда)

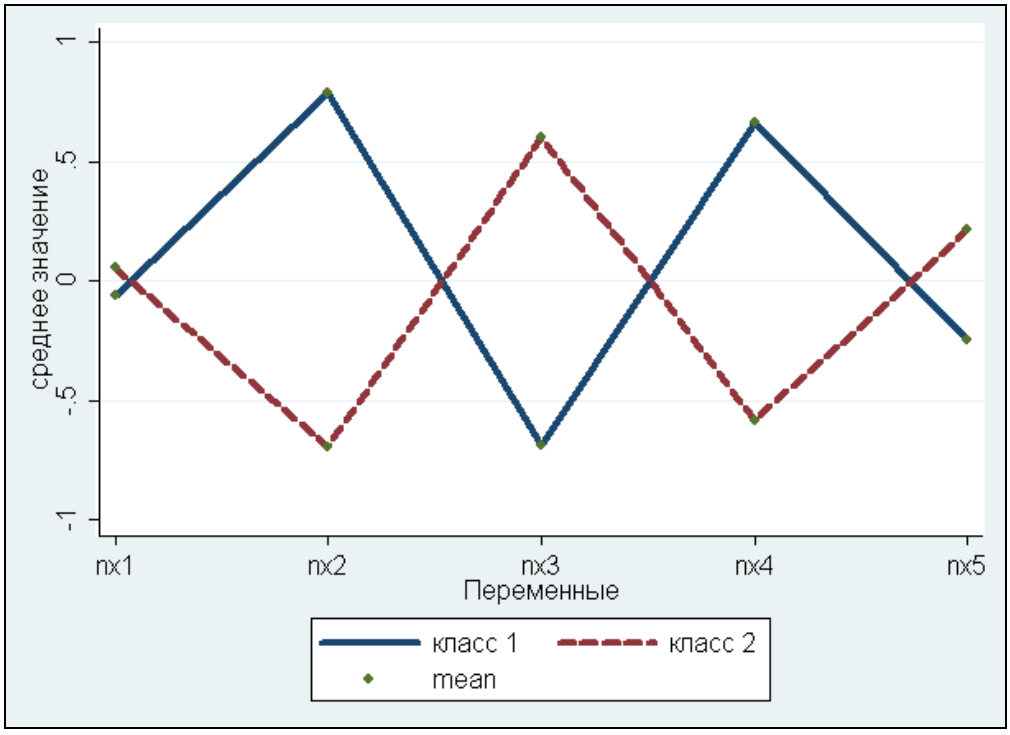


Рисунок 63– График средних значений признаков в классах (метод Уорда)

Таблица 7 – Результаты классификация муниципальных образований Оренбургской области методом Уорда (пакет Stata)

Номер кластера	Количество объектов в кластере	Состав класса
$\{S_{21}\}$	21	**Города:** Абдулино, Медногорск **Районы:**Абдулинский, Асекеевский, Бугурусланский, Бузулукский, Гайский, Грачевский, Илекский, Кваркенский, Кувандыкский, Курманаевский, Матвеевский, Новосергиевский, Октябрьский, Пономаревский, Саракташский, Северный, Сорочинский, Шарлыкский, Ясненский.
$\{S_{22}\}$	26	**Города:**, Бугуруслан, Бузулук, Гай, Кувандык, Новотроицк, Оренбург, Орск, Соль-Илецк, Сорочинск, Ясный. **Районы:** Адамовский, Акбулакский, Александровский, Беляевский**,** Домбаровский, Красногвардейский, Новоорский, Оренбургский, Первомайский, Переволоцкий, Сакмарский, Светлинский, Соль-Илецкий, Ташлинский, Тоцкий, Тюльганский.

Для реализации итерационного метода кластерного анализа выберем пункты меню **Statistics \ Multivariate analysis \ Cluster analysis \ Cluster data \ Kmeans**. В появившемся окне (рисунок 64) на вкладке **Main** в поле **Variables** укажем переменные (если оставить поле пустым, при классификация будет проведена по всем переменным, имеющимся в файле с данными). В списке **K (the number of groups)** укажем количество классов, на которые будет разбивать исследуемую совокупность объектов. В списке (Dis) similarity measure выберем тип анализируемых переменных и используемую метрику расстояния между объектами (в нашем случае евклидово расстояние). Отметим, что при реализации метода k-средних в Stata можно выбрать

любую из описанных выше метрик расстояния между объектами, в отличие от пакета Statistica, где может быть использовано только евклидово расстояние.

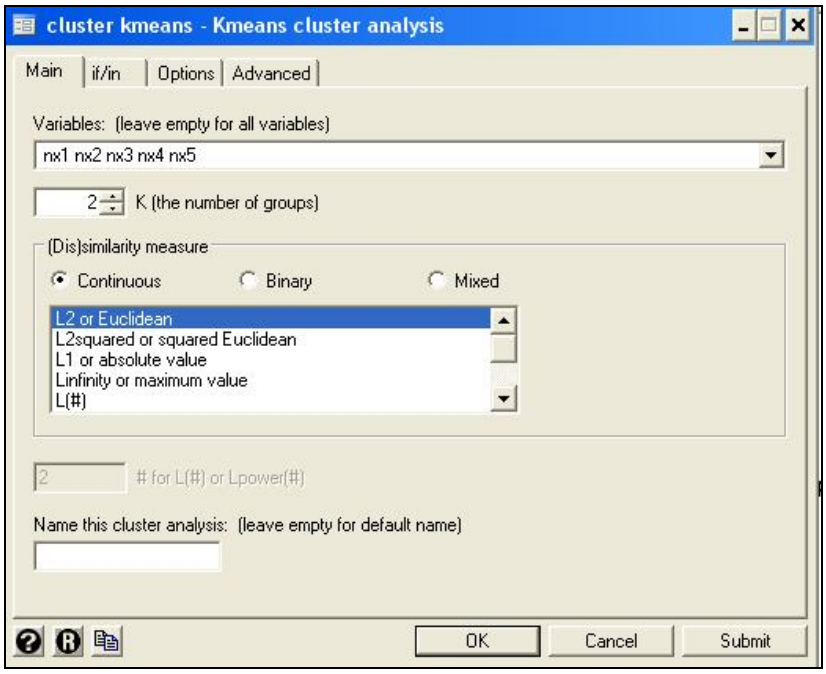


Рисунок 64 – Вид вкладки Main окна задания параметров классификации методом k-средних

На вкладке **Options** (рисунок 65) задается правило выбора объектов, которые будут начальными эталонами:

K unique random observations – случайным образом выбранные K объектов

First K observations – первые K объектов в файле с данными (при активации опции Exlude the K observations эти объекты не подлежат классификации, а используются только как начальные эталоны)

Last K observations – последние K объектов в файле с данными (при активации опции Exlude the K observations эти объекты не подлежат классификации, а используются только как начальные эталоны)

K random centers chosen from within the range of the data – эталонные значения центров классов получаются в результате генерации случайных чисел, равномерно распределенных на интервалах, соответствующих диапазонам изменения анализируемых данных

Group means from K random partitions of the data – все объекты случайным образом делятся на K групп, и средние значения признаков в каждой из групп берутся в качестве начальных эталонов.

Group means from K partitions formed by grouping every Kth observation – формируется K групп: объекты с номерами 1, 1+K, 1+2K и т.д. образуют первую группу, объекты с номерами 2, 2+K, 2+2K и т.д. образуют вторую группу и т.д. Средние значения признаков в каждой из групп берутся в качестве начальных эталонов.

Group means from K (nearly equal) contiguous partitions of the data - формируется K групп примерно одинакового объема: приблизительно n/K первых объектов образуют первую группу, следующие n/K объектов – вторую и т.д. Средние значения признаков в каждой из групп берутся в качестве начальных эталонов.

Group means from partitions defined by initial grouping variable – в выпадающем списке выбирается переменная, содержащая разбиение объектов на группы. Эта переменная может быть сформирована, например, после реализации какого-либо иерархического метода классификации. Средние значения признаков в каждой из групп берутся в качестве начальных эталонов.

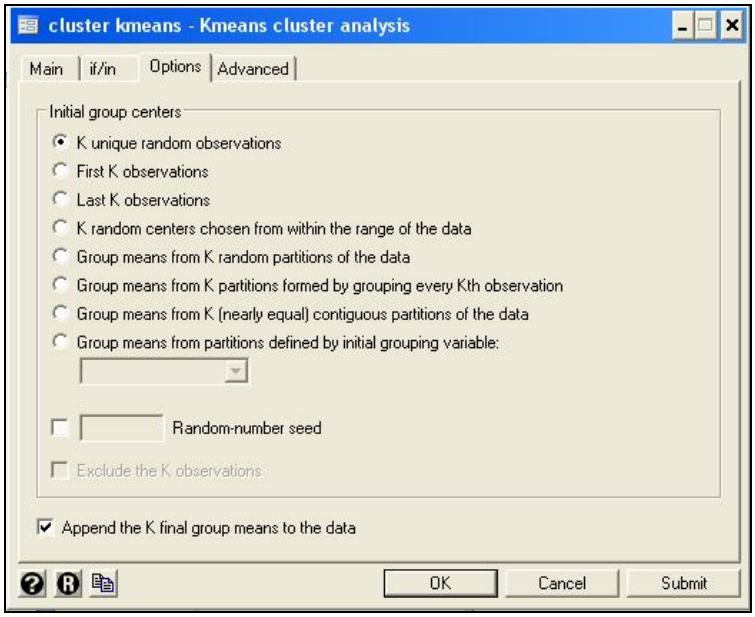


Рисунок 65 – Вид вкладки Options окна задания параметров классификации методом k-средних

Нажмем ОК. В файле с данными будет добавлена переменная с результатами классификации, по умолчанию названная _clus_3 (так как это третий по счету метод классификации, вызванный в текущей сессии). Аналогично вышеописанному, построим график средних значений признаков в классах (рисунок 66).

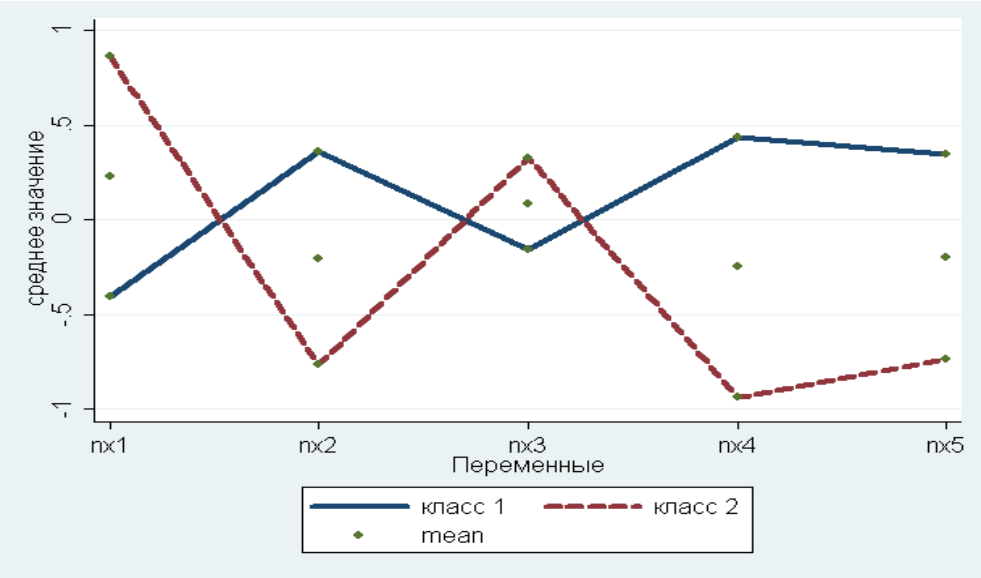


Рисунок 66 – График средних значений признаков в классах (метод k-средних)

По сравнению с классификацией иерархическими методами произошло существенное перераспределение объектов по классам, на наш взгляд, не слишком удачное: для объектов второго класса теперь характерна высокая смертность, высокий удельный вес населения старше трудоспособного возраста и одновременно положительный миграционный прирост. Попробуем провести классификацию методом k-средних, взяв в качестве начального разбиения результаты разбиения методом полных связей. В результате получаем график средних значений признаков в классе (рисунок 67), таблицу средних значений признаков в классах (рисунок 68) и таблицу с результатами классификации (таблица 8).

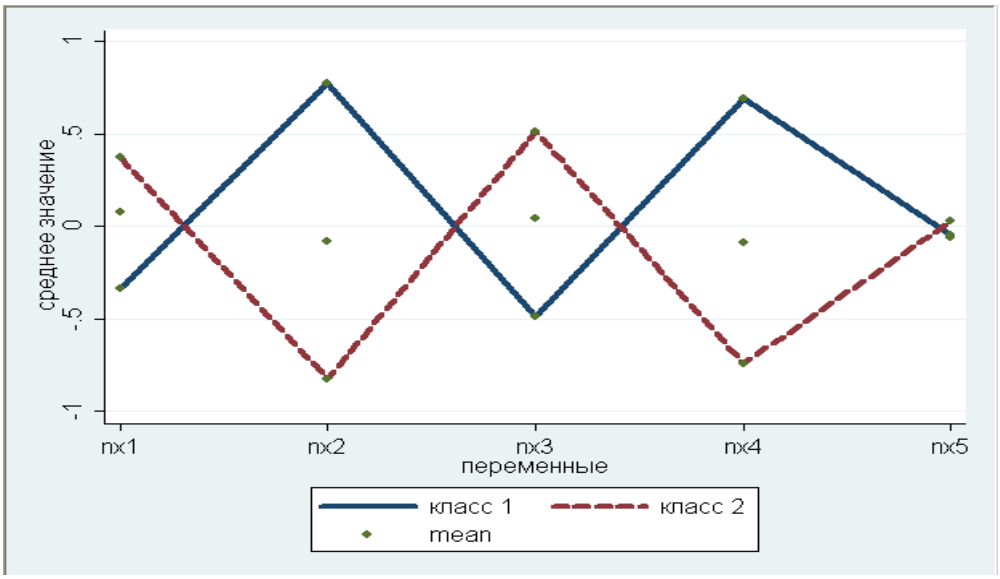
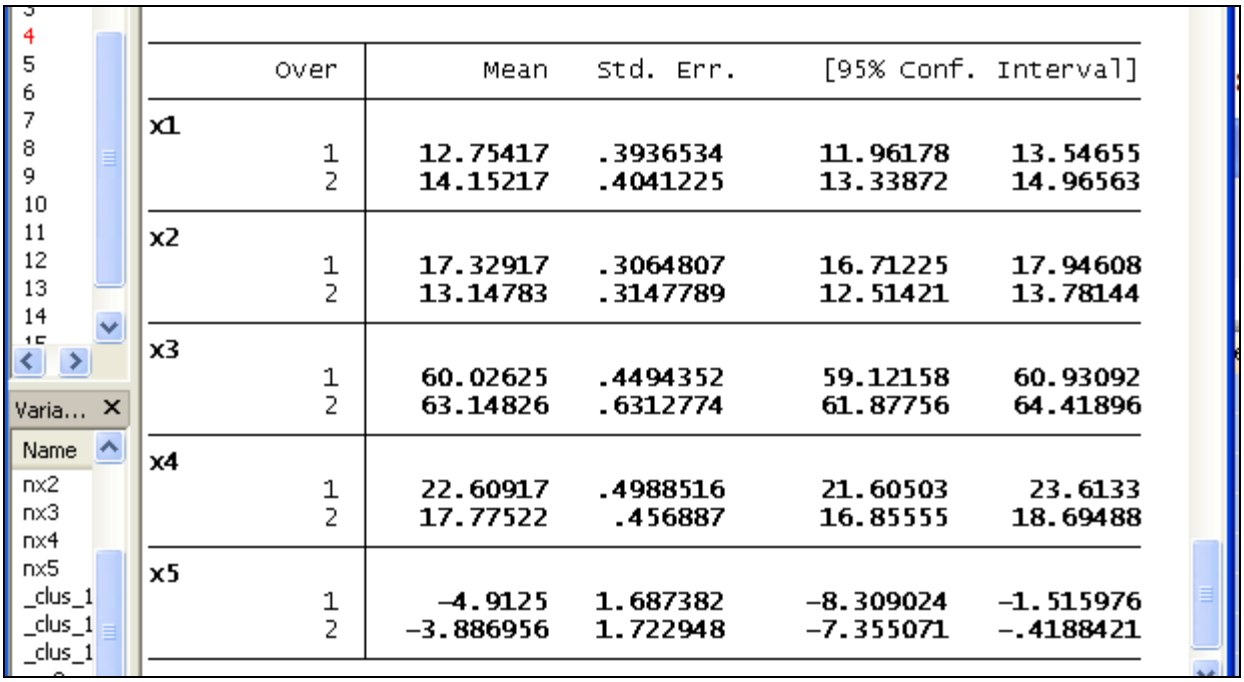


Рисунок 67 – График средних значений признаков в классах (метод k-средних, начальное разбиение по методу полных связей)



Varia...

Name

nx2
nx3
nx4
nx5
_clus_1
_clus_1
_clus_1

Over	Mean	Std. Err.	[95% Conf.	Interval]
x1				
1	12.75417	.3936534	11.96178	13.54655
2	14.15217	.4041225	13.33872	14.96563
x2				
1	17.32917	.3064807	16.71225	17.94608
2	13.14783	.3147789	12.51421	13.78144
x3				
1	60.02625	.4494352	59.12158	60.93092
2	63.14826	.6312774	61.87756	64.41896
x4				
1	22.60917	.4988516	21.60503	23.6133
2	17.77522	.456887	16.85555	18.69488
x5				
1	-4.9125	1.687382	-8.309024	-1.515976
2	-3.886956	1.722948	-7.355071	-.4188421

Рисунок 68 - Вид окна Stata после расчета средних значений признаков в классах (метод k-средних)

Таблица 8 - Результаты классификация муниципальных образований Оренбургской области методом k-средних с начальным разбиением по методу полных связей (пакет Stata)

Номер кластера	Количество объектов в кластере	Состав класса
$\{S_{31}\}$	25	**Города:** Абдулино, Гай, Кувандык, Медногорск, Новотроицк, Орск. **Районы:** Абдулинский, Адамовский, Александровский, Беляевский, Гайский,**,** Кваркенский, Кувандыкский, Курманаевский, Матвеевский, Новоорский, Новосергиевский, Оренбургский, Пономаревский, Переволоцкий, Саракташский, Северный, Шарлыкский, Ясненский, Тоцкий.
$\{S_{32}\}$	22	**Города:**, Бугуруслан, Бузулук, Оренбург, Соль-Илецк, Сорочинск, Ясный. **Районы:** Акбулакский, Асекеевский, Бугурусланский, Бузулукский, Грачевский, Илекский, Домбаровский, Красногвардейский, Первомайский, Октябрьский, Сакмарский, Светлинский, Соль-Илецкий, Ташлинский, Сорочинский, Тюльганский.

Интерпретация классов близка к интерпретации классов, выделенных методами полных связей и Уорда, за исключением нивелирования различий в уровне миграционного прироста.

Сравнение классификаций

С помощью метода «полных связей», метода Уорда и метода k-средних были получены различные классификации. Сводная таблица результатов классификаций муниципальных образований Оренбургской области, полученных различными методами кластерного анализа, приведена в приложении Б (таблица Б.4)

Для выбора лучшей классификации необходимо воспользоваться функционалами качества разбиения, например, $Q_3(S) = \sum_{l=1}^{p} \sum_{j=1}^{k} S_j^2(l) \to \min$.

Покажем, как оценить дисперсии признаков в каждом классе на примере разбиения, полученного методов полных связей (переменная rez2). Используем пункт меню **Statistics \ Summary, tables, and tests \ Tables \ Table of summary statistics (tablestat)** и в появившемся окне в поле Variables выберем анализируемые переменные \, в поле укажем группирующую переменную rez2, в одном из списков Statistics to display выберем Variance (рисунок 69).

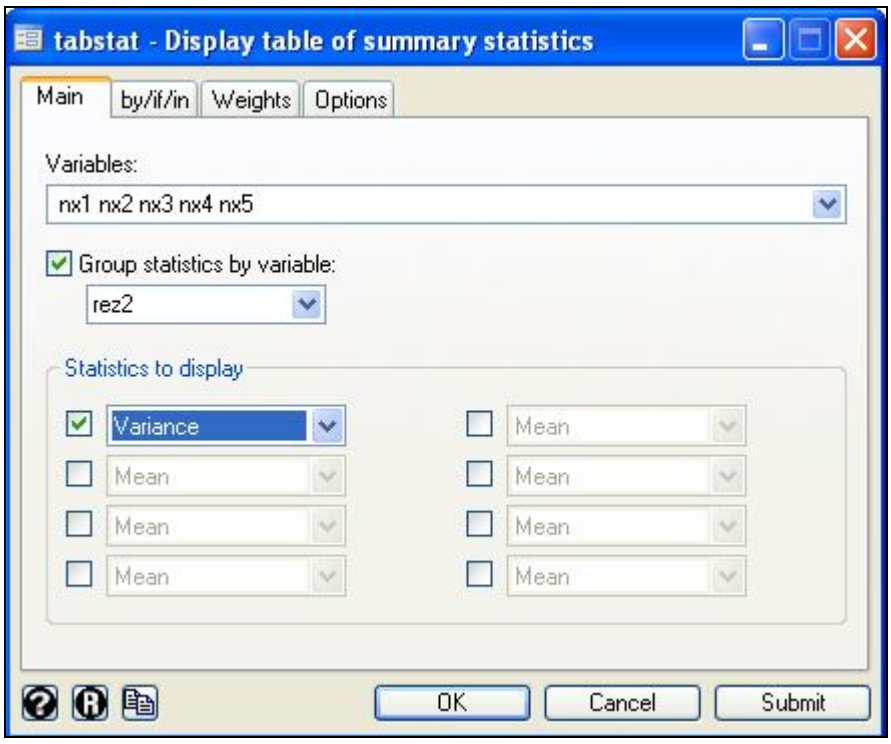


Рисунок 69 – Вид окна при оценке внутриклассовых дисперсий признаков (метод полных связей)

После нажатия ОК в основном окне программы появится таблица (рисунок 70).

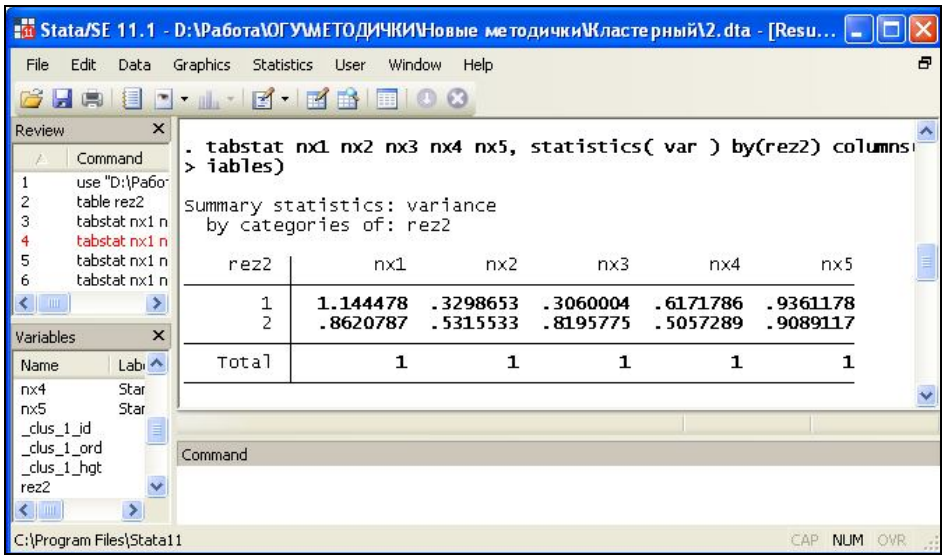


Рисунок 70 – Вид основного окна Stata после расчета дисперсий признаков внутри классов, выделенных методом полных связей.

Рассчитав дисперсии признаков в классах, выделенных методов Уорда и k-средних, сведем результаты в таблицу (таблица 9)

Таблица 9 – Дисперсии признаков в классах

Признак	Метод классификации					
	полных связей		Уорда		k-средних	
	1 класс	2 класс	1 класс	2 класс	1 класс	2 класс
nx1	1,144	0,862	1,261	0,806	0,858	0,876
nx2	0,330	0,532	0,408	0,491	0,331	0,326
nx3	0,306	0,820	0,298	0,843	0,505	0,945
nx4	0,617	0,506	0,772	0,482	0,512	0,410
nx5	0,936	0,909	1,189	0,773	0,983	1,000
Сумма дисперсий внутри класса	*3,334*	*3,628*	*3,929*	*3,396*	*3,188*	*3,557*
Сумма дисперсий по всем классам	*6,961*		*7,324*		*6,745*	

Получаем, что $Q(S_1) = 6{,}961$, $Q(S_2) = 7{,}324$ и $Q(S_3) = 6{,}745$. Таким образом, ориентируясь на данный критерий, при разделении на 2 класса наилучшей следует признать классификацию, полученную методом к-средних. Отметим, что такой подход к сравнению классификаций, полученных разными методами более обоснован, когда выбранное, например, по индексу Калински и Харабаза, оптимальное число классов одинаково для всех используемых методов. Так, для классификации по методу Уорда, число классов, равное 2, только близко к оптимальному. Поэтому, на наш взгляд, разделение муниципальных образований области на 2 класса, безусловно, довольно четко характеризует демографическую ситуацию в регионе, но лишь на довольно высоком уровне агрегирования.

4.2.2 Порядок создания do-файла

Выделим все использовавшиеся команды в окне Command и нажмем правую кнопку мыши. В контекстном меню выберем **Send to Do-file Editor** (рисунок 71).

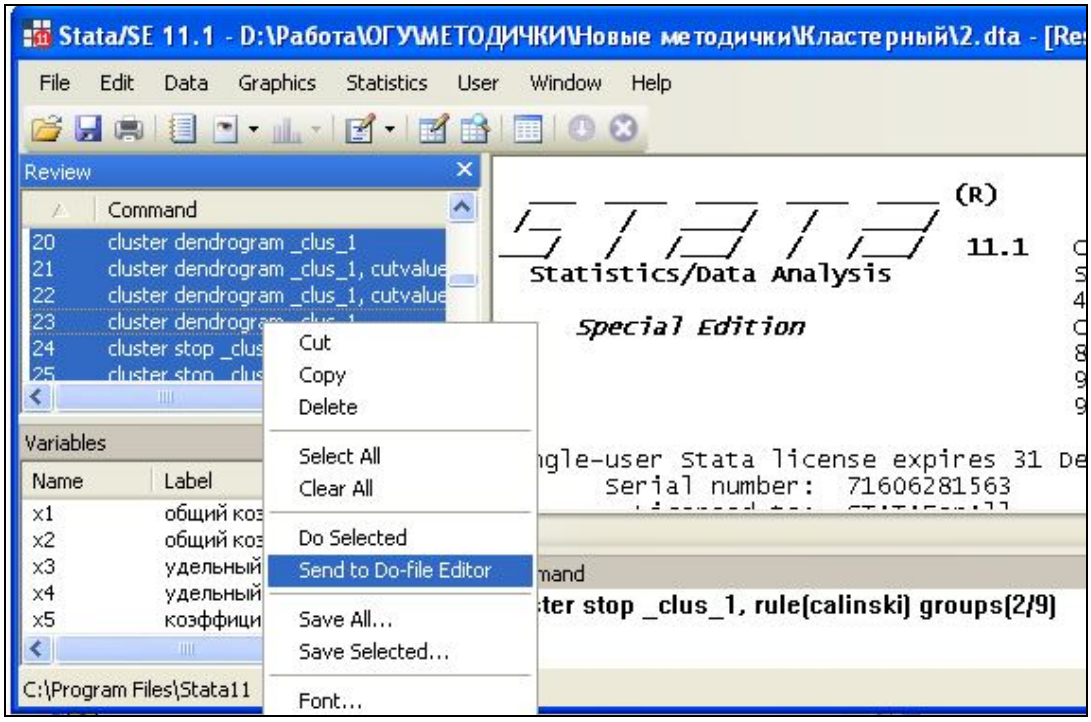


Рисунок 71 – Вид основного меню Stata перед копированием команд в редактор do-файлов

Откроется окно редактора do-файлов, в которое будут скопированы все выделенные команды (рисунок 72).

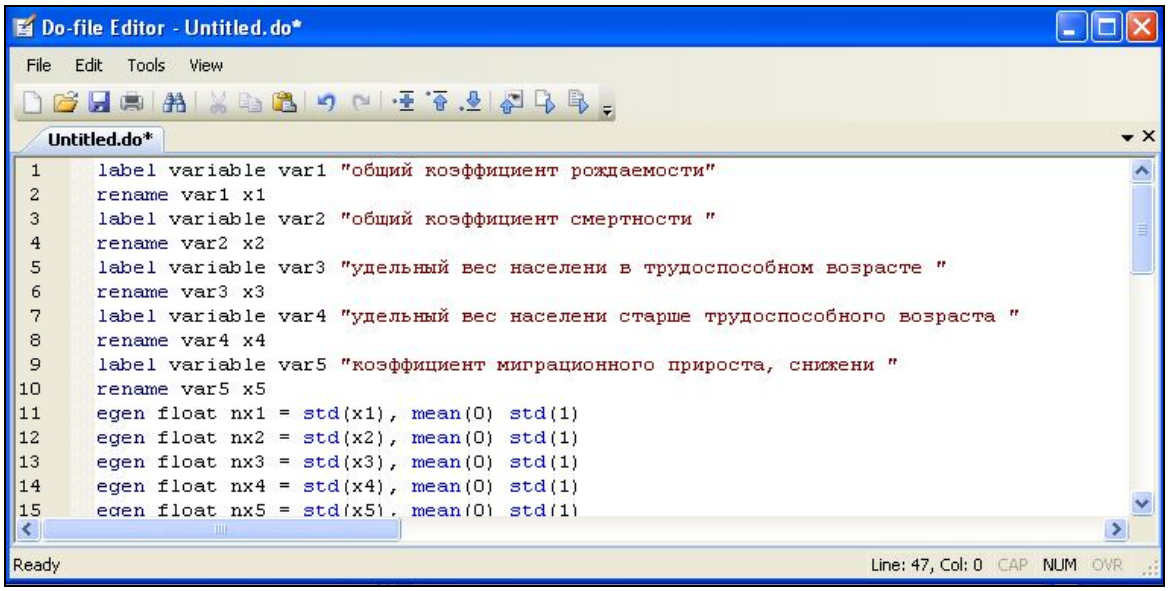


Рисунок 72 –Вид окна редактора do-файлов после вставки команд

При проведении анализа мы использовали однотипные операции, которые нужно было применить к нескольким переменным, например, стандартизация. Целесообразно в этом случае использовать цикл, который мог бы выглядеть, например, так:

```
forvalues i=1/5 {
egen float nx`i' = std(x`i'), mean(0) std(1)
}
```

Осуществляя любые операции со статистическими данными, нужно вести лог операций. Для этого используется команда **log using** *путь к файлу лога* **, replace**.

Например, **log using "C:\stata\cluster\cluster_log1.log", replace**.

Примерный вид do-файла для проведения кластерного анализа представлен на рисунке 73.

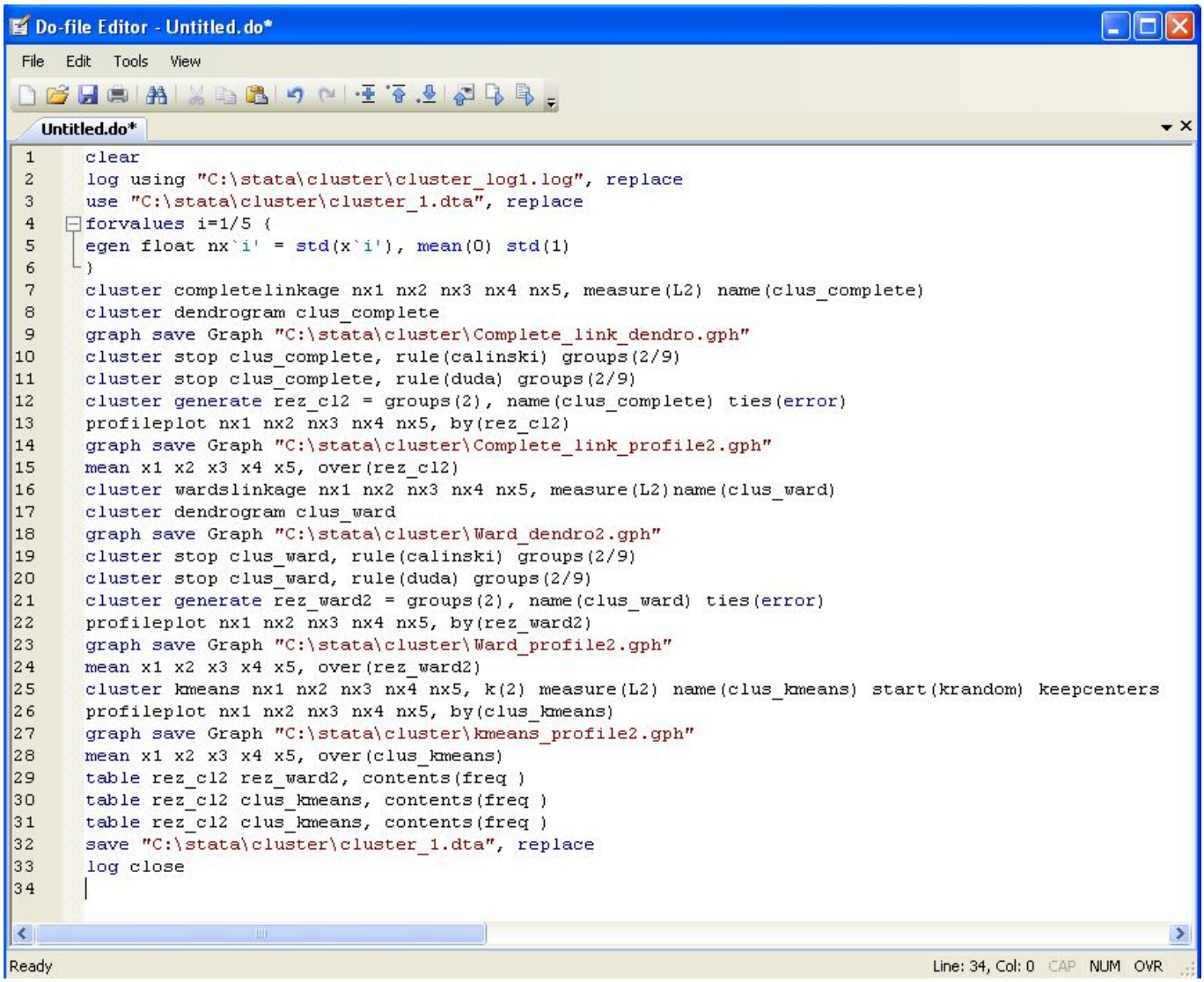


Рисунок 73 - Примерный вид do-файла для проведения кластерного анализа

Описание использованных команд представлено в таблице 10.

Таблица 10 – Описание команд в do-файле для проведения кластерного анализа

№ строки do-файла	Команда	Описание
1	2	3
1	clear	очистить память
2	log using "C:\stata\cluster\cluster_log1.log", replace	открыть или перезаписать существующий файл лога по указанному адресу
3	use "C:\stata\cluster\cluster_1.dta", replace	использовать для анализа файл с данными по указанному пути
4	forvalues i=1/5 {	начать цикл forvalues
5	egen float nx`i' = std(x`i'), mean(0) std(1)	создать новую переменную nxi, стандартизировав переменную xi
6	}	закончить цикл forvalues
7	cluster completelinkage nx1 nx2 nx3 nx4 nx5, measure(L2) name(clus_complete)	провести классификацию методом полных связей по переменным nx1 nx2 nx3 nx4 nx5, используя в качестве метрики обычное евклидово расстояние, задав имя clus_complete
8	cluster dendrogram clus_complete	построить полную дендрограмму, отражающую объединение классов по методу с именем clus_complete
9	graph save Graph "C:\stata\cluster\Complete_link_dendro.gph"	сохранить дендрограмму по указанному пути
10	cluster stop clus_complete, rule(calinski) groups(2/9)	вывести для кластерного анализа с именем clus_complete значения индекса Калински и Харабаза при разбиении на 2,3,…,9 классов
11	cluster stop clus_complete, rule(duda) groups(2/9)	вывести для кластерного анализа с именем clus_complete значения индекса Дуды и Харта при разбиении на 2,3,…,9 классов rez_cl2

Продолжение таблицы 10

1	2	3
12	cluster generate rez_cl2 = groups(2), name(clus_complete) ties(error)	создать переменную rez_cl2, которая каждому объекту ставит в соответствие число 1 или 2 в зависимости от того, к какому из двух классов относится объект в результате классификации методом кластерного анализа с именем clus_complete
13	profileplot nx1 nx2 nx3 nx4 nx5, by(rez_cl2)	построить график средних значений переменных nx1 nx2 nx3 nx4 nx5, взяв в качестве группирующей переменной rez_cl2
14	graph save Graph "C:\stata\cluster\Complete_link_profile2.gph"	сохранить график средних значений по указанному пути
15	mean x1 x2 x3 x4 x5, over(rez_cl2)	вывести на экран таблицу со средними значениями переменных x1 x2 x3 x4 x5, взяв в качестве группирующей переменной rez_cl2
16	cluster wardslinkage nx1 nx2 nx3 nx4 nx5, measure(L2) name(clus_ward)	-//-
17	cluster dendrogram clus_ward	-//-
18	graph save Graph "C:\stata\cluster\Ward_dendro2.gph"	-//-
19	cluster stop clus_ward, rule(calinski) groups(2/9)	-//-
20	cluster stop clus_ward, rule(duda) groups(2/9)	-//-
21	cluster generate rez_ward2 = groups(2), name(clus_ward) ties(error)	-//-
22	profileplot nx1 nx2 nx3 nx4 nx5, by(rez_ward2)	-//-
23	graph save Graph "C:\stata\cluster\Ward_profile2.gph"	-//-

Продолжение таблицы 10

1	2	3
24	mean x1 x2 x3 x4 x5, over(rez_ward2)	-//-
25	cluster kmeans nx1 nx2 nx3 nx4 nx5, k(2) measure(L2) name(clus_kmeans) start(group(rez_cl2)) keepcenters	провести кластерный анализ методом k-средних по переменным nx1 nx2 nx3 nx4 nx5, разбивать на 2 класса, в качестве метрики использовать обычно евклидово расстояние, а в качестве начальных эталонов – средние значения признаков в классах, определенных переменной rez_cl2 (в нашем случае это результаты разбиения методом полных связей)
26	profileplot nx1 nx2 nx3 nx4 nx5, by(clus_kmeans)	-//-
27	graph save Graph "C:\stata\cluster\kmeans_profile2.gph"	-//-
28	mean x1 x2 x3 x4 x5, over(clus_kmeans)	-//-
29	table rez_cl2 rez_ward2, contents(freq)	построить таблицу сопряженности переменных rez_cl2 и rez_ward2 (результатов классификации методом полных связей и методом Уорда)
30	table rez_cl2 clus_kmeans, contents(freq)	-//-
31	table rez_ward2 clus_kmeans, contents(freq)	-//-
32	save "C:\stata\cluster\cluster_1.dta", replace	сохранить внесенные в файл с данными изменения
33	log close	закрыть лог

Для запуска do-файла используется команда **File \ Do** .

5 Содержание письменного отчета

Отчет должен быть оформлен на листах формата А4 с титульным листом, оформленным соответствующим образом, и содержать следующее:

1) постановку задачи с исходными данными для анализа;
2) краткое изложение теории;
3) результаты компьютерной обработки данных;
4) анализ полученных результатов;
5) содержательную интерпретацию полученных результатов.

6 Вопросы к защите лабораторной работы

1) Сформулировать постановку задачи лабораторной работы.
2) Каким методом решалась задача классификации и чем обусловлен выбор этого метода?
3) Сформулировать, в чем суть выбранного метода решения задачи.
4) Какое программное средство использовалось для решения задачи?
5) Как решалась задача приведения признаков к одинаковым единицам измерения?
6) Из каких соображений задавалось расстояние между объектами?
7) Какие методы иерархических агломеративных кластер-процедур использовались при решении задачи?
8) Есть ли различия в результатах классификации муниципальных образований, полученных различными методами кластерного анализа? С чем это связано?
9) Как определялось оптимальное количество классов, на которые целесообразно разбить имеющуюся совокупность?
10) На основе какой информации была дана характеристика классам?
11) Привести наиболее и наименее типичные объекты для каждого класса, полученного методом к-средних?

12) Продемонстрировать, каким образом изменятся алгоритм работы с пакетами, выдаваемые результаты и их интерпретация в случае классификации не объектов, а признаков.

13) Как поступить в случае, если по результатам различных методов кластерного анализа один из объектов выделяется в отдельный класс? С чем это связано?

Список использованных источников

1 Айвазян, С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001. – Т. 1: Теория вероятностей и прикладная статистика. – 656 с.

2 Боровиков, В.П. STATISTICA – Статистический анализ и обработка данных в среде Windows / В.П. Боровиков, И.П. Боровиков. – М.: Инф. изд. дом «Филин», 1998. – 608 с.

3 Дубров, А.М. Многомерные статистические методы: учебник / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 1998. – 352 с.

4 Сошникова, Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г.Е. Уебе, М. Шефер. – М.: ЮНИТИ, 1999. – 598 с.

5 Тюрин, Ю.Н. Статистический анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров; под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528 с.

6 Большаков, А.А. Методы обработки многомерных данных и временных рядов: учебное пособие для вузов / А.А. Большаков, Р.Н. Каримов. – М.: Горячая линия - Телеком, 2008. – 522 с.

Приложение А
(обязательное)

Исходные данные для анализа

Таблица А.1 – Перечень социально-экономических показателей, характеризующих города и районы Оренбургской области

X1	удельный вес население в трудостпособном возрасте (%)
X2	удельный вес население старше трудостпособного возраста (%)
X3	доля женщин в общей численности (%)
X4	средний возраст (лет)
X5	общий коэффициент рождаемости (на 1000 человек)
X6	общий коэффициент смертности (на 1000 человек)
X7	коэффициент младенческой смертности (на 1000 человек)
X8	смертность от инфаркта (на 1000 человек)
X9	смертность от новообразований (на 1000 человек)
X10	смертность от отравлений алкоголем (на 1000 человек)
X11	смертность от самоубийств (на 1000 человек)
X12	смертность от убийств (на 1000 человек)
X13	обеспеченность населения врачами (на 10000 человек)
X14	общая заболеваемость (на 1000 человек)
X15	врожденные аномалии (на 1000 человек)
X16	травмы и отравления (на 1000 человек)
X17	уровень брачности населения (на 1000 человек)
X18	уровень разводимости населения (на 1000 человек)
X19	коэффициент миграционного прироста (на 1000 человек)
X20	среднемесячная номинальная начисленная заработная плата (руб.)
X21	число пострадавших с утратой трудоспособности (на 1000 человек)
X22	средний размер пенсий (руб.)
X23	площадь жилищ, приходящаяся в среднем на одного жителя (кв.м.)
X24	благоустройство жилищного фонда газом (%)
X25	благоустройство жилищного фонда отоплением (%)
X26	благоустройство жилищного фонда водопродом (%)
X27	число официально зарегистрированных безработных (на 1000 человек)
X28	охват детей дошкольными учреждениями (%)
X29	число дневных общеобразовательных учреждений
X30	удельный вес учащихся, занимающихся во II или III смену (%)
X31	инвестиции в основной капитал на душу населения (рублей)
X32	инвестиции, направленные в жилищное строительство
X33	удельный вес организаций, использующих электронную почту (%)
X34	удельный вес организаций, использующих сеть Интернет (%)
X35	удельный вес организаций, имеющих веб-сайт (%)

Продолжение таблицы А.1

X36	затраты на информационные и коммуникационные технологии (тыс.руб.)
X37	число учреждений культурно-досугового типа
X38	число общедоступных библиотек
X39	выбросы загрязняющих веществ в атмосферный воздух от стационарных источников (тысяч тонн/$км^2$))
X40	использование свежей воды (млн.куб.м.)
X41	число предприятий обрабатывающего производства
X42	число предприятий строительства
X43	число предприятий оптовой и розничной торговли
X44	наличие телефонных аппаратов
X46	ввод в действие жилых домов
X46	оборот розничной торговли (руб.)
X47	оборот общественного питания (тыс.руб.)
X48	объем платных услуг на душу населения (рублей)

Таблица А.2 – Значения социально-экономических показателей, характеризующих города и районы Оренбургской области, за 2008 год

Наименование района/города	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
Абдулинский	53,69	30,93	53,7	44,1	8,5	20,8	19,8	93,1	245,4	0	42,3	33,9	30	780,4	0,3	58,5
Адамовский	62,31	16,29	52,8	35,3	14,1	12,4	14,3	23,4	144	16,7	63,6	16,7	16,5	942,6	2,3	93
Акбулакский	62,71	16,74	50,5	35,6	15,5	11,9	4,5	23,6	175	26,9	43,7	16,8	21,9	625,8	0,5	39,8
Александровский	61,69	19,82	52	38	12,3	13,6	4,2	46,7	233,6	10,4	41,5	10,4	27	614,3	3,8	55
Асекеевский	59,44	23,32	52,3	42,2	12,5	17,2	3,5	44	184,9	13,2	79,3	26,4	20,9	905,5	1,7	63,5
Беляевский	61,76	19,21	51,9	37,6	14	14,7	7,4	25,8	211,7	10,3	36,1	15,5	23,7	733,3	0,9	50,8
Бугурусланский	60,42	22,54	52,9	39,4	13,5	19,4	6,7	27,3	213,5	36,3	45,4	36,3	19,2	626,1	0,7	62
Бузулукский	59,12	23,24	53,3	39,6	15,3	18,6	11,9	50,8	200,2	38,8	38,8	29,9	25,7	562,1	1,2	43,4
Гайский	59,43	20,59	53,1	38	15,8	17,7	23	9,1	145,9	27,4	91,2	27,4	33,2	918,9	5,4	75,3
Грачевский	60,03	22,17	52,7	39,9	13,1	15,2	5,1	47,2	195,6	47,2	40,5	6,7	28,6	835,3	1,3	70,4
Домбаровский	62,36	15,47	52,2	34,7	15,7	11,7	20,3	37,5	176,6	5,4	80,3	21,4	27,3	615,3	0,3	51,4
Илекский	59,89	21,06	52	38	15,1	15,9	2,3	59,4	216,7	28	59,4	14	23,8	1053,8	6,7	80,6
Кваркенский	60,41	19,20	52,2	37,4	14,5	16,7	19,7	56,9	241,9	23,7	56,9	19	19,1	860,7	0,1	105,3
Красногвардейский	60,37	19,17	52,1	37,4	15,2	13	2,8	21,6	155,6	8,6	51,9	13	27,3	728,7	1,6	58,2
Кувандыкский	58,77	21,23	52,6	38,2	14,7	15,9	3	66,6	191	4,4	66,6	8,9	28,4	873,2	4,3	53,2
Курманаевский	60,63	22,93	52,6	40,3	13	18,2	7,8	45,4	141,2	50,4	15,1	5	26,9	771,7	1,1	45,8
Матвеевский	58,95	23,67	53,1	40,4	10,4	17,5	0	48,3	262,4	13,8	82,9	13,8	23	849,4	1,6	38,3
Новоорский	61,11	19,27	53,2	37,2	14,6	16	8,6	63	198,4	9,5	103,9	28,3	27,5	558,3	0,3	48,9
Новосергиевский	59,47	21,70	52,9	38,6	14,3	19	11,4	75,9	233,2	21,7	54,2	35,3	25,5	689,1	4	81,2
Октябрьский	61,05	21,49	51,7	38,9	13,7	15,3	3,3	31,4	215	-	44,8	13,4	34,4	769,9	5,2	73,3
Оренбургский	64,18	17,94	52	36,8	13	11,4	7,3	33,4	168,5	12	45,5	9,4	39,8	806,9	2	76,2
Первомайский	62,17	16,11	51,6	35,5	16,8	11,8	12	35	136,5	14	63	10,5	22,1	1050,1	4,1	97,4
Переволоцкий	60,66	21,05	52,7	38,4	13,6	13	5,7	54	192,2	10,1	30,4	33,7	22,9	794,3	0,3	58,3
Пономаревский	58,45	25,70	53,1	41,5	10,4	17,5	23,2	48,1	210,4	6	54,1	12	27,8	672,2	0,7	64,4
Сакмарский	63,12	19,13	52,6	37,7	12,6	13,7	9,8	36,2	157,9	16,5	59,2	16,5	26,3	845,9	1,6	86,9
Саракташский	59,88	22,16	53	38,8	13,8	15,9	11,8	46	195,6	16,1	64,4	25,3	27,1	1058,6	1,7	88,4
Светлинский	61,71	17,88	52,7	36,5	12,9	14,2	9,2	53,4	255	5,9	65,2	11,9	23,3	676	0,6	33,1
Северный	59,67	23,77	52,2	40,7	9,8	19,6	0	46,8	117	5,9	41	17,6	26,6	884	3	98

Продолжение таблицы А.2

Наименование района/города	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
Соль-Илецкий	60,62	17,46	50,9	35,8	17,6	14,6	8,6	21,9	178,6	7,3	51	21,9	22,1	629,6	1,5	57,9
Сорочинский	57,23	23,30	52,8	39,2	14,8	19	19,2	39,2	137,3	0	78,5	13,1	30,4	852,6	1,9	98,9
Ташлинский	61,73	18,07	52,3	36,7	13,6	12,5	8,3	26,2	194,7	26,2	52,4	15	27,7	1024,2	7,2	71,8
Тоцкий	73,38	13,47	42,8	32,9	12	10,6	6,2	24,9	116,8	0	17,4	14,9	16,9	889,6	1,6	31,1
Тюльганский	62,51	19,45	52,6	38,1	12,7	14,2	3,4	25,6	213	12,8	55,4	12,8	23	1022,1	1,8	111
Шарлыкский	58,25	25,59	53,1	41,2	11,9	16,2	20,9	24,4	263	19,5	34,1	0	29,3	677,6	2,8	65,7
Ясненский	59,93	15,74	51,4	34,9	18,6	14,4	0	15,1	181,5	0	90,8	30,3	20	986,3	12,5	87,4
г.Абдулино	62,02	20,10	54,5	38,5	13,40	16,9	14,9	52,8	182,3	14,4	76,8	4,8	50,3	837,3	3,3	99,5
г.Бугуруслан	63,69	20,65	55,1	38,4	11,90	15,5	9,6	44	240,8	9,6	17,2	15,3	44,2	905,8	2,8	75,2
г.Бузулук	66,06	18,87	55	37,1	11,40	14,7	10,9	32,6	212,2	25,8	22,5	19,1	43,2	887,8	2	143,9
г.Гай	63,08	20,36	54,5	38,6	11,80	14,6	10,5	42,7	211,2	12,6	37,7	12,6	36,7	842,3	2,9	103,4
г.Кувандык	62,41	21,16	54,4	39,2	11,80	16,6	9	35,4	255,2	0	56,7	17,7	38,4	834,1	2,4	111,7
г.Медногорск	60,28	24,62	55,4	40,7	11,10	18,1	19,8	63,8	275,3	3,4	53,7	10,1	34,8	745,9	1,3	65,4
г.Новотроицк	64,41	20,57	52,7	38,8	10,70	16,5	8,6	53,1	229	33,4	44,2	27,5	38,4	979,3	7,7	123,5
г.Оренбург	65,61	19,21	54,4	37,9	11,90	12,6	6,1	81,5	212,2	19,2	23,8	14,5	104,5	838	8,6	143,9
г.Орск	63,04	20,93	55,8	38,3	12,10	16,7	8,6	43,6	224,6	32,6	44,4	30,2	37,6	665,5	1	104,6
г.Соль-Илецк	63,34	17,03	51,2	36,6	15,90	11,8	4,8	30,3	189,1	11,3	30,3	26,5	45,1	746,1	3,1	107,9
г.Сорочинск	62,90	19,09	53,4	37,8	14,60	14,7	1,7	27,4	201,9	17,1	37,6	17,1	43,6	792,1	3,4	109,5
г.Ясный	69,10	12,00	49,9	33,4	15,10	10,3	2,5	53,8	150	3,8	61,5	15,4	44,3	825,3	2,8	112,4

Продолжение таблицы А.2

Наименование района/города	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32
Абдулинский	5,7	0,2	-22	6844	2,4	3533	24,9	100	63,7	69	25	7	26	0,8	15626	4559,3
Адамовский	7,7	3,6	-9,6	8102	1,4	3643	19,9	98,6	97,8	180	69,6	41,1	33	9,2	23410	5265,6
Акбулакский	6,8	3,9	-3,4	6897	3,1	3725	17,4	95,4	66,4	224	37,8	37,4	42	4,1	9855	2853,3
Александровский	6,9	4,2	-2,3	7933	2,5	3882	18,9	100	100	116	94,2	34,6	38	7,4	4881	1033,6
Асекеевский	7,7	4,1	-11	6601	0,8	3883	20,4	93,9	94,3	172	37,3	38,3	40	7,1	15466	4415,8
Беляевский	7	4,2	1,7	7600	9,3	3713	18,4	97,9	99,9	163	57	44,7	29	2,7	12186	3576,2
Бугурусланский	8,8	5,2	-8,1	8359	0,9	3845	20,9	97,7	98,9	104	32,2	20	33	2,7	7621	1955,9
Бузулукский	11,1	5,6	2,8	8846	0,4	3962	20,2	97,8	82,3	142	44,6	30,1	38	0,4	9143	3129,6
Гайский	5,8	4,9	-16	10441	0	3763	21,4	98,5	91,9	77	36,3	43,3	28	5,2	28919	4539
Грачевский	8,2	5,2	-15,6	9173	0,5	4029	23,2	100	100	61	61,9	44	18	0	14864	4148,9
Домбаровский	8,5	5,1	-8,2	9890	1,5	3601	20,6	98,7	100	69	36,7	53,5	18	17,8	7589	2824,3
Илекский	7,8	4,3	1,3	7422	2,9	3836	18,9	97,3	97,6	187	65	28,4	23	7,9	13617	3993,6
Кваркенский	7,5	3,5	-15,4	7024	3,1	3669	20,2	97,2	49,4	219	77,1	38,4	36	3,6	13341	2613,9
Красногвардейский	8,3	4,2	-6,9	8905	0	3896	21,3	99,1	93,4	122	91,2	37,2	40	2,8	15138	4475
Кувандыкский	5,9	0,3	-18,3	6481	3,2	3557	18,3	100	99,1	105	61,6	29	44	0	17342	6072,6
Курманаевский	8,3	6,3	-5,6	9907	1,6	4042	22	99,2	97,1	115	44,4	52,3	24	0	9813	3164,7
Матвеевский	6,8	3,2	-8,6	8154	0,8	3918	21,5	99,6	99,6	128	62,4	45,2	19	0	6909	2934,5
Новоорский	9,9	5,6	-7	12340	1	3916	25,4	93,2	88,8	72	61,6	66,5	21	3,6	54648	6977,6
Новосергиевский	8,8	4,4	1,8	9065	2,9	3843	21,9	99,9	97,6	102	64	30,6	58	9,1	13043	4275,5
Октябрьский	7,7	3,9	8,6	9239	4,1	3968	22,3	98,2	99,2	137	70,9	33,3	25	5,7	15905	6648,2
Оренбургский	9,1	4,1	20,1	18439	1,9	3939	20,4	98,2	59	165	75,2	55,8	52	8,8	86533	16552,1
Первомайский	7,9	4,2	-9	9467	2,1	3698	18,3	99,1	97,5	196	59,9	53,2	49	2	16027	3837,1
Переволоцкий	10,3	2,9	2,7	7516	3,4	3947	19,9	99	96,2	170	71,6	39,9	38	11,4	7208	1630,9
Пономаревский	6,9	4,2	-2,9	8703	1	3930	23,3	99,4	90,9	206	41,3	43,2	18	18,2	10831	5551
Сакмарский	7,7	4,2	2,3	9007	3,5	3920	19,2	100	93,5	107	69,3	34,9	19	12,7	34435	6311
Саракташский	9,2	4,8	8,2	7889	3,3	3926	18,6	94,2	100	75	49,6	45,1	44	17,9	17738	6073,3
Светлинский	6,6	5,4	-12,8	9575	1,8	3868	21,4	98,9	99,8	131	80,8	41,3	12	15,4	6122	0
Северный	6,7	3,6	-6,8	8569	1,8	3896	20,9	97,5	99,8	37	42,3	53,2	31	0	10028	4393
Соль-Илецкий	6,8	0,3	-8,9	6731	1,6	3784	17	96,1	93,7	93	40,5	28,1	37	9,2	4463	493,1

Продолжение таблицы А.2

Наименование района/города	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30	X31	X32
Сорочинский	6,4	4,9	-6,1	7910	9,6	3874	21,9	99	99,3	94	51,1	48,9	26	0	7992	2166,5
Ташлинский	8,1	4,6	0,5	6223	3,1	3821	21,4	100	100	104	49,3	48,4	46	11,4	20077	6295,6
Тоцкий	9	4,9	-4,8	8513	3	3852	18,2	97,4	93,1	214	78,2	48,5	35	12,1	6730	2880,3
Тюльганский	8,2	6	-0,2	7713	3,8	3901	19,9	94,7	89,7	172	76,9	56,1	25	5,7	9785	3083,7
Шарлыкский	7	4,2	-0,7	7900	9	3850	21,2	98,5	95,3	129	47,2	30,2	34	7,7	14754	4949,7
Ясненский	7	0,3	-20,1	10116	1,3	3912	19,1	94	50,3	62	65,9	20,4	12	0	16464	1714,4
г.Абдулино	9,3	6,5	-3,9	11221	2,6	3973	22,3	99,9	99,8	162	51,6	49,7	8	20,8	7473	5016,9
г.Бугуруслан	7,6	5,3	-3,1	12772	1,6	4342	20,9	96,2	62,6	244	61	73	8	13,2	31121	3555,4
г.Бузулук	7,8	5,4	7,8	16348	2,3	4328	20,1	88,3	80,6	332	72	67,9	14	18,7	244895	10262,6
г.Гай	8,9	6	-5,2	14791	2,1	4431	21,4	95,8	99,9	176	95	75,2	9	0	107950	3436,2
г.Кувандык	8,2	8,9	-0,6	9829	2,2	3915	18,8	97,3	100	245	58	67,1	7	10,8	7844	5866,1
г.Медногорск	6,8	4,6	4,3	10516	2,3	4391	22,2	76,5	96,6	210	68,9	55,9	13	14,9	12107	2302,6
г.Новотроицк	8,1	5,1	-2,7	14167	1,3	4504	20,7	97,1	99,4	190	94,1	80	23	7,3	101680	2758,7
г.Оренбург	8,3	5	-3,1	15940	1,8	4409	21,3	90,3	98,8	1627	98,6	69,1	96	14,9	47374	7639,9
г.Орск	7,8	5,3	2,5	12428	3,4	4364	22,2	98,1	99,7	945	83,8	63,4	53	10,6	17891	2791,1
г.Соль-Илецк	9,4	7,8	-8,4	11261	2,2	4005	16,5	94,4	85	227	86,8	59,2	8	30,9	23894	11293,5
г.Сорочинск	9,6	6,7	0,8	12267	4,1	4020	20,6	100	100	221	84,4	69,9	7	48,1	17885	5730,8
г.Ясный	9,6	9,4	-15,4	11366	1	3923	18,6	95,4	100	144	100	85,5	4	6,4	13982	4417,1

Продолжение таблицы А.2

Наименование района/города	X33	X34	X35	X36	X37	X38	X39	X40	X41	X42	X43	X44	X45	X46	X47	X48
Абдулинский	0	0	0	742,3	35	25	0,1	0,3	2	1	5	91,5	268,5	9963	0	1547
Адамовский	74,3	77,1	0	6450,3	31	20	1,4	1,1	12	8	37	154,1	352	14363	18611	4552,6
Акбулакский	68,8	40,6	12,5	5640,7	29	21	0,7	0,8	18	11	35	135,4	80,8	10739	11317	3855,3
Александровский	92,3	61,5	11,5	3487,7	38	27	0,2	0,8	9	4	31	134,4	62,3	16051	17912	3826,7
Асекеевский	72,7	42,4	0	2877,2	43	28	5,3	0,5	12	5	24	156,5	266,3	10774	18464	4633,3
Беляевский	39,3	39,3	3,6	3994,3	28	21	0,1	0,7	8	6	21	136,5	213,1	9898	13844	3768
Бугурусланский	15	35	0	1970,9	36	30	2,1	3,1	4	14	22	92,9	226,3	16142	20018	2914,6
Бузулукский	54,5	15,2	0	1660,7	49	39	6,2	2,6	21	11	29	99,7	192,7	7242	15590	4258,6
Гайский	66,7	66,7	11,1	802,3	29	17	0,2	0,1	6	4	12	83,9	281,4	18897	3674	3871,3
Грачевский	64,3	57,1	3,6	2884	21	17	22,2	0,6	10	3	27	193,7	250,2	10971	14515	5850,7
Домбаровский	77,3	68,2	4,5	3483,5	21	17	7,9	0,9	12	11	31	133	180,1	12878	12650	4147,5
Илекский	62,1	44,8	10,3	3988,8	19	21	0	1,5	7	8	20	113,7	240,8	11408	17097	3966,6
Кваркенский	63	66,7	7,4	4991	29	21	5,9	1	3	6	21	167,6	163,4	9976	7053	3827,6
Красногвардейский	85,7	78,6	17,9	5349,1	43	34	74,4	2,6	14	9	20	151,2	269,8	13421	37536	4147,3
Кувандыкский	42,9	9,5	0	844,4	34	31	5,8	0,4	9	2	6	107,9	207,4	4331	1799	1779,1
Курманаевский	88,9	74,1	3,7	6775,2	33	27	22,8	1	11	4	28	154,4	181,2	9815	47800	4109,8
Матвеевский	53,8	53,8	7,7	2470,1	29	21	1,5	0,5	2	4	11	130,6	177	17208	16519	3143,4
Новоорский	68	64	16	46652,3	21	18	8,4	1550,5	27	21	53	157	521,1	19332	57697	8751,2
Новосергиевский	65,9	56,8	9,1	9457,3	52	33	38,4	3,4	22	14	60	146,5	253,2	28028	61786	4933,5
Октябрьский	70	46,7	6,7	9137,2	26	23	4,3	1,1	18	8	24	161,1	474,6	14967	24317	5179
Оренбургский	82,1	85,1	23,9	53979,9	40	36	8,1	8	170	317	559	88	924,2	28298	237451	16947
Первомайский	77,4	90,3	6,5	3616	38	30	17,1	3,3	16	27	26	107,8	231,4	13062	46358	3915,4
Переволоцкий	43,2	64,9	10,8	7944,5	34	27	93,9	2,9	16	13	48	132,2	98,4	14311	58780	4625,6
Пономаревский	41,4	51,7	3,4	3265	24	22	11	2,3	6	10	14	150,5	334,8	21257	25774	3684,3
Сакмарский	81,3	62,5	6,3	3605,3	18	20	2,5	2,4	18	25	58	106,3	380,6	13571	25973	4535
Саракташский	93,8	66,7	8,3	5959	48	35	7,6	1	30	20	57	147,5	370,5	14311	32759	5720,4
Светлинский	82,6	52,2	13	3998,2	12	11	0,6	2,6	16	7	24	192,6	145,4	17574	4240	6631,4
Северный	48,4	41,9	6,5	2866	35	30	5,2	0,4	21	5	22	191	263,2	18705	6904	4767,7
Соль-Илецкий	13	13	4,3	1181,7	35	29	0	0,5	7	3	14	85,5	23,2	18555	6192	1397,2

Продолжение таблицы А.2

Наименование района/города	X33	X34	X35	X36	X37	X38	X39	X40	X41	X42	X43	X44	X45	X46	X47	X48
Сорочинский	15,8	52,6	0	635,5	32	24	13,6	4,2	4	1	0	108,8	130,7	7341	20041	2282,4
Ташлинский	50	56,3	6,3	3715,9	41	28	7,6	1,5	11	11	30	129,5	379,7	14708	10499	4025,5
Тоцкий	88,9	41,7	5,6	4824,2	34	25	1,4	3,1	21	14	43	126,8	173,7	14467	49467	4323,4
Тюльганский	47,1	58,8	5,9	4901,9	22	22	0,1	0,9	19	5	30	167,1	185,5	14599	14065	5728,7
Шарлыкский	36,7	36,7	6,7	3438,9	35	29	3,3	0,5	9	14	23	190,8	269,1	20166	10140	3820
Ясненский	36,4	18,2	0	238,3	18	10	1	0,3	2	1	3	146,4	27,6	6757	1976	3707,2
г.Абдулино	84,6	80,8	7,7	9089,6	1	4	0,2	0,8	37	18	51	213,3	302,5	39047	41094	8441,5
г.Бугуруслан	94	90	16	62468	3	10	0,3	4,2	71	68	220	262,4	359,5	29235	91612	11370,2
г.Бузулук	92	88	25,3	178888,6	3	10	4,3	9,9	135	130	432	246,8	385,9	35606	179393	16549,6
г.Гай	90,3	80,6	22,6	30749,5	2	4	1,2	14,9	62	20	100	291,1	362,1	27158	104235	13356,9
г.Кувандык	75	75	50	7366,9	1	7	1,1	3,1	32	18	61	254,3	353,8	27173	49716	12733,9
г.Медногорск	88,9	83,3	33,3	14655,8	7	12	47,8	5,1	44	24	49	213	108,8	22986	17350	9440,7
г.Новотроицк	83,7	91,8	24,5	112232,3	8	11	87,3	35,2	150	195	287	282,4	151	34918	435156	14288,5
г.Оренбург	90,8	91,4	44,1	1510639,5	14	32	54,5	79,7	1327	1995	4714	274,7	600	158053	4773575	23086,9
г.Орск	93,2	85,6	23,7	169540,4	12	15	155,9	45,8	399	439	1203	243,9	206	50514	470875	15718,9
г.Соль-Илецк	85,2	92,6	11,1	12219,6	8	6	0,3	1,4	22	28	89	187,3	549,8	39541	34899	11761,7
г.Сорочинск	92	92	20	12701,8	2	5	1,2	1,2	25	19	61	155,2	469,2	29225	64294	15597,3
г.Ясный	76,9	73,1	23,1	12468,1	3	2	3,1	4,9	9	14	33	275,3	243,7	23697	56004	10261

Приложение Б
(справочное)

Результаты кластерного анализа

Таблица Б.1 – Средние значения признаков в кластерах, полученных методом «полных связей»

Показатель	Обозначение	Кластер 1	Кластер 2	Кластер 3
Общий коэффициент рождаемости	X1	-0,2217117	1,1640639	-0,3499324
Общий коэффициент смертности	X2	-0,357436	-0,9348287	1,0173155
Удельный вес населения в трудоспособном возрасте	X3	0,5827793	0,0558887	-0,832759
Удельный вес населения старше трудоспособного возраста	X4	-0,2751167	-1,0038977	0,9429779
Коэффициент миграционного прироста	X5	0,5839905	-0,6457664	-0,4397433

Таблица Б.2 – Средние значения признаков в кластерах, полученных методом Уорда

Показатель	Обозначение	Кластер 1	Кластер 2	Кластер 3
Общий коэффициент рождаемости	X1	-0,44396	0,805393	-0,06339
Общий коэффициент смертности	X2	-0,3674	-1,18114	0,787384
Удельный вес населения в трудоспособном возрасте	X3	0,514698	0,738928	-0,68681
Удельный вес населения старше трудоспособного возраста	X4	-0,16189	-1,21698	0,663552
Коэффициент миграционного прироста	X5	0,711036	-0,52856	-0,24454

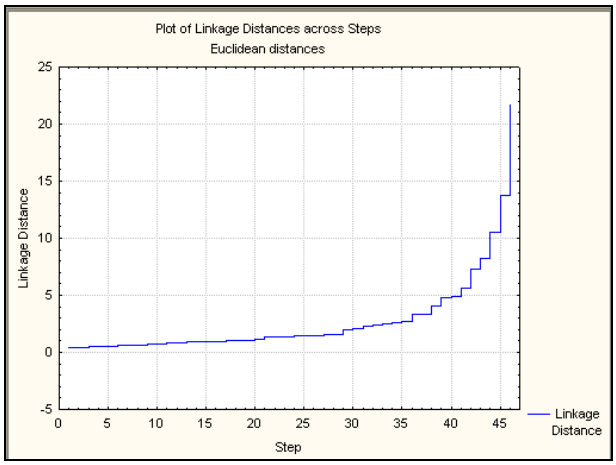


Рисунок Б.1 – График изменения расстояния между объединенными классами, полученными методом Уорда

Таблица Б.3 – Результаты классификаций муниципальных образований Оренбургской области, полученных различными методами кластерного анализа

Муниципальные образования	Методы кластерного анализа		
	полных связей	Уорда	*K*-средних
Абдулинский	3	3	3
Адамовский	2	2	2
Акбулакский	2	2	2
Александровский	1	1	1
Асекеевский	3	3	3
Беляевский	1	1	1
Бугурусланский	3	3	3
Бузулукский	3	3	3
Гайский	3	3	3
Грачевский	3	3	3
Домбаровский	2	2	2
Илекский	1	3	1
Кваркенский	3	3	3
Красногвардейский	2	2	2
Кувандыкский	3	3	3
Курманаевский	3	3	3
Матвеевский	3	3	3
Новоорский	1	3	1
Новосергиевский	3	3	3
Октябрьский	1	3	1
Оренбургский	1	1	1
Первомайский	2	2	2
Переволоцкий	1	1	1
Пономаревский	3	3	3
Сакмарский	1	1	1
Саракташский	1	3	1

Продолжение таблицы Б.3

Муниципальные образования	Методы кластерного анализа		
	полных связей	Уорда	*К*-средних
Светлинский	2	2	2
Северный	3	3	3
Соль-Илецкий	2	3	2
Сорочинский	3	2	3
Ташлинский	1	1	1
Тоцкий	1	2	2
Тюльганский	1	1	1
Шарлыкский	3	3	3
Ясненский	2	3	2
г.Абдулино	1	3	1
г.Бугуруслан	1	1	1
г.Бузулук	1	1	1
г.Гай	1	1	1
г.Кувандык	1	1	1
г.Медногорск	3	3	3
г.Новотроицк	1	1	1
г.Оренбург	1	1	1
г.Орск	1	1	1
г.Соль-Илецк	2	2	2
г.Сорочинск	1	1	1
г.Ясный	1	2	2