Министерство образования и науки Российской Федерации

Севастопольский государственный университет

Кафедра ИС

Отчет

по лабораторной работе №4

«КЛАСТЕРНЫЙ АНАЛИЗ»

Выполнил:

ст. гр. ИTб-33д

Лисянский А. И.

Проверил:

Токарев А.И.

Севастополь

2015

**1.Цель работы**

Закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа по экспериментальным данным.

По результатам наблюдений за функционированием объектов получены экспериментальные данные. Требуется провести кластерный анализ этих данных.

**2.Ход работы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Название монеты | Год выпуска | Номинал, грн | Диаметр, мм | Вес, г | Тираж, шт. |
| Рождество Христово | 1999 | 5 | 35,00 | 16,54 | 100000 |
| Крещение Руси | 2000 | 5 | 35,00 | 16,54 | 100000 |
| 10-летие Вооруженных Сил Украины | 2001 | 5 | 28,00 | 9,40 | 50000 |
| 300-летие Конституции Филиппа Орлика | 2010 | 10 | 38,60 | 31,10 | 7000 |
| 5 лет Конституции Украины | 2001 | 2 | 31,00 | 12,80 | 30000 |
| Независимость | 1996 | 2000000 | 39,00 | 31,10 | 10000 |
| Государственный Гимн Украины | 2005 | 10 | 38,61 | 31,10 | 3000 |
| Богдан Хмельницкий | 1995 | 200000 | 33,00 | 14,35 | 250000 |
| Юрий Кондратюк | 1997 | 2 | 33,00 | 14,35 | 20000 |
| Небесная сотня | 2015 | 5 | 35,00 | 16,54 | 50000 |
| Город-герой Севастополь | 1995 | 200000 | 33,00 | 14,35 | 75000 |
| Бандура | 2003 | 5 | 28,00 | 9,40 | 30 000 |
| Гончар | 2010 | 5 | 35,00 | 16,50 | 45000 |
| 50-летие Организации Объединенных Наций (ООН) | 1995 | 200000 | 38,61 | 28,28 | 100000 |
| Героическая оборона Севастополя 1854–1856 годов | 2004 | 10 | 38,61 | 31,10 | 8000 |

Таблица 1 – значение характеристик юбилейных монет Украины.

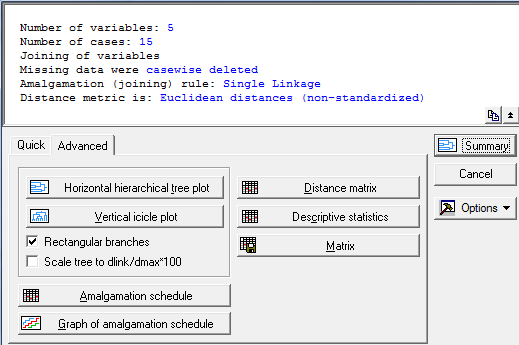


Рисунок 1 - Окно, содержащее результаты кластерного анализа "Joining Results" .

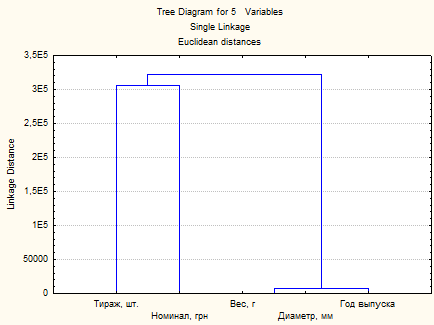


Рисунок 2 - Vertical icicle plot

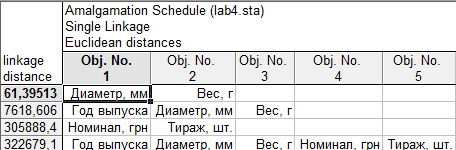


Рисунок 3 -Щелчком мыши раскроем строку Amalgamation schedule, содержащую протокол объединения кластеров.

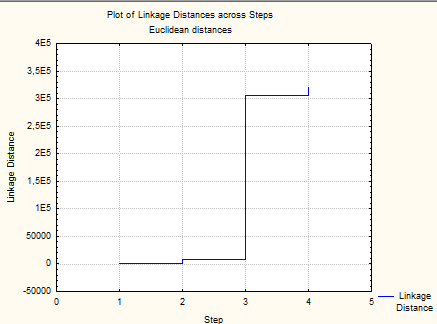


Рисунок 4 – Граф расстояний при объединении кластеров

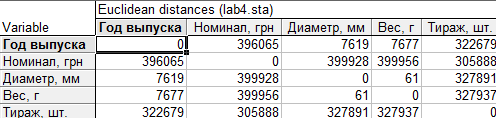


Рисунок 5 - Матрица расстояний.

**K - means clustering** (*метод К средних*).

Суть этого метода состоит в следующем: исследователь заранее определяет количество классов (k) на которые необходимо разбить имеющиеся наблюдения, и первые k - наблюдений становятся центрами этих классов. Для каждого следующего наблюдения рассчитываются расстояния до центров кластеров и данное наблюдение относится к тому кластеру, расстояние до которого было минимальным. После чего для этого кластера (в котором увеличилось количество наблюдений) рассчитывается новый центр тяжести ( как среднее по каждому показателю) по всем включенным в кластер наблюдениям.

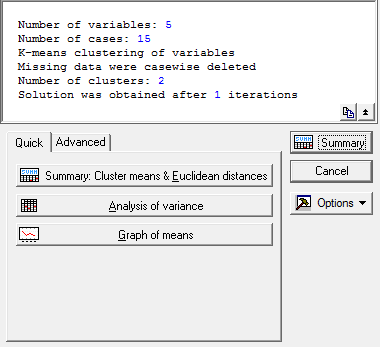


Рисунок 6 - K-Means Clustering Results

**Вывод результатов и их анализ**

В верхней части окна (в том же порядке, как они идут на экране):

* Количество переменных;
* Количество наблюдений;
* Классификация наблюдений (или переменных, зависит от установки в предыдущем окне в строке **Cluster**) методом K - средних;
* Наблюдения с пропущенными данными удаляются (или: изменяются средними значениями. Зависит от установки в предыдущем окне в строке Missing data).
* Количество кластеров;
* Решение достигнуто после : итераций.
* **Analysis of Variance** (*анализ дисперсии*). После нажатия появляется таблица (рис.7) , в которой приведена межгрупповая и внутригрупповая дисперсии. Где строки - переменные (наблюдения), столбцы - показатели для каждой переменной: дисперсия между кластерами, число степеней свободы для межклассовой дисперсии, дисперсия внутри кластеров, число степеней свободы для внутриклассовой дисперсии, F - критерий, для проверки гипотезы о неравенстве дисперсий. Проверка данной гипотезы похожа на проверку гипотезы в дисперсионном анализе, когда делается предположение о том, что уровни фактора не влияют на результат.

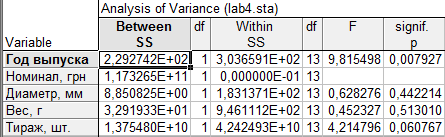


Рисунок 7 - Analysis of Variance (*анализ дисперсии*).

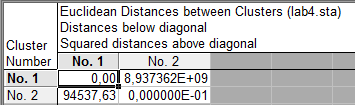


Рисунок 8 – Расстояние между кластерами

В таблице приведены расстояния между класстерами. И по вертикали и по горизонтали указаны номера кластеров. Таким образом при пересечении строк и столбцов указаны расстояния между соответствующими класстерами. Причем выше диагонали (на которой стоят нули) указаны квадраты, а ниже просто евклидово расстояние.

Graph of means представляет собой графическое изображение (рис.9) информации содержащейся в таблице, выводимой при нажатии кнопку Analysis of Variance (анализ дисперсии). На графике показаны средние значения переменных для каждого кластера.

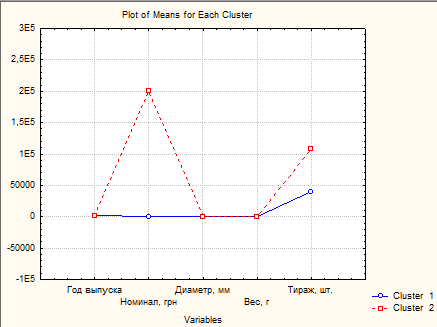


Рисунок 9 - Graph of means.

По горизонтали отложены учавствующие в классификации переменные, а по вертикали - средние значения переменных в разрезе получаемых кластеров.

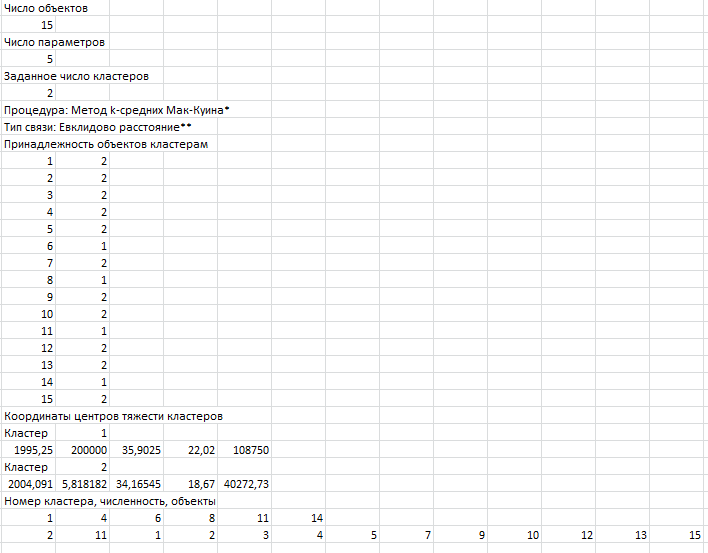


Рисунок 10 – Кластерный анализ в Exсel

Рисунок 11 – гистограмма центров тяжести переменных по кластерам

Выводы в ходе лабораторной работы был изучен кластерный анализ по экспериментальным данным. Обычно, когда результаты кластерного анализа методом K средних получены, можно рассчитать средние для каждого кластера по каждому измерению, чтобы оценить, насколько кластеры различаются друг от друга. (В нашем случае (рис.9), значения переменных пересекаются, но все же мы можем наблюдать достаточно четкие различия кластеров. Для более отчетливой группировки следует сократить число параметров.). Значения F-статистики, полученные для каждого измерения, являются другим индикатором того, насколько хорошо соответствующее измерение дискриминирует кластеры. Так как у нас решение найдено после одной итерации (меньше чем мы задали), то можно сделать выводод о том, что итоговая конфигурация является искомой.