**ПРАКТИЧЕСКАЯ РАБОТА №8**

Метод К-средних (K-Means Clustering)

### Цель работы: научить студентов методам группирования многомерных данных, показать возможности визуализации последовательного формирования кластеров сходных объектов, продемонстрировать вариативность методов кластеризации.

### Используем: Пакет Statistica 6.0

### Можно выполнить в Excel:

### [Excel кластерный анализ Excelka.ru - все про Ексель](https://excelka.ru/vba/excel-klasternyj-analiz.html#:~:text=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%20K%2D%D1%81%D1%80%D0%B5%D0%B4%D0%BD%D0%B8%D1%85%20%D0%B7%D0%B0%D0%BA%D0%BB%D1%8E%D1%87%D0%B0%D0%B5%D1%82%D1%81%D1%8F%20%D0%B2%20%D1%81%D0%BB%D0%B5%D0%B4%D1%83%D1%8E%D1%89%D0%B5%D0%BC%3A,%D0%B8%20%D0%BC%D0%B0%D0%BA%D1%81%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8%20%D0%B8%D0%B7%D0%BC%D0%B5%D0%BD%D1%87%D0%B8%D0%B2%D0%BE%D1%81%D1%82%D0%B8%20%D0%BC%D0%B5%D0%B6%D0%B4%D1%83%20%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B0%D0%BC%D0%B8)

### [Как сделать кластерный анализ в Excel: сфера применения и инструкция (exceltable.com)](https://exceltable.com/otchety/kak-sdelat-klasternyy-analiz)

### Можно выполнить онлайн:

### [Метод К-средних онлайн (semestr.ru)](https://axd.semestr.ru/upr/average.php)

### ХОД РАБОТЫ:

### *Иерархическая кластеризация*

1. Файл данных

Пример основан на выборке данных об автомобилях. Открыть файл Cars.sta (через директорию Datasets / Examples), в котором находятся следующие переменные:

* Цена машины (переменная Price);
* Время разгона автомашины (переменная Acceler);
* Время торможения (со скорости 80 миль/час до полной остановки – переменная Braking);
* Коэффициент управляемости (переменная Handling);
* Пробег на галлон топлива (переменная Mileage).
1. Шкала измерений

Все алгоритмы кластеризации вычисляют расстояние между объектами или кластерами, поэтому при использовании переменных, измеренных в различных единицах, необходимо привести данные к такой шкале, посредством которой можно производить действия с переменными. Обычно этой возможности достигают с помощью стандартизации данных (командой Standardize из меню Data), в результате которой каждая переменная имеет нулевое среднее значение и единичную дисперсию. Важно отметить, что значения переменных, которые используются при вычислении расстояний, должны иметь сравнимые величины, так как в противном случае результаты кластерного анализа окажутся смещенными. В файле Cars.sta операция стандартизации уже выполнена.

Файл данных приведен на рис. 1.

1. Цель анализа

Группирование автомобилей по набору переменных.



Рис. 1. Файл данных (стандартизованные величины)

1. Выполнение анализа

Выбрать модуль Cluster Analysis из меню Statistics - Multivariate Exploratory Techniques для отображения стартовой панели Clustering Method. Далее, выбрать опцию Joining (tree clustering) и нажать ОК. В появившемся окне Cluster Analysis: Joining (Tree Clustering) указать позицию Variables для отображения окна диалога по выбору переменных (select variables), в котором необходимо отметить все переменные, после чего, нажав ОК, возвратиться к окну кластерного анализа (рис. 2).

Вследствие того, что в рассматриваемом примере группируются автомобили (строки исходной матрицы данных), то при положении окна кластерного анализа в опции Advanced в позиции Cluster нужно выбрать Cases (rows). В зависимости от целей исследования могут группироваться и признаки (столбцы матрицы данных); в этом случае в окне Cluster выбирается Variables (columns). Кроме того, в раскрывающемся окне правил объединения (Amalgamation (linkage) rule) указать метод полной связи (Complete Linkage). Окно анализа примет вид, показанный на рис. 3.



Рис. 2. Окно кластерного анализа



Рис. 3. Окно кластерного анализа после выбора переменных

1. Меры расстояния

Методы кластеризации успешно работают на основе сходства (различия) между объектами. Естественной оценкой сходства является расстояние, которое рассчитывается в методах кластеризации. В качестве оценок расстояний могут быть выбраны следующие:

Евклидово расстояние: distance (x, y) = {Si (xi – yi)2}1/2;

Квадрат евклидова расстояния: distance (x, y) = Si (xi – yi)2;

Манхеттенское расстояние: distance (x, y) = Si |xi – yi|;

Чебышевское расстояние: distance (x, y) = Maximum |xi – yi|;

Степенное расстояние: distance (x, y) = (Si |xi – yi|p)1/r.

В данном примере выберем в качестве меры расстояния – евклидово и отметим этот выбор в окне расстояний.

1. Метод объединения

Выбор метода объединения определяется значениями коэффициентов в формуле Ланса-Уильямса, используемой в кластерном анализе и зависит от целей анализа. В любом случае полученная группировка объектов должна проверяться различными методами и оцениваться на основании логики и здравого смысла. Здесь вначале выберем правило группирования в виде полных связей (Complete Linkage).

1. Результаты кластеризации

Нажав ОК, приходим к окну результатов кластерного анализа (рис. 4).



Рис. 4. Окно результатов кластерного анализа

В этом окне можно выбрать вид полученной дендрограммы: горизонтальной (Horizontal hierarchical tree) или (Vertical icicle plot), указать тип ветвей этого дерева (прямоугольные или наклонные), получить и проанализировать матрицу расстояний между объектами, построить последовательный график объединения объектов и просмотреть ход группирования. На рис. 5 и 6 показаны, соответственно, в качестве примеров вертикальная дендрограмма и график объединения.



Рис. 5. Дендрограмма объектов



Рис. 6. График объединения

1. Задание

Провести такой анализ для других данных.

### *Кластеризация методом k-средних*

Сущность этого метода заключается в нахождении оптимального деления объектов на k классов. Эта процедура перемещает объекты из кластера в кластер с целью минимизации внутрикластерной дисперсии и максимизации межкластерной дисперсии.

1. Файл данных

Пример основан на выборке данных об автомобилях. Открыть файл Cars.sta (через директорию Datasets/Examples). Примем, что исходные данные (автомобили) нужно разделить на 3 кластера.

1. Проведение анализа

Выбрать модуль Cluster Analysis из меню Statistics - Multivariate Exploratory Techniques для отображения стартовой панели Clustering Method. Далее, выбрать опцию K-means clustering и нажать ОК. В появившемся окне Cluster Analysis: K-means Clustering указать позицию Variables для отображения окна диалога по выбору переменных (select variables), в котором необходимо отметить все переменные, после чего, нажав ОК, возвратиться к окну кластерного анализа. Так как мы хотим получить 3 кластера автомобилей, необходимо ввести это значение в окно Number of clusters, а в окне Cluster указать: Cases (rows) (рис. 7).



Рис. 7. Окно диалога кластерного анализа методом k-средних

**Число итераций *(Number of iterations).***Это окно используется для установления максимального числа итераций, которые должны быть выполнены. Метод k-средних представляет собой итеративную процедуру, и на каждом шаге объекты перемещаются в различные кластеры. Алгоритм, выполняющий этот процесс, в данном модуле достаточно эффективен, поэтому установленное по умолчанию число итераций, равное 10, обычно не требуется увеличивать.

**Начальные центры кластеров**. При выборе начальных центров кластеров возможны 3 ситуации, описанные ниже. Отметим, что результаты кластеризации методом k-средних зависят от начального выбора центров:

* **Выбор наблюдений для максимизации начальных межкластерных расстояний.**При выборе этой опции наблюдения или объекты рассматриваются как начальные центры кластеров. Подчеркнем особенности алгоритма в этом случае:
1. Алгоритм будет отбирать первые N строк (число кластеров) как центры кластеров.
2. Последующие строки будут заменять предыдущие центры кластеров, если их наименьшее расстояние до любого центра кластеров оказывается больше, чем наименьшее расстояние между кластерами.
3. В противном случае последующие строки будут заменять начальные центры кластеров, если их наименьшее расстояние от центра кластера больше, чем расстояние от этого центра кластера до любого другого кластерного центра.

Эффект такой процедуры отбора заключается в максимизации начальных расстояний между кластерами.

* **Упорядочение расстояний и отбор наблюдений при постоянных интервалах.**При выборе этой опции расстояния между всеми объектами вначале будут упорядочены, а затем объекты при постоянных интервалах будут выбраны как центры кластеров.
* **Выбор первых N наблюдений (числа кластеров).**При выборе этой опции первые N наблюдений рассматриваются как начальные центры кластеров. Таким образом, эта опция обеспечивает полный контроль над выбором начальной конфигурации.

В данном примере остановимся на второй опции: Упорядочение расстояний и отбор наблюдений при постоянных интервалах.

1. Результаты анализа

При завершении анализа на экране монитора появится окно результатов анализа (K - Means Clustering Results), в котором выделим опцию Advanced (Расширенная) для подробного рассмотрения итогов (рис. 8).



Рис. 8. Окно результатов анализа

#### Идентификация кластеров

Для рассмотрения состава каждого кластера необходимо нажать клавишу Members of each cluster & distances (Состав каждого кластера и расстояния), после чего появятся 3 таблицы, в которых указаны объекты, попавшие в определенный кластер (рис. 9).

Как видно из приведенных таблиц, каждый кластер содержит различное число объектов: в первом кластере находятся 13 автомобилей, во втором – 7, в третьем – 2.



Рис. 9. Состав кластеров

#### Описательные статистики для каждого кластера

Другой способ идентификации заключается в рассмотрении средних значений для каждого кластера. Этого можно достичь различными путями, например: отобразить статистики отдельно для каждого кластера (нажать клавишу Descriptive statistics for each cluster), вывести средние для всех кластеров и расстояния между ними в отдельную таблицу (нажать клавишу Summary: Cluster means & Euclidean distances) или построить графики этих средних (нажать клавишу Graph of means). Обычно графики дают наиболее наглядную информацию. В качестве примера на рис.10 приведены графики средних значений.



Рис. 10. Графики средних значений

## *Пакет Statgraphics. 5.1*

1. Файл данных

Пример основан на выборке данных об автомобилях. Открыть файл Cardata.sf.

1. Выполнение анализа

Из строки Menu последовательно выбрать SPECIAL … MULTIVARIATE METHODS … CLUSTER ANALYSIS для отображения окна кластерного анализа, показанного на рис. 11.



Рис. 11. Окно кластерного анализа

Двойным щелчком мыши выберем количественные переменные, которые автоматически появятся в поле данных этого окна. Для наглядности итоговых результатов в поле Select укажем первые 20 строк исходной матрицы данных, набрав first (20). Щелкнув ОК, переходим к анализу результатов.

1. Методы кластеризации

Наиболее часто используемыми процедурами при кластерном анализе являются иерархические и не иерархические способы. В первом случае образуются древоподобные структуры (дендрограммы), независимо от того, применяются ли агломеративные или дивизимные методы. При иерархических агломеративных процедурах каждое наблюдение вначале рассматривается как отдельный кластер, и последующие шаги формируют из этих одноместных объектов новые агрегированные кластеры, уменьшающие их число на каждом шаге. В противном случае кластерные процедуры определяют дивизимные методы (вначале все объекты рассматриваются как один кластер) или метод k -средних.

В данном пакете реализованы алгоритмы расчета 6 методов кластеризации, различающиеся способом объединения объектов. Кратко опишем каждый их них.

**Ближний сосед.**В этом методе определяются 2 объекта, находящиеся на наименьшем расстоянии, и помещаются в один кластер. Затем определяется следующее минимальное расстояние и к уже сформированному кластеру присоединяется третий объект или образуется новый кластер из двух объектов. Процедура продолжается до тех пор, пока все объекты не сформируют один кластер. Этот метод также носит название одиночной связи.

**Дальний сосед.**Здесь используется максимальное расстояние между двумя объектами. Все объекты в кластере соединяются один с другим при некотором максимальном расстоянии (минимальном сходстве). Этот метод известен под названием полной связи.

**Центроидный метод.**В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести. Каждый раз при группировке объектов образуются новые центроиды. Кластерные центроиды перемещаются при добавлении нового объекта (или группы объектов) к существующему кластеру.

**Медианный метод.**Этот метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т. е. числами объектов в них).

**Метод группового среднего.**В этом методе расстояние между кластерами вычисляется как среднее расстояние между всеми объектами в одном кластере и средним расстоянием между всеми объектами в другом кластере.

**Метод Уорда (Ward).**Метод минимизирует сумму квадратов для любых двух кластеров, которые могут быть сформированы на каждом шаге. Этот метод формирует кластеры, которые имеют малое число объектов, и стремится создать кластеры, содержащие примерно одинаковое число объектов.

1. Выбор параметров кластерного анализа

Нажав на правую кнопку мыши, получим окно диалога для выбора параметров анализа (рис. 12).



Рис. 12. Окно диалога для выбора параметров анализа

В рассматриваемом случае при 20 объектах рекомендуется выбрать метод Уорда, для чего в поле методов раскрывшегося окна укажем этот метод. Кроме того, установим число кластеров, равное 3, на которое желательно разбить исследуемую совокупность. Для этого необходимо ввести это число в поле в нижнем левом углу окна диалога. Нажав ОК, приходим к итоговым результатам.

1. Анализ результатов

Среди табличных опций, окно которых открывается второй кнопкой слева в строке меню, выделим следующие три позиции (см. рис. 13):



Рис. 13. Окно табличных опций

Эти таблицы показывают, соответственно, итоговый анализ, состав кластеров и расписание (последовательность) объединения.

Среди графических опций (третья кнопка слева) выделим такие позиции (см. рис. 14):



Рис. 14. Окно графических опций

Выделенные опции показывают дендрограмму анализа и график объединения. Наиболее интересной оказывается дендрограмма, приведенная на рис. 15.



Рис. 15. Дендрограмма объединения объектов