МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Дальневосточный федеральный университет»

ШКОЛА ЕСТЕСТВЕННЫХ НАУК

кафедра информатики, математического и компьютерного моделирования

РЕФЕРАТ

по дисциплине «Непрерывные математические модели»
Направление 01.04.02 «Прикладная математика и информатика»

на тему **«Кластерный анализ»**

Выполнила студентка группы М8103
Шонова Д.Д.

г. Владивосток
2016

Содержание

Введение

В наше время кластерный анализ является основой для многих, как научных, так и практических исследований в самых разных областях. Он пригоден для использования во всех случаях, когда надо разделить большой массив информации на группы (кластеры), пригодные для дальнейшей обработки. Существует много программ для проведения кластерного анализа, одной из которых является программный комплекс статистической обработки данных STATISTICA.

В данном реферате рассматриваются различные методы кластерного анализа, алгоритмы кластеризации, а также используемые в анализе меры расстояний; и приводятся два примера кластеризации практически значимых данных.

Методы кластерного анализа

Кластерный анализ – это способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп как “сгустков” этих точек. “Кластер” в переводе с английского означает “сгусток”.

Методы кластерного анализа относятся к так называемым многомерным методам: перед исследователем находится поле из множества объектов, каждый из которых описывается множеством переменных. Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы объектов.

Кластерный анализ делится на несколько этапов:

1) Выбор переменных, на основе которых будет производиться кластеризация;
2) Выбор меры расстояния между объектами;
3) Преобразование переменных таким образом, чтобы они имели близкую значимость;
4) Выбор метода кластеризации;
5) Задание количества кластеров;
6) Интерпретация полученных результатов;
7) Оценка эффективности кластерного анализа.

Большинство методов кластерного анализа являются объединительными. Их применение начинается с создания элементарных кластеров, состоящих только из одной точки. На каждом последующем шаге происходит объединение двух наиболее близких кластеров в один. Момент остановки этого процесса задается пользователем. Графически данных процесс может быть представлен в виде дендрограммы — дерева объединения кластеров.

Также существуют методы кластерного анализа, называемые дивизионными. Они пытаются разбивать объекты на кластеры непосредственно и в большинстве случаев требуют задать число кластеров.

Виды мер расстояния

Для определения близости элементов вводят так называемую меру расстояния. Для её измерения в кластерном анализе используются следующие способы:

1) Евклидово расстояние — наиболее общий тип расстояния, вычисляемый по формуле: $\rho(x,y)=\left[\sum_i (x_i-y_i)^2\right]^{\frac{1}{2}}$;

2) Квадрат евклидова расстояния — используется для придания веса более отдалённым друг от друга объектам. Вычисляется следующим образом: $\rho(x,y)=\sum_i (x_i-y_i)^2$;

3) Манхэттенское расстояние (или расстояние городских кварталов) — среднее разностей по координатным осям. В большинстве случаев приводит к таким же результатам, как и евклидова мера расстояния. Вычисляется по формуле: $\rho(x,y)=\sum_i |x_i-y_i|$;

4) Расстояние Чебышёва — полезно, когда нужно определить различными два объекта, отличающиеся только одним измерением. Формула: $\rho(x,y)=\max_i |(x_i-y_i)|$;

5) Степенное расстояние — используется, когда нужно прогрессивно уменьшить или увеличить вес, относящийся к размерности, для которой соответствующие объекты сильно различаются. Вычисляется следующим образом: $\rho(x,y)=\left(\sum_i |(x_i-y_i)|^p\right)^{\frac{1}{r}}$, где p и r — параметры, определяемые пользователем;

6) Процент несогласия — используется в случаях, когда данные являются категориальными и вычисляется как: $\rho(x,y)=\left(num(x_i\neq y_i)\right)/I$.

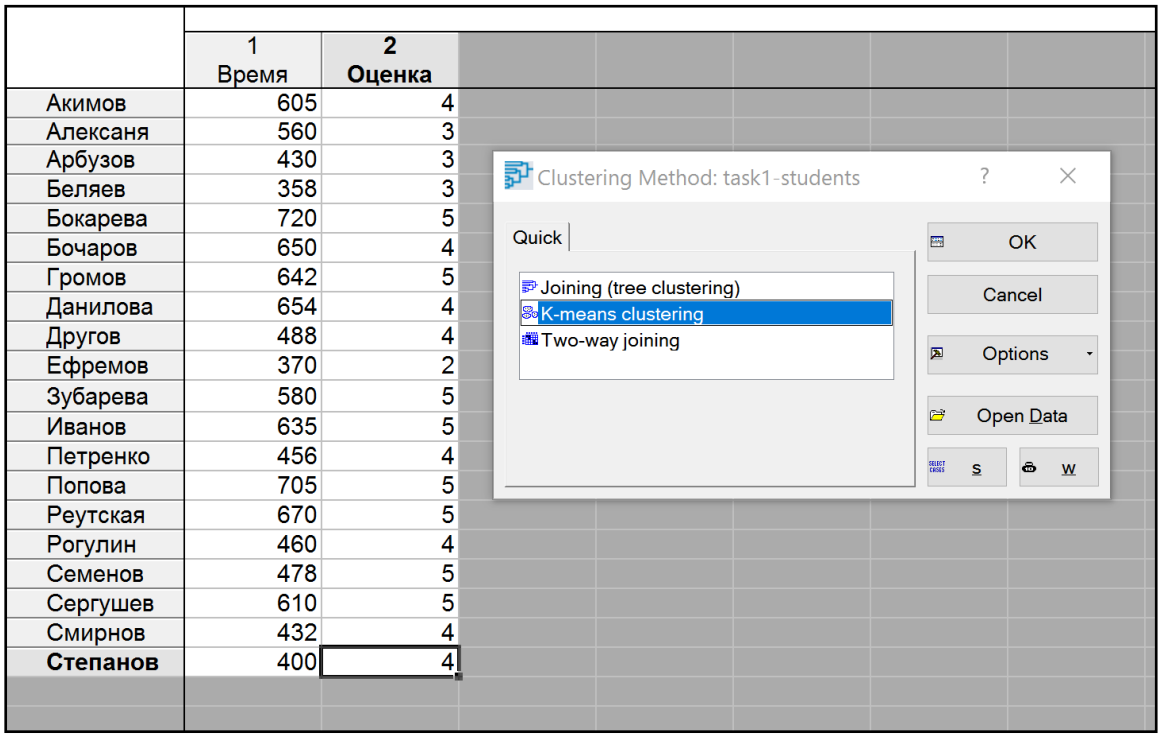
Алгоритмы объединения в кластеры

Существуют следующие методы объединения объектов в кластеры:

1. Метод ближайшего соседа (одиночная связь) — расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами в различных кластерах;
2. Метод наиболее удалённого соседа (полная связь) — расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах;
3. Невзвешенное попарное соединение — расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них;
4. Взвешенное попарное соединение — идентичен предыдущему за исключением того, что при вычислениях размер соответствующих кластеров используется в качестве весового коэффициента;
5. Невзвешенный центроидный метод — расстояние между двумя кластерами определяется как расстояние между их центрами тяжести;
6. Взвешенный центроидный метод — идентичен предыдущему, за исключением того, что при вычислении используются веса для учёта разницы между размерами кластеров;
7. Метод Варда — для оценки расстояния используются методы дисперсионного анализа.

Пример использования кластерного анализа в системе STATISTICA: образование

Рассмотрим пример с анализом зависимости оценки, полученной студентом во время экзамена и объёмом его самостоятельной работы. В таблицу данных программы STATISTICA 13.2 введём информацию о работе и оценках некоторого набора студентов, и используем модуль кластерного анализа. Начнём с метода k-средних, относящегося к дивизионным методам.



	1 Время	2 **Оценка**
Акимов	605	4
Алексаня	560	3
Арбузов	430	3
Беляев	358	3
Бокарева	720	5
Бочаров	650	4
Громов	642	5
Данилова	654	4
Другов	488	4
Ефремов	370	2
Зубарева	580	5
Иванов	635	5
Петренко	456	4
Попова	705	5
Реутская	670	5
Рогулин	460	4
Семенов	478	5
Сергушев	610	5
Смирнов	432	4
Степанов	400	4

Рисунок 1. Таблица данных и модуль кластерного анализа

Попробуем несколько вариантов разбивки на кластеры. Первоначальное количество кластеров установим равное двум.

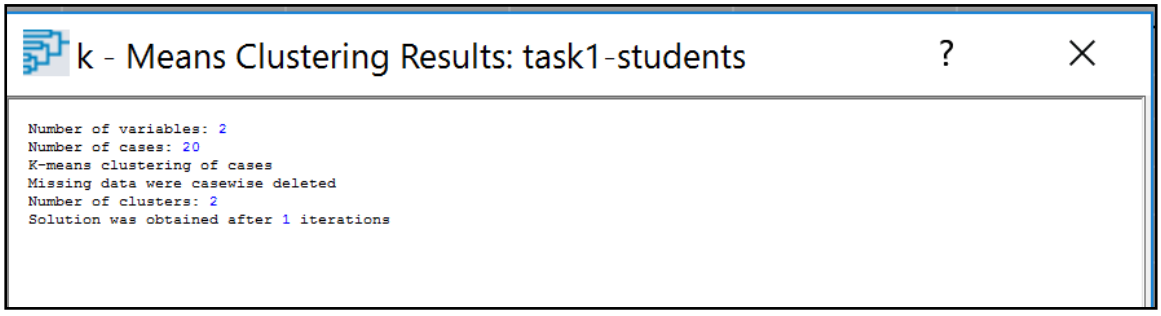


Рисунок 2. Результаты первого варианта анализа

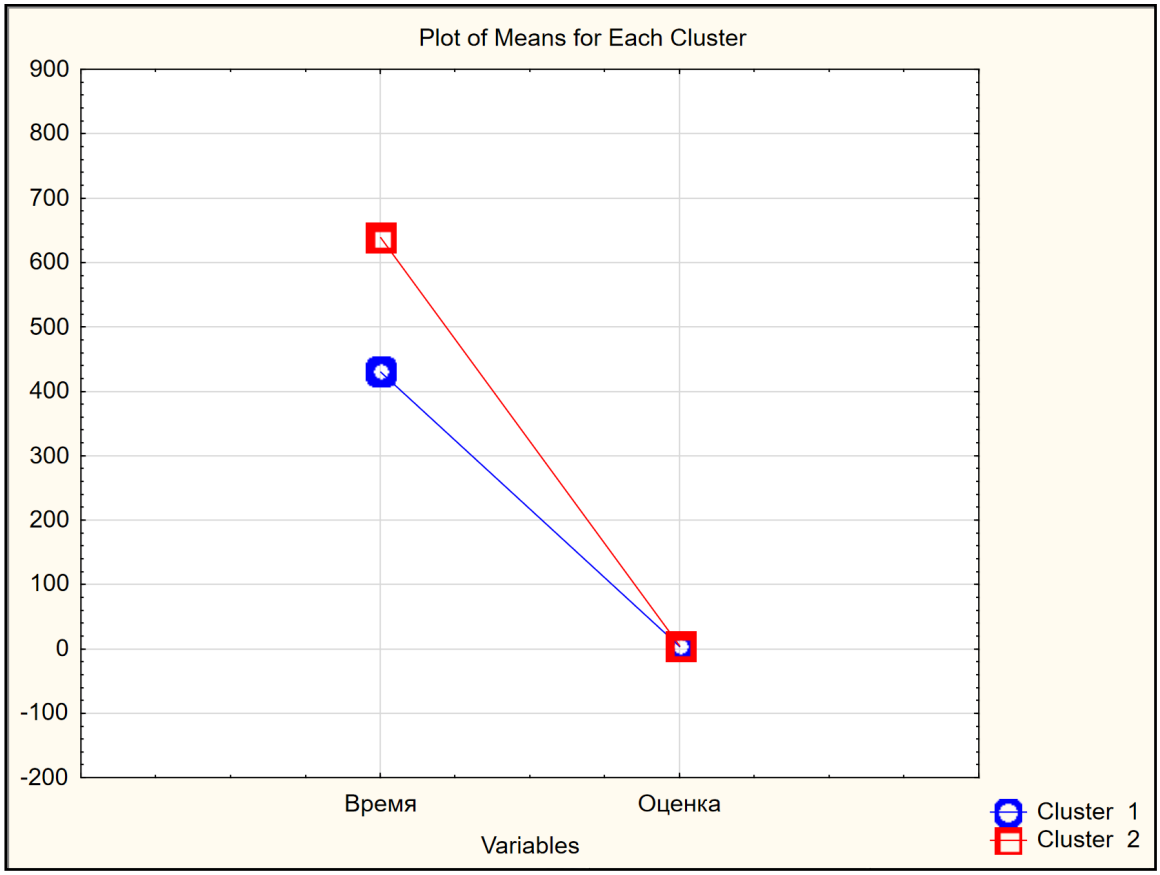


Рисунок 3. Графики средних значений переменных для двух групп кластеров

Рассмотрев графики средних на рис. 3, можно сказать, что кластеры заметно отличаются по переменной “время”, но очень мало — по переменной “оценка”. Это происходит потому что в первом случае используются трехзначные числа, и соответственно, разброс значений считается бóльшим. Чтобы сделать обе переменные равноправными в образовании кластеров, используется стандартизация данных. Её суть состоит в преобразовании каждой переменной в число, колеблющееся около нуля: отрицательные значения говорят, что данное наблюдение меньше среднего в выборке, а положительные — что больше.

Теперь при изучении графиков средних видно, что студентов можно разбить на две группы (рис. 5):

1) низкая посещаемость и низкая оценка на экзамене;
2) высокая посещаемость и соответственно высокий результат.

Изображения 6-7 отображают более подробные данные о новообразованных кластерах: дескриптивные статистики и евклидовы расстояния.

	1 Время	2 **Оценка**
Акимов	0,514568875	-0,171410161
Алексаня	0,127674984	-1,31414457
Арбузов	-0,990018479	-1,31414457
Беляев	-1,60904871	-1,31414457
Бокарева	1,50329771	0,971324246
Бочаров	0,901462767	-0,171410161
Громов	0,83268163	0,971324246
Данилова	0,935853335	-0,171410161
Другов	-0,491355242	-0,171410161
Ефремов	-1,505877	-2,45687897
Зубарева	0,299627825	0,971324246
Иванов	0,772498136	0,971324246
Петренко	-0,766479787	-0,171410161
Попова	1,37433308	0,971324246
Реутская	1,07341561	0,971324246
Рогулин	-0,732089219	-0,171410161
Семенов	-0,577331662	0,971324246
Сергушев	0,557557085	0,971324246
Смирнов	-0,972823195	-0,171410161
Степанов	-1,24794774	-0,171410161

Рисунок 4. Таблица после стандартизации

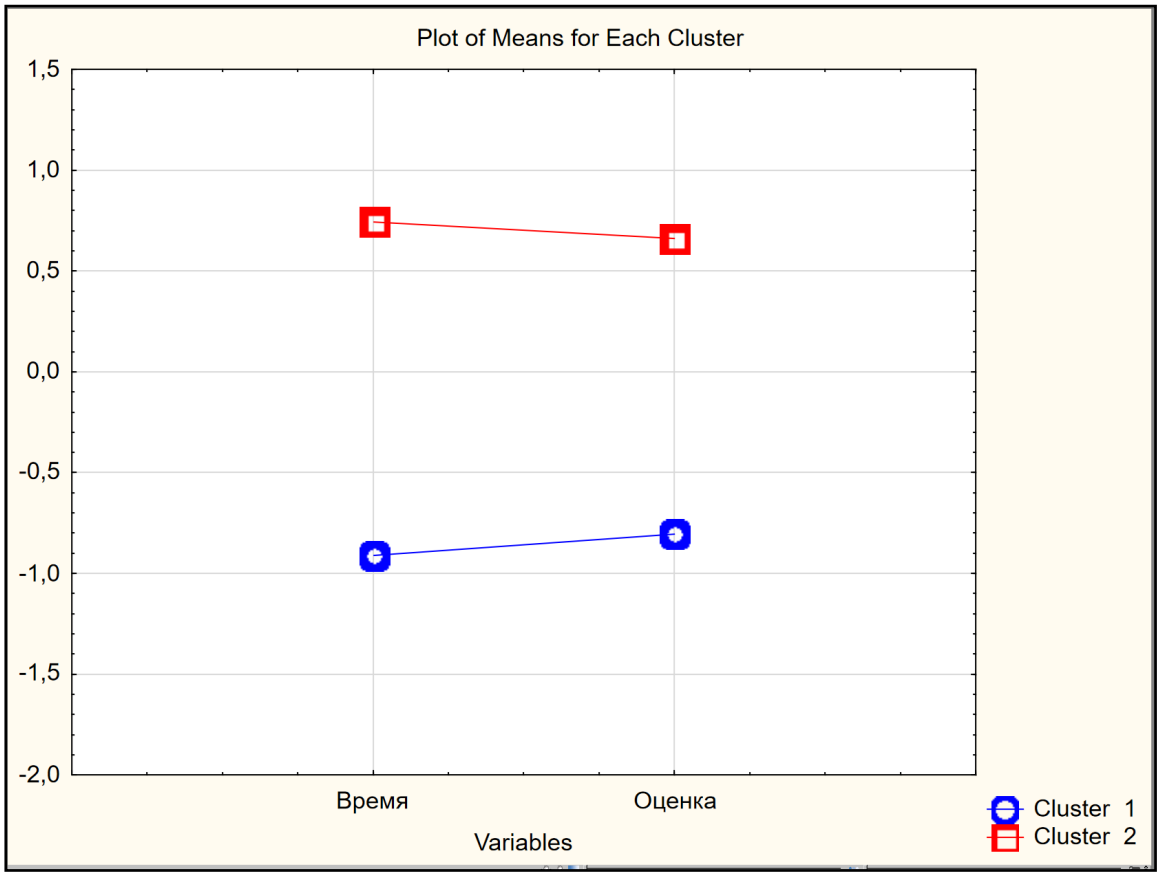


Рисунок 5. Графики средних для стандартизированных переменных

Cluster Number	Euclidean Distances between Clusters (task1-students) Distances below diagonal Squared distances above diagonal	
	No. 1	**No. 2**
No. 1	0,000000	2,442558
No. 2	1,562869	0,000000

Рисунок 6. Евклидовы расстояния между центрами кластеров

Variable	Descriptive Statistics for Cluster 2 (task1-students) Cluster contains 11 cases		
	Mean	Standard Deviation	Variance
Время	0,744360	0,564219	0,318343
Оценка	0,659669	0,533771	0,284911

Рисунок 7. Дескриптивные статистики кластеров

Числа на рисунках 8 и 9 показывают расстояние от каждого объекта (студента) до центра его кластера. Чем меньше расстояние, тем типичнее объект для данного кластера.

	Members of Cluster Number 1 (task1-students) and Distances from Respective Cluster Center Cluster contains 9 cases
	Distance
Алексаня	0,816775
Арбузов	0,363582
Беляев	0,611118
Другов	0,537639
Ефремов	1,240942
Петренко	0,460202
Рогулин	0,466160
Смирнов	0,451117
Степанов	0,508625

Рисунок 8. Объекты, принадлежащие первому кластеру

	Members of Cluster Number 2 (task1-students) and Distances from Respective Cluster Center Cluster contains 11 cases		
	Distance		
Акимов	0,609712		
Бокарева	0,580136		
Бочаров	0,598070		
Громов	0,229052		
Данилова	0,603060		
Зубарева	0,384003		
Иванов	0,221270		
Попова	0,496988		
Реутская	0,320473		
Семенов	0,960208		
Сергушев	0,256928		

Рисунок 9. Объекты, принадлежащие второму кластеру

На изображениях 10 и 11 отображены графики средних для разбиения на три и четыре кластера соответственно. В первом случае появляется третья группа студентов, которые при небольшом количестве часов самостоятельной работы всё же получили достаточно хорошие оценки на экзамене.

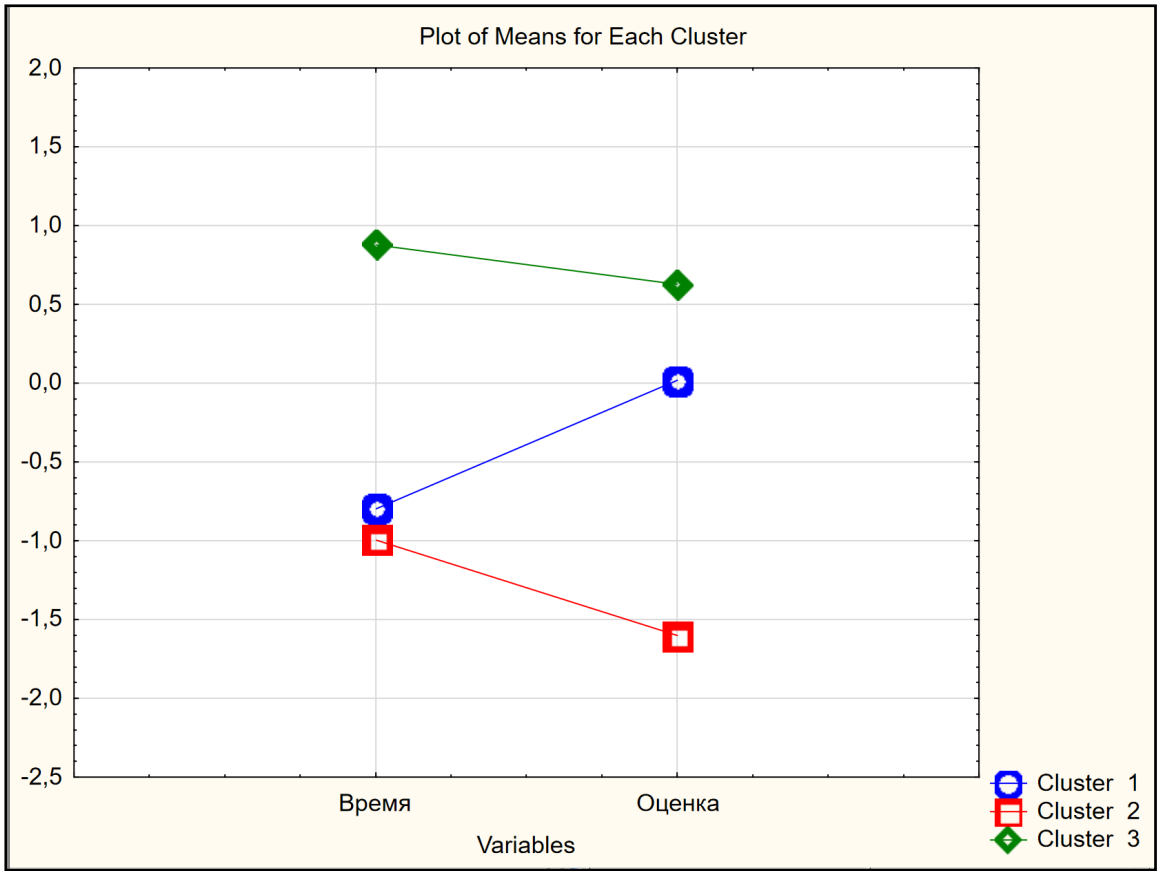


Рисунок 10. Графики средних для переменных при разбивке на три кластера

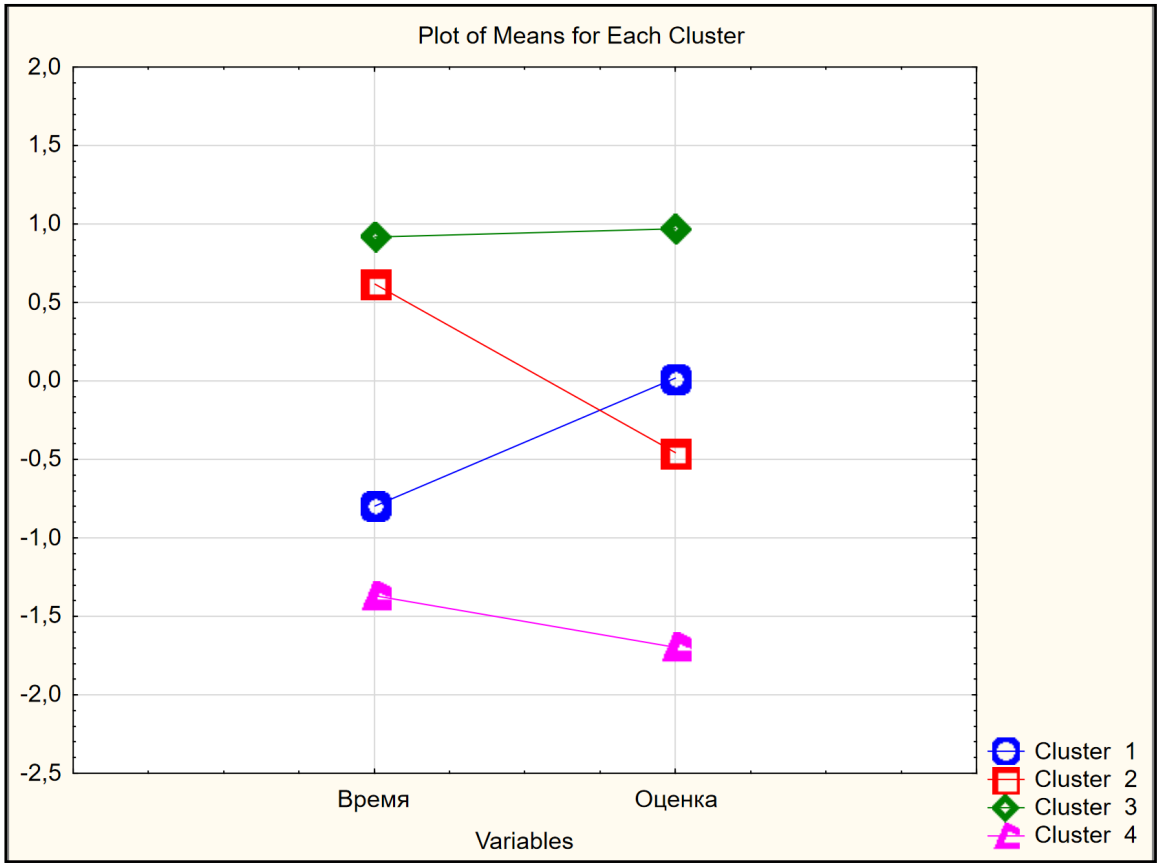


Рисунок 11. Графики средних для переменных при разбивке на четыре кластера

Во втором же случае появляется ещё одна любопытная группа: студенты, посещающие занятия, но всё же слабо проявившие себя на экзамене. Можно сказать, что вариант разбиения на четыре кластера наиболее полно характеризует реальную жизненную ситуацию.

Рассмотрим на этом же примере и объединительный метод анализа: построение дерева классификации. Используем последовательно алгоритмы одиночной и полной связи, а также евклидову и манхэттенскую метрики.

Анализ иерархических деревьев на рисунках с 12 по 15 подтверждает факт схожести результатов вычисления евклидовой и манхэттенской метрик, а также показывает, что полная связь даёт более плотно сформированные группы объектов для данной выборки.

Вообще, на дереве, построенном с помощью алгоритма полной связи, можно заранее примерно определить достаточное количество кластеров для качественного анализа, чем мы и воспользуемся в следующем примере.

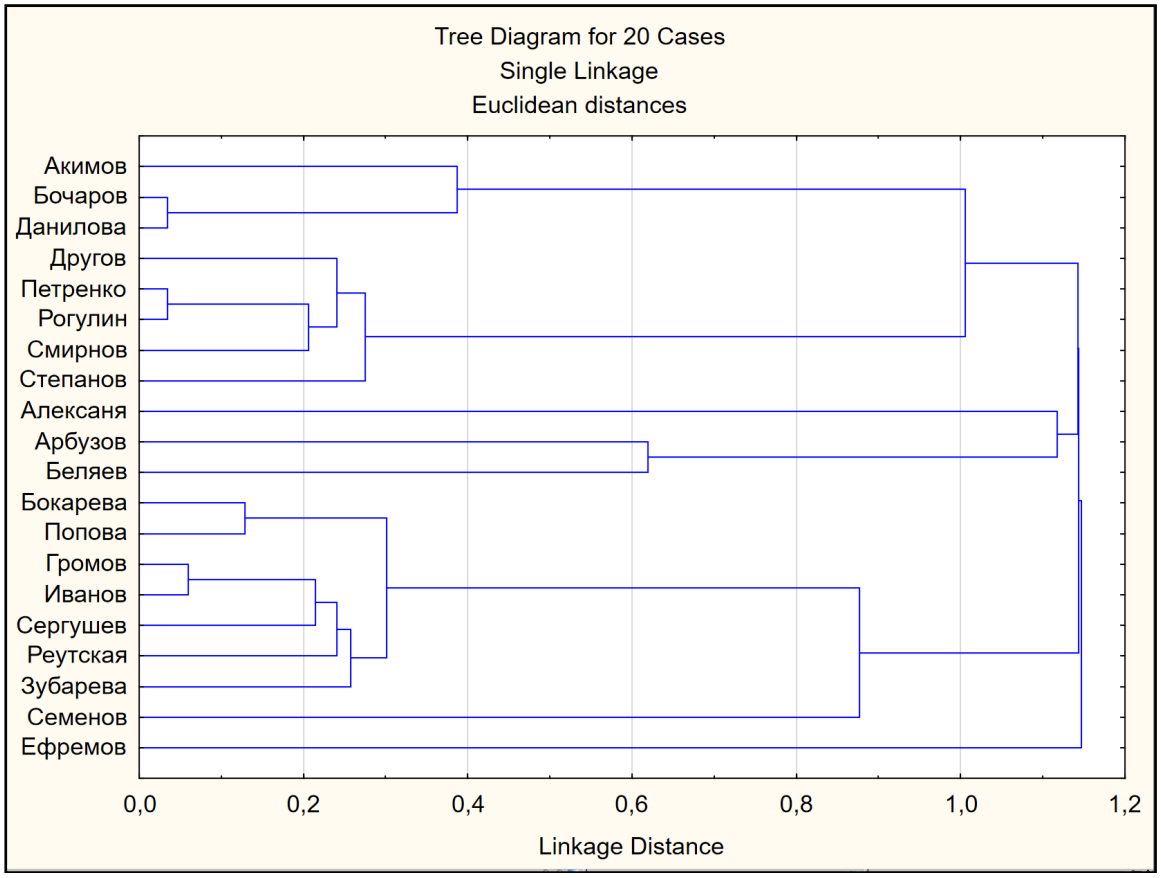


Рисунок 12. Дендрограмма: метод одиночной связи и евклидова метрика

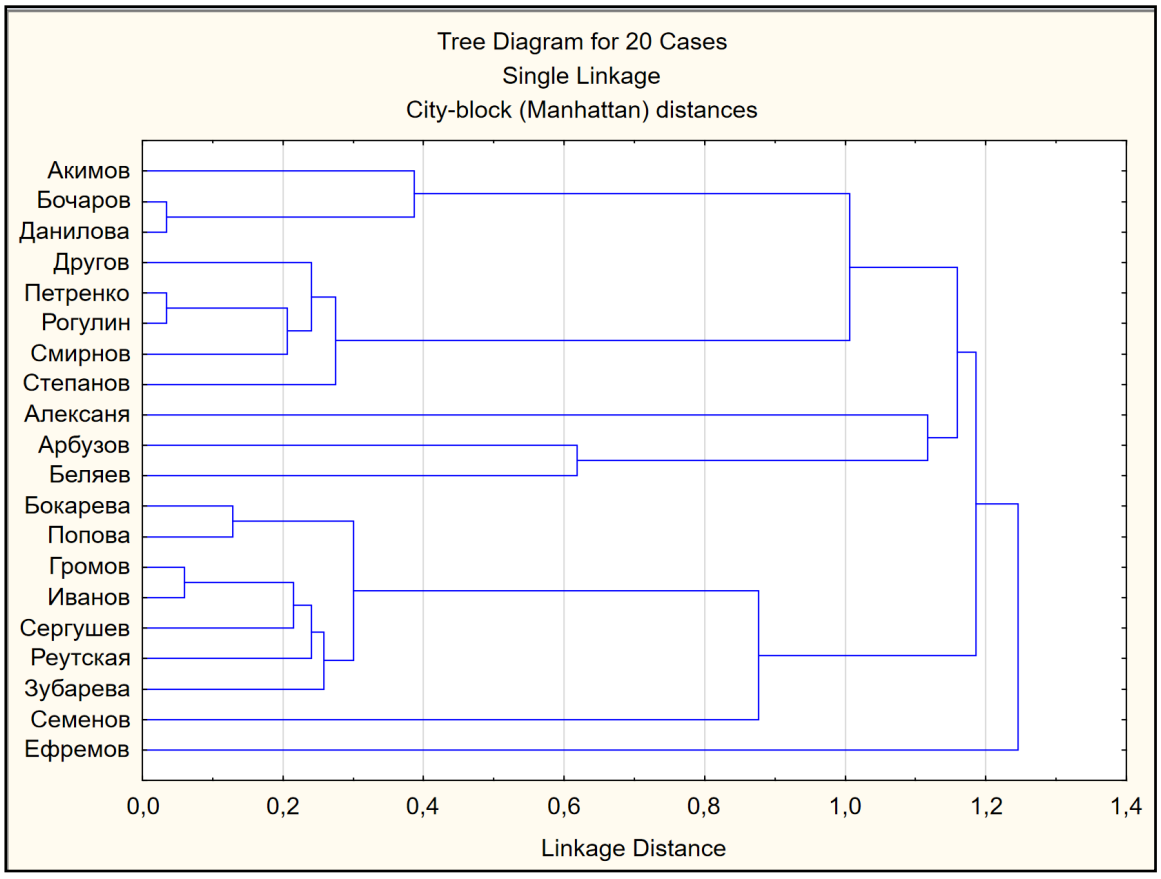


Рисунок 13. Дендрограмма: метод одиночной связи и манхэттенская метрика

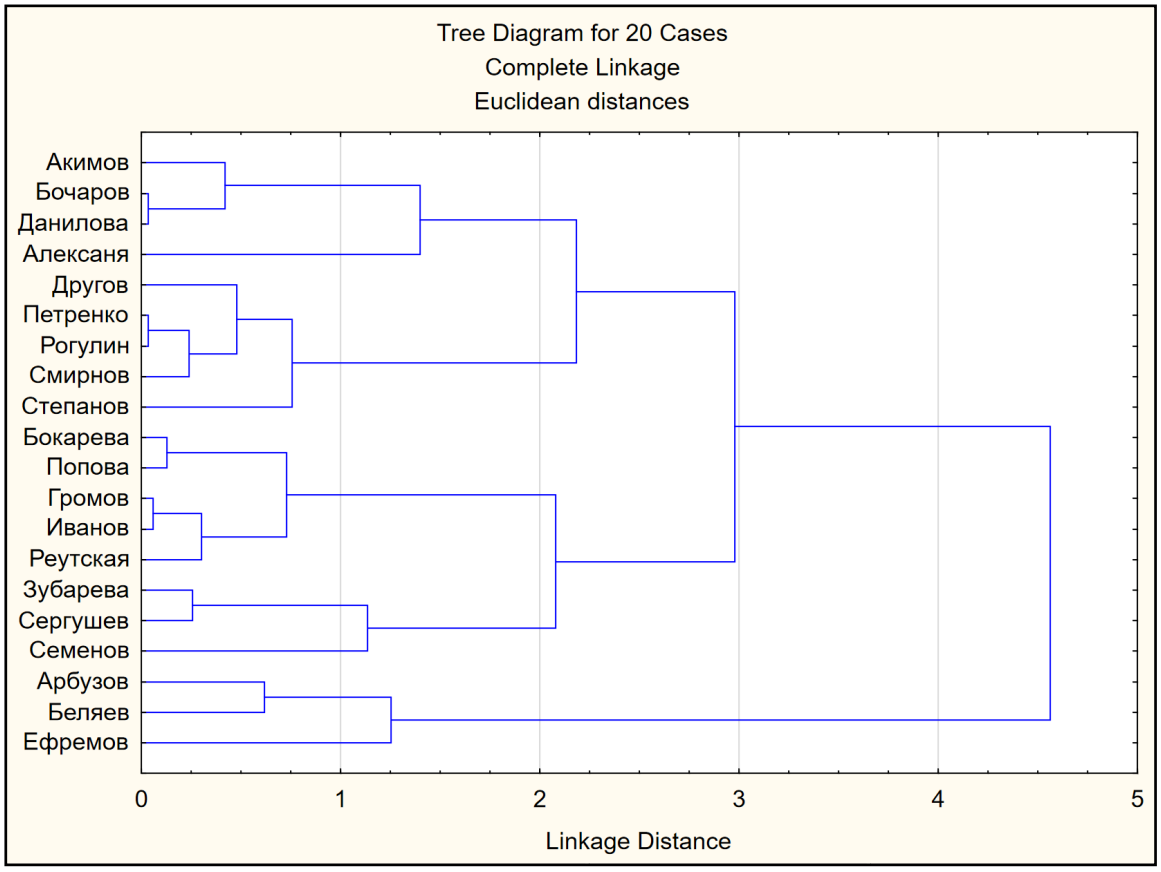


Рисунок 14. Дендрограмма: метод полной связи и евклидова метрика

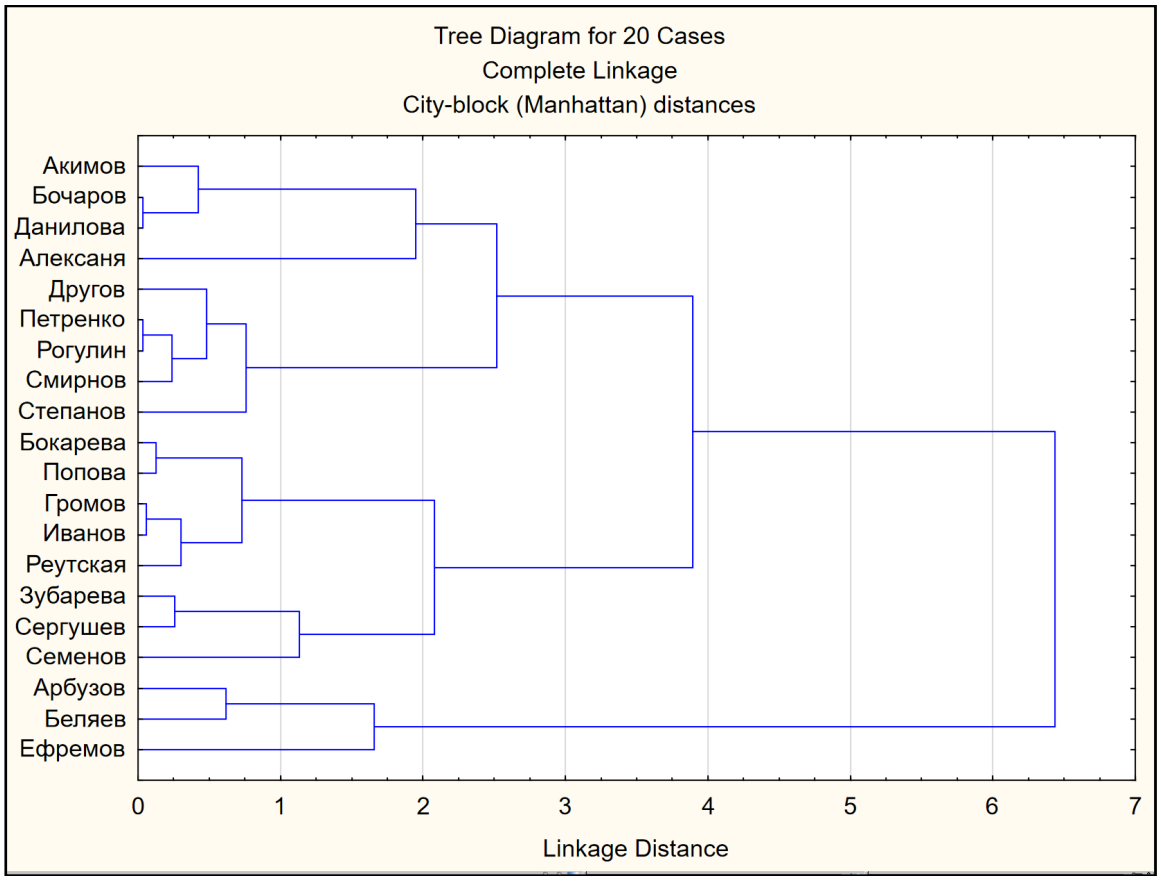


Рисунок 15. Дендрограмма: метод полной связи и манхэттенская метрика

Пример использования кластерного анализа в системе STATISTICA: автострахование

Целью данного анализа является разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной группе риска. Наблюдения, попавшие в одну группу, характеризуются одинаковой вероятностью наступления страхового случая, которая впоследствии оценивается страховщиком. Для решения данной задачи наиболее эффективно использование кластерного анализа.

	1 Цена атомобиля	2 Возраст владельца	3 Стаж вождения	**4** **Возраст автомобиля**
Audi	0,866	25	3	1
BMW	0,493	26	8	2
Corvette	1,254	53	31	15
Chrysler	0,691	44	15	7
Dodge	0,751	20	1	12
Eagle	0,652	29	5	3
Ford	0,703	35	17	5
Honda	0,425	60	37	10
Isuzu	0,729	42	22	9
Mazda	0,126	39	15	7
Mersedes	1,105	51	32	5
Mitsubisi	0,623	28	7	6
Nissan	6478	21	1	4
Pontiac	0,614	33	11	1
Porsche	3,456	46	27	1
Toyota	0,059	41	19	8

Рисунок 16. Данные об автомобилях и их владельцах

Поскольку здесь, как и в предыдущем примере, используются измерения различных типов, данные необходимо стандартизовать.

	1 Цена атомобиля	2 Возраст владельца	3 Стаж вождения	**4** **Возраст автомобиля**
Audi	-0,249981734	-1,01203982	-1,09892428	-1,21506294
BMW	-0,250212082	-0,928140144	-0,665850671	-0,972050354
Corvette	-0,249742123	1,33715105	1,32628792	2,1871133
Chrysler	-0,250089806	0,582053989	-0,059547621	0,243012588
Dodge	-0,250052753	-1,43153819	-1,27215372	1,45807553
Eagle	-0,250113891	-0,676441122	-0,925694835	-0,729037765
Ford	-0,250082395	-0,173043078	0,113681822	-0,243012588
Honda	-0,250254075	1,92444877	1,84597625	0,972050354
Isuzu	-0,250066339	0,41425464	0,546755429	0,729037765
Mazda	-0,250438724	0,162555618	-0,059547621	0,243012588
Mersedes	-0,249834138	1,16935171	1,41290264	-0,243012588
Mitsubisi	-0,2501318	-0,760340796	-0,752465392	0
Nissan	3,74999959	-1,34763851	-1,27215372	-0,486025177
Pontiac	-0,250137358	-0,340842426	-0,406006507	-1,21506294
Porsche	-0,248382269	0,749853337	0,979829036	-1,21506294
Toyota	-0,2504801	0,330354966	0,286911265	0,486025177

Рисунок 17. Переменные после процедуры стандартизации

Для начала выясним, формируют ли автомобили естественные кластеры, для этого построим вертикальное иерархическое дерево (рис. 18).

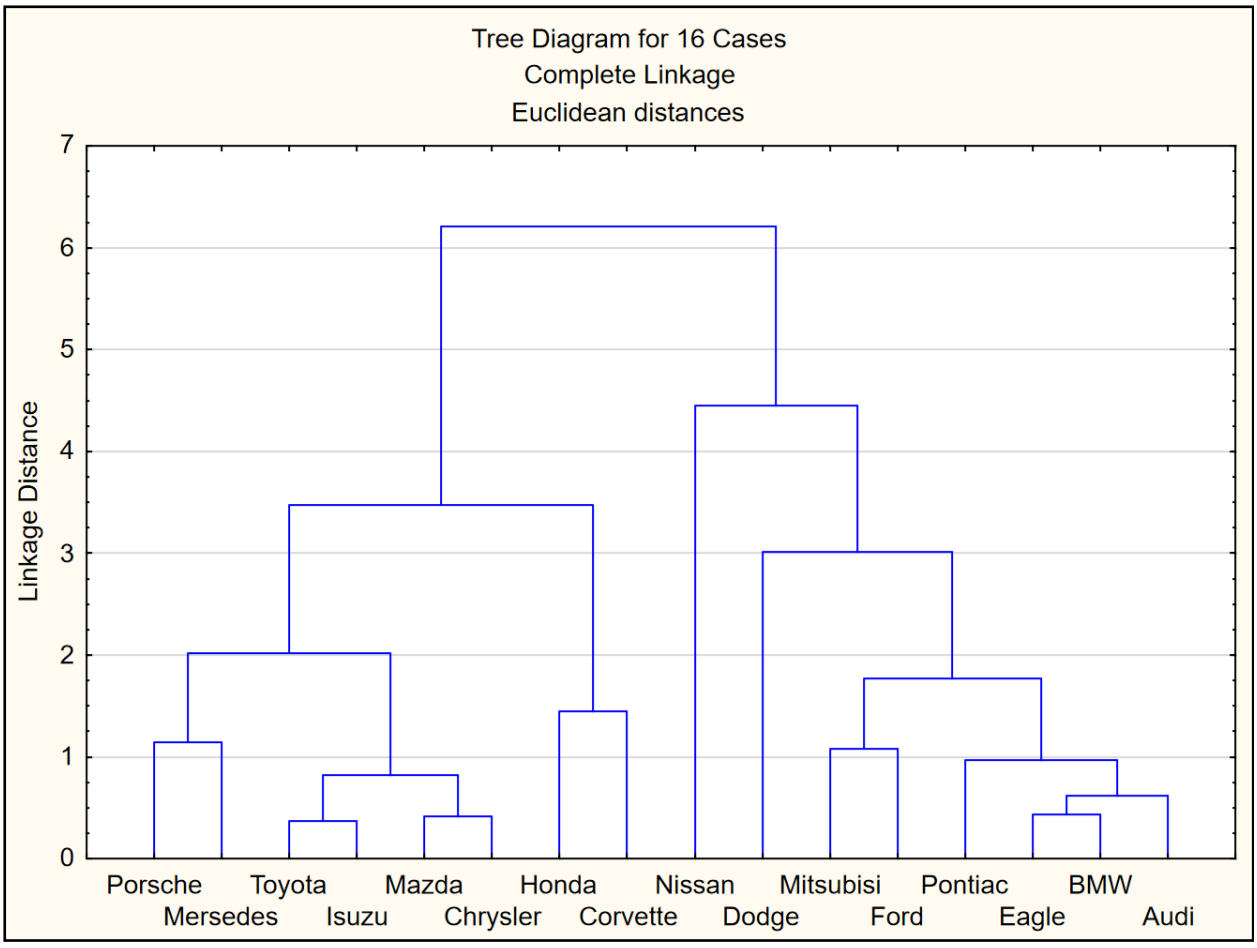


Рисунок 18. Дендрограмма объединения автомобилей в естественные кластеры

Исходя из визуального представления результатов на рис. 18, можно попробовать разбить автомобили на три, пять или семь кластеров методом k-средних и проверить значимость различия между полученными группами.

После классификации рассчитаем с помощью дисперсионного анализа среднее значение показателей по каждому кластеру, чтобы оценить, насколько они различаются между собой. Для значимого различия необходимо, чтобы значение p было меньше чем 0.05.

	Analysis of Variance (task2-automobiles)					
Variable	**Between SS**	df	Within SS	df	F	signif. p
Цена атомобиля	1,28499	2	13,71501	13	0,60900	0,558703
Возраст владельца	12,84451	2	2,15549	13	38,73334	0,000003
Стаж вождения	12,35283	2	2,64717	13	30,33175	0,000013
Возраст автомобиля	6,41592	2	8,58408	13	4,85823	0,026572

Рисунок 19. Дисперсионный анализ для разбиения на три кластера

Variable	Analysis of Variance (task2-automobiles)					
	Between SS	df	Within SS	df	F	signif. p
Цена атомобиля	15,00000	4	0,000000	11		
Возраст владельца	13,74257	4	1,257428	11	30,05506	0,000007
Стаж вождения	14,02072	4	0,979276	11	39,37297	0,000002
Возраст автомобиля	7,90551	4	7,094489	11	3,06437	0,063457

Рисунок 20. Дисперсионный анализ для разбиения на пять кластеров

Variable	Analysis of Variance (task2-automobiles)					
	Between SS	df	Within SS	df	F	signif. p
Цена атомобиля	15,00000	6	0,000000	9		
Возраст владельца	14,20282	6	0,797185	9	26,72431	0,000030
Стаж вождения	13,85968	6	1,140322	9	18,23127	0,000145
Возраст автомобиля	13,43504	6	1,564960	9	12,87736	0,000572

Рисунок 21. Дисперсионный анализ для разбиения на семь кластеров

Как видно, значение p недостаточно мало в случае разделения объектов на три и пять кластеров. Таким образом, можно посчитать, что разбиение на семь кластеров даст наилучшую оценку.

Наконец, выведем отдельно объекты, входящие в каждый из кластеров, а также их расстояния от центров кластеров (рис. 22-28).

	Members of Cluster Number 1 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Ford	0,268583
Mitsubisi	0,268583

Рисунок 22. Объекты и евклидовы расстояния первого кластера

	Members of Cluster Number 2 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 4 cases
	Distance
Chrysler	0,182995
Isuzu	0,239552
Mazda	0,182995
Toyota	0,065523

Рисунок 23. Объекты и евклидовы расстояния второго кластера

	Members of Cluster Number 3 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Corvette	0,361540
Honda	0,361540

Рисунок 24. Объекты и евклидовы расстояния третьего кластера

	Members of Cluster Number 4 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 4 cases
	Distance
Audi	0,230797
BMW	0,112970
Eagle	0,172633
Pontiac	0,286158

Рисунок 25. Объекты и евклидовы расстояния четвёртого кластера

	Members of Cluster Number 5 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 1 cases
	Distance
Nissan	0,00

Рисунок 26. Объекты и евклидовы расстояния пятого кластера

	Members of Cluster Number 6 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Mersedes	0,285965
Porsche	0,285965

Рисунок 27. Объекты и евклидовы расстояния шестого кластера

	Members of Cluster Number 7 (task2-automobiles) and Distances from Respective Cluster Center Cluster contains 1 cases
	Distance
Dodge	0,00

Рисунок 28. Объекты и евклидовы расстояния седьмого кластера

Заключение

Методы кластерного анализа, совместно со статистическими программными комплексами, позволяют быстро и эффективно разбить массив данных на группы, обладающие общими признаками и пригодные для дальнейшего исследования. Более того, как показывает практика, комбинирование разных методов анализа: как объединительных, так и дивизионных, позволяет добиться более точной кластеризации объектов при меньших трудозатратах.

При проведении кластерного анализа исследователю необходимо выбрать метод анализа, провести сегментацию рассматриваемых объектов и проверить результаты решения на статистическую адекватность. Что в совокупности даёт значимый практический результат и делает кластерный анализ незаменимым во многих сферах деятельности, где необходимы статистические исследования.

Список литературы

1. Кластерный анализ [Электронный ресурс] // Википедия. — Режим доступа: https://ru.wikipedia.org/wiki/Кластерный_анализ, свободный.
2. Кузнецов Д. Ю., Трошина Т. Л. Кластерный анализ и его применение // Ярославский педагогический вестник. — №4. — 2006. — Режим доступа: http://vestnik.yspu.org/releases/uchenue_praktikam/33_4/, свободный.
3. Пример использования кластерного анализа STATISTICA в автостраховании [Электронный ресурс] // StatSoft. — Режим доступа: http://statsoft.ru/solutions/ExamplesBase/branches/detail.php?ELEMENT_ID=1573, свободный.