XX. Кластерный анализ

В результате кластерного анализа при помощи предварительно заданных переменных формируются группы наблюдений. Под наблюдениями здесь понимаются отдельные личности (респонденты) или любые другие объекты. Члены одной группы (одного кластера) должны обладать схожими проявлениями переменных, а члены разных групп различными.

Наряду с кластеризацией наблюдений в SPSS предусмотрена кластеризация переменных. Здесь на основе заданных наблюдений образовываются группы переменных. Так как в принципе то же самое делает и факторный анализ (см. гл. 19), то в этой главе мы ограничимся рассмотрением только кластеризации наблюдений.

20.1. Принцип кластерного анализа

Для рассмотрения принципа кластерного анализа выберем сначала очень простой пример.

- Откройте файл bier.sav, который содержит некоторые данные о 17 сортах пива (см. рис. 20.1).

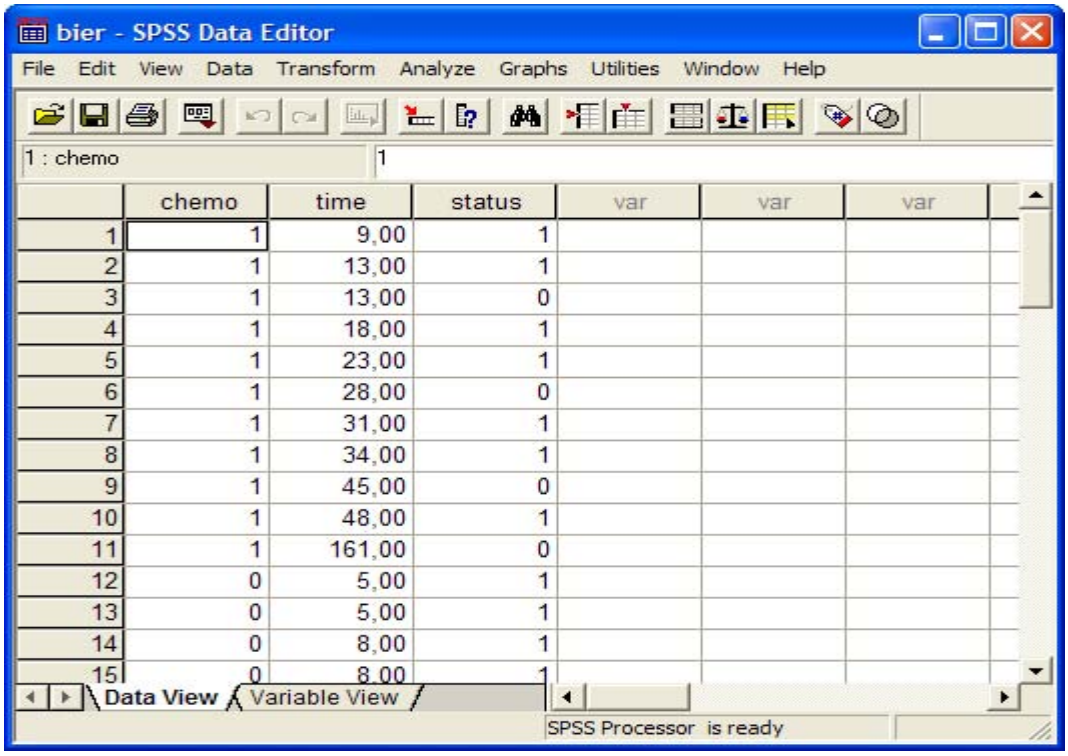

Рис. 20.1: *Данные файла bier.sav в редакторе данных*

Переменная herkunft (производитель) указывает на страну-производителя пива, где США закодированы с помощью единицы. Расходы (kosten) приведены в долларах США для ёмкости равной 12 унциям для жидкости (примерно одна треть литра); калорийность указана для одинакового количества пива. Содержание алкоголя приводится в процентах.

Возьмём переменные kalorien (калории) и kosten (расходы) и представим их при помощи простой диаграммы рассеяния.

- Выберите в меню Graphs (Графики) Scalier... (Диаграмма рассеяния)
- Переменную kalorien (калории) поместите в поле оси x, а переменную kosten (расходы) в поле оси y, и для обозначения наблюдения используйте переменную bier (пиво).
- Через кнопку Options... (Опции) активируйте опцию Display Chart with case labels (Показывать график с метками наблюдений).

Вы получите диаграмму рассеяния, представленную на рисунке 20.2.

Вы увидите четыре отдельных отчётливых группировки точек, три из них в нижней половине диаграммы и одну в верхнем правом углу. Следовательно, переменные kalorien (калории) и kosten (расходы), явно распадаются на четыре различных кластера по сортам пива.

Сорта пива, которые по значениям двух рассмотренных переменных похожи друг на друга, принадлежат к одному кластеру; сорта пива, находящиеся в различных кластерах, не похожи друг на друга. Решающим критерием для определения схожести и различия двух сортов пива является расстояние между точками на диаграмме рассеяния, соответствующими этим сортам.

Самой распространенной мерой для определения расстояния между двумя точками на плоскости, образованной координатными осями x и y, является евклидова мера:

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

где x_1:, и x_n — координаты первой точки, y: и уг — координаты второй точки.

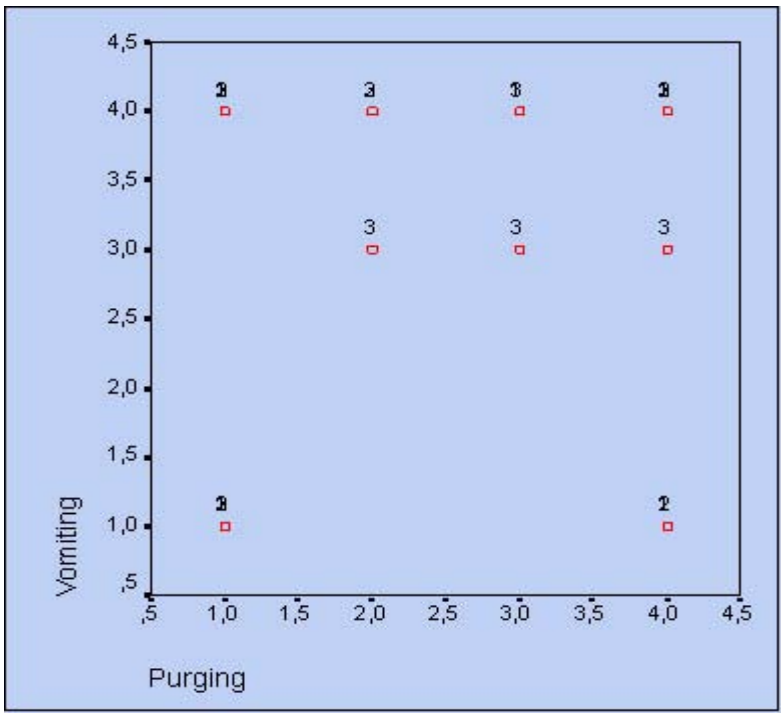

Рис. 20.2: *Диаграмма рассеяния переменных kalorien (калории) и kosten (расходы)*

В соответствии с этой формулой расстояние между сортами пива Budweisei Heineken составляет:

$$\sqrt{(144-152)^2+(0,43-0,77)^2}=8,007$$

Это расстояние лишь незначительно превосходит то, которое получилось бы, если бы для расчета была взята только одна переменная — kalorien (калории):

|144 - 152 | = 8

Данный эффект можно объяснить тем, что уровни значений переменных kalorien (калории) и kosten (расходы) очень сильно отличаются друг от друга: у переменной kosten (расходы) значения меньше 1, а у переменной kalorien (калории) больше 100. Согласно формуле евклидовой меры, переменная, имеющая большие значения, практически полностью доминирует над переменной с малыми значениями.

Решением этой проблемы является рассмотренное в главе 19.1 z-преобразование (стандартизация) значений переменных. Стандартизация приводит значения всех преобразованных переменных к единому диапазону значений, а именно от —3 до +3.

Если Вы произведёте такое преобразование для переменных kalorien (калории) и kosten (расходы), то для пива Budweiser получите стандартизованные значения равные 0,400 и —0,469 соответственно, а для пива Heineken стандартизированные значения 0,649 и 1,848 соответственно.

Тогда расстояние между двумя сортами пива получится равным

$$\sqrt{(0,400-0,649)^2+(-0,469-1,848)^2}=2,330$$

Таким образом, при помощи диаграммы рассеяния для двух переменных: kalorien (калории) и kosten (расходы), мы провели самый простой кластерный анализ. Мы выбрали такой вид графического представления, с помощью которого можно было бы отчётливо распознать группирование в кластеры (четыре в нашем случае).

К сожалению, столь отчётливая картина отношений между переменными, как в приведенном примере, встречается очень редко. Во-первых, структуры кластеров, если вообще таковые имеются, не так чётко разделены, особенно при наличии большого количества наблюдений. Скорее наоборот, кластеры размыты и даже проникают друг в друга. Во-вторых, как правило, кластерный анализ проводится не с двумя, а с намного большим количеством переменных.

При кластерном анализе с тремя переменными можно ввести ещё одну ось — ось z и рассматривать размещение наблюдений, а также проводить расчёт расстояния по формуле евклидовой меры в трёхмерном пространстве.

При наличии более трёх переменных определение расстояния между двумя точками x и y в любом n-мерном пространстве для математиков не представляет особого труда. Формула Евклида в таких случаях приобретает следующий вид:

$$\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

Наряду с евклидовой мерой расстояния, SPSS предлагает и другие дистанционные меры, а также меры подобия. Так что кластерный анализ можно проводить не только с переменными,

относящимися к интервальной шкале, как в приведенном случае, но и с дихотомическими переменными, к примеру. В таком ситуации применяется уже другие дистанционные меры и меры подобия (см. разд. 20.3).

При проведении кластерного анализа отдельные кластеры могут формироваться при помощи пошагового слияния, для которого существует ряд различных методов (см. разд. 20.4). Важную роль играют иерархические и партиционные методы, причём последние применяются в подавляющем большинстве случаев. Оба эти метода можно задействовать, если пройти через меню Analyze (Анализ) Classify (Классифицировать)

Они помещены в этом меню под именами Hierarchical Cluster... (Иерархический кластер) и K-Means Cluster... (Кластерный анализ методом к-средних).

Рассмотрим сначала иерархический кластерный анализ, причём начнём с простого примера с 17 сортами пива.

20.2. Иерархический кластерный анализ

В иерархических методах каждое наблюдение образовывает сначала свой отдельный кластер. На первом шаге два соседних кластера объединяются в один; этот процесс может продолжаться до тех пор, пока не останутся только два кластера. В методе, который в SPSS установлен по умолчанию (Between-groups linkage (Связь между группами)), расстояние между кластерами является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров.

20.2.1. Иерархический кластерный анализ с двумя переменными

Соберём заданные 17 сортов пива в кластеры при помощи параметров kalorien (калории) и kosten (расходы).

- Выберите в меню Analyze (Анализ) Classify (Классифицировать) Hierarchical Cluster... (Иерархический кластерный анализ)

Вы увидите диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ) (см. рис. 20.3).

- Переменные kalorien (калории) и kosten (расходы) поместите в поле тестируемых переменных, а текстовую переменную bier (пиво) в поле с именем Label cases by: (Наименования (метки) наблюдений:).

- Щелчком по выключателю Statistics... (Статистики) откройте диалоговое окно Hierarchical Cluster Analysis: Statistics (Иерархический кластерный анализ: Статистики) и наряду с выводом последовательности слияния (Agglomeration schedule) активируйте вывод показателя принадлежности к кластеру для каждого наблюдения. Хотя на основании графического представления на диаграмме рассеяния (см. рис. 20.2) и ожидается результат в виде четырёх кластеров, но не можем быть полностью уверены в достижении этого результата. Поэтому, для верности активируйте Range of solutions: (Область решений) и введите числа 2 и 5 в качестве границ области.

- Вернувшись в главное диалоговое окно, щёлкните по выключателю Plots... (Диаграммы). Активируйте опцию вывода древовидной диаграммы (Dendrogram) и посредством опции None (Нет) отмените вывод накопительной диаграммы.

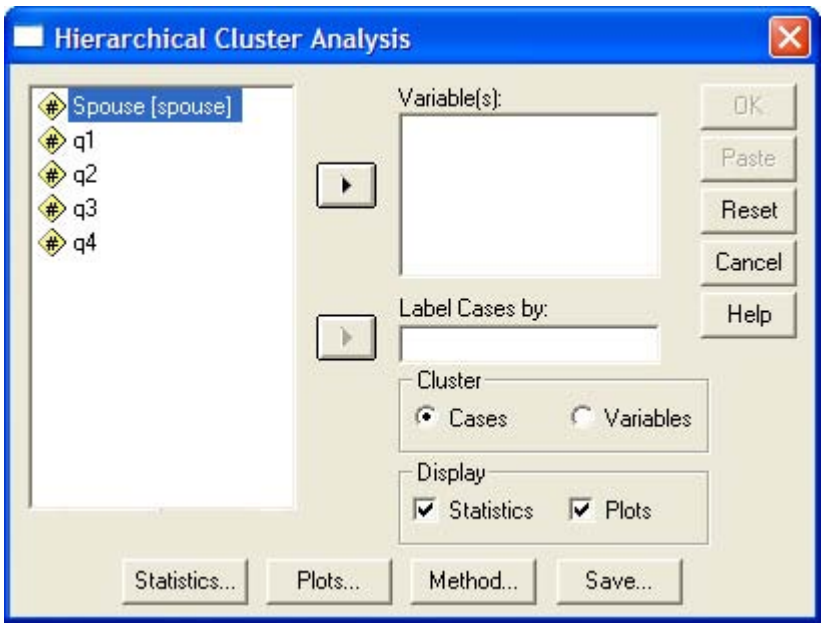

Рис. 20.3: *Диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ)*

- С помощью кнопки Method... (Метод) Вы получаете возможность выбрать метод образования кластеров, а также метод расчета дистанционной меры и меры подобия соответственно.

SPSS предлагает, в общей сложности, семь различных методов объединения, которые будут рассмотрены в главе 20.4. Метод Between-groups linkage (Связь между группами) устанавливается по умолчанию.

Дистанционные меры и меры подобия зависят от вида переменных, участвующих в анализе, то есть выбор меры зависит от типа переменной и шкалы, к которой она относится: интервальная переменная, частоты или бинарные (дихотомические) данные. В рассматриваемом примере фигурируют данные, относящиеся к интервальной шкале, для которых по умолчанию в качестве дистанционной меры устанавливается квадрат евклидового расстояния (Squared Euclidean distance). Некоторые дистанционные меры и меры подобия будут рассмотрены в главе 20.3.

- Оставьте предварительные установки и в поле Transform Values (Преобразовывать значения) установите z-преобразование (стандартизацию) значений; необходимость этой опции была уже рассмотрена в главе 20.1. Другие предлагаемые возможности стандартизации играют скорее второстепенную роль.

- Вернитесь назад в главное диалоговое окно и начните расчёт нажатием OK.

После обычной обшей статистической сводки итогов по наблюдениям, в окне просмотра сначала приводится обзор принадлежности, из которого можно выяснить очерёдность построения кластеров, а также их оптимальное количество. По двум колонкам, расположенным под общей шапкой Cluster Combined (Объединение в кластеры), можно увидеть, что на первом шаге были объединены наблюдения 5 и 12 (т.е. Heineken и Becks); эти две марки максимально похожи друг на друга и отдалены друг от друга очень малое расстояние. Эти два наблюдения образовывают кластер с номером 5, в то время как кластер 12 в обзорной таблице больше не появляется. На следующем шаге происходит объединение наблюдений 10 и 17 (Coors Light и Schlitz Light), затем 2 и 3 (Lowenbrau и Michelob) и т.д.

Agglomeration Schedule

(Порядок агломерации)						
Stage (Шаг)	**Cluster Combined (Объединение в кластеры)**		**Coefficients (Коэффициенты)**	**Stage Cluster First Appears (Шаг, на котором кластер появляется впервые)**		**Next Stage (Следующий шаг)**
	Cluster 1 (Кластер 1)	Cluster 2 (Кластер 2)		Cluster 1 (Кластер 1)	Cluster 2 (Кластер 2)	
1	5	12	8.508E-03	0	0	9
2	10	17	2.880E-02	0	0	4
3	2	3	4.273E-02	0	0	13
4	8	10	6.432E-02	0	2	7
5	7	13	8.040E-02	0	0	8
6	1	15	,117	0	0	8
7	8	9	,206	4	0	14
8	1	7	,219	6	5	12
9	5	11	,233	1	0	11
10	14	16	,313	0	0	14
11	4	5	,487	0	9	16
12	1	6	,534	8	0	13
13	1	2	,820	12	3	15
14	8	14	1,205	7	70	15
15	1	8	4,017	13	14	16
16	1	4	6,753	15	11	0

Для определения, какое количество кластеров следовало бы считать оптимальным, решающее значение имеет показатель, выводимый под заголовком "коэффициент". По этим коэффициентом подразумевается расстояние между двумя кластерами, определенное на основании выбранной дистанционной меры с учётом предусмотренного преобразования значений. В нашем случае это квадрат евклидового расстояния, определенный с использованием стандартизованных значений. На этом этапе, где эта мера расстояния между двумя кластерами увеличивается скачкообразно, процесс объединения в новые кластеры необходимо остановить, так как в противном случае были бы объединены уже кластеры, находящиеся на относительно большом расстоянии друг от друга.

В приведенном примере — это скачок с 1,205 до 4,017. Это означает, что после образования трёх кластеров мы больше не должны производить никаких последующих объединений, а результат с тремя кластерами является оптимальным. Визуально же мы ожидали результат с четырьмя кластерами. Оптимальным считается число кластеров равное разности количества наблюдений (здесь: 17) и количества шагов, после которого коэффициент увеличивается скачкообразно (здесь: 14).

В пояснении нуждаются ещё и три последние колонки вышеприведенной таблицы, отражающей порядок агломерации; для этого в качестве примера мы рассмотрим строку, соответствующую 14 шагу. Здесь объединяются кластеры 8 и 14. Перед этим кластер 8 уже участвовал в объединениях на шагах 4 и 7, последний раз, стало быть, на шаге 7. Строго говоря, название колонки Stage Cluster First Appears (Шаг, на котором кластер появляется впервые) можно считать ошибочным и вместо этого её следовало назвать Cluster Last Appears (Последнее появление кластера). Кластер 14 последний раз участвовал в объединении кластеров на шаге 10. Новый

кластер 8 затем примет участие в объединении кластеров на шаге 15 (колонка: Next Stage (Следующий шаг)).

Далее по отдельности для результатов расчёта содержащих 5, 4, 3 и 2 кластеров, приводится таблица с информацией о принадлежности каждого наблюдения к кластеру.

Cluster Membership (Принадлежность к кластеру)

Case (Случай)	5 Clusters (5 кластеров)	4 Clusters (4 кластера)	3 Clusters (3 кластера)	2 Clusters (2 кластера)
1:Budweiser	1	1	1	1
2:Lowenbrau	2	1	1	1
3:Michelob	2	1	1	1
4:Kronenbourg	3	2	2	2
5:Heineken	3	2	2	2
6:Schmidts	1	1	1	1
7:Pabst Blue Ribbon	1	1	1	1
8:Miller Light	4	3	3	1
9:Budweiser Light	4	3	3	1
10:Coors Light	4	3	3	1
11:Dos Equis	3	2	2	2
12:Becks	3	2	2	2
13:Rolling Rock	1	1	1	1
14:Pabst Extra Light	5	4	3	1
15:Tuborg	1	1	1	1
16:OlympiaGold Light	5	4	3	1
17:Schlitz Light	4	3	3	1

Таблица показывает, что два наблюдения 14 и 16 (Pabst Extra Light и Olympia Gold Light) при переходе к 3-х кластерному решению были включены в кластеры, соседствующие на диаграмме рассеяния; эти марки пива при оптимальном кластерном решении рассматриваются как принадлежащие к одному кластеру. Если посмотреть на 2-х кластерное решение, то оно группирует наблюдения 4, 5, 11 и 12 (Kronenbourg, Heineken, Dos Equis, Becks), то есть марки верхних правых кластеров диаграммы рассеяния; это марки иностранного производства.

В заключение приводится затребованная нами дендрограмма, которая визуализирует процесс слияния, приведенный в обзорной таблице порядка агломерации. Она идентифицирует объединённые кластеры и значения коэффициентов на каждом шаге. При этом отображаются не исходные значения коэффициентов, а значения приведенные к шкале от 0 до 25. Кластеры, получающиеся в результате слияния, отображаются горизонтальными пунктирными линиями.

******HIERARCHICAL CLUSTER ANALYSIS*** Dendrogram usinc (Average Linkage (Between Groups)	
Rescaled Distantce	Cluster Combine
CASE	0 5 10 15 20 25
Label	Hum +---- + ---- + ---- + _--_-- + ------ +
Heineken	5
Becks	12 - -
Dos Equis	11----------

Krcnenbourg	4 --
LcMBribrau	2 ----
Michelcb	3 - -
Pabst Blue Ribbon	7 ----------
Rolling Rode	13 -- -
Budweiser	4 ----------
Tuborg	15 - -
Schmdts	6 ----------
Coors Light	10 -
Schlitz Light	17 -
Miller Light	8 - - - - -
Budweiser Light	9 ----------
Pabst Extra Light	14 ----------
Olynpia Gold Light	16 - ----------

В то время как дендрограмма годится только для графического представления процесса слияния, по диаграмме накопления можно проследить деление кластеров. Так как начиная с 7 версии SPSS графическое представление диаграммы накопления оставляет желать лучшего, мы отказались от активирования ее вывода.

Для вводного рассмотрения мы выбрали довольно простой пример, включающий только две переменных. В этом случае конфигурация кластеров поддается представлению в графическом виде.

20.2.2. Иерархический кластерный анализ с более чем двумя переменными

Рассмотрим пример из области кадровой политики некоего предприятия. 18 претендентов прошли 10 различных тестов в кадровом отделе предприятия. Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 балов. Список тестов был следующим:

№ теста	**Предмет теста**
1	Память на числа
2	Математические задачи
3	Находчивость при прямом диалоге
4	Тест на составление алгоритмов
5	Уверенность во время выступления
6	Командный дух
7	Находчивость
8	Сотрудничество
9	Признание в коллективе
10	Сила убеждения

Результаты теста хранятся в файле assess.sav в переменных t1-t10. В файле находится также и текстовая переменная для характеристики тестируемых. С использованием результатов теста соответствия, мы хотим провести кластерный анализ, целью которого является обнаружение групп кандидатов, близких по своим качествам.

- Откройте файл assess.sav.
- Выберите в меню Analyze (Анализ) Classify (Классифицировать) Hierarchical Cluster... (Иерархический кластерный анализ)
- В диалоговом окне Hierarchical Cluster Analysis (Иерархический кластерный анализ) переменные tl-tlO поместите в поле тестируемых переменных, а текстовую переменную name (имя) используйте для обозначения (маркировки) наблюдений.
- Для начала должно быть достаточно вывода обзорной таблицы порядка агломерации; не делайте больше запроса на какие-либо данные и деактивируйте вывод диаграмм. Так как все переменные в этом примере имеют одинаковые пределы значений, стандартизация переменных является излишней.

Обзорная таблица порядка агломерации выглядит следующим образом:

Agglomeration Schedule

(Порядок агломерации)						
Stage (Шаг)	**Cluster Combined (Объединение в кластеры)**		**Coefficients (Коэффициенты)**	**Stage Cluster First Appears (Шаг, на котором кластер появляется впервые)**		**Next Stage (Следующий шаг)**
	Cluster 1 (Кластер 1)	**Cluster 2 (Кластер 2)**		**Cluster 1 (Кластер 1)**	**Cluster 2 (Кластер 2)**	
1	1	4	,000	0	0	6
1	14	18	2,000	0	0	4
3	12	15	2,000	0	0	6
4	9	14	2,000	0	2	8
5	2	10	2,000	0	0	13
6	1	12	3,000	1	3	15
7	13	16	4,000	0	0	12
8	9	11	4,000	4	0	11
9	5	7	5,000	0	0	14
10	6	17	6,000	0	0	13
11	3	9	6,000	0	8	15
12	8	13	7,000	0	7	14
13	2	6	7,500	5	10	16
14	5	8	12,833	9	12	16
15	1	3	194,000	6	11	17
16	2	5	198,500	13	14	17
17	1	2	219,407	15	16	0

Значительный скачок коэффициента наблюдается после 14-го шага; как указанно в разделе 20.1, это означает, что для данных, включающих 18 наблюдений, оптимальным является решение с четырьмя кластерами. Авторы в этом месте добавляют следующее: данный пример является искусственным, и из дидактических соображений мы предварительно скомпоновали данные таким образом, чтобы получился однозначный результат. После определения оптимального количества кластеров организуем для каждого наблюдения вывод информации о принадлежности к кластеру.

- Для этого вновь откройте диалоговое окно Hierarchical Cluster Analysis (Иерархический кластерный анализ) и щёлкните по выключателю Statistics... (Статистики). В разделе

Cluster Membership (Принадлежность к кластеру) активируйте опцию Single solution (Одно решение) и укажите желаемое количество кластеров 4.

Информацию о принадлежности каждого наблюдения к определённому кластеру вы можете сохранить в новой переменной.

- Пройдите выключатель Save... (Сохранить), активируйте опцию Single solution (Одно решение) и для указания желаемого количества кластеров введите 4. Теперь помимо таблицы порядка агломерации для каждого наблюдения будет выводиться и информация о принадлежности к кластеру.

Из следующей таблицы видно, что в первый кластер входят четыре человека, во второй кластер — опять четыре человека, в третий кластер — пять человек и в четвёртый кластер — снова пять человек. Неясно ещё, что означают эти четыре кластера, то есть о чём говорят результаты 10 тестов, соответственно относящиеся к этим кластерам. Разобраться в значении кластеров нам помогут кластерные профили; они представляют собой средние значения переменных, которые включены в анализ, распределённые по кластерной принадлежности.

Cluster Membership (Принадлежность к кластеру)

Case (Случай)	**4 Clusters (4 кластера)**
1:VolkerR	1
2:Sigrid K	2
3:Elmar M	3
4:Peter B	1
5:Otto R	4
6:Elke M	2
7:Sarah K	4
8:PeterT	4
9:Gudrun M	3
10:Siglinde P	2
1 1 :Werner W	3
12:Achim Z	1
13:DieterK	4
14:Boris P	3
15:Silke W	1
16:ClaraT	4
17:Manfred K	2
18:Richard M	3

Если Вы рассмотрите данные в редакторе данных, то заметите, что добавилась переменная clu4_l; эта переменная указывает на кластерную принадлежность каждого наблюдения и может быть использована для расчёта кластерного профиля.

- Выберите в меню Analyze (Анализ) Compare Means (Сравнить средние значения) Means... (Средние значения)

Переменным t1-t10 присвойте статус зависимых переменных, а переменной clu4_1 статус независимой переменной, и начните расчёт. В качестве результатов расчёта выводятся средние значения и стандартные отклонения итогов десяти тестов для четырёх кластеров. Для удобства поместим средние значения в отдельную таблицу.

	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Память на числа	10,00	10,00	4,20	4,80
Математические задачи	10,00	10,00	4,80	4,40
Находчивость при прямом диалоге	9,00	4,25	10,00	4,00
Тест на составление алгоритмов	10,00	10,00	4,40	4,00
Уверенность во время выступления	10,00	4,75	10,00	4,20
Командный дух	9,50	4,50	4,40	10,00
Находчивость	9,25	3,75	10,00	4,40
Сотрудничество	9,75	4,25	4,00	10,00
Признание в коллективе	10,00	4,25	3,80	10,00
Сила убеждения	9,50	4,25	10,00	5,00

Тестируемые, входящие в первый кластер имеют очень хорошие показатели во всех тестах. Это те конкурсанты, которые наверняка прошли бы на завершающий отборочный тур. Во второй кластер включены те, кто имеет хорошие показатели по математическим тестам (память на числа, математические задачи, тест на составление алгоритмов), но со слабыми оценками в социальной компетентности и уверенности при выступлениях. В третий кластер вошли те, кто уверенно себя чувствует во время выступления, но имеют слабые показатели в математических тестах и социальной компетентности. В конце концов, в четвёртом кластере, собраны люди с высоким уровнем социальной компетентности, но со слабыми результатами в тестах на решение математических задач и на силу убеждения.

В примерах, подобных этому, перед проведением кластерного анализа рекомендуется сократить количество переменных. Подходящим методом для этого является факторный анализ (см. гл. 19), который большое количество переменных заменяет меньшим количеством факторов. Продемонстрируем данный процесс на следующем примере.

20.2.3. Иерархический кластерный анализ с предварительным факторным анализом

Рассмотрим пример из области географии. В 28 европейских странах в 1985 году были собраны следующие данные, выступающие здесь в качестве переменных:

Переменная	Значение
land	Страна
sb	Процент городского населения
lem	Средняя продолжительность жизни мужчин
lew	
ks	Детская смертность на 1000 новорожденных
so	Количество часов ясной погоды в году
nt	Количество дней пасмурной погоды в году
tjan	Средняя дневная температура в январе
tjul	Средняя дневная температура в июле

Эти данные вы увидите, если откроете файл europa.sav. Переменная land является текстовой переменной, предназначенной для обозначения страны.

Целью нашего кластерного анализа является нахождение стран с похожими свойствами. При самом общем рассмотрении переменных (от непосредственного указания стран мы здесь воздержимся) становится заметным, что данные, содержащиеся в файле связаны

исключительно с ожидаемой продолжительностью жизни или с климатом. Лишь процентный показатель населения, проживающего в городах, не вписывается в эти рамки. Стало быть, сходства, которые возможно будут найдены между некоторыми странами, основываются на продолжительности жизни и климате этих стран.

Исходя из вышесказанного, в данном случае перед проведением кластерного анализа рекомендуется сократить количество переменных. Подходящим методом для этого является факторный анализ (см. гл. 19), который вы можете провести, выбрав в меню Analyze (Анализ) Data Reduction (Преобразование данных) Factor... (Факторный анализ)

Если Вы проведёте факторный анализ и примените, к примеру, вращение по методу варимакса, то получите два фактора. В первый фактор войдут переменные: lem. lew, ks и sb, а во второй фактор - переменные: so, nt, tjan и tjul. Первый фактор однозначно характеризует продолжительность жизни, причём высокое значение фактора означает высокую продолжительность жизни, а второй отражает климатические условия; здесь высокие значения означают тёплый и сухой климат. Вместе с тем, Вы наверняка заметили, что в первый фактор интегрирована и переменная sb, что очевидно указывает на высокую ожидаемую продолжительность жизни при высоких процентных долях городского населения. Вы можете рассчитать факторные значения для этих двух факторов и добавить их к файлу под именами fac1_1 и fac2_1. Чтобы Вам не пришлось самостоятельно проводить факторный анализ на этом этапе, указанные переменные уже включены в файл europa.sav. Вы можете видеть, к примеру, что высокой продолжительностью жизни обладают северные страны (высокие значения переменной fac1_1) или южные страны с тёплым и сухим климатом (высокие значения переменной fac2_1). Факторные значения можно вывести с помощью меню Analyze (Анализ) Reports (Отчёты) Case Summaries... (Итоги по наблюдениям)

Они выглядят следующим образом:

Case Summaries a (Итоги по наблюдениям)

	LAND (Страна)	**Lebenserwartung (Ожидаемая продолжительность жизни)**	**Klima (Климат)**
1	ALBA	-1,78349	,57155
2	BELG	,55235	-,57937
3	BULG	-,43016	-,13263
4	DAEN	,97206	-,23453
5	DDR	,26961	-,3351 1
6	DEUT	,19121	-,44413
7	FINN	-,30226	-1,28467
8	FRAN	1,05511	1,04870
9	GRIE	,12794	2,65654
10	GROS	,75443	-,05221
11	IRLA	,16370	-,66514
12	ISLA	1,75315	-,97421
13	ITAL	,40984	1,68933
14	JUGO	-2,63161	-,44127
15	LUXE	-.16469	-,98618
16	NIED	1,31001	-,29362
17	NORW	,96317	-,46987
18	OEST	-,20396	-,31971
19	POLE	-,65937	-,92081

20	PORT	-1,10510	1,59478
21	RUMA	-1,32450	,09481
22	SCHD	1,22645	-,20543
23	SCHZ	, 56289	-,45454
24	SOWJ	-,67091	-1,32517
25	SPAN	, 83627	1,91193
26	TSCH	-,59407	-,40632
27	TUER	-,52049	1,04424
28	UNGA	-,75761	-,08695
Total N	28	28	28

a. Limited to first 100 cases (Ограничено первыми 100 наблюдениями).

Распределим эти 28 стран по кластерам при помощи двух факторов: ожидаемая продолжительность жизни и климат.

- Выберите в меню Analyze (Анализ) Classify (Классифицировать) Hierarchical Cluster... (Иерархический кластерный анализ)
- Переменные fac1_1 и fac2_1 поместите в поле тестируемых переменных, а переменную land (страна) — в поле с именем Label cases by: (Наименование (маркировка) наблюдений).
- После прохождения выключателя Statistics... (Статистики), наряду с таблицей порядка агломерации сделайте запрос на вывод информации о принадлежности к кластеру для наблюдений. Активируйте Range of solutions: (Область решений) и введите граничные значения 2 и 5.
- Для сохранения информации о принадлежности отдельных наблюдений к кластеру в виде дополнительных переменных, воспользуйтесь выключателем Save... (Сохранить). В соответствии с установками, произведенными в диалоговом окне статистики, активируйте и здесь Range of solutions: (Область решений) и введите граничные значения 2 и 5.
- Деактивируйте вывод дендрограмм. Так как переменные, используемые в данном кластерном анализе, являются факторными значениями с одинаковыми областями допустимых значений, то стандартизация (z-преобразование) значений является излишней.

Agglomeration Schedule

(Порядок агломерации)						
Stage (Шаг)	**Cluster Combined (Объединение в кластеры)**		**Coefficients (Коэффициенты)**	**Stage Cluster First Appears (Шаг, на котором кластер появляется впервые)**		**Next Stage (Следующий шаг)**
	Cluster 1 (Кластер 1)	**Cluster 2 (Кластер 2)**		**Cluster 1 (Кластер 1)**	**Cluster 2 (Кластер 2)**	
1	16	22	1,476	0	0	8
2	2	23	1,569	0	0	10
3	5	6	1,803	0	0	5
4	4	17	5,546	0	0	8
5	5	11	8,487	3	0	10
6	3	18	8,617	0	0	12

7	7	15	,108	0	0	15
8	4	16	,118	4	1	13
9	26	28	,129	0	0	12
10	2	5	,148	2	5	18
11	19	24	,164	0	0	15
12	3	26	,183	6	9	20
13	4	10	,228	8	0	18
14	13	25	,231	0	0	19
15	7	19	,254	7	11	20
16	1	21	,438	0	0	22
17	20	27	,645	0	0	22
18	2	4	,648	10	13	21
19	8	13	,810	0	14	23
20	3	7	,939	12	15	24
21	2	12	1,665	18	0	24
22	1	20	1,793	16	17	25
23	8	9	1,839	19	0	27
24	2	3	2,229	21	20	26
25	1	14	4,220	22	0	26
26	1	2	5,925	25	24	27
27	1	8	6,957	26	23	0

Сначала приводятся самые важные результаты. В таблице порядка агломерации Вы можете проследить последовательность образования кластеров; объяснения по этому поводу приводились в разделе 20.1. Скачкообразное изменение коэффициентов наблюдается при значениях 2,229 и 4,220; это означает, что после образования четырёх кластеров больше не должно происходит ни каких объединений и решение с четырьмя кластерами является оптимальным.

Принадлежность наблюдений к кластерам можно взять из нижеследующей таблицы, которая содержит также и информацию о принадлежности к кластерам для других вариантов решения (пять, три и два кластера).

Если Вы посмотрите на четырёхкластернное решение на нижеследующей таблице, то заметите, к примеру, что к третьему кластеру относятся следующие страны: Франция, Греция, Италия и Испания. Это страны с высокой продолжительностью жизни и тёплым климатом и поэтому не зря они являются предпочтительными для отдыха.

Cluster Membership (Принадлежность к кластеру)

Case (Случай)	**5 Clusters (5 кластеров)**	**4 Clusters (4 кластера)**	**3 Clusters (3 кластера)**	**2 Clusters (2 кластера)**
1:ALBA	1	1	1	1
2:BELG	2	2	2	1
3:BULG	3	2	2	1
4:DAEN 5:DEUT	2	2	2	1
6:DDR	2	2	2	1
7:FINN	3	2	-3	2

8:FRAN	4	3	-3	2
9:GRIE	4	2	2	1
10:iGROS	2		2	1
11:IRLA	2	2	2	1
12:ISLA	2	3	o	2
13:ITAL	4	4	1	1
14:JUGO	5	2	2	1
1 5:LUXE	3	2	2	1
16:NIED	2		2	1
17:NORW	2	2	2	1
18:OEST	3	2	2	1
19:POLE	3	2 1	1	1
20:PORT	1	1	1	1
21:RUMA	1	2		1
22:SCHD 23:SCHZ	2	2	2	1
24:SOWJ	3	1	i	2
!25:SPAN	4	1		
26:TSCH	3	1	1	1
27:TUER 28:UNGA	1	2	1	1

20.3. Меры расстояния и меры сходства

Основой кластеризации (образования групп) наблюдений является дистанционная матрица и матрица подобия наблюдений. Так как расстояние (дистанция) также применяется и для оценки подобия, то разница между этими двумя матрицами не велика. В зависимости от того, к какой шкале измерений относятся переменные, участвующие в анализе, SPSS предлагает различные дистанционные меры и меры подобия.

20.3.1. Переменные, относящиеся к интервальной шкале (метрические переменные)

Для переменных такого рода на выбор предлагается восемь различных мер расстояния и мер сходства, которые мы и рассмотрим далее. Примером расчёта послужат два наблюдения из файла assess.sav (см. гл. 20.3), для которых расстояние и подобие должны быть рассчитаны с использованием переменных t3 и t4:

	t3	t4
Отто Р.	5	4
Эльке М.	4	10

Евклидова дистанция (расстояние)

Евклидова дистанция между двумя точками x и y — это наименьшее расстояние между ними. В двух- или трёхмерном случае — это прямая, соединяющая данные точки. Общей формулой для n-мерного случая (л переменных) является: 1

$$dist = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Сокращение dist, как и в следующей формуле, соответствует слову дистанция. Для ! приведенного примера получим

$$dist = \sqrt{(5-4)^2 + (4-10)^2} = 6,0828$$

Квадрат евклидового расстояния

Этот вариант устанавливается по умолчанию. Благодаря возведению в квадрат при расчёте лучше учитываются большие разности. Эта мера должна всегда использоваться при построении кластеров при помощи центроидного и медианного методов, а также метода Варда (Ward-Method) (см. разд. 20.5).

$$dist = \sum_{i=1}^{n}(x_i - y_i)^2$$

Для приведенного примера имеем cfot=(5-4)2 + (4-10)2 =37

Косинус

Как и для корреляционных коэффициентов Пирсона, область значений этой меры находится между -1 и +1.

$$Подобие = \frac{\sum_{i=1}^{n}(x_i, y_i)}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2)}}$$

Для приведенного примера имеем

$$Подобие = \frac{5*4+4*10}{\sqrt{(5^2+4^2)*(4^2+10^2)}} = 0,8700$$

Корреляция Пирсона

Если кластеризация наблюдений осуществляется только на основании двух переменных, то корреляционный коэффициент Пирсона (см. разд. 15.1) со значениями находящимися в пределах от -1 до +1 не годится для использования в качестве меры подобия; он будет давать только значения -1 или +1.

Чебышев (Chebychev)

Разностью двух наблюдений является абсолютное значение максимальной разности последовательных пар переменных, соответствующих этим наблюдениям.

В приведенном примере абсолютная разность значений первой переменной равна 1, а второй переменной — 6. Поэтому разность Чебышева равна 6.

Блок (Block)

Эта дистанционная мера, называемая также дистанцией Манхэттена или в шутку — дистанцией таксиста, определяется суммой абсолютных разностей пар значений. Для двумерного пространства это не прямолинейное евклидова расстояние между двумя точками, а путь, который должен преодолеть Манхэттенский таксист, чтобы проехать от одного дома к другому по улицам, пересекающимся под прямым углом.

$$dist = \sum_{i=1}^{n} | x_i - y_i |$$

Для нашего примера имеем dfst = |5-4| + |4-10| = 7

Минковский (Minkowski)

Расстояние Минковского равно корню r-ой степени из суммы абсолютных разностей пар значений взятых в r-ой степени:

$$dist = (\sum_{i=1}^{n} | x_i - y_i |^{r})^{1/r}$$

В SPSS при расчете этого расстояния допускается применение только квадратного корня, в то время как степень разности значений можно выбрать в пределах от 1 до 4. Если эту степень взять равной 2, то получим евклидово расстояние.

Пользовательская мера

Это обобщенный вариант расстояния Минковского. Это расстояние, называемое также степенным расстоянием, равно корню r-ой степени из суммы абсолютных разностей пар значений взятой в p-ой степени:

$$dist = (\sum_{i=1}^{n} | x_i - y_i |^{p})^{1/r}$$

Здесь как для корня, так и для степени суммы можно выбирать значения от 1 до 4.

20.3.2. Частоты

В качестве примера возьмём файл laender.sav, в котором значения переменных отображают частоты. В файле находится текстовая переменная land (федеральная земля) и три переменные cdu, spd и andere (другие). Для шестнадцати земель Федеративной Республики Германия в 1994 году эти переменные отображают количество мест в земельном парламенте, принадлежащих двум основным партиям — CDU и SPD, а также места, относящиеся к другим партиям.

- Откройте файл laender.sav.
- На основании трёх переменных cdu, spd и andere проведите иерархический кластерный анализ, текстовую переменную land примените для обозначения наблюдений.
- Через выключатель Method... (Метод) активируйте опцию Counts (Частоты). У Вас появится возможность выбора между двумя дистанционными мерами.

Мера хи-квадрат

Для того, чтобы найти расстояние между двумя наблюдениями, сравнивают частоты выпадения переменных, относящихся к этим наблюдениям. В качестве примера рассмотрим две федеративные земли: Хессен и Тюринген:

	CDU	**SPD**	**Andere (Другие)**
Хессен	46	46	18
Тюринген	43	21	25

Для такой таблицы долей присутствия разных партий подходит статистика хи-квадрат (см. разд. 11.3.1). Квадратный корень из значения хи-квадрат будет применяться в качестве дистанционной меры.

В приведенном примере значение хи-квадрат получилось равным 8,447 значит дистанционная мера равна 2,9064.

Мера фи-квадрат

Эта мера представляет собой попытку нормализации меры хи-квадрат. Для этого она делится на квадратный корень общей суммы частот.

В рассматриваемом примере сумма частот для двух земель Хессен и Тюринген равна 199, так что мера фи-квадрат получается равной 0,2060.

Если Вы в качестве дистанционной меры выберите меру хи-квадрат, то получите результат, в котором оптимальным решением окажется решение с пятью кластерами. Два самых больших кластера образовываются землями, в которых CDU или SPD имеют большинство мест, один кластер — землями Бранденбург и Бремен, в управлении которых относительно велико представительство других партий, один кластер образует Бавария, в связи с абсолютно доминирующей ролью CDU и один кластер — Саксония, тоже в связи с доминирующей ролью CDU, но с некоторой долей других партий, которая больше доли SPD.

20.3.3. Бинарные переменные

Здесь, как правило, речь идёт о переменных, которые указывают на факт осуществления некоторого события или выполнения определённого критерия. В файле данных это обстоятельство должно быть закодировано при помощи двух численных значений, причём в соответствии с установками по умолчанию, SPSS для кодировки осуществления события ожидает цифру 1.

Если сопоставить друг с другом две переменные, то все возможные сочетания наблюдений дают четыре различные частоты, которые называются a, Ь, c, d и имеют следующий смысл:

		Переменная 2	
		сбылось	**не сбылось**
Переменная 1	Сбылось	a	b
	Не сбылось	c	d

На основании этих частот, можно рассчитать множество различных дистанционных мер, 27 из которых применяются в SPSS. Двадцать разновидностей мер, называемых мерами подобия, рассмотрены в разделе 15.4. Остальные приводятся ниже.

Квадрат евклидовою расстояния

Бинарное евклидово расстояние, возведенное в квадрат, представляет собой количество наблюдений, для которых, по крайней мере, один из критериев присутствует и один отсутствует. Эта мера является установкой по умолчанию.

dist= b + c

Евклидово расстояние

Бинарное евклидово расстояние представляет собой корень из числа наблюдений, для которых, по крайней мере, один из критериев присутствует и один отсутствует.

$$dist = \sqrt{b+c}$$

Разность длин

Эта мера имеет минимальное значение равное 0 и не имеет верхнего придела.

$$dist = \frac{(b-c)^2}{(a+b+c+d)^2}$$

Образцовая разность

Образцовая разность может принимать значения от 0 до 1.

$$dist = \frac{bc}{(a+b+c+d)^2}$$

Дисперсия

Дисперсия имеет минимальное значение равное 0 и не имеет верхнего придела.

$$dist = \frac{b+c}{4(a+b+c+d)}$$

Форма

У этой дистанционной меры нет ни нижнего ни верхнего предела

$$dist = \frac{(a+b+c+d)(b+c)-(b-c)^2}{(a+b+c+d)^2}$$

Мера Ланса и Уильямса (Lance and Williams)

Эта мера может принимать значения от 0 до 1.

$$dist = \frac{b+c}{2a+b+c}$$

Приведенные меры отличаются друг от друга присутствием в соответствующей формуле различных наборов из четырёх частот: a, b, c и d

Так, для евклидовой меры в расчёт включают только те наблюдения, для которых имеется один признак и отсутствует другой, а в других дистанционных формулах учитываются все частоты. Исключением является дистанционная мера по Лансу и Уильямсу, в которой в расчет не берутся те наблюдения, для которых отсутствуют оба признака.

На какой мере Вы остановите свой выбор, зависит от того, какую роль вы отводите частотам a, b, c и d.

20.4. Методы объединения

SPSS предлагает, в общей сложности, семь методов объединения. Из них метод Связь между группами (Between-groups linkage) устанавливается по умолчанию.

Связь между группами

Дистанция между кластерами равна среднему значению дистанций между всеми возможными парами наблюдений, причём один наблюдения берётся из одного кластера, а другой из другого. Информация, необходимая для расчёта дистанции, находится на основании всех теоретически возможных пар наблюдений. По этой причине данный метод и устанавливается по умолчанию.

Связь внутри групп

Это вариант связи между группами, а именно, здесь дистанция между двумя кластерами рассчитывается на основании всех возможных пар наблюдений, принадлежащих обоим кластеров, причём учитываются также и пары наблюдений, образующиеся внутри кластеров.

Близлежащий сосед

Дистанция между двумя кластерами определяется, как расстояние между парой значений наблюдений, расположенных друг к другу ближе всего, причём каждое наблюдение берётся из своего кластера.

Дальний сосед

Дистанция между двумя кластерами определяется как расстояние между самыми удалёнными друг от друга значениями наблюдений, причём каждое наблюдение берётся из своего кластера.

Центроидная кластеризация

В обоих кластерах рассчитываются средние значения переменных относящихся к ним наблюдений. Затем расстояние между двумя кластерами рассчитывается как дистанция между двумя осредненными наблюдениями.

Медианная кластеризация

Этот метод похож на центроидную кластеризацию. Однако в предидущем методе центроид нового кластера получается как взвешенное среднее центроидов обоих исходных кластеров, причём количества наблюдений исходных кластеров образовывают весовой коэффициент. В медианном же методе оба исходных кластера берутся с одинаковым весом.

Метод Варда (Ward-Method)

Сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных. Затем вычисляются квадраты евклидовых расстояний от

отдельных наблюдений каждого кластера до этого кластерного среднего значения. Эти дистанции суммируются. Потом в один новый кластер объединяются те кластера, при объединении которых получается наименьший прирост общей суммы дистанций. Так как некоторые из предлагаемых методов имеют явные недостатки (Близлежащий сосед, Дальний сосед), а другие очень мало наглядны и плохо поддаются последующему анализу, рекомендуется применять устанавливаемый по умолчанию и наиболее понятный метод Between-groups linkage (Связь между группами).

20.5. Кластерный анализ при большом количестве наблюдений (Кластерный анализ методом к-средних)

Иерархические методы объединения, хотя и точны, но трудоёмки: на каждом шаге необходимо выстраивать дистанционную матрицу для всех текущих кластеров. Расчётное время растёт пропорционально третьей степени количества наблюдений, что при наличии нескольких тысяч наблюдений может утомить и серьёзные вычислительные машины.

Поэтому при наличии большого количества наблюдений применяют другие методы. Недостаток этих методов заключается в том, что здесь необходимо заранее задавать количество кластеров, а не так как в иерархическом анализе, получить это в качестве результата Эту проблему можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров. Если количество кластеров указать предварительно, то появляется следующая проблема: определение начальных значений центров кластеров. Их также можно взять из предварительно проведённого иерархического анализа, в котором для каждого наблюдения рассчитывают средние значения переменных, использовавшихся при анализе, а потом в определённой форме сохраняют их в некотором файле. Этот файл может быть затем прочитан методом, который применяется для обработки больших количеств наблюдений. Если нет желания проходить весь этот длинный путь, то можно воспользоваться методом, предлагаемым для данного наблюдения программой SPSS. Если количество кластеров k, которое необходимо получить в результате объединения, задано заранее, то первые k наблюдений, содержащихся в файле, используются как первые кластеры. На последующих шагах кластерный центр заменяется наблюдением, если наименьшее расстояние от него до кластерного центра больше расстояния между двумя ближайшими кластерами. По этому правилу заменяется тот кластерный центр, который находится ближе всего к данному наблюдению. Таким образом получается новый набор исходных кластерных центров. Для завершения шага процедуры рассчитывается новое положение центров кластеров, а наблюдения перераспределяются между кластерами с изменённым центрами. Этот итерационный процесс продолжается до тех пор, пока кластерные центры не перестанут изменять свое положение или пока не будет достигнуто максимальное число итераций.

В качестве примера расчёта по этому алгоритму, рассмотрим выборку из результатов исследований Института социологии Марбургского Университета им. Филипса, в котором проводился опрос 1000 студентов относительно использования ими компьютера и их отношения к современным информационным и телекоммуникационным технологиям. В разделе "Пользование компьютерными программами" были представлены следующие вопросы с различным количеством подпунктов, на которые необходимо было ответить в соответствии с пятибальной шкалой (от отлично до абсолютно не использую):

1. Насколько свободно вы можете работать в следующих приложениях?

Обработка текста, Графические программы, обработка звука или видео, монтаж Базы данных и табличные расчёты

2. Насколько хорошо вы владеете следующими языками программирования?

BASIC, Paskal ,C, Машинные языки, Программирование для Интернета(к примеру, HTML), Java

3. Насколько хорошо Вы можете работать в следующих операционных системах?

DOS , Windows ,UNIX

4. Насколько хорошо Вы разбираетесь в следующих возможностях Интернета?

E-mail, группы новостей, почтовая рассылка, Путешествие по всемирной сети Интернет, Chat, IRC, ICQ, Предложение собственных услуг(к примеру, домашней страницы)

5. Насколько хорошо Вы разбираетесь в играх?

Как часто Вы играете в компьютерные игры, Насколько хорошо Вы ориентируетесь в сценах компьютерных игр?

Ответы на эти вопросы хранятся в переменных vla-v5b в файле computer.sav. В этом файле также находятся и другие переменные, использовавшиеся при исследовании (пол, возраст, место жительства, профессия). На основании вопросов об использовании программных продуктов попытаемся определить группы (кластеры) пользователей. Для начала рекомендуется сократить количество переменных при помощи факторного анализа, как описано в разделе 20.2.3.

- Откройте файл computer.sav
- Выберите в меню Analyze (Анализ) Data Reduction (Преобразование данных) Factor... (Факторный анализ)
- Переменные vla-v5b внесите в список целевых переменных.
- Через выключатель Extraction... (Отбор) деактивируйте вывод неповёрнутого факторного решения.
- Через выключатель Rotation... (Вращение) для осуществления вращения активируйте метод варимакса.
- Минуя выключатель Options... (Опции) в разделе Coefficient Display Format (Формат отображения коэффициентов) (подразумеваются факторные нагрузки) активируйте Sorted by Size (Отсортированные по размеру). Затем активируйте опцию Suppress absolute values less then: (Не выводить абсолютные значения меньше чем:) и введите значение ,40.
- В заключение щёлкните по выключателю Scores... (Значения), чтобы значения факторов сохранить в виде новых переменных.

В результате расчёта было отобрано четыре фактора и добавлено в файл четыре переменные от (fac1_1 до fac4_1), которые и отображают эти четыре фактора. Среди результатов присутствует повёрнутая факторная матрица (см. следующую таблицу).

Факторная матрица красноречиво демонстрирует, что отобранные факторы могут быть расположены в следующей смысловой последовательности (по убыванию значимости):

- Приложение
- Программирование
- Использование Интернета
- Игры

Rotated Component Matrix

(Повёрнутая матрица компонентов)				
	Component (Компонент)			
	1	2	3	4
Textverarbeitung (Обработка текста)	,848			
Windows	,840			
DOS	,653			

WWW	,619			
Datenbanken (Базы данных и табличные расчёты)	,611			
Multimedia (Мультимедиа)	,535			
C		,771		
Maschinensprache (Машинные языки)		,741		
PASCAL		,729		
BASIC		,612		
Java		,606	,474	
UNIX		,587	,504	
Chat			,699	
eigene Dienste (Предложение собственных услуг)			,696	
Internetsprachen (Программирование для Интернет)		,468	,670	
Email	,584		,609	
ICQ			,601	
Szene (Сцены компьютерных игр)				,881
Intensitaet (Интенсивность)				,850

Extraction Method: Principal Component Analysis (Метод отбора: Анализ главных компонентов).

Rotation Method: Varimax with Kaiser Normalization (Метод вращения: варимакс с нормализацией Кайзера).

a. Rotation converged in 11 iterations (Вращение осуществлено за 11 итераций).

Теперь используем сохранённые нами значения этих четырёх факторов для проведения кластерного анализа для студентов. Так как количество наблюдений равное 1085 слишком велико для иерархического кластерного анализа, выберем метод анализа кластерных центров.

- Присвойте переменным fac1_1-fac4_1 метки: "Приложения", "Программирование", "Использование Интернет" и "Игры" соответственно.
- Выберите в меню Analyze (Анализ) Classify (Классифицировать) K-Means Cluster... (Кластерный анализ методом к-средних)

Откроется диалоговое окно K-Means Cluster Analysis (Кластерный анализ методом к-средних).

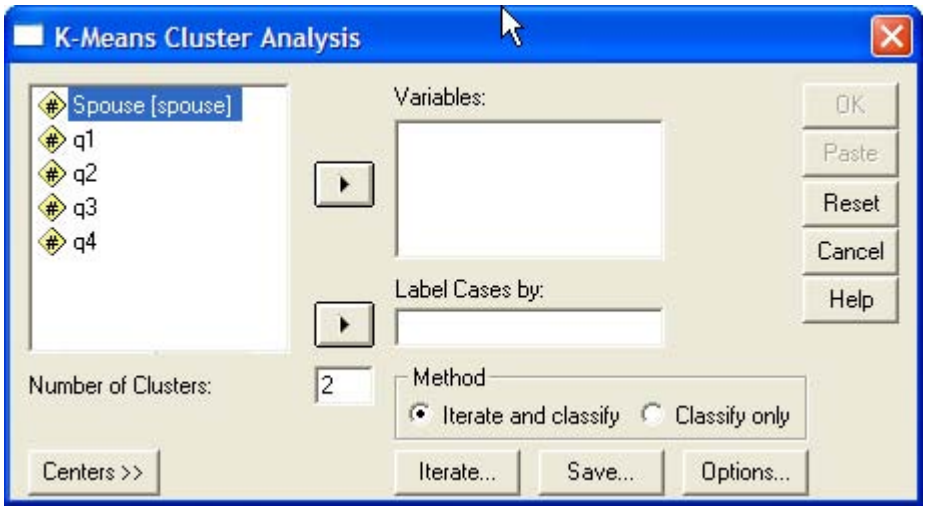

Рис. 20.4: *Диалоговое окно K-Means Cluster Analysis (Анализ кластерных центров)*

- Переменные от fac1_1 до fac4_1 поместите в поле тестируемых переменных. Теперь Вы подошли к тому месту, где нужно указывать количество кластеров. Подходящим вариантом было бы сперва провести иерархический кластерный анализ для произвольно выбранных наблюдений и получившееся количество кластеров принять за оптимальное. Вы, конечно же, можете провести и несколько опытных, пробных расчётов с различным количеством кластеров и после этого определиться с подходящим вариантом решения.
- Мы остановимся на четырёх кластерах; введите это значение в поле Number of Clusters (Количество кластеров).
- Через выключатель Iterate... (Итерации) укажите число итераций равное 99; установленное по умолчанию количество итераций равное 10, оказалось бы недостаточным.
- Щёлкните по выключателю Save... (Сохранить), чтобы при помощи дополнительных переменных зафиксировать принадлежность наблюдений к кластеру.
- Щёлкните на ОК, чтобы начать расчёт.

Сначала приводятся первичные кластерные центры и обобщённые данные итерационного процесса (30 итераций); затем выводятся окончательные кластерные центры и информация о количестве наблюдений.

Final Cluster Centers

(Кластерные центры окончательного решения)				
	Cluster (Кластер)			
	1	2	3	4
Приложение	-,15219	-,62362	-,23459	1,16856
Программирование	-2,91321	,232223	,23371	,05918
Использование Интернет	-1,71057	,7232	-.02994	,25268
Игры	,04717	,51053	-1,51014	,26081

При оценке кластерных центров следует в первую очередь обратить внимание на то, что здесь речь идёт о средних значениях факторов, которые находятся в пределах примерно от -3 до +3. К тому же, надо помнить, что в соответствии с кодировкой ответов (1 = отлично, 5 = абсолютно не использую) большое отрицательное значение фактора означает его большую степень его проявления, то есть сигнализирует о высокой компетентности, и наоборот, большое положительное значение фактора подразумевает низкую степень его проявления.

Если учесть всё вышесказанное, то наши четыре кластера можно интерпретировать следующим образом:

Кластер1: Программисты, Интернет-эксперты

Кластер2: Пользователи стандартного программного обеспечения

Кластер3: Игроки

Кластер4: Начинающие пользователи

В заключение выводятся показатели количества наблюдений, относящихся к каждому из кластеров. Группа пользователей (кластер 2) наиболее многочисленна.

Number of Cases in each Cluster

(Количество наблюдений в каждом кластере)		
Cluster (Кластер)	1	63,000
	2	488,000
	3	221,000
	4	313,000
Valid (Действительные)		1085,000
Missing (Отсутствующие)		,000

К исходному файлу была добавлена переменная qc1_1, отражающая принадлежность к определённому кластеру. Эту переменную можно использовать для обнаружения возможных связей между кластерной принадлежностью и полом, возрастом, профессией и происхождением (западные земли Германии, восточные земли Германии, зарубежные страны).

Наряду с количеством кластеров можно так же, как было упомянуто в начале главы, задать и первичные кластерные центры. Для этого их необходимо определённым образом ввести в файл данных SPSS. Изучим процесс создания такого файла на рассмотренном примере,

- После щёлка в диалоговом окне K-Means Cluster Analysis (Кластерный анализ методом k-средних) по выключателю Centers» (Центры), диалоговое окно примет расширенный вид (см. рис. 20.5).
- Активируйте Read initial from (Читать первичные значения из) и щёлкните на выключателе File... (Файл). Откроется диалоговое окно K-Means Cluster Analysis: Read initial from (Кластерный анализ методом К-средних: Читать первичные значения из).
- Откройте файл zentren.sav.

Файл содержит

- количественную переменную с именем cluster_
- одну строку для каждого кластера
- первичные значения для каждой кластерной переменной.

То, как выглядит этот файл в редакторе данных, Вы можете увидеть на рисунке 20.6. Аналогично тому, как Вы смогли считать из файла первичные кластерные центры, при помощи выключателя Write final as (Сохранить окончательные результаты как), Вы можете сохранить окончательные кластерные центры в отдельном файле для дальнейших расчётов.

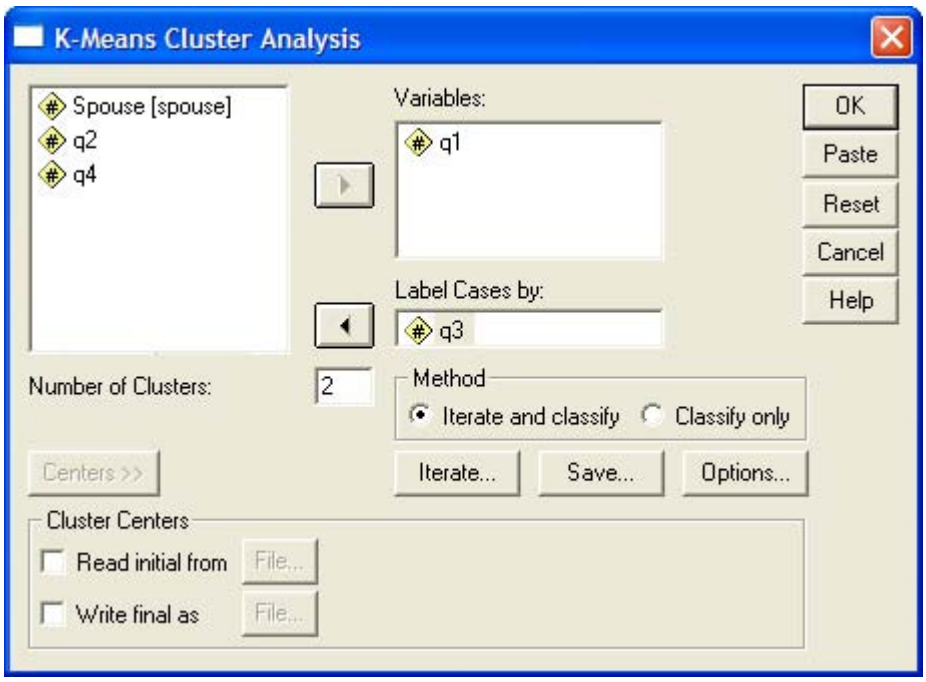

Рис. 20.5: *Диалоговое окно K-Means Cluster Analysis (Анализ кластерных центров)*

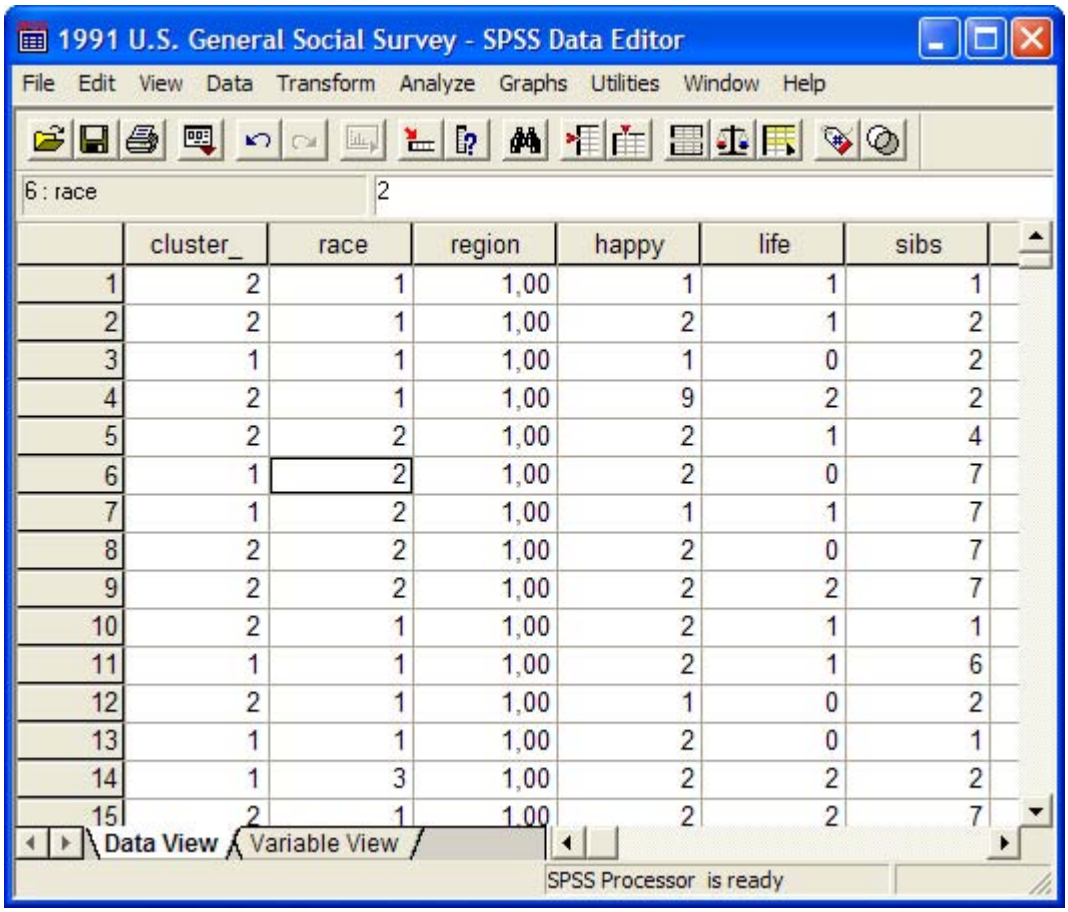

Рис. 20.6: *Файл с первичными кластерными центрами*

Мы надеемся, что при помощи приведенных примеров нам удалось пробудить у Вас интерес к кластерному анализу и облегчить понимание интереснейших статистических методов.