Министерство образования и науки Республики Казахстан

Костанайский государственный университет им. А. Байтурсынова

Кафедра информационных систем

Н.Н. Вардиашвили, Ю.А. Жуненко

АНАЛИЗ ДАННЫХ И ПРОГНОЗИРОВАНИЕ ЭКОНОМИКИ

Практикум
(для студентов экономических специальностей)
Часть 2

Костанай, 2011

УДК 330.46
ББК 65в631
В 18

Авторы: Вардиашвили Н.Н., канд. эконом. наук, профессор
Жуненко Ю.А., преподаватель

Рецензенты:
А. Байманкулов, доцент кафедры «Программное обеспечение» КГУ им. А. Байтурсынова, к.э.н.
Б. Жаналинов, доцент кафедры «Банковское дело» КГУ им. А. Байтурсынова, к. т. н.
В. Бутко, доцент кафедры «Техника и технологии» КСТУ им. З. Алдамжар, к. т. н.

Вардиашвили Н.Н., Жуненко Ю.А.
В 18 Анализ данных и прогнозирование экономики. Практикум по дисциплине для студентов экономических специальностей. Часть 2. - Костанай: КГУ им. А. Байтурсынова, 2011.- 120 с.
ISBN 978-601-7322-15-1

Практикум по дисциплине «Анализ данных и прогнозирование экономики» включает в себя основы теории, типовые задачи, задания для самостоятельной работы по разделу: методы многомерного анализа статистических данных. Предназначен для обучения студентов экономических специальностей постановке, решению экономических задач с использованием многомерных методов статистического анализа, соответствующего программного обеспечения экономических задач и интерпретации результатов решения этих задач.

Может быть использован преподавателями, магистрантами, аспирантами в учебной и научно-исследовательской работе, а также экономистами-практиками.

Утвержден УМС Костанайского государственного университета им. А.Байтурсынова протокол от 16.06.2011 № 7.

ISBN 978-601-7322-15-1

© КГУ им. А. Байтурсынова, 2011
© Вардиашвили Н.Н.
© Жуненко Ю.А.

Содержание

Введение……..... 4
Раздел 1. Многомерные методы статистического анализа 5
Тема 1. Дисперсионный анализ 5
Тема 2. Факторный анализ 29
Тема 3. Кластерный анализ 50
Тема 4. Дискриминантный анализ. 69
Тема 5. Многомерное шкалирование 85
Раздел 2. Пример использования многомерных методов статистического анализа данных при сегментировании рынка. 103
Список источников 109
Приложения 110

Введение

Для полноты анализа экономических явлений и прогнозирования их развития необходимо использовать многомерные методы анализа статистических данных.

Вторая часть практикума по дисциплине «Анализ данных и прогнозирование экономики» призвана решить эту задачу. В практикуме содержится материал последующим многомерным методам статистического анализа экономики: дисперсионному, факторному, кластерному, дискриминантному и многомерному шкалированию. По этим методам имеется достаточное количество учебной литературы, содержащей как теорию, так и примеры решения задач. Однако, основным недостатком такой литературы является то, что в ней не ставятся четко задачи, не приводятся детальные схемы решения задач, недостаточно полно интерпретируются результаты решения и не поясняется, как их использовать в анализе, планировании и прогнозировании, упор больше делается на математическую сторону, нежели на содержательную трактовку.

В этой связи в данном практикуме сделана попытка рассмотреть решение экономических задач с использованием многомерных методов статистического анализа с акцентом на содержательную интерпретацию результатов полученного решения. Поскольку с точки зрения их практического использования не только в области экономики, но и в психологии, социологии, медицине, биологии, археологии и т.д. при наличии соответствующих программных продуктов и умении ими пользоваться от исследователя не требуется досконального знания математико-статистического аппарата, наиболее важным является интерпретация результатов решения задач и определение сфер их приложения.

Каждая тема практикума содержит необходимый теоретический материал, типовую задачу с детальной технологией ее решения с использованием определенного пакета прикладных программ, предназначенного для решения подобных задач, содержательный анализ полученного решения и выводы, задания для самостоятельной работы.

Темы лабораторных работ относительно автономны и могут изучаться и выполняться независимо друг от друга.

Практикум будет полезен не только студентам, но и всем желающим освоить многомерные методы анализа статистических данных с использованием современных программных средств.

Раздел 1. Многомерные методы статистического анализа
Тема 1. Дисперсионный анализ

Цель: изучить основы дисперсионного анализа, приобрести умения и навыки проведения исследования с помощью дисперсионного анализа и интерпретации полученных результатов.

Дисперсионный анализ – это раздел математической статистики, посвященный методам выявления влияния одной или нескольких независимых переменных (факторов) на одну зависимую переменную (одномерный анализ) или на несколько зависимых переменных (многомерный анализ). Факторами могут быть внешние условия, влияющие на результаты физического, производственного, экономического эксперимента, - например, температура, атмосферное давление, тип оборудования, качество сырья, местоположение фирмы и др. В зависимости от количества факторов, включенных в анализ, различают однофакторный анализ, двухфакторный анализ и многофакторный анализ.

Однофакторный дисперсионный анализ.

Однофакторный дисперсионный анализ позволяет проверить гипотезу о значимости влияния какого-либо фактора на значения выборки наблюдений.

Задание 1.

Некоторые профессиональные ассоциации не любят проводить свои ежегодные собрания в Нью-Йорке из-за дороговизны цен проживания в отелях и больших налогов. Действительно ли цены в отелях Нью-Йорка гораздо выше, чем в других крупных городах (США).

При анализе таких данных, мы можем выбирать из двух гипотез:

- Основная гипотеза считает, что данные во всех группах (цены в отелях) идентичны, то есть средние цены равны;
- Альтернативная гипотеза считает, что имеется статистически значимая разница между значениями групп (ценами в разных отелях разных городов).

Используемая в однофакторном дисперсионном анализе модель называется *моделью средних* и выражается следующей формулой:

$$y = \mu_i + \varepsilon,$$

где μ_i - среднее *i*-й группы, ε - случайная ошибка, удовлетворяющая нормальному распределению со средним 0 и дисперсией σ^2. При наличии P групп в модели средних используются следующие формулировки нулевой и альтернативной гипотез:

- Нулевая гипотеза H_0: $\mu_1 = \mu_2 = ... = \mu_p$;
- Альтернативная гипотеза H_a: не все μ_i равны.

Обратите внимание, что предположения в модели средних аналогичны следующим предположениям в анализе регрессии:

- Ошибки нормально распределены;

- Ошибки независимы;
- Ошибки характеризуются постоянством дисперсии.

Сходство с анализом регрессии не случайно. Как показано далее в этом разделе, дисперсионный анализ можно рассматривать как частный случай регрессии (он фактически предваряет регрессионный анализ).

Итак, рассмотрим вопрос о том, действительно ли цены в отелях Нью-Йорка выше, чем в других крупных городах. Для ответа на этот вопрос рассмотрим цены проживания в отелях (доллары США) в четырех крупных городах США: Лос-Анджелесе, Нью-Йорке, Сан-Франциско и Вашингтоне. Для каждого города были случайным образом выбраны данные о восьми отелях из путеводителя. Исходные данные представлены на рисунке 1.

	A	B	C	D
1	Вашингтон	Лос-Анджелес	Нью-Йорк	Сан-Франциско
2	115	119	170	99
3	185	150	135	185
4	166	110	185	265
5	189	79	250	109
6	125	145	250	169
7	64	140	170	99
8	120	165	210	175
9	119	175	215	110

Рисунок 1 – Исходные данные для однофакторного дисперсионного анализа

Для проведения дисперсионного анализа выполним следующие действия.

Технология выполнения:

1. Запустим программу MS Excel.
2. Выберем команду меню *Сервис – Анализ данных*, в открывшемся диалоговом окне *Анализ данных* выберем пункт *Однофакторный дисперсионный анализ*.
3. Заполним диалоговое окно Однофакторный дисперсионный анализ:
   - Входной интервал – диапазон ячеек *A1:D9*;
   - Группирование – *по столбцам*;
   - Отметим флажок – *Метки в первой строке*, поскольку первая строка содержит заголовки столбцов;
   - В разделе *Параметры вывода* – опция *Новый рабочий лист* и введем название – *Дисперсионный анализ цен*.
4. Заполненное окно представлено на рисунке 2.
5. На рисунке 3 представлены результаты дисперсионного анализа.

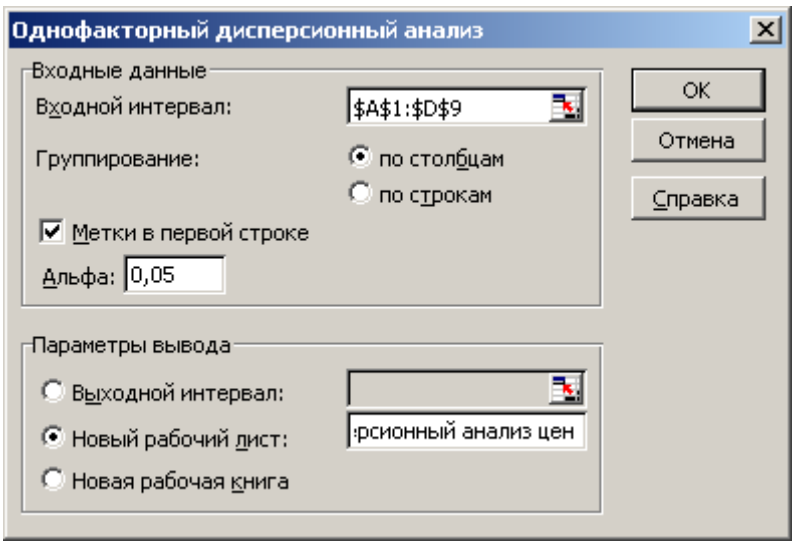


Рисунок 2 – Окно Однофакторный дисперсионный анализ

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	Вашингтон	8	1083	135,375	1771,125		
6	Лос-Анджелес	8	1083	135,375	980,8392857		
7	Нью-Йорк	8	1585	198,125	1649,553571		
8	Сан-Франциско	8	1211	151,375	3414,839286		
9							
10							
11	Дисперсионный анализ						
12	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
13	Между группами	21145,38	3	7048,458	3,607029825	0,025489555	2,946685269
14	Внутри групп	54714,5	28	1954,089			
15							
16	Итого	75859,88	31				

Рисунок 3 – Результаты дисперсионного анализа

Интерпретация результатов дисперсионного анализа

При выполнении дисперсионного анализа следует определить, какая часть изменчивости объясняется случайностью, а какая – другими факторами. Для этого сумма квадратов (т.е. сумма квадратов отклонений от среднего) разбивается на две части: одна объясняется разницей между группами, а другая – случайной ошибкой. Сначала следует напомнить формулу суммы квадратов отклонений от среднего:

Итоговая сумма $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$, (1.1)

Где n – общая сумма наблюдений, $\bar{y}$ - среднее для *всех* наблюдений. Итак, итоговая сумма квадратов отклонений от среднего равна 75859,88 (она

указана в ячейке B16) на рисунке 3, а среднее для всех наблюдений – 155,0625.

Попробуем выразить итоговую сумму квадратов отклонений от среднего в другом виде. Для этого разобьем данные на группы. При наличии *P* групп, где *i*-ая группа содержит n_i наблюдений (группы не обязательно должны быть одинакового размера), подсчитаем общую сумму квадратов отклонений от среднего для каждой группы отдельно:

Итоговая сумма $SS = \sum_{i=1}^{P}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$ (1.1.1).

Здесь переменная y_{ij} обозначает *j*-ое наблюдение в *i*-ой группе (цену проживания в отеле определенного города). Например, y_{23} обозначает третье наблюдение во второй группе (цену в третьем отеле Лос-Анджелеса). Обратите внимание, что сами значения не изменились, а изменился только порядок вычисления общей суммы. Сначала вычисляется сумма квадратов отклонений от среднего в первой группе, затем сумма во второй группе и т.д., после чего суммы по каждой группе складываются для получения общей суммы.

Теперь перейдем к вычислению с использованием выборочного среднего $\bar{y}_i$ для *каждой i*-ой группы. В примере об отелях выборочные средние для городов равны приведенным ниже значениям (ячейки D5:D8).

Лос-Анджелес	135,375
Сан-Франциско	151,375
Вашингтон	135,375
Нью-Йорк	198,125

С помощью выборочных средних для групп можно вычислить общую сумму квадратов внутри каждой группы. Она равна сумме квадратов отклонений, где отклонение вычисляется как разница между средним для группы и значением наблюдения. Сумма квадратов отклонений называется *суммой квадратов ошибок* (*error sum of squares* - *SSE*) и выражается формулой:

$$SSE = \sum_{i=1}^{P}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \qquad (1.1.2)$$

Эта сумма иногда называется внутригрупповой суммой квадратов, так как в ней сумма квадратов подсчитывается отдельно внутри каждой группы. Значение SSE в данном примере равно 54714,5 (в ячейке B14).

Заключительной частью дисперсионного анализа является вычисление суммы квадратов отклонений групповых средних от общего среднего. Эта величина называется *межгрупповой суммой квадратов* или *сумой квадратов по условиям испытаний* (*treatment of squares* - *SST*) и выражается формулой:

$$SST = \sum_{i=1}^{P} n_i(\bar{y}_i - \bar{y})^2 \qquad (1.1.3).$$

Обратите внимание: здесь вычисляется квадрат разницы группового среднего от общего среднего, который умножается на количество наблюде-

ний в группе. В данном примере каждая группа содержит восемь наблюдений, т.е. n_i = 8, а межгрупповая сумма равна 21145,38 (в ячейке B13).

Заметьте также, что общая сумма квадратов равна внутригрупповой сумме квадратов + межгрупповая сумма квадратов, так как 75859,88 = 21145,38 + 54714,50, т.е. итоговая сумма SS = SSE + SST.

Объясним смысл этой формулы на рассматриваемом примере с ценами проживания в отелях разных городов. Если средние цены в разных городах существенно отличаются, то межгрупповая сумма квадратов будет большой, а если они практически одинаковы – будет близка к нулю. И наоборот: большая величина межгрупповой суммы квадратов означает существенную разницу средних цен проживания в разных городах, а малая величина – небольшую разницу.

Большая величина межгрупповой суммы квадратов также может объясняться большим количеством групп, поэтому в таких случаях рекомендуется изучить количество групп в наборе данных, т.е. количество степеней свободы в столбце df (в ячейках C13:C16) в таблице *Дисперсионный анализ*. Количество степеней свободы для фактора Город (который в данном случае является условием разделения на группы) равняется количеству групп минус 1, т.е. 4 – 1 = 3 (в ячейке C13). А количество степеней свободы для общей суммы квадратов равняется общему количеству наблюдений минус 1, т.е. 32 – 1 = 31 (ячейка C16). Остальные степени свободы относятся к ошибке и равны 31 – 3 = 28 (в ячейке C14).

В ячейках D13:D14 столбца *MS* (*среднее квадратическое – mean square - MS*) приводятся значения суммы квадратов, деленной на количество степеней свободы, которые можно рассматривать как дисперсии. Первое значение 7048,458 (в ячейке D13) – это дисперсия цен в разных городах, а второе значение 1954,089 (в ячейке D14) – дисперсия цен в одном городе. Среднее внутригрупповой суммы квадратов также дает оценку σ^2, т.е. дисперсии ошибки ε, которая упоминается при описании модели средних. Если изменчивость цен для разных городов выше изменчивости цен в одном городе, можно предположить, что средняя стоимость проживания в разных городах не одинакова.

Для проверки этого предположения вычислим отношение двух дисперсий. Если это значение больше F-критического, то нулевая гипотеза отвергается.

В примере с данными о стоимости проживания в отелях разных городов F-значение равно 3,61 (в ячейке E13) и удовлетворяет распределению *F(3, 28)* с p-значением 0,025 (в ячейке F13), которое меньше 0,05. Это значение больше табличного, равного 2,95. Таким образом, можно отвергнуть нулевую гипотезу и принять альтернативную гипотезу о том, что средние цены проживания в отелях разных городов отличаются.

Хотя в таблице *Дисперсионный анализ* не приводятся данные о регрессии, их можно вывести. Например, отношение межгрупповой суммы квадратов отклонений к общей сумме квадратов дает значение параметра R^2 (коэф-

фициент детерминации). В данном примере R^2 = 21145,38 / 75859,88 = 0,2787. Таким образом, около 27,9 % изменчивости проживания в отелях объясняется особенностями городов, в которых они находятся.

Однофакторный дисперсионный анализ и анализ регрессии

Дисперсионный анализ можно представить как особую форму анализа регрессии, но с дискретными, а не с непрерывными предикторами. С помощью этой аналогии можно получить дополнительное представление о данных, но для этого нужно переформулировать модель.

Ранее использовалась *модель средних*:

$$y = \mu_i + \varepsilon,$$

для *i*-ой группы. Эквивалентный способ выражения этой взаимосвязи представляет собой *модель влияния*:

$$y = \mu + \alpha_i + \varepsilon,$$

где μ - среднее, α_i - влияние *i*-ой группы, ε - случайная ошибка, удовлетворяющая нормальному распределению со средним 0 и дисперсией σ^2.

Применим это уравнение к примеру с данными об отелях, которые разбиты на четыре группы. Каждая группа содержит данные об одном городе, поэтому согласно модели влияния у нас имеется среднее μ и четыре члена: $\alpha_1, \alpha_2, \alpha_3$ и α_4. Однако в данном случае возникает проблема *перепараметризации*, поскольку модель содержит пять неизвестных параметров при наличии только четырех известных средних (рис. 4). В результате уравнению модели влияния удовлетворяет бесконечное множество решений. Для исправления этого недостатка обычно сокращают количество параметров одним из двух способов. Первый заключается в ограничении значений членов влияния α_i так, чтобы их сумма равнялась нулю. Второй состоит в задании нулевого значения для одного из членов влияния. Применим второй подход для примера с данными о стоимости проживания в отелях разных городов и выполним дисперсионный анализ с помощью анализа регрессии.

Индикаторы.

Для выполнения дисперсионного анализа с помощью анализа регрессии можно создать переменные-индикаторы для данных. Переменные-индикаторы принимают значения 1 или 0 в зависимости от того, относятся данные к определенной группе или нет. Например, можно создать переменную-индикатор, которая принимает значение 1, если наблюдение относится к отелям в Сан-Франциско, или значение 0, если наблюдение не относится к ним.

Создадим переменные-индикаторы как показано на рисунке 4.

	A	B	C	D	E	F
1	Город	Цена	Индикатор (В)	Индикатор (ЛА)	Индикатор (СФ)	Индикатор (Н)
2	Вашингтон	115	1	0	0	0
3	Вашингтон	185	1	0	0	0
4	Вашингтон	166	1	0	0	0
5	Вашингтон	189	1	0	0	0
6	Вашингтон	125	1	0	0	0
7	Вашингтон	64	1	0	0	0
8	Вашингтон	120	1	0	0	0
9	Вашингтон	119	1	0	0	0
10	Лос-Анджелес	119	0	1	0	0
11	Лос-Анджелес	150	0	1	0	0
12	Лос-Анджелес	110	0	1	0	0
13	Лос-Анджелес	79	0	1	0	0
14	Лос-Анджелес	145	0	1	0	0
15	Лос-Анджелес	140	0	1	0	0
16	Лос-Анджелес	165	0	1	0	0
17	Лос-Анджелес	175	0	1	0	0
18	Нью-Йорк	170	0	0	0	1
19	Нью-Йорк	135	0	0	0	1
20	Нью-Йорк	185	0	0	0	1
21	Нью-Йорк	250	0	0	0	1
22	Нью-Йорк	250	0	0	0	1
23	Нью-Йорк	170	0	0	0	1
24	Нью-Йорк	210	0	0	0	1
25	Нью-Йорк	215	0	0	0	1
26	Сан-Франциско	99	0	0	1	0
27	Сан-Франциско	185	0	0	1	0
28	Сан-Франциско	265	0	0	1	0
29	Сан-Франциско	109	0	0	1	0
30	Сан-Франциско	169	0	0	1	0
31	Сан-Франциско	99	0	0	1	0
32	Сан-Франциско	175	0	0	1	0
33	Сан-Франциско	110	0	0	1	0

Рисунок 4 – Представление исходных данных с переменными-индикаторами

Значения в столбце C с подписью *Индикатор (В)* (Вашингтон) равняются 1, если значение в строке относится к отелю в Вашингтоне, или 0 в противном случае. Аналогично, значения в следующих трех столбцах равны 1, если значение в строке относится к отелю в Лос-Анджелесе, Нью-Йорке или Сан-Франциско, или 0 в противном случае.

Подгонка модели влияния.

Попробуем теперь с помощью столбцов с индикаторами подогнать модель влияния в примере с данными о стоимости проживания в отелях разных городов.

Для подгонки модели влияния с помощью анализа регрессии выполним следующие действия:

1. Выберем в меню *Сервис* команду *Анализ данных*.
2. В открывшемся диалоговом окне *Анализ данных* выберем элемент *Регрессия*, щелкнем на кнопке *ОК*.
3. В открывшемся окне Регрессия заполним следующие параметры (рис. 5):

– Входной интервал Y – диапазон `B1:B33`;
– Входной интервал X – `C1:E33`. Следует напомнить, что во избежание перепараметризации модели нужно исключить один член влияния, например с данными об отелях Нью-Йорка (в нашем примере можно было бы исключить данные для любого города).
– Установим флажок *Метки*.
– Выберем переключатель *Новый рабочий лист* и введем название *Модель влияния*.
– Снимем флажки всех параметров в разделе *Остатки*. Щелкнем на кнопке *ОК*.

4. Результаты регрессионного анализа представлены на рисунке 6.

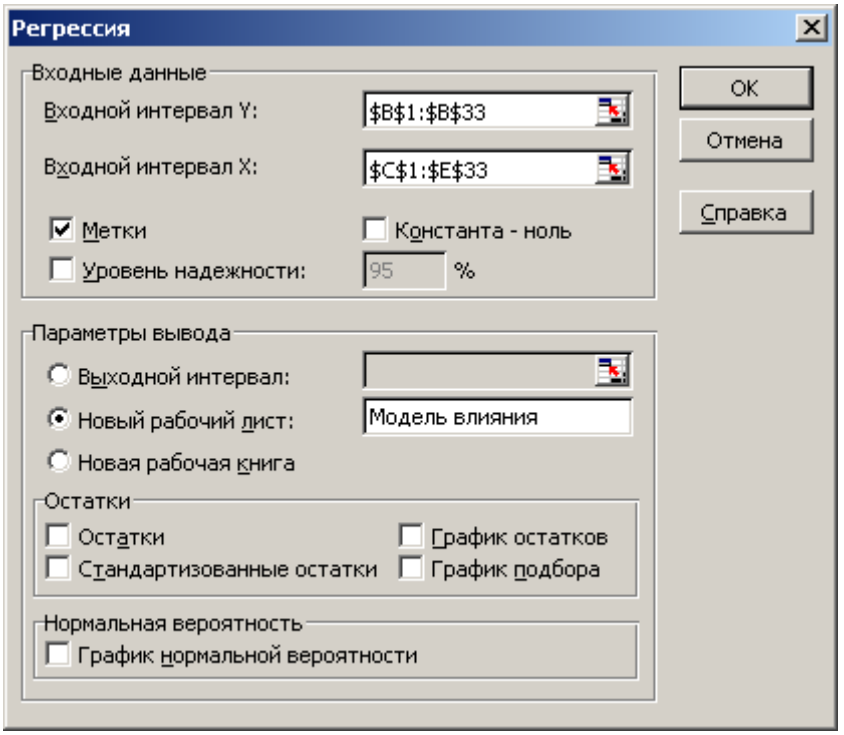


Рисунок 5 – Окно Регрессия

Таблица с результатами дисперсионного анализа (в ячейках `A10:F14`), приведенная на рисунке 6, получена с помощью регрессии, поэтому она эквивалентна таблице дисперсионного анализа (рис. 3), за исключением того, что прежняя строка *Между группами* в данной таблице называется *Регрессия*, а прежняя строка *Внутри групп* называется *Остатки*.

Коэффициент пересечения в ячейке `B17` равняется средней стоимости проживания в отелях Нью-Йорка. Значения в ячейках `B18:B20` содержат отличия от средней стоимости проживания в отелях Нью-Йорка для отелей в других городах.

F-отношение соответствует показанному в дисперсионном анализе

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,527960747							
5	R-квадрат	0,27874255							
6	Нормированный R-квадрат	0,201464966							
7	Стандартная ошибка	44,20508213							
8	Наблюдения	32							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	3	21145,375	7048,458333	3,60702982	0,025489555			
13	Остаток	28	54714,5	1954,089286					
14	Итого	31	75859,875						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	198,125	15,62885667	12,67687101	4,0381E-13	166,1107388	230,1393	166,1107	230,1393
18	Индикатор (В)	-62,75	22,10254106	-2,839040082	0,00833117	-108,0250024	-17,475	-108,025	-17,475
19	Индикатор (ЛА)	-62,75	22,10254106	-2,839040082	0,00833117	-108,0250024	-17,475	-108,025	-17,475
20	Индикатор (СФ)	-46,75	22,10254106	-2,115141416	0,04344743	-92,02500236	-1,475	-92,025	-1,475

Среднее для Нью-Йорка

Отличия от средней стоимости проживания в отелях Нью-Йорка

Рисунок 6 – Результаты регрессионного анализа

Как с помощью переменных-индикаторов создать модель влияния? Рассмотрим значения в столбце *Индикатор (В)* и данными для Вашингтона. Для любых отелей за пределами Вашингтона значение индикатора равно 0, поэтому член влияния умножается на 0 и, следовательно, не имеет никакого влияния на оценку стоимости проживания. Член влияния имеет ненулевое значение только для данных об отелях Вашингтона.

Как видите, с помощью регрессии можно подогнать модель влияния таким образом, чтобы получить практически ту же информацию, которую дает однофакторный дисперсионный анализ.

В рассматриваемой модели предполагается, что средняя цена проживания в одноместном номере отеля в Нью-Йорке значительно выше, чем в одноместном номере отеля в Лос-Анджелесе или Вашингтоне, но не обязательно выше, чем в Сан-Франциско. Получается, что в Нью-Йорке средняя стоимость проживания в одноместном номере отеля составляет 198 долларов, что на 63 доллара выше, чем в Лос-Анджелесе или Вашингтоне, и на 47 долларов выше, чем в Сан-Франциско.

Примечание. Инструмент *Регрессия* модуля *Анализ данных* можно использовать для вычисления модели средних вместо модели влияния. Для этого в текстовом поле *Входной интервал X* диалогового окна *Регрессия* нужно указать все переменные-индикаторы, а затем установить флажок *Константа – ноль* для удаления постоянного члена из уравнения модели. Полученные оценки будут соответствовать средним значениям разных групп.

Задание 2.

Проведите регрессионный анализ для вычисления модели средних (см. примечание).

Как можно использовать результаты однофакторного дисперсионного анализа для прогнозирования?

Если наблюдается существенное различие между средними значениями групп, то лучший прогноз для каждой группы – это просто внутригрупповое среднее значение.

Так, для нашей задачи по анализу цен номеров в отелях разных городов можно спрогнозировать, что цена на одноместный номер в отелях Лос-Анджелесе составит 135,38 долларов, в Сан-Франциско – 151,38 долларов, в Вашингтоне – 135,38 долларов, в Нью-Йорке – 198,13 долларов.

Если существенного различия между средними значениями групп нет, наилучший прогноз – это просто общее среднее.

Помимо проведения однофакторного дисперсионного анализа в MS Excel такой анализ можно проводить в программе Stadia[1].

Задание 3.

Провести однофакторный дисперсионный анализ рассмотренной выше задачи о ценах на номера в отелях с помощью ППП Stadia.

Технология выполнения:

1. Запустим программу Stadia .
2. В открывшемся окне, своим строением похожим на таблицы MS Excel, введем исходные данные как показано на рисунке 7[2].

Таблица данных

	x1	x2	x3	x4
1	DC	LA	NY	SF
2	115	119	170	99
3	185	150	135	185
4	166	110	185	265
5	189	79	250	109
6	125	145	250	169
7	64	140	170	99
8	120	165	210	175
9	119	175	215	110

Рисунок 7 – Пример ввода исходных данных

3. После ввода данных выберем пункт меню *F9 – Статистика*, появится следующее окно (рис. 8) , в котором предоставляются возможные статистические методы для анализа.

[1] STADIA - Пакет прикладных программ, содержащий набор эффективных методов анализа: описательная статистика, критерии различия, категориальный, дисперсионный, корреляционный и спектральный анализ, сглаживание, фильтрация, прогнозирование, простая, множественная, пошаговая и нелинейная регрессия, дискриминантный, кластерный и факторный анализ, шкалирование, методы контроля качества, вычисление и согласие распределений, анализ и замена пропущенных значений и т. д.; полный комплект деловой и научной графики: функции, зависимости, распределения, диаграммы рассеяния, многомерные диаграммы, карты, поверхности, вращения, сплайны, прогнозы, гистограммы, столбиковые, башенные и круговые диаграммы, дендрограммы, установка размеров, надписей по осям и под рисунком, графический архив и проч.

[2] Условные обозначения городов США: DC – Вашингтон, LA – Лос-Анджелес, NY – Нью-Йорк, SF – Сан-Франциско.

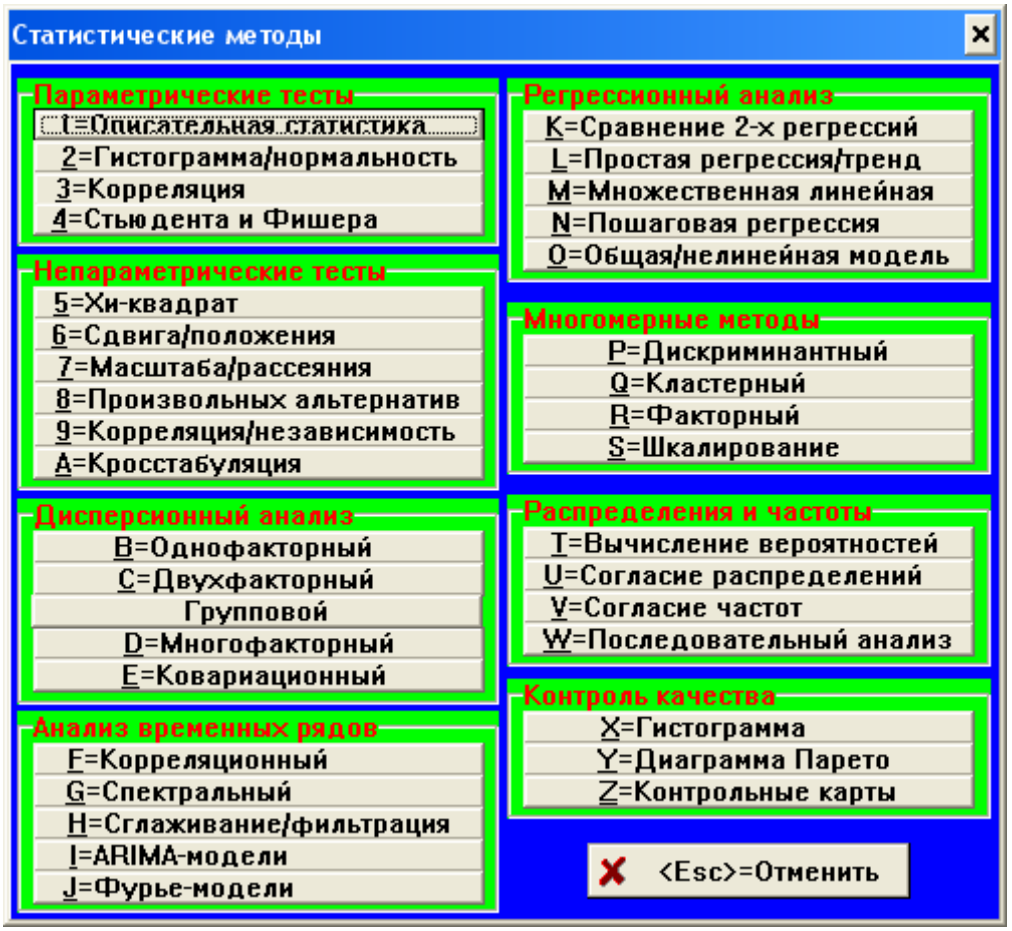


Рисунок 8 – Окно *Статистические методы*

4. В разделе *Дисперсионный анализ* выберем пункт – *Однофакторный*, далее укажем тип анализа - *Параметрический* и в открывшемся окне *Анализ переменных* (рис. 9) укажем переменные, предназначенные для анализа. Нам требуются все четыре переменные, поэтому добавим их по одной с помощью кнопки или все сразу, нажав кнопку Все.

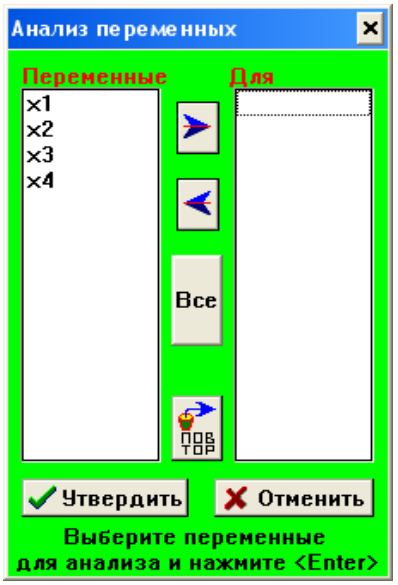


Рисунок 9 – Окно Анализ переменных

5. Нажав на кнопку *Утвердить*, получим результаты, представленные на рисунке 10. Сравните эти результаты с результатами, полученными при расчетах в MS Excel.

Результаты

```
1-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл:
Пропущн=4 4
          параметрический

Источник   Сум.квадр  Ст.своб Ср.квадр      F   Значимость Сила влияния
  Факт.1     2,115E4        3     7048    3,607    0,025 -0,09251
  Остат.     5,471E4       28     1954
  Общая      7,586E4       31     2447

F(фактор1)=3,607, Значимость=0,025, степ.своб = 3,28
   Гипотеза 1: <Есть влияние фактора на отклик>

          Параметры модели:
Среднее = 155,1, доверит.инт.=158,1
Эффект1 = -19,69, доверит.инт.=193,9
Эффект2 = -19,69, доверит.инт.=193,9
Эффект3 = 43,06, доверит.инт.=193,9
Эффект4 = -3,688, доверит.инт.=193,9

             Парные сравнения Шеффе
Переменные  Разность  Интервал    Значим  Гипотеза H1
    1-2           0    65,52         1
    1-3       62,75    65,52    0,0647
    1-4          16    65,52    0,9122
    2-3       62,75    65,52    0,0647
    2-4          16    65,52    0,9122
    3-4       46,75    65,52    0,2375
    3-4       46,75    65,52    0,2375        0    37,83        1
    3-4       46,75    65,52    0,2375        0    37,83        1        0    37,83        1
```

Рисунок 10 – Окно Результаты

В данном окне результатов первая часть таблицы не вызывает трудностей, так как там представлены уже знакомые нам расчеты (см. рис. 3) суммы квадратов отклонений, степеней свободы, значение F-критерия Фишера. В таблице *Параметры модели* в первой строке рассчитано общее среднее (по всем данным), а далее показаны отклонения среднего каждой группы, т.е. среднее цен на проживание в отелях в каждом городе, от общего среднего. Так, например, в строке Эффект1 значение вычислено так: 135,37 – 155,1 = -19,73 (среднее цен в Вашингтоне – общее среднее) и т.д. В таблице *Парные сравнения Шеффе* рассчитаны разности между средними значениями каждой группы. Например, в строке *1-2* значение разности: 0, так как средние цены проживания в отелях первого указанного нами города при вводе данных (Вашингтон) и второго (Лос-Анджелес) – равны; и там и там 135,38.

Двухфакторный дисперсионный анализ

Когда на зависимую переменную могут влиять два фактора, двухфакторный дисперсионный анализ позволяет определить, какой из факторов оказывает, если вообще оказывает, значительное влияние на зависимую пере-

менную. При двухфакторном дисперсионном анализе также можно определить, есть ли значимая взаимосвязь между двумя факторами.

Для двухфакторного дисперсионного анализа можно применить следующее уравнение модели средних:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

где y_{ijk} - переменная отклика, μ_{ij} - среднее на *i*-м уровне одного фактора и *j*-м уровне другого фактора. Для каждой комбинации двух факторов может быть несколько наблюдений, которые называются повторениями. Кроме того, ε_{ijk} - это случайная ошибка на *i*-м уровне одного фактора и *j*-м уровне другого фактора для *k*-го повторения, удовлетворяющая нормальному распределению со средним 0 и дисперсией σ^2.

Обычно для двухфакторного дисперсионного анализа применяется следующая модель влияния:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

где y_{ijk} - переменная отклика, μ - общее среднее, α_i - влияние i-й обработки для одного фактора, β_j - влияние j-й обработки для другого фактора, γ_{ij} - степень взаимодействия двух факторов, т.е. степень их взаимного влияния. Например, разные средства массовой информации (телевидение, журналы и т.д.) могут одинаково или по-разному влиять на объем продаж рекламируемого товара. Если увеличение объема продаж не зависит от средства массовой информации, то взаимодействие факторов равно 0, в противном случае между факторами (затраты на рекламу и вид массовой информации) есть взаимодействие.

Задание 4.

Определить, зависят ли продажи товара от торговых представителей и района, в котором он работает. Предположительно, имеется 4 торговых представителя и пять районов; каждый из торговых представителей отработал по одному месяцу в каждом из пяти разных районов. Исходные данные расположены на рисунке 11.

	A	B	C	D	E
1		ТП1	ТП2	ТП3	ТП4
2	Район 1	1	3	10	12
3	Район 2	17	12	16	14
4	Район 3	17	21	22	25
5	Район 4	20	10	17	23
6	Район 5	22	22	37	32

Рисунок 11 – Исходные данные

Эта модель называется двухфакторным дисперсионным анализом, поскольку на продажи потенциально могут влиять два фактора (район и торговый представитель) и каждый продавец работал в каждом районе только по одному разу.

Технология выполнения:

1. Запустим программу MS Excel.
2. Выберем в меню *Сервис* команду *Анализ данных* и в открывшемся окне выберем пункт *Двухфакторный дисперсионный анализ без повторений*. После этого заполним открывшееся диалоговое окно как показано на рисунке 12.

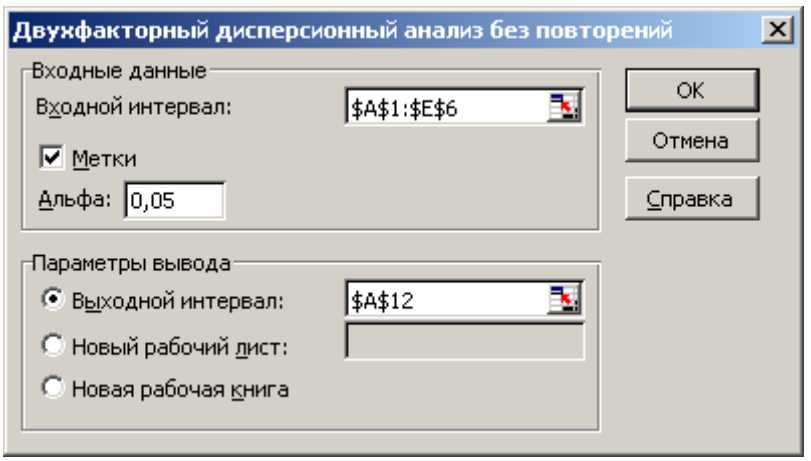


Рисунок 12 – Диалоговое окно для настройки модели без повторений

   - Входной интервал – диапазон ячеек `A1:E6`;
   - Установим флажок *Метки*;
   - Выходной лист – ячейка `A12` в качестве левого верхнего угла выходного интервала.
3. Итоговый вывод результатов показан на рисунке 13. Результаты в ячейках `G12`, `F15:F24` не сгенерированы средством анализа данных Excel. В эти ячейки введены формулы с определенной целью, разъяснения будут приведены ниже.

Итак, для того чтобы определить, оказывает ли значительное влияние на продажи фактор, указанный в строках (район) или в столбцах (торговый представитель), рассмотрим показатели в таблице *Дисперсионный анализ*.

Главным образом, нас интересуют значения в колонке F, которые свидетельствуют о значимости либо незначимости влияющего фактора.

Сначала разберемся со степенями свободы, так как их количественное значение требуется знать для определения табличного значения F-критерия Фишера. Количество степеней свободы для каждого фактора равняется количеству уровней фактора минус 1. Так, например, для фактора *Район*, количество степеней свободы равно 4 (5 районов - 1) (см. ячейку на пересечении строки `Строки` и столбца `df`), для фактора *Торговый представитель* – 3 (4-1) (см. ячейку на пересечении строки `Столбцы` и столбца `df`). Для члена взаимодействия между районами и торговыми представителями количество степеней равно произведению степеней свободы двух факторов, т.е. $4 \times 3 = 12$ (пересечение строки `Погрешность` и столбца `df`). Наконец, существует n-1 или 19 степеней свободы для общей суммы квадратов.

	A	B	C	D	E	F	G
1		ТП1	ТП2	ТП3	ТП4		
2	Район 1	1	3	10	12		
3	Район 2	17	12	16	14		
4	Район 3	17	21	22	25		
5	Район 4	20	10	17	23		
6	Район 5	22	21	37	32		
7							
8							
9							
10							
11							
12	Двухфакторный дисперсионный анализ без повторений						17,6
13							
14	*ИТОГИ*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
15	Район 1	4	26	6,5	28,3333333	-11,1	
16	Район 2	4	59	14,75	4,91666667	-2,85	
17	Район 3	4	85	21,25	10,9166667	3,65	
18	Район 4	4	70	17,5	31	-0,1	
19	Район 5	4	112	28	60,6666667	10,4	
20							
21	ТП1	5	77	15,4	69,3	-2,2	
22	ТП2	5	67	13,4	59,3	-4,2	
23	ТП3	5	102	20,4	104,3	2,8	
24	ТП4	5	106	21,2	67,7	3,6	
25							
26							
27	Дисперсионный анализ						
28	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
29	Строки	1011,3	4	252,825	15,8759812	9,74468E-05	3,259166727
30	Столбцы	216,4	3	72,13333333	4,52956567	0,02409531	3,490294821
31	Погрешность	191,1	12	15,925			
32							
33	Итого	1418,8	19				

Рисунок 13 – Вывод модели двухфакторного дисперсионного анализа без повторений

Фактор *Район* (строки) имеет F-значение 15,87, что больше табличного значения 3,26 (со степенями свободы 4, 12) или это же значение представлено в колонке F-критическое по строке данного фактора.

Фактор *Торговый представитель* имеет расчетное F-значение равным 4,53, что также больше F-критического.

P-значения и того и другого фактора меньше заданного уровня значимости (0,05), т.е. оба фактора оказывают влияние на уровень продаж товара.

Таким образом, по полученным F и P-значениям можно сделать вывод о значимости влияния обоих факторов на результат, т.е. на продажи влияет и район и торговый представитель.

Прогнозирование продаж на основе имеющейся информации о торговом представителе и районе.

Для того чтобы спрогнозировать продажи на месяц, можно использовать следующее уравнение 1:

Спрогнозированные продажи = Общее среднее + (Влияние торгового представителя) + (Влияние района)

В данном уравнении фактор *Влияние торгового представителя* = 0, если он не значим. Если же этот фактор значим, *влияние торгового представителя* вычисляется по формуле: «Среднее значение для конкретного ТП – общее среднее». По аналогии фактор Влияние района = 0, если он не значим. Если же этот фактор значим, его величина вычисляется по формуле: «Среднее значение для конкретного района – Общее среднее».

1. Для вычисления общего среднего, например, в ячейке G12, введем формулу СРЗНАЧ(B2:E6).
2. Влияние торгового представителя и района вычислим путем копирования формулы D15-$G$12 из ячейки F15 в диапазон ячеек F16:F24.
3. Например, спрогнозируем продажи для торгового представителя 4 в районе 2 (согласно уравнению 1): 17,6 + 3,6 – 2,85 = 18,35.
4. Это значение вычислим в ячейке C38 по формуле: G12+F24+F16.
5. Если бы, например, влияние района оказалось бы значимым, а торгового представителя – нет, то формула бы выглядела следующим образом: 17,6 – 2,85 =14,75.

Задание 5.

Имеются данные, показанные на рисунке 14, отображающие сведения о продажах видеоигры за месяц. Для каждой комбинации «цена-реклама» имеется три наблюдения.

Определить, влияет ли изменение цены и затрат на рекламу на продажи видеоигры. Определить, значима ли взаимосвязь между ценой и рекламой.

	A	B	C	D	E	F
2				Цена		
3				Низкая	Средняя	Высокая
4		Расходы на рекламу	Низкие	41	21	10
5				25	20	11
6				23	16	8
7			Средние	28	28	11
8				30	22	22
9				32	18	18
10			Высокие	35	26	21
11				45	40	26
12				47	32	20

Рисунок 14 – Данные о продажах видеоигры

Технология выполнения:

1. Запустим программу MS Excel.
2. Для анализа этих данных выберем в меню *Сервис* команду *Анализ данных*.

3. В открывшемся окне выберем пункт *Двухфакторный дисперсионный анализ с повторениями*.
4. Заполним диалоговое окно (рис. 15).

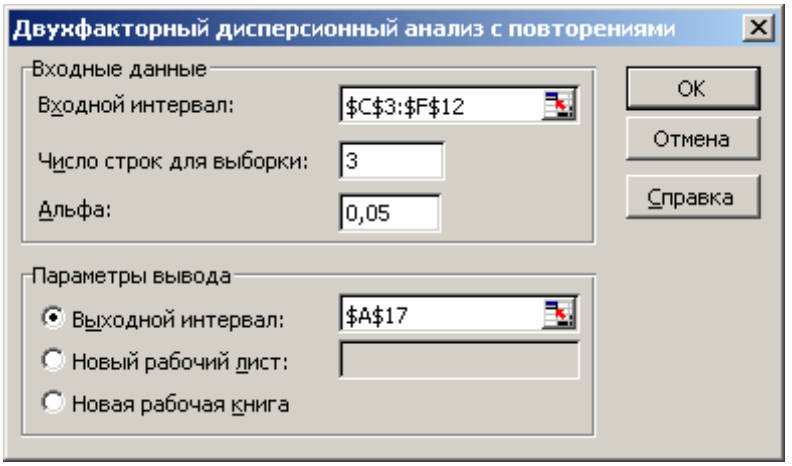


Рисунок 15 – Параметры диалогового окна для проведения процедуры двухфакторного дисперсионного анализа с повторениями

– Входной интервал – диапазон ячеек `C3:F12`, включая заголовки. При двухфакторном дисперсионном анализе с повторениями Excel требует, чтобы для каждого уровня эффекта столбцового фактора в первой строке каждого столбца входного диапазона был определен заголовок. Поэтому мы ввели в ячейки `D3:F3` заголовки *Низкая*, *Средняя* и *Высокая*, чтобы указать возможный уровень цены. Excel также требует, чтобы для каждого уровня эффекта строкового фактора был указан заголовок в первом столбце входного диапазона. Эти заголовки должны находиться в начале строки набора данных, соответствующего каждому уровню. Поэтому мы поместили заголовки, поясняющие низкий, средний и высокий уровень расходов на рекламу в ячейки `C4`, `C7` и `C10`.
– В поле *Число строк для выборки* вводим число 3, поскольку мы располагаем тремя наблюдениями для каждой комбинации уровня цены и рекламы.
– Выходной интервал – укажем ячейку `A17` как левый верхний угол выходного диапазона.
– Нажмем *OK*.

Анализ результатов.

Главным образом, при отображении результатов нас интересует таблица дисперсионный анализ (рис. 16).

	A	B	C	D	E	F	G
44							
45	Дисперсионный анализ						
46	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
47	Выборка	804,962963	2	402,4815	13,51617	0,00026045	3,554557146
48	Столбцы	1405,407407	2	702,7037	23,59826	9,3204E-06	3,554557146
49	Взаимодействие	50,59259259	4	12,64815	0,424751	0,78877694	2,927744173
50	Внутри	536	18	29,77778			
51							
52	Итого	2796,962963	26				

Рисунок 16 – Вывод двухфакторного дисперсионного анализа с повторениями

Количество степеней свободы определяется также как и в дисперсионном анализе без повторений (см. задание 4). То есть для первого фактора (расходы на рекламу) число степеней свободы равно количество уровней фактора минус 1: 3 – 1 = 2. Аналогичным образом рассчитывается количество степеней свободы для фактора *цена*. Для члена взаимодействия между ценой и рекламой количество степеней свободы равно произведению степеней свободы двух факторов, т.е. 2 × 2 = 4. Количество степеней свободы для общей суммы квадратов равно *n* – 1, т.е. 27 – 1 = 26. Для ошибки остается 27 – (2 + 2 + 4) = 18 степеней свободы.

По количеству степеней свободы 2, 18 определяем табличное значение F-критерия Фишера для фактора *расходы на рекламу* (оно равно 3,55, это же значение видно в колонке F-критическое) и сравниваем его с расчетным, оно равно 13,52. Аналогичную процедуру осуществляем со вторым фактором. F-расчетное для фактора *цена*, равное 23,59, также больше табличного значения 3,55. Следовательно, оба рассматриваемых нами фактора влияют на продажи видеоигр.

Теперь выясним, как между собой взаимодействуют факторы, то есть цена и расходы на рекламу. Сравниваем значение F-критерия, полученного при расчете, с табличным и видим, что он меньше (0,42 < 2,93). Вывод: между собой факторы не взаимодействуют.

Эти же выводы можно сделать, глядя на значения параметров в колонке *P-значение* по каждому фактору. Если это значение меньше 0,15, то фактор значим, и наоборот.

Прогнозирование продаж при отсутствии взаимодействия.

При отсутствии значимого взаимодействия между независимыми факторами можно прогнозировать продажи при помощи двухфакторного дисперсионного анализа с повторениями аналогично тому, как мы это делали посредством двухфакторного дисперсионного анализа без повторений. Вот уравнение, которое мы будем использовать (уравнение 2):

Прогнозируемые продажи = Общее среднее + Эффект строки, или рекламы (если значим) + Эффект столбца, или цена (если значим)

Для расчета по данной формуле необходимо вычислить следующие параметры: общее среднее, эффект строки, эффект столбца:

1. Для вычисления общего среднего введем следующую формулу в ячейку, например, D1: `СРЗНАЧ(D4:F12)`.
2. Эффект строки. В ячейках G4, G7 и G10 вычислим эффект для каждого уровня расходов на рекламу, введя формулу (для эффекта низких расходов): `СРЗНАЧ(D4:F6)-$D$1`. Для эффектов средних и высоких расходов формулы аналогичны, введите их самостоятельно.
3. Эффект столбца. В ячейках D13, D14 и D15 вычислим эффект каждого уровня цены по формуле (эффект низкой цены, ячейка D13): `СРЗНАЧ(D4:D12)-$D$1`. Для эффектов средней и высокой цен проделайте аналогичные вычисления самостоятельно.

Теперь предположим, нам требуется спрогнозировать продажи при высоком уровне цены и средних расходах на рекламу (используем уравнение 2): 25,037 + (-1,814) + (-8,704) = 14,52 единиц (то есть общее среднее равно 25,04, эффект среднего уровня расхода на рекламу равен -1,814, эффект высокой цены равен -8,704).

Теперь рассмотрим ***пример***, в котором будет наблюдаться взаимодействие между независимыми факторами, то есть между ценой и расходами на рекламу, и спрогнозируем продажи. Изменим наши исходные данные следующим образом (рис. 17).

	A	B	C	D	E	F
2				Цена		
3				Низкая	Средняя	Высокая
4		Расходы на рекламу	Низкие	41	21	15
5				25	20	14
6				23	16	13
7			Средние	28	28	14
8				30	22	13
9				32	18	12
10			Высокие	50	34	13
11				51	40	13
12				52	32	13

Рисунок 17 – Измененные данные

Проведем процедуру двухфакторного дисперсионного анализа с повторениями, его результат приведен на рисунке 18.

	A	B	C	D	E	F	G
44							
45	Дисперсионный анализ						
46	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
47	Выборка	828,962963	2	414,4815	24,22294	7,8567E-06	3,554557146
48	Столбцы	2498,740741	2	1249,37	73,01515	2,3075E-09	3,554557146
49	Взаимодействие	509,9259259	4	127,4815	7,450216	0,00100645	2,927744173
50	Внутри	308	18	17,11111			
51							
52	Итого	4145,62963	26				

Рисунок 18 – результат вывода двухфакторного дисперсионного анализа с повторениями при наличии взаимодействия

Как видно по значениям колонки F, все расчетные показатели больше табличных, следовательно, наблюдается как влияние цены и расходов на рекламу на продажи, так и взаимодействие между собой цены и расходов на рекламу.

Для прогноза продаж для любой комбинации цены и рекламы необходимо просто вычислить среднее значение трех наблюдений для этой комбинации. Так, например, наилучший прогноз продаж в месяце с высокими расходами на рекламу и средней ценой равен:

$$\frac{34+40+32}{3}=\frac{106}{3}=35{,}33 \text{ единиц.}$$

Теперь рассмотрим проведение двухфакторного дисперсионного анализа без повторений на примере условия задания 4 с использованием программы Stadia.

Технология выполнения:

1. Запустим программу Stadia.
2. Введем исходные данные так, как показано на рисунке 19.

	x1	x2	x3	x4
1	1	3	10	12
2	17	12	16	14
3	17	21	22	25
4	20	10	17	23
5	22	21	37	32

Рисунок 19 – Исходные данные

Обратите внимание, что столбцы соответствуют значениям первого фактора, строки – второго. На пересечении находятся измерения, сделанные при каждом сочетании уровней исследуемых факторов.

3. Выберем пункт меню *F9 – Статистика*, в появившемся окне в разделе *Дисперсионный анализ* выберем *Двухфакторный*, в следующем окне *– Неповторяемый*.
4. Получим следующий результат (см. рис. 20):
5. Сравните результат решения с полученным ранее (см. решение задания 4).

Результаты

```
2-ФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ.  Файл:
          Факторный план: неповторяемый

Источник    Сум.квадр  Ст.своб Ср.квадр      F    Значимость Сила влияния
  Факт.1        216,4         3    72,13     4,53    0,0238    0,1237
  Факт.2         1011         4    252,8    15,88    0,0002     0,931
  Остат.        191,1        12    15,93
  Общая          1419        19    74,67

F(фактор1)=4,53, Значимость=0,0238, степ.своб = 3,12
   Гипотеза 1: <Есть влияние фактора на отклик>
F(фактор2)=15,88, Значимость=0,0002, степ.своб = 4,12
   Гипотеза 1: <Есть влияние фактора на отклик>

          Параметры модели:
Среднее = 17,6, доверит.инт.=2,433
Эффект1-1 = -2,2, доверит.инт.=6,277
Эффект1-2 = -4,2, доверит.инт.=6,277
Эффект1-3 = 2,8, доверит.инт.=6,277
Эффект1-4 = 3,6, доверит.инт.=6,277
Эффект2-1 = -11,1, доверит.инт.=6,297
Эффект2-2 = -2,85, доверит.инт.=6,297
Эффект2-3 = 3,65, доверит.инт.=6,297
Эффект2-4 = -0,1, доверит.инт.=6,297
Эффект2-5 = 10,4, доверит.инт.=6,297
```

Рисунок 20 – Окно вывода результатов двухфакторного дисперсионного анализа без повторений

Задание 6.

Книжное издательство хотело бы знать, влияет ли расположение его книг в отделах компьютерной литературы книжных магазинов на продажи. Определить, действительно ли важно, что книги стоят в начале, в середине или в конце отдела компьютерной литературы, если наблюдались еженедельные продажи книг (в сотнях экземпляров) в 12 магазинах. В 5 магазинах книги размещались в передней части отдела, в 4 – в конце и в 3 – в середине (исходные данные приведены на рисунках 21, 22).

Вариант 1.

	A	B	C	D
1				
2				
3		Начало отдела	Конец отдела	Середина
4		7	12	10
5		10	13	11
6		8	15	12
7		9	18	
8		11		

Рисунок 21 –Данные о продажах книг

Вариант 2.

	A	B	C	D
1				
2				
3		Начало отдела	Конец отдела	Середина
4		7	2	3
5		20	16	19
6		8	25	11
7		8	13	
8		2		

Рисунок 22 – Данные о продаже книг

Задание 7.

Три группы продавцов продавали один и тот же штучный товар, расфасованный в различные упаковки. После окончания срока распродажи был произведен тестовый контроль над случайно отобранными продавцами из каждой группы. При этом были получены следующие результаты (рис. 23). Определить, влияет ли вид упаковки на уровень продаж.

	A	B	C	D	E	F	G	H	I
1									
2		Номер группы	Число продаж, которые сделали продавцы						
3		1	1	3	2	1	0	2	1
4		2	2	3	2	1	4		
5		3	4	5	3				

Рисунок 23 – Исходные данные к заданию 7

Задание 8.

В течение шести лет использовались пять различных технологий выращивания свеклы. Установите, влияет ли применение различных технологий на урожайность свеклы. Данные о достигнутой урожайности (тыс. ц/га) приведены на рисунке 24.

	A	B	C	D	E	F
1		Технология				
2	Год	A1	A2	A3	A4	A5
3	1	1,2	0,6	0,9	1,7	1
4	2	1,1	1,1	0,6	1,4	1,4
5	3	1	0,8	0,8	1,3	1,1
6	4	1,3	0,7	1,1	1,5	0,9
7	5		0,7		1,2	1,2
8	6		0,7		1,3	

Рисунок 24 – Исходные данные по урожайности свеклы при разных технологиях возделывания

Задание 9.

Предполагается, что на результат производственного процесса влияют давление (высокое, среднее или низкое) и температура (высокая, средняя или низкая). На основе данных, приведенных на рисунке 25, определите, как температура и / или давление влияют на результат.

	A	B	C	D	E	F
1						
2					Давление	
3			Температура	Низкое	Среднее	Высокое
4			Низкая	15	13	8
5			Средняя	18	16	11
6			Высокая	7	4	1

Рисунок 25 – Исходные данные

Задание 10.

Определить, как конкретный торговый представитель и число деловых звонков докторам влияют на продажи лекарства (тыс.долл.), прописанного каждым доктором (рис. 26).

	A	B	C	D	E	F	G
1							
2				ТП 1	ТП 2	ТП 3	ТП 4
3			1	2	3	4	3
4		Число звонков		4	5	5	12
5			3	4	5	4	6
6				6	7	4	6
7			5	8	10	5	5
8				10	13	5	5

Рисунок 26 – Исходные данные

Задание 11.

Имеются данные об урожайности (ц/га) четырех сортов пшеницы, достигнутой при использовании пяти типов удобрений. Данные получены на 20 участках одинакового размера аналогичного почвенного состава. Необходимо определить, влияет ли сорт пшеницы и тип удобрения на урожайность пшеницы. Исходные данные представлены на рисунке 27.

	A	B	C	D	E
1		Сорт пшеницы			
2	Тип удобрения	A1	A2	A3	A4
3	1	19	25	17	21
4	2	22	19	19	18
5	3	26	23	22	25
6	4	18	26	20	23
7	5	21	22	21	24

Рисунок 27 – Исходные данные к заданию 11

Задание 12.

Требуется исследовать влияние артикула материала, из которого изготавливалась партия товара, и фасона изделия на сбыт продукции с целью определения вкусов покупателей в трех магазинах. Выручка от реализации продукции (тыс. усл. ден. ед.) приведена в таблице на рисунке 28.

	A	B	C	D	E	F
10						
11			Артикул			
12	Номер магазина	Фасон изделия	х/б	шерсть	синтетика	х/б и шерсть
13	1	Цельнокроенный	2,9	9,8	8,9	7,8
14	2		4,1	12,2	3,1	8,8
15	3		4,9	10,1	5,1	7,9
16	1	Прилегающий	23,9	7,9	8,7	3,1
17	2		7,9	8,8	8,1	3,2
18	3		22	9,1	16,8	3,2
19	1	Прямой	28,8	15,9	10,8	3,1
20	2		8,8	5,8	6,4	9,7
21	3		27,7	10,7	16,3	6,9

Рисунок 28 – Данные для задания 12

Тема 2. Факторный анализ

Цель: изучить основы факторного анализа, приобрести умения и навыки проведения исследования с помощью факторного анализа и интерпретации полученных результатов.

Факторный анализ - это процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводится к меньшему количеству независимых влияющих величин, называемых факторами. При этом в один фактор объединяются переменные, сильно коррелирующие между собой. Переменные из разных факторов слабо коррелируют между собой. Таким образом, целью факторного анализа является нахождение таких комплексных факторов, которые как можно более полно объясняют наблюдаемые связи между переменными, имеющимися в наличии.

Или, другими словами, главными целями факторного анализа являются: (1) сокращение числа переменных (редукция данных) и (2) определение структуры взаимосвязей между переменными, т.е. классификация переменных. Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации.

Модель факторного анализа имеет вид:

$$X_i = \sum_{j=1}^{p} \lambda_{ij} F_j + e_i \ ,$$

где λ_{ij} – постоянные величины, называемые факторными нагрузками,

F_j – общие факторы, используемые для представления всех *p* исходных переменных,

e_i – специфические факторы, уникальные для каждой переменной, $p \leq m$.

Задачами факторного анализа являются: определение числа общих факторов, определение оценок λ, определение общих и специфических факторов.

Факторный анализ как метод редукции данных

Предположим, что мы измеряем рост ста людей в дюймах и сантиметрах. Следовательно, имеем две переменные x_1 и x_2. Если далее необходимо исследовать, например, влияние различных пищевых добавок на рост, то очевидно, что для этого обе переменные не нужны, так как рост является одной характеристикой человека, независимо от того, в каких единицах он измеряется, то есть выбирается одна из этих переменных: x_1 или x_2.

Теперь предположим, что необходимо измерить удовлетворенность людей жизнью, для чего составляется вопросник. Добавим в перечень вопросов следующие: 1. Имеете ли вы хобби (x_1). 2. Удовлетворены ли вы своим хобби (x_2). Результаты преобразуются так, что средние ответы (например, для удовлетворенности) соответствуют значению 100, в то время как ниже и вы-

ше средних ответов расположены меньшие и большие значения, соответственно. Здесь также, как и в предыдущем примере, эти две переменные (x_1 и x_2) сильно коррелированы между собой. Из высокой коррелированности следует их избыточность, то есть вместо x_1 и x_2 необходимо ввести одну переменную, например, x_3 – есть ли у вас хобби?

Объединение двух переменных в один фактор.

Зависимость между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная путем подгонки линия регрессии дает графическое представление зависимости. Если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме, то такая переменная будет включать в себя наиболее существенные черты обеих переменных. Таким образом, сокращается число переменных и, в нашем случае, две заменяются одной. Новый фактор (переменная) в действительности является линейной комбинацией двух исходных переменных.

Анализ главных компонент. Выделение главных компонент.

Пример, в котором две коррелированные переменные объединены в один фактор, показывает главную идею факторного анализа или, более точно, анализа главных компонент. Если пример с двумя переменными распространить на большее число переменных, то вычисления становятся сложнее, однако основной принцип представления двух или более зависимых переменных одним фактором остается в силе.

В основном процедура выделения главных компонент подобна вращению, максимизирующему дисперсию (варимакс) исходного пространства переменных. Например, на диаграмме рассеяния можно рассматривать линию регрессии как ось X, повернув ее так, чтобы она совпала с прямой регрессии. Этот тип вращения называется вращением, максимизирующим дисперсию, так как критерий (цель) вращения заключается в максимизации дисперсии (изменчивости) «новой» переменной (фактора) и минимизации разброса вокруг нее.

Посмотрим на следующий рисунок (рис. 29). На нем изображено двумерное пространство наблюдений в осях *X* и *Y*, соответствующих двум измеряемым параметрам.

В новой системе координат на оси OX′ и OY′ отложены отрезки OA и OB соответственно единичной длины. Их проекции на старой координатной оси OX и OY будут соответственно:

$$\begin{cases} a_1 = \cos\alpha, \\ b_1 = \sin\alpha, \end{cases} \qquad \begin{cases} a_2 = -\sin\alpha \\ b_2 = \cos\alpha \end{cases}$$

где α – угол поворота против часовой стрелки.

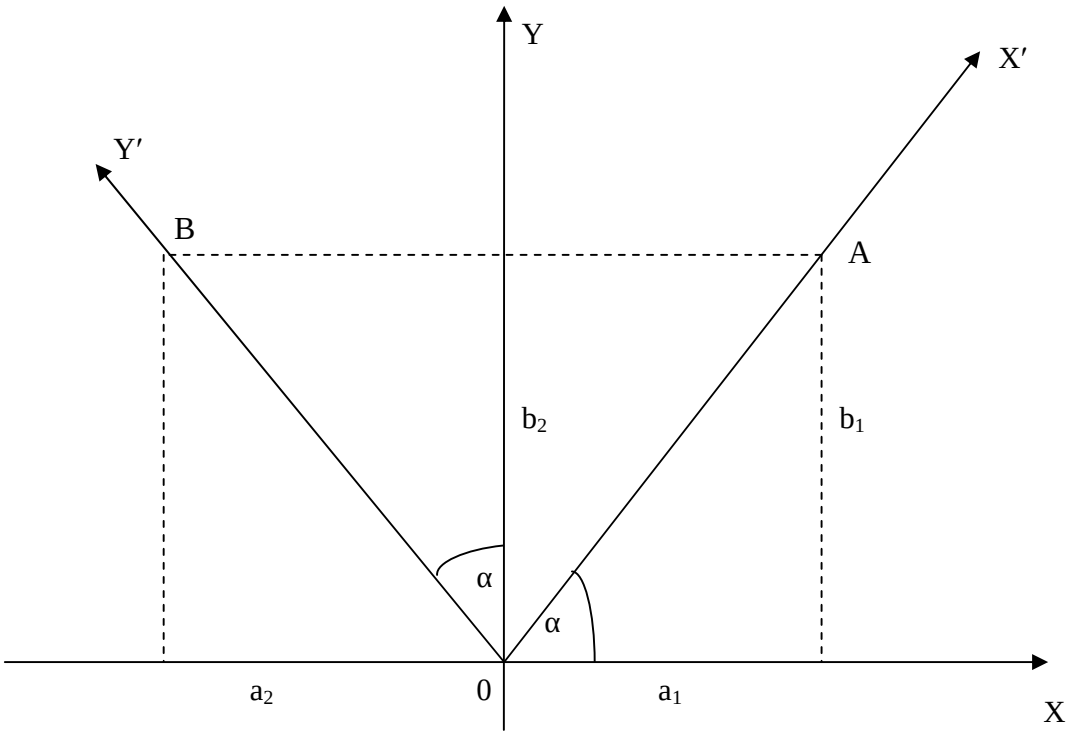


Рисунок 29 – Пример вращения системы координат

В том случае, когда имеются более двух переменных, можно считать, что они определяют трехмерное «пространство» точно так же, как две переменные определяют плоскость. Для случая более трех переменных становится невозможным представить точки на диаграмме рассеяния, однако логика вращения осей с целью максимизации дисперсии нового фактора остается прежней.

После того, как найдена линия, для которой дисперсия максимальна, вокруг нее остается некоторый разброс данных. И процедуру естественно повторить. В анализе главных компонент именно так и делается: после того, как первый фактор выделен, то есть, после того, как первая линия проведена, определяется следующая линия, максимизирующая остаточную вариацию (разброс данных вокруг первой прямой), и т.д. Таким образом, факторы последовательно выделяются один за другим. Так как каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, то факторы оказываются независимыми друг от друга. Другими словами, некоррелированными или *ортогональными*.

При выделении факторов возникает вопрос: сколько факторов выделять? Отметим, что в процессе последовательного выделения факторов они включают в себя все меньше и меньше изменчивости. Решение о том, когда следует остановить процедуру выделения факторов, главным образом зависит от точки зрения на то, что считать малой «случайной» изменчивостью. Это решение достаточно произвольно, однако имеются некоторые рекомендации, позволяющие рационально выбрать число факторов.

Посмотрим теперь на некоторые стандартные результаты анализа главных компонент. При повторных итерациях выделяются факторы с все меньшей и меньшей дисперсией. Для простоты изложения считаем, что обычно работа начинается с матрицы, в которой дисперсии всех переменных равны *1.0*. Поэтому общая дисперсия равна числу переменных. Например, если имеется 10 переменных, каждая из которых имеет дисперсию *1*, то наибольшая изменчивость, которая потенциально может быть выделена, равна 10 раз по 1. Предположим, что при изучении степени удовлетворенности жизнью включено 10 пунктов для измерения различных аспектов удовлетворенности домашней жизнью и работой. Дисперсия, объясненная последовательными факторами, представлена в таблице 1.

Таблица 1 – Факторный анализ (Дисперсия)

Наименование	**Собственные значения Выделение: Главные компоненты**			
Значение	**Собственные значения**	**% общей дисперсии**	**Кумулятивные собственные значения**	**Кумулятивная дисперсия, %**
1	**2**	**3**	**4**	**5**
1	6.118369	61.18369	6.11837	61.1837
2	1.800682	18.00682	7.91905	79.1905
3	0.472888	4.72888	8.39194	83.9194
4	0.407996	4.07996	8.79993	87.9993
5	0.317222	3.17222	9.11716	91.1716
6	0.293300	2.93300	9.41046	94.1046
7	0.195808	1.95808	9.60626	96.0626
8	0.170431	1.70431	9.77670	97.7670
9	0.137970	1.37970	9.91467	99.1467
10	0.085334	0.85334	10.00000	100.0000

Собственные значения

Во втором столбце (*Собственные значения)* таблицы результатов можно найти дисперсию нового (выделенного) фактора. В третьем столбце для каждого фактора приводится процент от общей дисперсии (в данном примере она равна *10*) для каждого фактора. Как можно видеть, первый фактор (значение 1) объясняет 61 процент общей дисперсии, фактор 2 (значение 2) - 18 процентов, и т.д. Четвертый столбец содержит накопленные или кумулятивные собственные значения, рассчитанные как последовательная сумма собственных значений второго столбца. Дисперсии, выделяемые факторами, названы *собственными значениями*. Это название происходит из использованного способа вычисления.

Собственные значения и задача о числе факторов

Как только получена информация о том, сколько дисперсии выделил каждый фактор, можно возвратиться к вопросу о том, сколько факторов следует оставить. Как говорилось выше, по своей природе это решение произ-

вольно. Однако имеются некоторые общеупотребительные рекомендации, и на практике следование им дает наилучшие результаты.

Критерий Кайзера. Сначала можно отобрать только факторы с собственными значениями, большими *1*. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является, вероятно, наиболее широко используемым. В приведенном выше примере на основе этого критерия следует сохранить только первые 2 фактора (две главные компоненты).

Критерий каменистой осыпи. Критерий каменистой осыпи является графическим методом, впервые предложенным Кэттелем (Cattell, 1966). Можно изобразить собственные значения, представленные в таблице ранее, в виде простого графика (рис. 30).

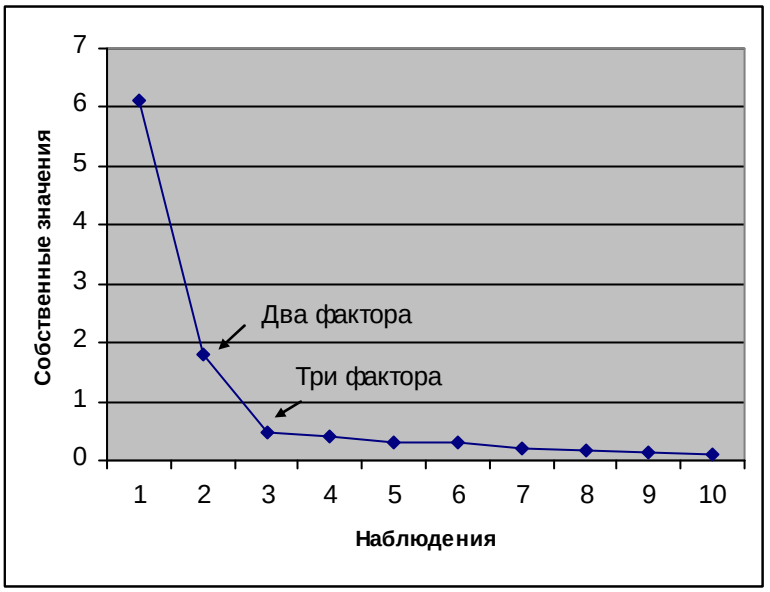


Рисунок 30 – График, отображающий графически использование критерия каменистой осыпи

Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь" - "осыпь" является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. В соответствии с этим критерием можно оставить в этом примере 2 или 3 фактора.

Какой критерий следует использовать. Оба критерия были изучены подробно Брауном (Browne, 1968), Кэттелем и Джасперсом (Cattell, Jaspers, 1967), Хакстианом, Рожерсом и Кэттелем (Hakstian, Rogers, Cattell, 1982), Линном (Linn, 1968), Тюкером, Купманом и Линном (Tucker, Koopman, Linn, 1969). Теоретически, можно вычислить их характеристики путем генерации случайных данных для конкретного числа факторов. Тогда можно увидеть, обнаружено с помощью используемого критерия достаточно точное число существенных факторов или нет. С использованием этого общего метода первый критерий (*критерий Кайзера*) иногда сохраняет слишком много фак-

торов, в то время как второй критерий (*критерий каменистой осыпи*) иногда сохраняет слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный дополнительный вопрос, а именно: когда полученное решение может быть содержательно интерпретировано. Поэтому обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее «осмысленное».

Анализ главных факторов

Прежде, чем продолжить рассмотрение различных аспектов вывода анализа главных компонент, введем анализ главных факторов. Вернемся к примеру вопросника об удовлетворенности жизнью, чтобы сформулировать другую «мыслимую модель». Можно представить себе, что ответы субъектов зависят от двух компонент. Сначала выбираем некоторые подходящие общие факторы, такие как, например, "удовлетворение своим хобби", рассмотренные ранее. Каждый пункт измеряет некоторую часть этого общего аспекта удовлетворения. Кроме того, каждый пункт включает уникальный аспект удовлетворения, не характерный для любого другого пункта.

Главные факторы в сравнении с главными компонентами.

Основное различие двух моделей факторного анализа состоит в том, что в анализе главных компонент предполагается, что должна быть использована *вся* изменчивость переменных, тогда как в анализе главных факторов используется только изменчивость переменной, общая и для других переменных. В большинстве случаев эти два метода приводят к весьма близким результатам. Однако анализ главных компонент часто более предпочтителен как метод сокращения данных, в то время как анализ главных факторов лучше применять с целью определения структуры данных.

Факторный анализ как метод классификации

Возвратимся к интерпретации результатов факторного анализа. Таким образом, термин *факторный анализ* будет включать как анализ главных компонент, так и анализ главных факторов. Предположим, что в выбранной точке анализа, когда в целом известно, сколько факторов следует выделить, необходимо узнать значимость факторов, то есть, можно ли интерпретировать их разумным образом и как это сделать. Чтобы проиллюстрировать, каким образом это может быть сделано, производятся действия «в обратном порядке», то есть, начинают с некоторой осмысленной структуры, а затем смотрят, как она отражается на результатах. Вернемся к примеру об удовлетворенности людей жизнью. Допустим, в результатах расчета получена корреляционная матрица для переменных, относящихся к удовлетворенности на работе и дома (табл. 2).

Таблица 2 – Факторный анализ (Корреляции).

Переменная	РАБОТА_1	РАБОТА_2	РАБОТА_3	ДОМ_1	ДОМ_2	ДОМ_3
РАБОТА_1	1.00	0.65	0.65	0.14	0.15	0.14
РАБОТА_2	0.65	1.00	0.73	0.14	0.18	0.24
РАБОТА_3	0.65	0.73	1.00	0.16	0.24	0.25
ДОМ_1	0.14	0.14	0.16	1.00	0.66	0.59
ДОМ_2	0.15	0.18	0.24	0.66	1.00	0.73
ДОМ_3	0.14	0.24	0.25	0.59	0.73	1.00

Переменные, относящиеся к удовлетворенности на работе, более коррелированы между собой, а переменные, относящиеся к удовлетворенности домом, также более коррелированы между собой. Корреляции между этими двумя типами переменных (переменные, связанные с удовлетворенностью на работе, и переменные, связанные с удовлетворенностью домом) сравнительно малы. Поэтому кажется правдоподобным, что имеются два относительно независимых фактора (два типа факторов), отраженных в корреляционной матрице: один относится к удовлетворенности на работе, а другой к удовлетворенности домашней жизнью.

Факторные нагрузки. Теперь проведем анализ главных компонент и рассмотрим решение с двумя факторами. Для этого рассмотрим корреляции между переменными и двумя факторами (или «новыми» переменными), как они были выделены по умолчанию; эти корреляции называются *факторными нагрузками* (табл. 3).

Таблица 3 – Факторный анализ (Факторные нагрузки).

Наименование	Факторные нагрузки (Нет вращения) Главные компоненты	
Переменная	**Фактор 1**	**Фактор 2**
РАБОТА_1	0.654384	0.564143
РАБОТА_2	0.715256	0.541444
РАБОТА_3	0.741688	0.508212
ДОМ_1	0.634120	-0.563123
ДОМ_2	0.706267	-0.572658
ДОМ_3	0.707446	-0.525602
Общая дисперсия	2.891313	1.791000
Доля общей дисп.	0.481885	0.298500

Каждая из переменных взаимодействует с каждым фактором. Из таблицы 3 видно, что все переменные более коррелируют с первым фактором. Этого следовало ожидать, потому что, как было сказано выше, факторы выделяются последовательно и содержат все меньше и меньше общей дисперсии. Однако при таких значениях факторов нельзя классифицировать переменные. Для получения более понятного результата проводится вращение факторов.

Вращение факторной структуры. Можно изобразить факторные нагрузки в виде диаграммы рассеяния. На этой диаграмме каждая переменная представлена точкой. Можно повернуть оси в любом направлении без изменения *относительного* положения точек; однако действительные координаты точек, то есть факторные нагрузки, должны, без сомнения, меняться. Если построить диаграмму для этого примера, то можно увидеть, что если повернуть оси относительно начала координат на 45 градусов, то можно достичь ясного представления о нагрузках, определяющих переменные: удовлетворенность на работе и дома (рис. 31).

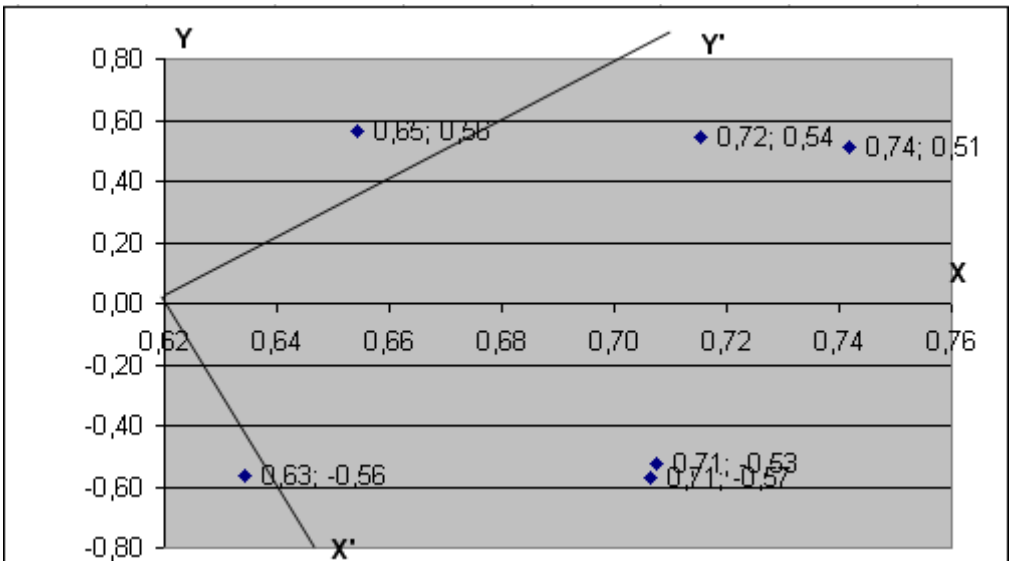


Рисунок 31 – Диаграмма рассеяния

Методы вращения. Существуют различные методы вращения факторов. Целью этих методов является получение понятной (интерпретируемой) матрицы нагрузок, то есть факторов, которые ясно отмечены высокими нагрузками для некоторых переменных и низкими - для других. Эту общую модель иногда называют *простой структурой*. Типичными методами вращения являются стратегии *варимакс*, *квартимакс*, и *эквимакс*.

Ниже приведена таблица нагрузок на факторы, возникшие в результате поворота осей (табл. 4).

Таблица 4 – Факторный анализ (Вращение факторов).

Наименование	**Факторные нагрузки (Варимакс нормализ.) Выделение: Главные компоненты**	
Переменная	**Фактор 1**	**Фактор 2**
РАБОТА_1	0.862443	0.051643
РАБОТА_2	0.890267	0.110351
РАБОТА_3	0.886055	0.152603
ДОМ_1	0.062145	0.845786
ДОМ_2	0.107230	0.902913
ДОМ_3	0.140876	0.869995
Общая дисперсия	2.356684	2.325629
Доля общей дисп.	0.392781	0.387605

Интерпретация факторной структуры. Теперь картина становится более ясной. Как и ожидалось, первый фактор отмечен высокими нагрузками на переменные, связанные с удовлетворенностью на работе, а второй фактор - с удовлетворенностью домом. Из этого можно заключить, что удовлетворенность, измеренная вопросником, составлена из двух частей: удовлетворенность домом и работой, следовательно, произведена *классификация* переменных.

Процедуру проведения факторного анализа рассмотрим на примере.

Задание 1.

Рассматриваются показатели, характеризующие уровень жизни населения в Казахстане за ряд лет. Они представлены в таблице 5.

Требуется, осуществив процедуру факторного анализа, определить число факторов, которые более полно объясняют взаимосвязи между переменными, сократив при этом общее количество переменных. Рассмотреть несколько вариантов выбора числа факторов, дать окончательно выбранным факторам семантически[1] значимые названия. Интерпретировать получаемые результаты.

Технология выполнения:

1. Запустим ППП Stadia. Факторный анализ будем проводить с использованием данного пакета прикладных программ.
2. В открывшееся окно введем исходные данные из таблицы 5, начиная с ячейки B3.
3. После ввода данных выберем в меню пункт *Статист=F9*. Откроется окно, предлагающее выбрать статистический метод (рис. 8, см. тему 1). Выберем *Факторный анализ*.
4. Откроется следующее окно (рис. 32)
5. С помощью кнопки *Все* перенесем переменные в правую часть окна.
6. Нажмем кнопку *Утвердить*.
7. Появится окно, изображенное на рисунке 33. В данном окне уточняется характер матрицы исходных данных: 1 – данные введены в виде m-переменных с n-измерениями или объектами, 2 – данные введены в виде матрицы корреляции между m-переменными. Здесь необходимо выбрать метод анализа: корреляций или ковариаций. Укажем характер матрицы исходных данных – 1, метод анализа – корреляций.

[1] Семантика - наука о понимании определённых знаков, последовательностей символов и других условных обозначений. Дать семантически значимое название – определить смысловое значение, характеризующее набор переменных.

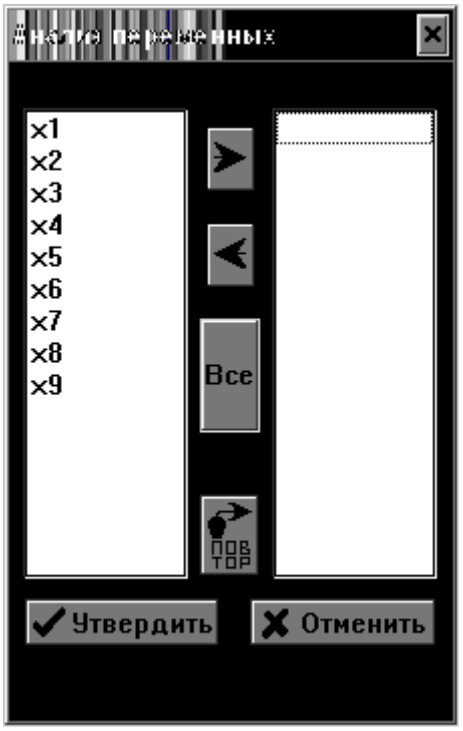


Рисунок 32 – Окно Анализ данных

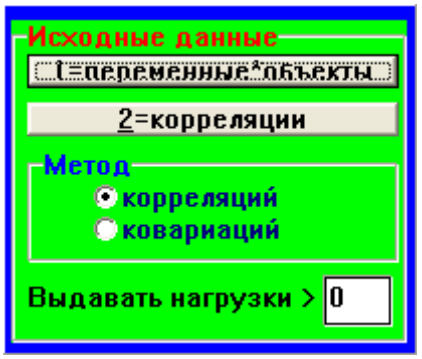
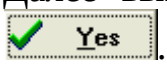


Рисунок 33 – Окно выбора исходных данных

8. Далее выводится запрос, представленный на рисунке 34. Нажмем 

.

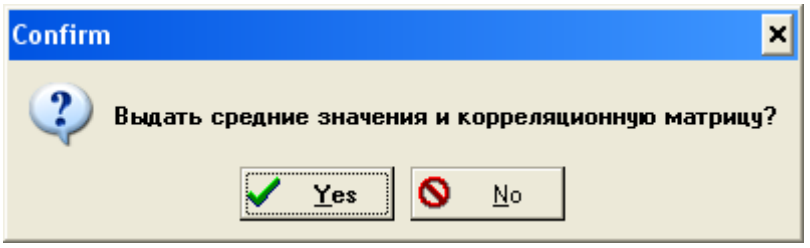


Рисунок 34

Таблица 5 – Исходные данные для проведения факторного анализа[1]

	A	B	C	D	E	F	G	H	I	J
1	Годы	Среднемесячная номинальная з/п, тенге	Величина прожит_минимума (в ср.на душу населения), тенге	Стоимость продовольств_ корзины, тенге	Сред_размер назначенной месячной пенсии, тенге	Ожидаемая продолжит_жизни с даты рождения, лет	Общий коэф-т рожд-ти на 1000 человек	Общий коэф-смертности на 1000 человек	Общий коэф-т брачности на 1000 человек	Общий коэф-т разводимости на 1000 человек
2		x1	x2	x3	x4	x5	x6	x7	x8	x9
3	1996	6841	3000	2000	3283	63,6	16,3	10,7	6,6	2,6
4	1997	8541	3120	2454	3554	64	15,2	10,4	6,6	2,3
5	1998	9683	3336	2601	4213	64,5	14,8	10,2	6,4	2,4
6	1999	11864	3394	2376	4270	65,5	14,2	9,8	5,8	1,7
7	2000	14374	4007	2805	4462	65,4	14,6	10	6,1	1,8
8	2001	17303	5655	3393	4947	65,8	14,91	9,95	6,25	1,99
9	2002	20323	6003	3602	5818	66	15,29	10,05	6,66	2,1
10	2003	23128	6457	3874	8198	65,83	16,63	10,41	7,41	2,13
11	2004	28329	6785	4071	8628	66,18	18,19	10,14	7,64	2,1
12	2005	34060	7618	4571	9061	65,91	18,42	10,37	8,12	2,14
13	2006	40790	8410	5046	9898	66,19	19,71	10,27	8,96	2,34
14	2007	52479	9653	5792	10654	66,34	20,79	10,22	9,45	2,33
15	2008	60805	12364	7419	13418	67,11	22,81	9,75	8,65	2,3
16	2009	67333	12660	7597	17090	68,6	22,45	8,97	8,84	2,48

[1] Источник данных: Уровень жизни населения в Казахстане. Статистический сборник. Астана, 2010. – 184 с.

9. Появится следующий запрос-уточнение (рис. 35), снова нажмем Yes.

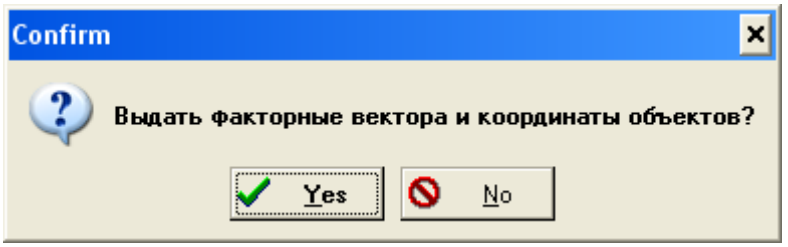


Рисунок 35

10. После вывода соответствующей информации будет выведено окно с предложением построения графика, нажмем *Отмена*.
11. Далее будет выведено окно, показанное на рисунке 36. В этом окне можно выбрать метод вращения факторов. Выберем метод *Варимакс* и оставим число факторов, которое указано по умолчанию. Появится уточняющее диалоговое окно (рис.37), в котором предлагается принять метод отбора факторов – критерий Кайзера. Нажав на кнопку Yes, получаем следующее окно (рис. 38). Снова нажмем Yes. Следующие два окна связаны с построением графиком и определением координат объектов, нажмем *Отменить*.
12. Далее снова появится окно, изображенное на рисунке 36. Выберем метод *Варимакс* и укажем количество факторов – 2, затем (после вывода информации и вывода этого же окна) – 3. После этого нажмем <Esc>=Отменить.

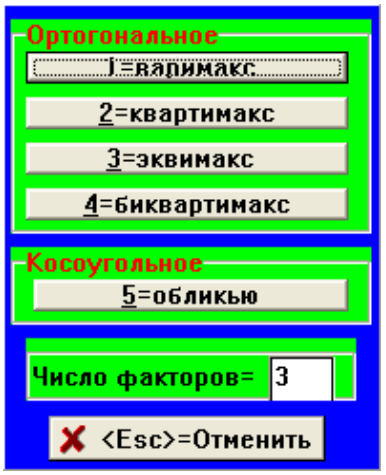


Рисунок 36

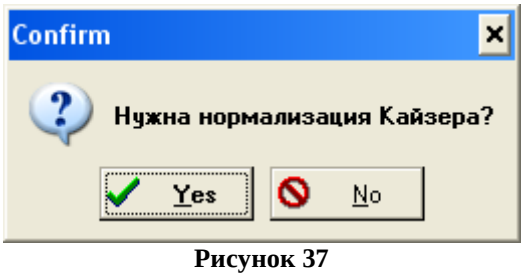


Рисунок 37

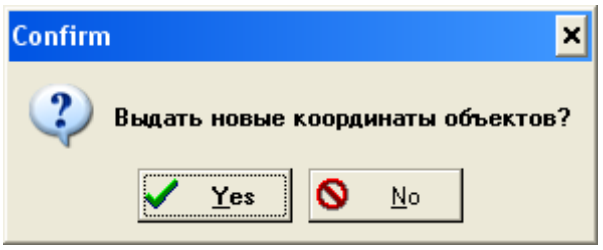


Рисунок 38

13. После выполнения всех действий должна быть выведена следующая информация.

```
ФАКТОРНЫЙ АНАЛИЗ.  Файл:
Переменная   Среднее  Ст.отклон.
       x1  2,828E4  2,005E4
       x2     6604     3254
       x3     4114     1798
       x4     7678     4110
       x5    67,71     7,69
       x6    17,45    2,979
       x7    10,09    0,411
       x8    7,177    1,417
       x9    2,194   0,2507

                  Корреляционная матрица
     x1       x2       x3       x4       x5       x6       x7    x8
x2  0,987
x3  0,991    0,995
x4  0,975    0,973    0,975
x5  0,58     0,623    0,64     0,524
x6  0,959    0,943    0,946    0,936    0,606
x7 -0,586   -0,58    -0,605   -0,624   -0,351   -0,413
x8  0,769    0,737    0,745    0,701    0,363    0,76    -0,198
x9  0,332    0,312    0,329    0,341    0,113    0,505    0,12   0,489


 Критическое значение=0,5248
 Число значимых коэффициентов = 23 (82%)
Собственные значения и процент объясняемой дисперсии факторов

Фактор:        1        2        3        4        5        6        7
Собств.зн     6,34     1,29   0,6463   0,4367   0,2329
Дисперс%     70,44    14,34    7,181    4,852    2,588
Накоплен%    70,44    84,78    91,96    96,82     99,4

     Переменная   <-- Собственные вектора (коэффициенты поворота фак-
торных осей) -->
        Фактор        1        2        3        4        5
            x1   0,3927  0,03922  0,08894  -0,1158   0,1361
            x2   0,3913  0,06403  0,01866 -0,09609   0,1894
            x3   0,3943  0,06617  0,01404 -0,04662  0,09802
            x4   0,3854  0,06198   0,1838  0,01042   0,2784
            x5   0,2566   0,1874  -0,8888   0,1786  -0,2324
            x6   0,3835  -0,1375  -0,0571 0,007403   0,3252
            x7  -0,2298  -0,5719  -0,3871  -0,4695   0,4601
```

x8	0,3115	-0,3334	0,1026	-0,5328	-0,6884
x9	0,1612	-0,7027	0,06479	0,6624	-0,1147

Объекты в координатах факторов

Фактор	1	2	3	4	5
1	-2,222	-2,159	-0,3448	0,7672	0,009783
2	-2,203	-0,8139	-0,1411	0,2992	-0,2592
3	-1,963	-0,7164	0,04544	0,8627	-0,4175
4	-2,36	1,994	0,1045	-0,2913	-0,3393
5	-2,058	1,379	-0,005783	-0,4128	-0,1711
6	-1,392	0,9435	0,09783	0,02049	-0,1844
7	-1,003	0,4204	0,08733	0,01749	-0,1754
8	-0,4606	-0,3468	-0,07213	-0,6276	0,2225
9	-0,5004	0,7316	-0,05075	0,611	1,533
10	0,5448	-0,484	0,07515	-0,9294	0,2595
11	1,499	-1,097	0,3003	-0,6651	-0,04521
12	2,385	-1,075	0,4408	-0,9341	0,02237
13	4,744	0,6061	-2,242	0,3111	-0,2298
14	4,987	0,6178	1,705	0,9711	-0,2259

Вращение: варимакс, число факторов=5

Переменная	Общность	Специфичность
x1	0,9949	0,005046
x2	0,9886	0,0114
x3	0,9945	0,005542
x4	0,9867	0,01334
x5	0,9998	0,0002697
x6	0,9836	0,01644
x7	0,9992	0,0007856
x8	0,9996	0,0003669
x9	0,9995	0,0005289

Корреляции факторов

Фактор	1	2	3	4	5
1	1				
2	0,08313	1			
3	0,1226	0,03135	1		
4	-0,1203	-0,1775	0,08462	1	
5	0,109	-0,1215	-0,01601	-0,03554	1

Координаты объектов в системе факторов

N	1	2	3	4	5
1	1,641E4	-5580	-1,578E4	2,11E4	-2,651E4
2	1,915E4	-6513	-1,842E4	2,462E4	-3,125E4
3	2,15E4	-7331	-2,049E4	2,769E4	-3,498E4
4	2,375E4	-8070	-2,25E4	3,044E4	-3,907E4
5	2,779E4	-9402	-2,647E4	3,55E4	-4,612E4
6	3,39E4	-1,139E4	-3,255E4	4,314E4	-5,64E4
7	3,872E4	-1,304E4	-3,69E4	4,931E4	-6,447E4
8	4,515E4	-1,532E4	-4,231E4	5,783E4	-7,43E4
9	5,182E4	-1,755E4	-4,85E4	6,616E4	-8,607E4
10	5,993E4	-2,024E4	-5,623E4	7,629E4	-1,003E5
11	6,949E4	-2,344E4	-6,518E4	8,831E4	-1,17E5
12	8,51E4	-2,863E4	-7,995E4	1,078E5	-1,446E5
13	1,018E5	-3,426E4	-9,584E4	1,292E5	-1,722E5
14	1,134E5	-3,836E4	-1,055E5	1,444E5	-1,905E5

Переменная <-------- Факторные нагрузки после вращения---------->

Фактор	1	2	3	4	5
x1	2,725				
x2	2,722				
x3	2,653				
x4	2,768				
x5		-3,244			
x6	2,686				
x7			-3,013		
x8				-2,707	
x9		-3,043			

Вращение: варимакс, число факторов=2

Переменная	Общность	Специфичность
x1	0,9797	0,02033
x2	0,976	0,02401
x3	0,9912	0,008857
x4	0,9467	0,05327
x5	0,4627	0,5373
x6	0,9568	0,0432
x7	0,7568	0,2432
x8	0,7585	0,2415
x9	0,802	0,1979

Корреляции факторов

Фактор	1	2
1	1	
2	0,178	1

Координаты объектов в системе факторов

N	1	2
1	1,656E4	-1,598E4
2	1,933E4	-1,87E4
3	2,169E4	-2,098E4
4	2,394E4	-2,327E4
5	2,802E4	-2,728E4
6	3,42E4	-3,326E4
7	3,904E4	-3,803E4
8	4,546E4	-4,424E4
9	5,216E4	-5,092E4
10	6,034E4	-5,902E4
11	6,996E4	-6,856E4
12	8,569E4	-8,419E4
13	1,025E5	-1,005E5
14	1,141E5	-1,119E5

Переменная <-------- Факторные нагрузки после вращения---------->

Фактор	1	2
x1	2,74	
x2	2,764	
x3	2,787	
x4	2,721	
x5	2,01	
x6	2,436	
x7	-2,349	
x8		-2,314
x9		-3,122

Вращение: варимакс, число факторов=3

Переменная	Общность	Специфичность
x1	0,9848	0,01521
x2	0,9762	0,02379
x3	0,9913	0,008729
x4	0,9686	0,03144
x5	0,9732	0,02678
x6	0,9589	0,04109
x7	0,8536	0,1464
x8	0,7653	0,2347
x9	0,8048	0,1952

Корреляции факторов

Фактор	1	2	3
1	1		
2	-0,01658	1	
3	0,1903	-0,02907	1

Координаты объектов в системе факторов

N	1	2	3
1	1,508E4	-1,448E4	-2,394E4
2	1,761E4	-1,694E4	-2,795E4
3	1,978E4	-1,902E4	-3,119E4
4	2,184E4	-2,108E4	-3,434E4
5	2,554E4	-2,471E4	-4,034E4
6	3,113E4	-3,011E4	-4,949E4
7	3,557E4	-3,443E4	-5,625E4
8	4,152E4	-4,009E4	-6,486E4
9	4,764E4	-4,613E4	-7,441E4
10	5,508E4	-5,345E4	-8,623E4
11	6,386E4	-6,207E4	-9,998E4
12	7,819E4	-7,621E4	-1,226E5
13	9,355E4	-9,102E4	-1,469E5
14	1,043E5	-1,013E5	-1,623E5

Переменная <-------- Факторные нагрузки после вращения---------->

Фактор	1	2	3
x1	2,5		
x2	2,453		
x3	2,469		
x4	2,59		
x5			-3,267
x6	2,038		
x7	-2,602		
x8			
x9		-2,54	

Рассмотрим полученные данные и проанализируем их.

Рассмотрим полученные собственные значения.

Собственные значения и процент объясняемой дисперсии факторов

Фактор:	1	2	3	4	5	6	7
Собств.зн	6,34	1,29	0,6463	0,4367	0,2329		
Дисперс%	70,44	14,34	7,181	4,852	2,588		
Накоплен%	70,44	84,78	91,96	96,82	99,4		

Из девяти рассматриваемых факторов-переменных выделено пять новых факторов. Их дисперсия приведена в строке *Собственные значения*. Далее приводится процент от общей дисперсии (она равна 9, то есть 9 первоначальных переменных). Первый фактор объясняет 70,44 % общей дисперсии, второй – 14,34 % и т.д. Согласно критерию Кайзера отбора количества факторов, идея которого рассмотрена выше, следует оставить только те факторы, собственные значения которых больше единицы. У нас таковых два первых фактора. Пока не будем окончательно решать этот вопрос, а рассмотрим таблицу *Собственные вектора* (см. с. 41).

Переменная <-- Собственные вектора (коэффициенты поворота факторных осей) -->

Фактор	1	2	3	4	5
x1	0,3927	0,03922	0,08894	-0,1158	0,1361
x2	0,3913	0,06403	0,01866	-0,09609	0,1894
x3	0,3943	0,06617	0,01404	-0,04662	0,09802
x4	0,3854	0,06198	0,1838	0,01042	0,2784
x5	0,2566	0,1874	-0,8888	0,1786	-0,2324
x6	0,3835	-0,1375	-0,0571	0,007403	0,3252
x7	-0,2298	-0,5719	-0,3871	-0,4695	0,4601
x8	0,3115	-0,3334	0,1026	-0,5328	-0,6884
x9	0,1612	-0,7027	0,06479	0,6624	-0,1147

В данной таблице отображена зависимость между исходными переменными и вновь созданными факторами. Это матрица коэффициентов корреляции, которая в терминологии факторного анализа называется матрицей *факторных нагрузок*.

Каждая из девяти заданных переменных взаимодействует с каким-либо из новых факторов. Если посмотреть на первый фактор, то наблюдается слабая взаимосвязь с переменными x_1, x_2, x_3, x_4, x_6. Со вторым фактором наблюдается более тесная взаимосвязь переменных x_7 и x_9 и т.д. (наиболее существенные коэффициенты корреляции отмечены серым цветом).

В конечном итоге требуется разобраться, что из себя представляют новые факторы, какие компоненты в них собраны и какое смысловое название им можно дать. Прежде чем это сделать, проведем процедуру вращения факторов с целью получения возможно лучшего результата (более лучшей, по сравнению с полученной, матрицы факторных нагрузок). В данном случае использовался метод вращения факторов – варимакс. Было оставлено 5 факторов. После вращения получилась следующая матрица факторных нагрузок (см. с. 43):

Переменная <-------- Факторные нагрузки после вращения---------->

Фактор	1	2	3	4	5
x1	2,725				
x2	2,722				
x3	2,653				
x4	2,768				
x5			-3,244		
x6	2,686				
x7				-3,013	

```
x8                                          -2,707
x9          -3,043
```

Из нее видно, какие переменные с какими факторами связаны. Обратите внимание, что наше первое распределение практически совпало с результатами вращения. Итак, в данной матрице видны взаимодействия переменных x_1 – x_4, x_6 с первым фактором, переменной x_9 – со вторым, x_5 – с третьим, x_7 – с четвертым, x_8 – с пятым.

Теперь рассмотрим более детально это взаимодействие. С первым фактором взаимодействуют переменные, обозначающие: среднемесячную номинальную заработную плату, тенге (x_1), величину прожиточного минимума, тенге (x_2), стоимость продовольственной корзины, тенге (x_3), средний размер назначенной месячной пенсии, тенге (x_4), общий коэффициент рождаемости на 1000 человек (x_6) то есть в целом эти переменные можно охарактеризовать как показатели уровня жизни населения. Остальные факторы взаимодействуют с какой-либо одной переменной. Так, фактор 2 связан с переменной x_9, обозначающей общий коэффициент разводимости на 1000 человек, фактор 3 – с переменной x_5 - продолжительность жизни при рождении (лет), фактор 4 – с переменной x_7 – общий коэффициент смертности на 1000 человек, фактор 5 – с переменной x_8 – общий коэффициент брачности на 1000 человек. Если первый фактор еще можно интерпретировать, то остальные факторы довольно раздроблены по взаимодействию с переменными, относящимися по смыслу к определенной категории основных индикаторов уровня жизни – к демографическим показателям.

Поэтому используем результат, полученный еще при одном вращении факторов с выбором двух факторов (см. с. 43-44).

```
Переменная <-------- Факторные нагрузки после вращения---------->
   Фактор         1         2
       x1      2,74
       x2     2,764
       x3     2,787
       x4     2,721
       x5      2,01
       x6     2,436
       x7    -2,349
       x8              -2,314
       x9              -3,122
```

По данной матрице видно, что факторы сгруппированы довольно тесно. Итак, первые семь переменных взаимодействуют с первым фактором, последние два (коэффициенты брачности и разводимости) – со вторым фактором.

Для «чистоты эксперимента» осуществим еще одно вращение факторов, определив количество факторов – 3. Получим следующую матрицу:

```
Переменная <-------- Факторные нагрузки после вращения---------->
   Фактор         1         2         3
       x1       2,5
       x2     2,453
```

x3	2,469		
x4	2,59		
x5			-3,267
x6	2,038		
x7	-2,602		
x8			
x9		-2,54	

Теперь с первым фактором взаимодействуют переменные $x_1 - x_4$, x_6, x_7, со вторым - переменная x_9, с третьим фактором – x_5. Заметьте, переменная x_8 не взаимодействует ни с одним из выделенных факторов.

Думается, наилучшее количество факторов – 2 (это предположение уже было высказано при анализе собственных значений). Остановимся на предыдущем варианте и дадим семантические названия полученным факторам. Итак, переменные, отображающие доходы населения, и переменные, связанные с демографическими показателями (продолжительность жизни, коэффициенты рождаемости и смертности), взаимодействуют с первым фактором. В целом, эти показатели взаимосвязаны и между собой достаточно сильно, о чем свидетельствует матрица парных корреляций, полученная в начале решения. Назовем этот фактор «*Уровень жизни и развития*». Второй фактор, включающий такие переменные как общие коэффициенты брачности и разводимости на 1000 человек, назовем фактором «*воспроизводства населения*».

Итак, факторный анализ проведен. Целью его было сокращение числа переменных. В нашей задаче получено два фактора, представляющих собой линейную комбинацию значений исходных переменных. В дальнейшем можно использовать значения этих факторов, например, для проведения регрессионного анализа (значения факторов приведены ниже и на с.43):

Координаты объектов в системе факторов

N	1	2
1	1,656E4	-1,598E4
2	1,933E4	-1,87E4
3	2,169E4	-2,098E4
4	2,394E4	-2,327E4
5	2,802E4	-2,728E4
6	3,42E4	-3,326E4
7	3,904E4	-3,803E4
8	4,546E4	-4,424E4
9	5,216E4	-5,092E4
10	6,034E4	-5,902E4
11	6,996E4	-6,856E4
12	8,569E4	-8,419E4
13	1,025E5	-1,005E5
14	1,141E5	-1,119E5

Задание 2.

По исходным данным таблицы 6 провести факторный анализ, выделить значимое количество факторов, дать им семантически значимые названия.

Задание 3.

По данным по вариантам, представленным в приложении А настоящего практикума, выполнить факторный анализ, с помощью которого определить число факторов, которые наиболее полно объясняют взаимосвязи между переменными, сократив при этом общее количество переменных. Рассмотреть несколько вариантов выбора числа факторов, интерпретировать получаемые результаты. Дать окончательно выбранным факторам семантически значимые названия.

Таблица 6[1]

Годы	Среднемесячная номинальная з/п, тенге	Величина прожиточного минимума (в ср.на душу населения), тенге	Стоимость продовольственной корзины, тенге	Сред. размер назначенной месячной пенсии, тенге	Ожидаемая продолжительность жизни при рождении, лет	Численность больничных учреждений	Численность врачей всех специальностей (без зубных), тыс. чел.	Численность санаториев, пансионатов с лечением	Выброс вредных веществ в атмосферу от стационарных источников, тыс. тонн
1996	6841	3000	2000	3283	63,6	1244	57,9	80	2100
1997	8541	3120	2454	3554	64	1006	54,5	80	2368,7
1998	9683	3336	2601	4213	64,5	991	53,2	83	2327,6
1999	11864	3394	2376	4270	65,5	917	50,6	80	2308,6
2000	14374	4007	2805	4462	65,4	938	48,9	90	2429,4
2001	17303	5655	3393	4947	65,8	981	49	90	2582,7
2002	20323	6003	3602	5818	66	1005	51,3	99	2529,3
2003	23128	6457	3874	8198	65,83	1029	53,7	106	2884,3
2004	28329	6785	4071	8628	66,18	1042	54,6	107	3016,5
2005	34060	7618	4571	9061	65,91	1063	54,8	106	968,8
2006	40790	8410	5046	9898	66,19	1086	57,5	128	2921,2
2007	52479	9653	5792	10654	66,34	1055	59,4	127	915
2008	60805	12364	7419	13418	67,11	1041	58,9	101	2643,1
2009	67333	12660	7597	17090	68,6	1020	60,7	110	2320

[1] Данные для этого задания взяты из статистического сборника: Уровень жизни населения в Казахстане. Астана, 2010. – 184 с.

Тема 3. Кластерный анализ

Цель: изучить основы кластерного анализа, приобрести умения и навыки проведения исследования с помощью кластерного анализа и интерпретации полученных результатов.

«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» [5].

Методы кластерного анализа позволяют решать следующие задачи: классификации однородных объектов (например, сельскохозяйственных предприятий, имеющих разную специализацию) *с учетом признаков*, отражающих сущность, природу объектов; проверка выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов; построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности [7].

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.

С помощью кластерного анализа можно решать проблему выделения, например, групп предприятий, действующих в сходных условиях или с похожими результатами, однородных групп населения по различным аспектам жизнедеятельности или образа жизни в целом, в области биологии – разбиение животных на различные виды с целью содержательного описания различий между ними, в области медицины кластеризация заболеваний, лечения заболеваний или симптомов заболеваний. В области психиатрии правильная диагностика кластеров симптомов, таких как паранойя, шизофрения и т.д., является решающей для успешной терапии. В археологии с помощью кластерного анализа исследователи пытаются установить таксономии каменных орудий, похоронных объектов и т.д. Довольно часто данный метод многомерного статистического анализа применяется в маркетинговых исследованиям для выделения, например, групп потребителей какого-либо товара, обладающих сходными предпочтениями при выборе этого товара.

В основе понимания кластерного анализа лежит понятие таксономии. Таксономия (греч. taxis – расположение по порядку, nomos - закон) – теория классификации и систематизации сложноорганизованных областей действительности, имеющих иерархическое строение.

В общей постановке проблема классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов, статистически представленную в виде матрицы, разбить на сравнительно небольшое число однородных групп или классов. Полученные в результате разбиения классы часто называют кластерами (таксонами, образами), а методы их построения

соответственно кластерным анализом, численной таксономией, распознаванием образов с самообучением.

Важным моментом классификации, тесно связанным с определением понятия однородности объектов, является в общем случае задание правил вычисления величины расстояний между объектами. В каждой конкретной задаче этот выбор производится по-своему. Поэтому решение задач классификации многомерных объектов (например, сельскохозяйственных предприятий) предусматривает предварительный анализ, позволяющий существенно сократить размерность исходного пространства факторов (показателей) и отбор сравнительно небольшого их числа, наиболее существенных, наиболее информативных. На этом этапе важную роль играют заключения высококвалифицированных специалистов (экспертов) и экспертных комиссий. После установления m важнейших факторов задача сводится к построению дерева классификации из n объектов, объединяемых в классы более высокой общности поэтапно на основе критерия минимума расстояния в пространстве m-измерений, описывающих эти объекты, или к разбиению объектов на непересекающиеся классы.

Геометрической интерпретацией кластерного анализа в двумерном пространстве является поэтапное объединение в кластеры точек $A_i(x_i;y_i)$. Как известно, расстояние между двумя точками (объектами) в этом случае определяется по формуле:

$$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} .$$

В случае трехмерного пространства расстояние между точками $A_i(x_i;y_i;z_i)$ определяется по формуле:

$$d_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$$

В m-мерном пространстве для определения расстояния между объектами используются различные методы, так, если все переменные представлены в одних единицах измерения, то целесообразно использовать евклидовы расстояния, которые определяются по следующей формуле:

$$d[i, j] = \sqrt{(x[i,k] - x[j,k])^2} ,$$

где:

i - номер объекта, до которого определяется расстояние;

j - номер объекта, от которого определяется расстояние;

k - порядковый номер объекта, формально необходимый для исключения повторения вычисления расстояний;

x[i,k], x[j,k] – координаты точки.

Если же переменные даны в различных единицах измерения или сильно различаются по величине, то используют нормализованные евклидовы расстояния, которые находят по формуле:

$$d[i, j] = \sqrt{\frac{(x[i,k] - x[j,k])^2}{\sigma^2(k)}}$$

где:

$\sigma^2(k)$ - дисперсия выборки;

i,j,k – имеют тот же смысл, что и в простой евклидовой метрике.

Результаты по определению расстояния между объектами в двумерном пространстве для удобства дальнейшей обработки полученных данных можно записывать в виде диагональной матрицы (табл. 7). Диагональность матрицы позволяет исключить повторение одних и тех же комбинаций объектов, т.е. имеющих один и тот же набор номеров, записанных в разной последовательности. Например: $d_{42}=d_{24}$.

Таблица 7 - Диагональная матрица расстояния между объектами

i \ j	1	2	3	4	5	…	n-1
1	d_{11}						
2	d_{21}	d_{22}					
3	d_{31}	d_{32}	d_{33}				
4	d_{41}	d_{42}	d_{43}	d_{44}			
5	d_{51}	d_{52}	d_{53}	d_{54}	d_{55}		
…	…	…	…	…	…	…	
n	d_{n1}	d_{n2}	d_{n3}	d_{n4}	d_{n5}	…	d_{nn-1}

Следующим этапом кластерного анализа, после расчета расстояния между объектами, является выбор стратегии объединения в кластеры. Общим принципом объединения объектов в классы является определение числа кластеров и комплектование их по минимальным расстояниям между объектами. Стратегии подразделяются на *агломеративную*, которая допускает возможность включения одного и того же объекта в разные кластеры, и *дивизивную*, которая исключает такую возможность.

Агломеративная стратегия по способам определения расстояния между кластерами подразделяется на стратегии: «ближайшего соседа», «дальнего соседа», «группового соседа» и «гибкую». В зависимости от выбора стратегии расстояния вычисляются по следующим формулам:

$d[i+j,k]=MIN(d[i,k],d[j,k])$ – для стратегии ближайшего соседа;

$d[i+j,k]=MAX(d[i,k],d[j,k])$ – для стратегии дальнего соседа;

$$d[i+j,k]=\frac{d[i,k]n_i+d[j,k]n_j}{n_i+n_j}$$ - для стратегии группового соседа;

$$d[i+j,k]=\frac{(1-B)(d[i,k]+d[j,k])}{2-Bd[i,j]}$$ - для гибкой стратегии,

где:

d[i+j,k] – расстояние между кластером i+j и всеми другими кластерами k;

B – коэффициент гибкой стратегии, который изменяется от –1 до 1, при B>0 – пространство сжимается, при B<0 – пространство

растягивается, при B=0 метрика не изменяется.

После выше указанных расчетов определяются расстояния, на уровне которых произошло объединение каждого кластера. При дивизивной стратегии классификации объектов в каждый класс попадают различные объекты, при агломеративной стратегии один и тот же объект может попадать в различные классы, т.е. можно построить дерево классов.

В заключение можно построить дендрограмму (греч. dendron - дерево) расстояний объединения объектов по агломеративной стратегии и графы проекций кластеров на координатные плоскости по дивизивной стратегии с целью наглядного (графического) представления объектов.

Дендрограмма - дерево объединенных кластеров с порядковыми номерами объектов по горизонтальной оси и со шкалой расстояний по вертикальной оси (примером может служить рисунок 42).

В случае дивизивного метода выдаются найденные кластеры с центральным объектом для каждого, отмеченным звездочкой. Под центральным объектом (центром тяжести кластера) понимается тот объект (точка), от которого суммарное расстояние до других объектов (точек) того же кластера является минимальным. Если же в кластере содержится только два объекта, то его центр тяжести определяется по минимальному суммарному расстоянию до центров тяжести других кластеров (рис. 43).

Задание 1.

Провести классификацию (разделение на группы) накопительных пенсионных фондов Республики Казахстан по следующим характеристикам: размер собственного капитала, тыс. тенге, размер нераспределенного дохода (убытка), тыс. тенге, размер обязательств, активов, кредитного риска, тыс. тенге, значение коэффициента достаточности собственного капитала К1 с целью выделения наиболее предпочтительных пенсионных фондов. Исходные данные представлены в таблице 8[1].

Технология выполнения:

1. Проведем кластерный анализ с помощью ППП Stadia.
2. В открывшемся окне введем исходные данные из таблицы 8.
3. После этого с помощью меню *Статист=F9* вызовем окно, в котором представлены различные методы анализа данных, и выберем *Кластерный анализ*.

[1] Данные для задачи взяты с сайта Комитета по контролю и надзору финансового рынка и финансовых организаций Национального Банка Республики Казахстан (www.afn.kz), раздел «Накопительная пенсионная система», на 01.01.2011.

Таблица 8 – Исходные данные для проведения кластерного анализа

№ п/п	НПФ	Собственный капитал, тыс. тенге	Нераспределенный доход (непокрытый убыток) отчетного периода (январь-декабрь 2010 г.), тыс. тенге	Обязательства, тыс. тенге	Активы, тыс. тенге	Кредитный риск, тыс. тенге	Коэффициент достаточности собственного капитала К1
1	2	3	4	5	6	7	8
1	АО "Накопительный пенсионный фонд "Ұлар Үміт"	6 369 944	-651 475	701 677	7 071 621	119 277 024	0,020
2	АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН"	1 270 914	-454 656	66 271	1 337 185	24 244 496	0,035
3	АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк"	1 694 544	24 526	78 287	1 772 831	39 884 509	0,032
4	АО Накопительный пенсионный фонд "ГНПФ"	17 883 039	3 285 772	511 713	18 394 752	175 588 952	0,078
5	АО "Накопительный пенсионный фонд БТА Казахстан"дочерняя организация АО "БТА Банк"	10 324 936	-1 250 805	1 280 287	11 605 223	132 861 124	0,054
6	АО "Накопительный пенсионный фонд Народного Банка Казахстана, дочерняя организация АО "Народный Банк Казахстана"	26 937 742	6 211 129	888 523	27 826 265	165 424 685	0,091
7	АО "Накопительный пенсионный Фонд Казахмыс"	2 175 357	-915 885	186 900	2 362 257	17 224 624	0,046

Продолжение табл. 8

1	2	3	4	5	6	7	8
8	АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ"	1 748 805	323 216	39 931	1 788 736	17 040 214	0,061
9	АО "Евразийский накопительный пенсионный фонд"	3 197 064	520 398	33 341	3 230 405	38 031 716	0,058
10	АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк")	5 628 741	605 836	137 569	5 766 310	52 862 002	0,0802
11	АО "Открытый накопительный пенсионный фонд "Отан"	2 738 149	63 407	49 103	2 787 252	32 934 714	0,060
12	АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит"	3 224 582	500 156	21 356	3 245 938	34 905 097	0,074
13	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА"	1 005 226	-67 994	24 828	1 030 054	12 742 237	0,049

4. Откроется следующее окно (рис.39).

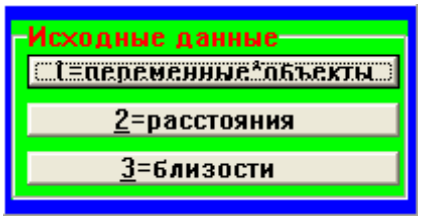


Рисунок 39

5. В качестве исходных данных может быть использована матрица значений m переменных для n объектов, квадратная матрица взаимных расстояний или квадратная матрица близостей между n объектами. Наши данные представлены в виде матрицы значений, состоящих из 6 переменных (характеристик) для 13 объектов (пенсионных фондов). Исходя из этого выберем пункт *1 = переменные*объекты*.
6. После выбора данного способа представления данных откроется окно с перечнем методов вычисления расстояний между объектами в многомерном пространстве (рис. 40).

Рисунок 40 – Окно Метрика вычисления расстояний

7. В окне на рисунке 40 предлагаются следующие методы:

1 = Эвклид – евклидовы расстояния применимы для переменных, измеренных в одних единицах;

2 = Норм.Эвклид – нормализованные евклидовы расстояния; более подходят для переменных, измеренных в различных единицах или сильно различающихся по величине;

3 = Квадр.Эвклид – суммирование квадратов можно использовать, чтобы стратегии объединения были аддитивны по отношению к переменным;

4 = Взвеш.Эвклид – если переменные обладают существенно различной значимостью, то целесообразно взвешенное суммирование квадратов, но при этом последняя строка матрицы данных должна содержать веса переменных;

5 = Корреляция – расстояния, вычисляемые на основе коэффициента корреляции, применимы к упорядоченным переменным типа временных серий;

7 = манхеттен – манхеттеновская метрика, как правило, применяется для

номинальных или качественных переменных;
8 = Брей-Картис – расстояния Брея-Картиса имеют значения между 0 и 1, однако, переменные с большими значениями оказывают большее влияние на результат;
9 = Канберра – канберровы расстояния также лежат между 0 и 1, но обладают нечувствительностью к сильно асимметричным данным. Однако применение этого метода при наличии нулевых и отрицательных значений нежелательно, поэтому рекомендуется предварительное преобразование данных посредством каждой переменной с некоторым значением.

8. Для наших данных наиболее подходящим методом вычисления расстояний являются нормированные евклидовы расстояния, поскольку количественные значения данных значительно различаются и имеют разные измерения. Так значения коэффициентов достаточности капитала достаточно малы и измеряются в долях единиц по сравнению с количественными значениями остальных данных. Следовательно, выбираем пункт *2 = Норм.Эвклид*.
9. Далее необходимо выбрать стратегию выделения кластеров: объединяющую (агломеративную) или разделяющую (дивизивную) (рис. 41).

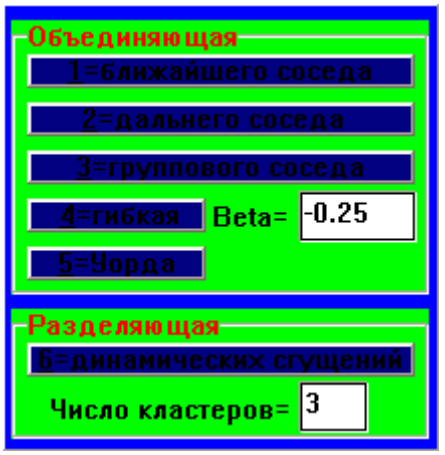


Рисунок 41 – Окно выбора стратегии выделения кластеров

Выберем объединяющую (*агломеративную*) стратегию.
Пояснение к окну выбора стратегии.
1 – Стратегия ближайшего соседа очень сильно сжимает пространство кластеров и рекомендуется для получения минимального дерева взамен групповой классификации;
2 – Стратегия дальнего соседа сильно растягивает пространство. В этом методе расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Этот метод обычно работает очень хорошо, когда объекты происходят на самом деле из реально различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочным", то этот метод непригоден.
3 – Стратегия группового соседа сохраняет метрику пространства.

4 – Гибкая стратегия универсальна и зависит от значения коэффициента бета, ввод которого запрашивается при выборе этой стратегии (бета должно быть меньше 1): при бета = 0 метрика не меняется, при бета > 0 пространство сжимается, при бета < 0 – растягивается.

5 – Уорда - Этот метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (SS) для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

10. Выделим кластеры на основе *объединяющей стратегии* с помощью метода *1 = ближайшего соседа*.
11. После определения стратегии выдается вопрос о необходимости вывода диагональной матрицы расстояний между объектами. Нажмем *Yes*.
12. При использовании агломеративного метода производится выдача последовательности кластеров, включающих объекты все *возрастающей общности* с указанием номеров входящих в них объектов и расстояния, на уровне которого произошло объединение этих объектов в каждый кластер.
13. Получим следующий результат (табл. 9).
14. После этого может быть выдана дендрограмма – дерево объединенных кластеров с порядковыми номерами объектов по горизонтальной оси и со шкалой расстояний по вертикальной оси (рис. 42).

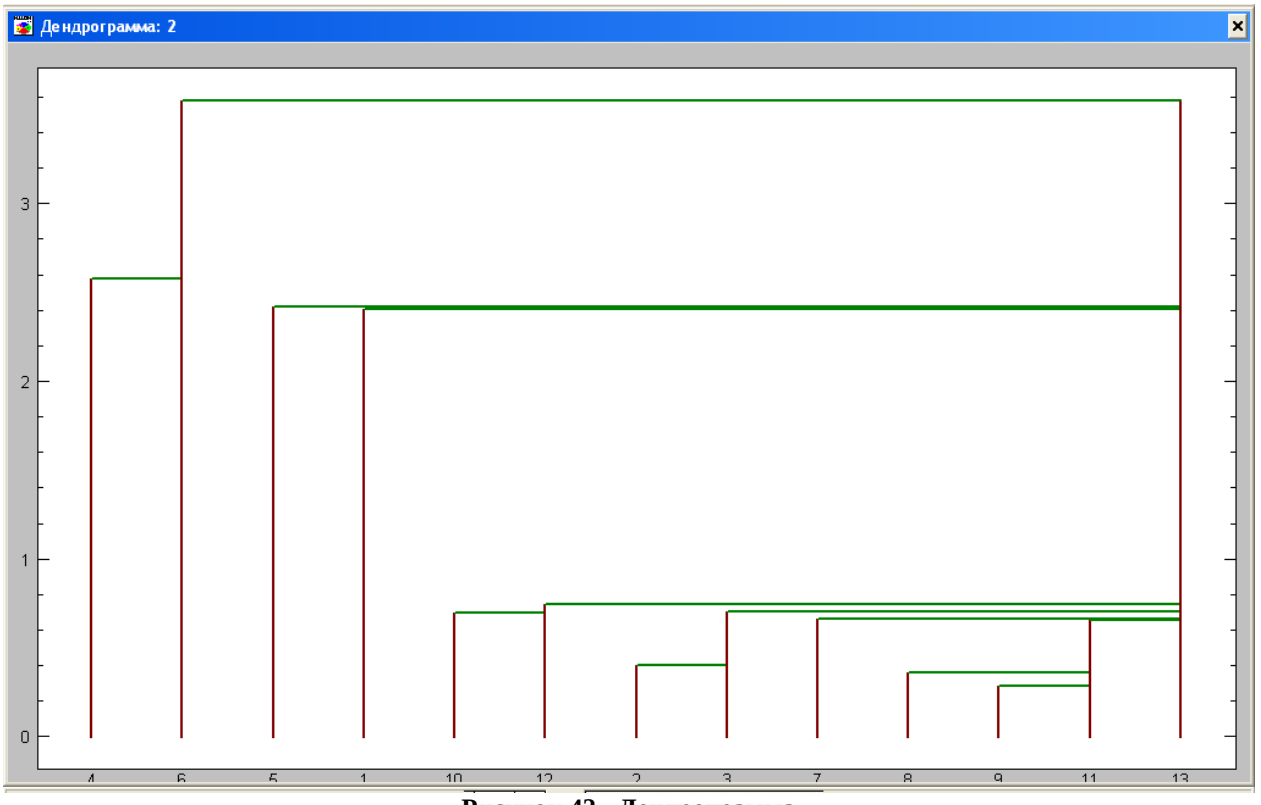


Рисунок 42 - Дендрограмма

Таблица 9 – Результаты кластерного анализа по агломеративной стратегии

КЛАСТЕРНЫЙ АНАЛИЗ. Файл:
Норм.Эвклид.+Ближ.сосед

Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)		
(2)	2,644											
(3)	2,41	0,4091										
(4)	4,297	5,162	4,948									
(5)	2,425	4,183	4,078	3,632								
(6)	6,42	7,354	7,15	2,58	5,403							
(7)	2,71	0,7114	0,99	4,951	3,83	7,121						
(8)	3,38	1,381	1,526	4,647	4,22	6,773	1,065					
(9)	3,063	1,338	1,368	4,257	3,975	6,444	1,113	0,4918				
(10)	3,616	2,539	2,571	3,587	3,749	5,715	2,1	1,397	1,267			
(11)	3,129	1,326	1,431	4,415	3,972	6,607	0,9815	0,3652	0,2883	1,26		
(12)	3,656	2,07	2,155	4,183	4,138	6,317	1,684	0,7828	0,8086	0,7056	0,7502	
(13)	3,132	0,7697	0,9994	5,005	4,303	7,16	0,6686	0,6584	0,8148	1,983	0,7396	1,413

К л а с т е р ы:
(список объектов) -> расстояние

(11,9) --> 0,2883
(11,8,9) --> 0,3652
(3,2) --> 0,4091
(13,11,8,9) --> 0,6584
(13,7,11,8,9) --> 0,6686
(12,10) --> 0,7056
(13,3,2,7,11,8,9) --> 0,7114
(13,12,10,3,2,7,11,8,9) --> 0,7502
(13,1,12,10,3,2,7,11,8,9) --> 2,41
(13,5,1,12,10,3,2,7,11,8,9) --> 2,425
(6,4) --> 2,58
(13,6,4,5,1,12,10,3,2,7,11,8,9) --> 3,587

15. Результаты расчетов можно представить в виде следующей таблицы (табл. 10)

Таблица 10

№ кластера	Накопительные пенсионные фонды	Расстояние объединения
1	АО «Открытый накопительный пенсионный фонд «Отан», АО «Евразийский накопительный пенсионный фонд»	0,2883
2	АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	0,3652
3	АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк", АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН"	0,4091
4	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	0,6584
5	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	0,6686
6	АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит", АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк")	0,7056
7	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк", АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	0,7114
8	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит", АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк"), АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк", АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	0,7502
9	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "Накопительный пенсионный фонд "Ұлар Үміт", АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит", АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк"), АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк",	2,41

Продолжение табл. 10

	АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	
10	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "Накопительный пенсионный фонд БТА Казахстан"дочерняя организация АО "БТА Банк" , АО "Накопительный пенсионный фонд "Ұлар Үміт", АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит", АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк"), АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк", АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	2,425
11	АО "Накопительный пенсионный фонд Народного Банка Казахстана, дочерняя организация АО "Народный Банк Казахстана", АО Накопительный пенсионный фонд "ГНПФ"	2,58
12	АО "Накопительный пенсионный фонд "РЕСПУБЛИКА", АО "Накопительный пенсионный фонд Народного Банка Казахстана, дочерняя организация АО "Народный Банк Казахстана", АО Накопительный пенсионный фонд "ГНПФ", АО "Накопительный пенсионный фонд БТА Казахстан"дочерняя организация АО "БТА Банк" , АО "Накопительный пенсионный фонд "Ұлар Үміт", АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит", АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк"), АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк", АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН", АО "Накопительный пенсионный Фонд Казахмыс", АО "Открытый накопительный пенсионный фонд "Отан", АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ", АО "Евразийский накопительный пенсионный фонд"	3,587

16. Проверим устойчивость полученного кластерного решения. Проверка устойчивости кластеризации сводится к проверке ее достоверности. Существует эмпирическое правило - устойчивая типология сохраняется при изменении методов кластеризации. Если сравниваемые классификации групп объектов наблюдений имеют долю совпадений более 70%, кластерное решение принимается.
17. Для проверки устойчивости полученного результата проведем кластерный анализ, используя другие методы кластеризации, например, метод «*дальнего соседа*» и метод «*Уорда*».
18. Получим при этом следующие результаты (табл. 11).

Таблица 11 – Результаты кластерного анализа по методам «дальнего соседа» и «Уорда»

КЛАСТЕРНЫЙ АНАЛИЗ. Файл:
Норм.Эвклид.+Дальн.сосед

Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)		
(2)	2,644											
(3)	2,41	0,4091										
(4)	4,297	5,162	4,948									
(5)	2,425	4,183	4,078	3,632								
(6)	6,42	7,354	7,15	2,58	5,403							
(7)	2,71	0,7114	0,99	4,951	3,83	7,121						
(8)	3,38	1,381	1,526	4,647	4,22	6,773	1,065					
(9)	3,063	1,338	1,368	4,257	3,975	6,444	1,113	0,4918				
(10)	3,616	2,539	2,571	3,587	3,749	5,715	2,1	1,397	1,267			
(11)	3,129	1,326	1,431	4,415	3,972	6,607	0,9815	0,3652	0,2883	1,26		
(12)	3,656	2,07	2,155	4,183	4,138	6,317	1,684	0,7828	0,8086	0,7056	0,7502	
(13)	3,132	0,7697	0,9994	5,005	4,303	7,16	0,6686	0,6584	0,8148	1,983	0,7396	1,413

К л а с т е р ы:
(список объектов) -> расстояние

(11,9) --> 0,2883
(3,2) --> 0,4091
(11,8,9) --> 0,4918
(13,7) --> 0,6686
(12,10) --> 0,7056
(13,3,2,7) --> 0,9994
(12,11,8,9,10) --> 1,397
(5,1) --> 2,425
(13,12,11,8,9,10,3,2,7) --> 2,571
(6,4) --> 2,58
(13,5,1,12,11,8,9,10,3,2,7) --> 4,303
(13,6,4,5,1,12,11,8,9,10,3,2,7) --> 7,354

Продолжение табл. 11

КЛАСТЕРНЫЙ АНАЛИЗ. Файл:
Норм.Эвклид.+Уорда

Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
(2)	2,644										
(3)	2,41	0,4091									
(4)	4,297	5,162	4,948								
(5)	2,425	4,183	4,078	3,632							
(6)	6,42	7,354	7,15	2,58	5,403						
(7)	2,71	0,7114	0,99	4,951	3,83	7,121					
(8)	3,38	1,381	1,526	4,647	4,22	6,773	1,065				
(9)	3,063	1,338	1,368	4,257	3,975	6,444	1,113	0,4918			
(10)	3,616	2,539	2,571	3,587	3,749	5,715	2,1	1,397	1,267		
(11)	3,129	1,326	1,431	4,415	3,972	6,607	0,9815	0,3652	0,2883	1,26	
(12)	3,656	2,07	2,155	4,183	4,138	6,317	1,684	0,7828	0,8086	0,7056	0,7502
(13)	3,132	0,7697	0,9994	5,005	4,303	7,16	0,6686	0,6584	0,8148	1,983	0,7396 1,413

К л а с т е р ы:
(список объектов) -> расстояние

(11,9) --> 0,2883
(3,2) --> 0,4091
(11,8,9) --> 0,4716
(13,7) --> 0,6686
(12,10) --> 0,7056
(13,3,2,7) --> 1,11
(12,11,8,9,10) --> 1,539
(5,1) --> 2,425
(6,4) --> 2,58
(13,12,11,8,9,10,3,2,7) --> 3,014
(13,5,1,12,11,8,9,10,3,2,7) --> 5,944
(13,6,4,5,1,12,11,8,9,10,3,2,7) --> 9,697

19. Если сравнить результаты кластеризации по трем методам, то видно, что они практически одинаковы в части попадания объектов в кластеры. Исходя из этого, выберем первый вариант (результат) для более детального анализа кластеризации (так как он, вместе с тем, предпочтительнее по расстояниям).
20. Проанализируем расстояния между объектами по кластерам. Чем меньше величина расстояния, тем однороднее объекты, - накопительные пенсионные фонды – в данном кластере. Согласно этой логике, можно сказать, что наиболее однородные (похожие) НПФ находятся в первом кластере, это АО «Открытый накопительный пенсионный фонд «Отан» и АО «Евразийский накопительный пенсионный фонд».

Не следует забывать, что похожесть пенсионных фондов оценивается по совокупности переменных (факторов), которые были априорно выбраны для их характеристики (табл. 8).

21. При увеличении расстояний между объектами в кластере «собираются» все более разнородные (менее сходные) объекты (пенсионные фонды).
22. Для выбора (определения) наилучшего варианта кластеризации проведем кластерный анализ по *дивизивной стратегии*. Для этого повторим действия, начиная с шага 3, выбрав на 9 шаге в окне на рисунке 41 пункт *6 = динамических сгущений*.
23. Получим следующие результаты (табл. 12).
24. При использовании дивизивного метода выдаются кластеры с *центральным объектом* для каждого, отмеченным звездочкой. Графический вывод возможен только в случае матрицы данных, содержащей исходные переменные. При этом запрашиваются индексы двух переменных для определения плоскости проекции, на которой указываются объекты, соединенные прямыми с центральным объектом соответствующего кластера (рис. 43). Запросы плоскости проекции и выдача графиков повторяются до «пустого» ответа.

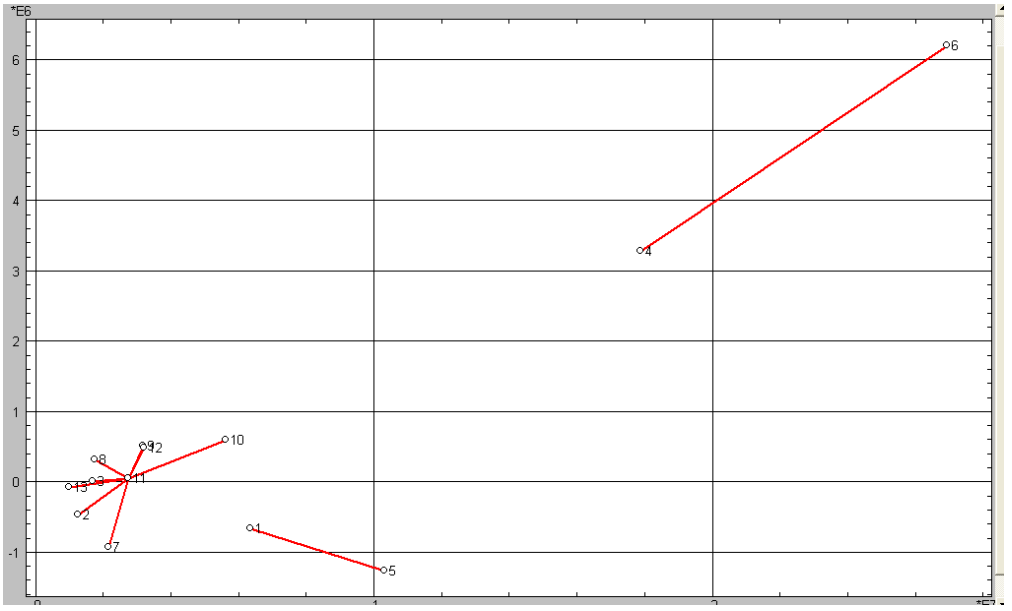


Рисунок 43

Таблица 12 – Результаты кластерного анализа при дивизивной стратегии

КЛАСТЕРНЫЙ АНАЛИЗ. Файл:
Норм.Эвклид.+Дивизивная

Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)		
(2)	2,644											
(3)	2,41	0,4091										
(4)	4,297	5,162	4,948									
(5)	2,425	4,183	4,078	3,632								
(6)	6,42	7,354	7,15	2,58	5,403							
(7)	2,71	0,7114	0,99	4,951	3,83	7,121						
(8)	3,38	1,381	1,526	4,647	4,22	6,773	1,065					
(9)	3,063	1,338	1,368	4,257	3,975	6,444	1,113	0,4918				
(10)	3,616	2,539	2,571	3,587	3,749	5,715	2,1	1,397	1,267			
(11)	3,129	1,326	1,431	4,415	3,972	6,607	0,9815	0,3652	0,2883	1,26		
(12)	3,656	2,07	2,155	4,183	4,138	6,317	1,684	0,7828	0,8086	0,7056	0,7502	
(13)	3,132	0,7697	0,9994	5,005	4,303	7,16	0,6686	0,6584	0,8148	1,983	0,7396	1,413

К л а с т е р ы:

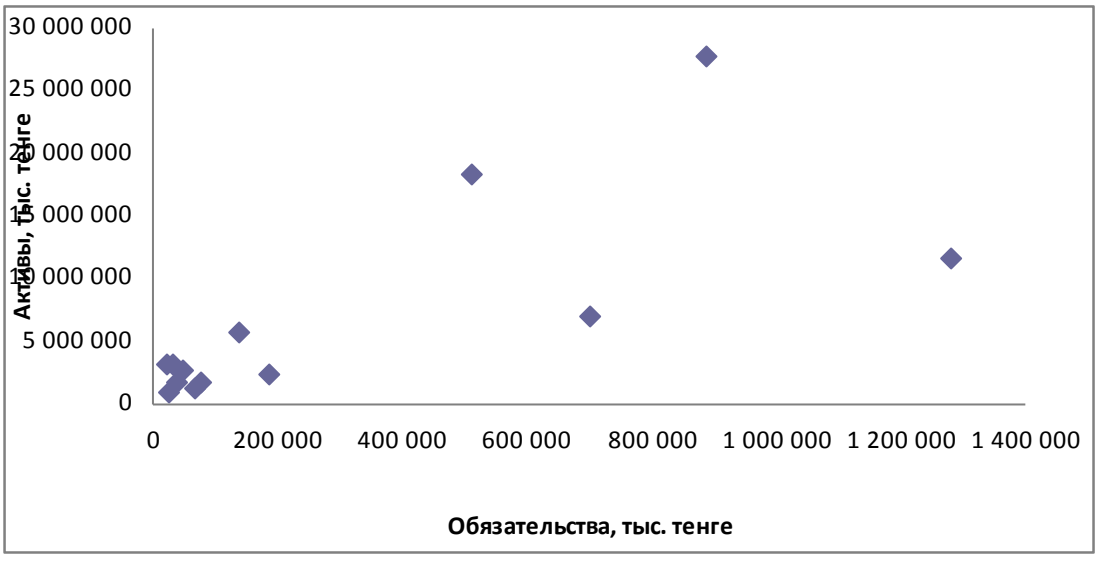
Среднее внутрикластерное расстояние=0,9343
1= (2,3,7,8,9,10,11*,12,13)
2= (1*,5)
3= (4*,6)

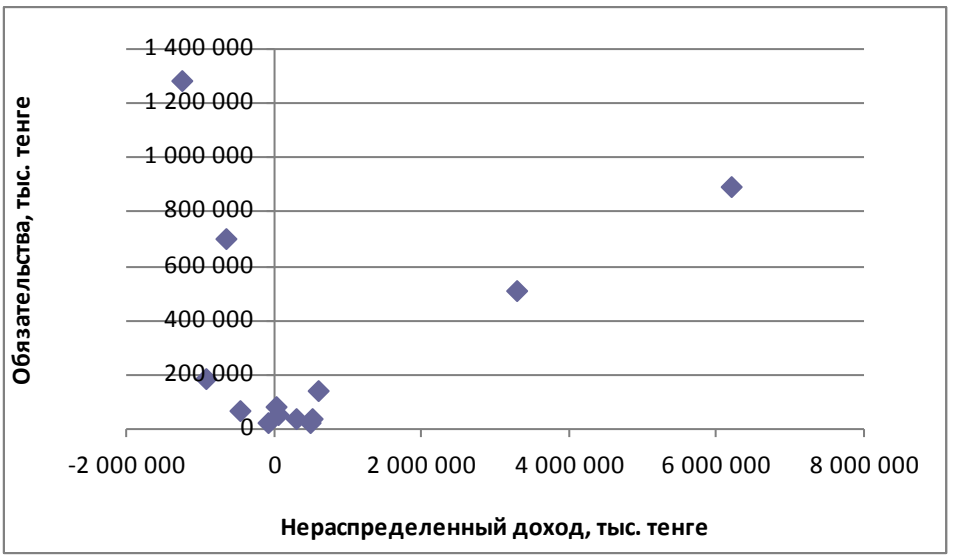
25. Как видно из результатов кластерного анализа по дивизивной стратегии, выделено три кластера, «содержимое» которых можно представить в виде следующей таблицы (табл. 13).

Таблица 13

№ кластера	Наименование объекта
1	АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН" АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк" АО "Накопительный пенсионный Фонд Казахмыс" АО "Накопительный пенсионный фонд "Нефтегаз - ДЕМ" АО "Евразийский накопительный пенсионный фонд" АО "Накопительный пенсионный фонд "РЕСПУБЛИКА" АО "Открытый накопительный пенсионный фонд "Отан"* АО "НПФ "Капитал" - Дочерняя организация АО "Банк ЦентрКредит" АО "Накопительный пенсионный фонд "ГРАНТУМ" (Дочерняя организация АО "Казкоммерцбанк")
2	АО "Накопительный пенсионный фонд "Ұлар Үміт"* АО "Накопительный пенсионный фонд БТА Казахстан" дочерняя организация АО "БТА Банк"
3	АО "Накопительный пенсионный фонд "ГНПФ"* АО "Накопительный пенсионный фонд Народного Банка Казахстана, дочерняя организация АО "Народный Банк Казахстана"

26. Как следует из принципа дивизивной стратегии, в каждый кластер попадают те объекты, которые не принадлежат другому кластеру, то есть происходит четкое распределение объектов по кластерам.
27. Чтобы лучше понять, по какому принципу объединены пенсионные фонды, построим в MS Excel диаграммы зависимости некоторых факторов друг от друга (тип диаграммы - точечная) (рис. 44).





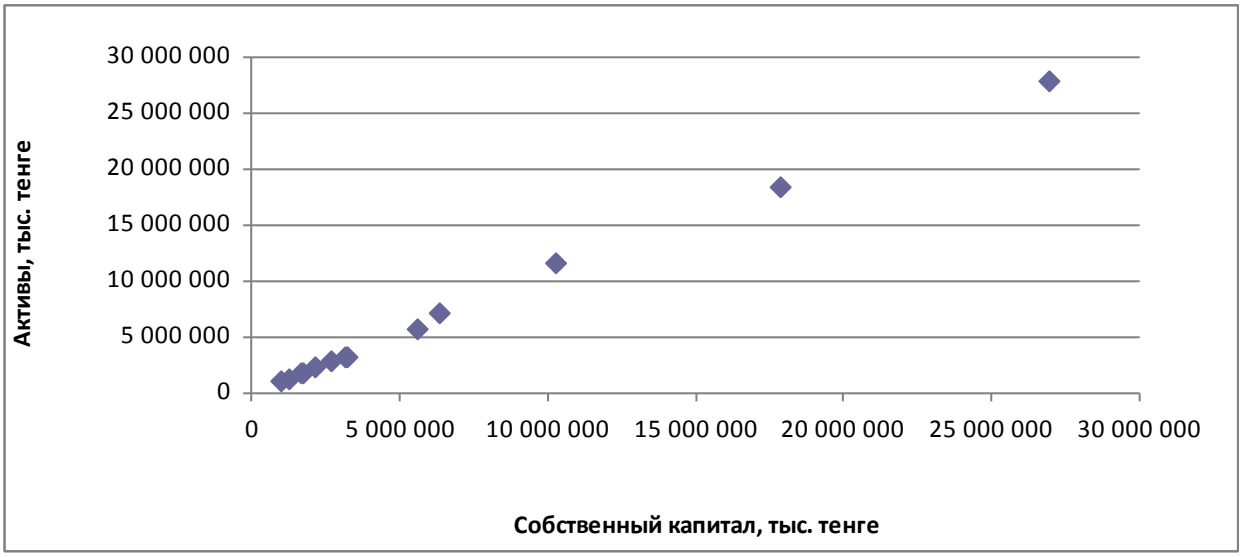


Рисунок 44

28. По графикам можно видеть распределение объектов по кластерам. Целая группа пенсионных фондов (сгустки точек у начала оси координат), это наиболее «многочисленный» первый кластер. Объекты этого кластера характеризуются средними значениями характеристик, по которым проводилась классификация.
29. Во второй кластер НПФ входят пенсионные фонды, имеющие так сказать, наихудшие показатели: АО "Накопительный пенсионный фонд "Ұлар Үміт" АО "Накопительный пенсионный фонд БТА Казахстан" дочерняя организация АО "БТА Банк". Если смотреть на характеристики по отдельности, то одни из них, например, величина активов, собственного капитала не хуже, чем у других фондов, отрицательная величина нераспределенного дохода также наблюдается у ряда других фондов (например, АО "Открытый накопительный пенсионный фонд "АМАНАТ КАЗАХСТАН"), однако в совокупности все эти параметры выделяют указанные пенсионные фонды в отдельный кластер.
30. Два пенсионных фонда: АО Накопительный пенсионный фонд "ГНПФ"

АО "Накопительный пенсионный фонд Народного Банка Казахстана, дочерняя организация АО "Народный Банк Казахстана" выделены по критерию наибольших значений по всем показателям, по которым проводилась кластеризация, то есть по величине активов, обязательств, дохода, собственного капитала, кредитного риска, коэффициента достаточности собственного капитала (третий кластер).

Итак, кластеризация накопительных пенсионных фондов с целью определения предпочтительности выбора лучшего фонда (фондов) с помощью разных методов (агломеративного и дивизивного) проведена. Думается, в данном случае более приемлемы результаты дивизивной стратегии, поскольку произошло четкое разделение пенсионных фондов. Полученную классификацию можно использовать для оценки деятельности и выбора пенсионных фондов.

Задание 2.

По данным, приведенным по вариантам в приложении А настоящего практикума, содержащем показатели деятельности некоторых коммерческих банков Республики Казахстан, провести процедуру кластеризации с целью определения наиболее прибыльных банков. Использовать агломеративную и дивизивную стратегии. Сделать выводы и рекомендации.

Тема 4. Дискриминантный анализ.

Цель: изучить основы дискриминантного анализа, приобрести умения и навыки проведения исследования с помощью дискриминантного анализа и интерпретации полученных результатов.

При наблюдении больших статистических совокупностей часто появляется необходимость разделить неоднородную совокупность на *однородные группы (классы)*. Такое расчленение позволяет получить лучшие результаты при дальнейшем моделировании зависимостей между отдельными признаками. Например, разбиение предприятий разной специализации или отрасли на несколько однородных групп по показателям производственно-хозяйственной деятельности; подбор кандидатов на определенную должность; оценка кредитоспособности заемщика в банковской деятельности и т.п.

Группа (класс) – генеральная совокупность, описываемая одномодальной функцией плотности *f(x)*.

Классификация – разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы, либо отнесение каждого из заданного множества объектов к одному из заранее известных классов.

Методы дискриминантного анализа разрабатывались П.Ч. Махаланобисом, Р. Фишером, Г. Хотеллингом и др.

К понятию дискриминантного анализа можно отнести задачи по разделению множества объектов на два непересекающихся класса. Множество объектов может представляться группой предприятий, характеризующихся одинаковыми показателями производственно-финансовой деятельности. Это могут быть сельскохозяйственные предприятия, заводы, фабрики, банки и т.д., среди которых нужно выделить класс эффективно функционирующих и класс объектов неэффективно функционирующих. Это могут быть группы семей, которые обследуются по качеству жилищных условий, которые характеризуются несколькими показателями, это могут быть группы животных, продуктивность которых и племенные особенности характеризуются многими показателями.

Результаты решения выше указанных задач с успехом могут быть использованы в принятии управленческих решений. Например, в решении вопроса о закрытии класса неэффективно функционирующих предприятий или предоставления кредитных ресурсов эффективно работающим предприятиям, а также для уточнения решения экспертных оценок, т.е. проверить гипотезу о верности экспертного заключения.

Постановка задачи.

В m-мерном векторном пространстве заданы n векторов:

$$A_1 = (a_{11}, a_{12}, ..., a_{1m})$$

$$\overrightarrow{A_2} = (a_{21}, a_{22}, ..., a_{2m})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$A_n = (a_{n1}, a_{n2}, ..., a_{nm})$$

Требуется проверить статистическую гипотезу о возможности разбиения n-объектов на два непересекающихся класса с заданным уровнем значимости.

Исходные данные (входная информация) задается в виде матрицы:

$$\begin{bmatrix} a_{11}, ..., a_{1m} \\ ..... \\ a_{n1}, ..., a_{nm} \end{bmatrix}$$

Теоретической основой метода дискриминантного анализа является регрессионный анализ, сущность которого сводится к построению следующей многофакторной линейной функции регрессии, называемой иначе дискриминирующей.

Дискриминирующая функция (решающее правило, процедура классификации) – статистика, служащая для построения правила классификации объектов по группам:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m,$$

где:

$$x_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ .... \\ a_{n1} \end{pmatrix}, \; x_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ .... \\ a_{n2} \end{pmatrix}, \; \ldots, \; x_m = \begin{pmatrix} a_{1m} \\ a_{2m} \\ .... \\ a_{nm} \end{pmatrix},$$

а $y = \begin{pmatrix} 1 \\ ... \\ 1 \\ 2 \\ ... \\ 2 \end{pmatrix}$,количество единиц в матрице-столбце у определяет количество объектов, относящихся к первому классу, а количество двоек – количество объектов относящихся ко второму классу.

На «входе» задачи классификации имеем:

а) априорные[1] сведения о классах: число классов, общий вид или свойства распределения *X* внутри каждого из классов, диапазон изменения анализируемых показателей;

б) обучающие выборки $X_{j1}, X_{j2}, \ldots, X_{jn_j}$, $j=1,2, \ldots, k$, k – число классов.

[1] *Априо́ри* (лат. *a priori* — буквально «от предшествующего») — знание, полученное до опыта и независимо от него (*знание априори*, *априорное знание*). Противоположностью априори является *апостериори* (лат. *a posteriori* — «от последующего») — знание, полученное из опыта (*опытное знание*).

Выборка x_{ji} называется *обучающей*, если известно, что все ее наблюдения X_i извлечены из одной и той же генеральной совокупности D_j. Можно использовать также понятие «обучение с учителем», под которым имеется в виду, что класс (с определенными априори характеристиками) задан. Тогда как к кластерному анализу более применимо понятие «обучение без учителя», так как количество классов и их характеристики в каком-то смысле неопределенны (формируются в процессе расчета).

На «выходе» задачи классификации:

а) набор наиболее информативных объясняющих переменных (типообразующих признаков), которые либо отбираются по определенному правилу из числа исходных признаков x_1, …, x_p, либо строятся в качестве некоторых их комбинаций.

б) правило отнесения (дискриминантная функция, классификатор) каждого классифицируемого объекта, заданного значениями описательных признаков X_i к одному из классов.

Если на «входе» есть обучающие выборки, то возникает задача *классификации с обучением*, если нет обучающих выборок – *классификация без обучения*.

На рисунке 45 показаны два множества S_1 и S_2 объектов, характеризующихся двумя признаками x_1 и x_2. По каждой переменной отдельно объекты разных множеств имеют сходные характеристики – это видно по проекциям объектов на каждую ось. Чтобы наилучшим образом разделить два множества, нужно построить линейную комбинацию переменных x_1 и x_2: $D(x) = b_1x_1 + b_2x_2$. Функция $D(x)$ называется *канонической дискриминантной функцией*, а величины b_1 и b_2 – *коэффициентами дискриминантной функции*. Фактически для двумерного случая это означает определение новой системы координат с осями L и C. Причем, проекции объектов разных множеств на ось L должны быть максимально разделены.

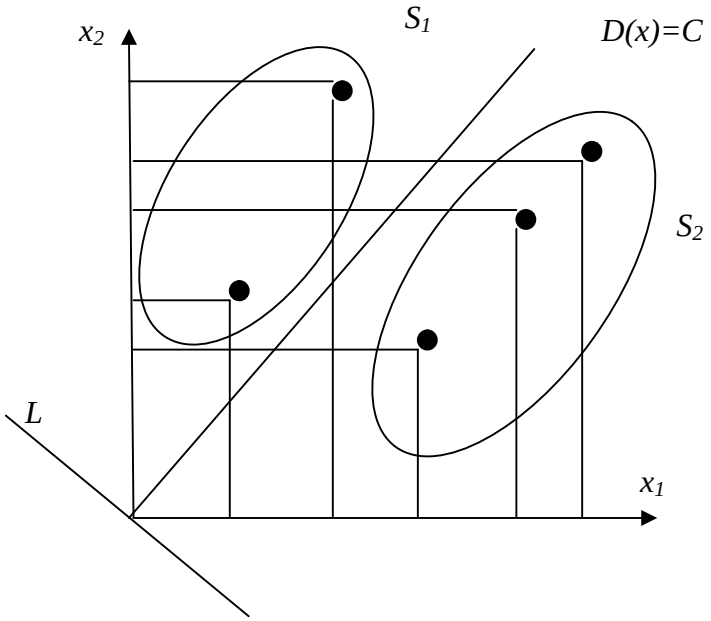


Рисунок 45 – Геометрическая интерпретация дискриминантной функции

Дискриминантная функция может быть как линейной, так и нелиней-

ной. Выбор ее вида зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных.

Дискриминантный анализ выполняется в предположении, что объекты принадлежат одному из двух (или более) классов.

Существуют ограничения, касающиеся свойств дискриминантных переменных:

- Количество дискриминантных переменных не должно превышать число объектов за вычетом двух, т.е. $0 < p < (m\text{-}2)$;
- Переменные должны быть измерены в метрической шкале;
- Дискриминантные переменные должны быть линейно независимы.

Для проверки качества построенных дискриминантных функций используется статистическая проверка гипотезы: H_0: $a_1 = a_2 = \ldots = a_k$ о равенстве математических ожиданий внутри классов. Для проверки такой гипотезы может быть использовано обобщенное расстояние Махаланобиса: $D^2 = \frac{R^2 n(n-2)}{(1-R^2)n_1 n_2}$, где R – коэффициент корреляции, n_1 и n_2 – число объектов в каждом классе. Если нулевая гипотеза верна, то D^2 аппроксимируется F-распределением Фишера.

Другим способом проверки гипотезы H_0 является использование статистики Уилкса (лямбда Уилкса):

$$\lambda = \left| \frac{\sum}{T} \right|, \text{ где } \hat{\sum} = \{\hat{\sigma}_{lq}\},\ \hat{T} = \{\hat{t}_{lq}\}$$

$$\text{и } \hat{t}_{lq} = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji}^{(l)} - \hat{a}^{(l)})(x_{ji}^{(q)} - \hat{a}^{(q)}),$$

$$\hat{\sigma}_{\partial\check{u}} = \frac{1}{n-k} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji}^{(l)} - \hat{a}_j^{(l)})(x_{ji}^{(q)} - \hat{a}_j^{(q)}),\ \ l, q = 1,2,..., p.$$

где $\sum$ - внутригрупповая ковариация,

T – полная матрица,

$x_{\text{ji}}^{(\text{l})}, x_{ji}^{(q)}$ - значение параметров, из которых состоит первоначальная матрица,

$a^{(l)}, a^{(q)}$ - общее среднее,

$a_j^{(l)}, a_j^{(q)}$ - выборочное среднее по каждой группе.

Используются оценки матрицы внутригрупповой ковариации $\sum$ и полной матрицы T. Статистика Уилкса также аппроксимируется F-распределением. Если $\lambda=0$, то дискриминация считается идеальной, если $\lambda=1$, то она считается плохой.

Расстояние каждого объекта по дискриминирующей функции находится подстановкой координат каждого вектора в дискриминирующую функцию:

$$Z_i = f(x_{i1}, x_{i2}, ..., x_{im})$$

Кроме задачи классификации наблюдений в один из классов, интерес

представляет задача прогнозирования. Насколько хорошо построенные дискриминирующие фукнции могут предсказывать, к какому классу принадлежит конкретное наблюдение, анализируют таблицу сопряженности «Факт - прогноз» (табл.14) и оценивают вероятность ошибочной классификации каждого класса. Если процент ошибок высок, то причиной этого является либо нарушение предположений о нормальности многомерного распределения и равенстве ковариационных матриц, либо использование плохих дискриминантных переменных.

Таблица 14 – Таблица сопряженности

	Фактически	
Модель	положительно	отрицательно
положительно	TP	FP
отрицательно	FN	TN

- TP (*True Positives*) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);
- TN (*True Negatives*) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);
- FN (*False Negatives*) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый "ложный пропуск" – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);
- FP (*False Positives*) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах:

Доля истинно положительных примеров (True Positives Rate):

$$TPR = \frac{TP}{TP + FN} \cdot 100\%$$

Доля ложно положительных примеров (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \cdot 100\%$$

Также можно использовать следующий прием (так называемая «перекрестная проверка»): один из элементов обучающей выборки исключается и строится классификация по оставшимся элементам и удаленный элемент используется в качестве экзаменующего. Затем удаленный возвращается в выборку, удаляется второй элемент и процедура повторяется. Алгоритм заканчивает свою работу, когда удалению

подвергнутся все элементы выборки по очереди. Ошибка дискриминации складывается из ошибок дискриминации каждого экзаменующего элемента.

Рассмотрим примеры проведения дискриминантного анализа.

Задание 1.

Имеются данные[1], характеризующие кредитоспособность заемщика (табл. 15). Требуется на основании этих данных провести дискриминацию (классификацию) клиентов (заемщиков) с целью построения модели прогнозирования кредитоспособности нового заемщика, позволяющей принимать решения о выдаче кредита либо отказе.

Для построения модели оценки потенциальных заемщиков используются данные из анкет клиентов. Используем сформированную выборку в 27 человек, взявших потребительский кредит на срок – 1 год. Для фактора Пол примем значения: 0-женский, 1-мужской, для фактора Состояние в браке: 0-не состоящие в браке; 1 – состоящие в браке. Фактор благонадежности также бинарный (имеет два значения 0- не вернувшие кредит, 1 – вернувшие).

Технология выполнения:

1. Запустим программу Stadia и введем исходные данные из таблицы 15.
2. Теперь необходимо определиться, на сколько классов будем делить заемщиков. В данной задаче клиенты разделены на 2 класса – благонадежные (1 класс), неблагонадежные (2 класс). Для того чтобы программа поняла наше деление на классы, внесем числа, обозначающие класс, в пустой столбец, следующий за введенными данными. Соответственно, первые 17 строк (заемщиков) будут иметь первый класс, так как по фактору благонадежности они имеют значение 1, то есть они вернули кредит (поставим цифру 1), последующие 10 заемщиков будут отнесены ко второму классу, поставим цифру 2 (согласно исходным данным они не вернули кредит).
3. Далее выберем в меню пункт *Статист=F9*, в открывшемся окне в разделе *Многомерные методы* выберем *Дискриминантный*.
4. Получим следующий результат (табл. 16).

Рассмотрим полученные результаты.

1. В первой строке рассчитано расстояние Махаланобиса и значимость.
2. Расстояние Махаланобиса, как уже говорилось в разделе теории к данной теме, показывает, насколько каждый объект (заемщик) в n-мерном пространстве независимых переменных (факторов) отклоняется от центра статистической совокупности (т.е. от общего среднего). Чем эта величина больше, тем больше разница между выделяемыми априорно классами. Чем эта величина меньше, т.е. чем ближе она к нулю, тем разница между рассматриваемыми объектами (заемщиками) меньше, следовательно, делить на классы их не следует, либо нами выбраны не те признаки, по которым возможно четкое разделение на классы. В нашем случае объекты однородны, и их не следует подвергать делению на классы.

[1] Данные сформированы по результатам деятельности одного из банков второго уровня Республики Казахстан.

Таблица 15 – Исходные данные

№ п/п	Возраст	Пол	Состояние в браке	Иждивенцы	Стаж работы	Жилая недвижимость в собственности, $	Наличие кредита	Личный доход в месяц после налогообложения, тенге	Сумма кредита, тенге	Месячный платеж, тенге	Благонадежность
1	33	1	0	0	12	55000	0	50000	100000	16898	1
2	37	0	0	0	17	0	0	55000	70000	11829	1
3	25	0	0	0	3	0	0	35000	80000	13518	1
4	23	1	0	0	1	0	0	30000	50000	8449	1
5	26	1	1	1	3	33000	0	40000	100000	16899	1
6	28	0	0	0	4	0	0	40000	80000	13518	1
7	29	0	0	1	5	25000	0	50000	75000	12673	1
8	27	1	1	1	7	30000	1	80000	50000	8499	1
9	29	1	1	1	3	75000	0	70000	120000	20278	1
10	30	1	1	0	7	40000	0	60000	100000	16898	1
11	31	1	1	1	8	33000	0	55000	101000	17067	1
12	32	1	1	2	10	75000	0	62000	102000	17235	1
13	30	1	0	0	8	15000	0	61000	100000	16898	1
14	35	1	1	2	13	25000	0	58000	120000	20278	1
15	38	1	1	3	16	40000	0	57000	150000	25347	1
16	40	0	1	0	18	38000	0	75000	250000	42246	1
17	31	0	0	0	9	0	0	35000	50000	8449	1
18	50	1	1	1	28	45000	0	42000	65000	10984	0
19	24	1	1	1	2	0	0	20000	50000	8449	0
20	29	1	1	2	5	50000	0	60000	200000	33797	0
21	27	1	1	2	8	50000	0	60000	120000	20278	0
22	28	0	1	0	10	40000	1	40000	100000	16898	0
23	29	1	0	0	6	0	0	90000	80000	13518	0
24	26	1	0	0	3	35000	0	40000	90000	15208	0
25	39	0	1	2	17	55000	0	39000	200000	33797	0
26	30	1	0	0	8	0	0	28000	80000	13519	0
27	37	1	1	2	15	62000	0	100000	500000	87493	0

Таблица 16 – Результаты дискриминантного анализа

ДИСКРИМИНАНТНЫЙ АНАЛИЗ. Файл:

Расстояние Махаланобиса=3,066, значимость=0,9898

Класс	<--- Коэффициенты дискриминантной функции: a[0],a[1],... --->								
1	-205	17,58	10,37	-10,59	2,634	-15,42-0,0002657	73,82-0,0001598	0,001385-0,007757	0
2	-201,7	17,41	12,53	-11,14	2,697	-15,2-0,0002772	76,18-0,0002227	0,001524-0,008453	0

Объект	Класс	D^2	Значим	Вероят.отнесения
1	2	10,12	0,5198	0,6189
2	1	7,911	0,7212	0,8299
3	1	4,13	0,9659	0,7492
4	2	4,303	0,9602	0,6573
5	2	3,678	0,9784	0,5319
6	1	4,955	0,9333	0,8487
7	1	8,47	0,6706	0,9334
8	1	15,75	0,1505	0,5915
9	1	10,85	0,4555	0,9137
10	1	6,486	0,839	0,7496
11	1	2,083	0,9981	0,6393
12	1	7,051	0,7949	0,7586
13	1	4,263	0,9615	0,5256
14	2	4,127	0,966	0,5494
15	2	8,764	0,6436	0,7151
16	1	16,44	0,1256	0,5808
17	1	4,057	0,9682	0,8122
18	2	17,38	0,097	0,6012
19	2	10,14	0,5176	0,6992
20	2	7,794	0,7316	0,6824
21	2	14,05	0,2302	0,5252
22	2	14,72	0,1954	0,8021
23	1	10,56	0,4807	0,9203
24	2	5,616	0,8977	0,5926
25	2	8,153	0,6995	0,6823
26	2	5,985	0,8743	0,8485
27	2	22,5	0,0207	0,9601

3. Показатель значимость можно сравнить с показателем лямбды Уилкса. Дискриминация считается идеальной (как говорилось выше), если лямбда Уилкса равна нулю и считается плохой, если этот показатель равен 1. В нашем примере показатель значимости равен 0,9898, то есть очень близок к единице. Следовательно, дискриминация проведена неверно.
4. Далее представлены коэффициенты дискриминирующей функции. Дискриминирующая функция представляет собой уравнение регрессии, состоящее из свободного члена и коэффициентов при переменных. Интерпретация коэффициентов здесь такая же, как и в регрессионном анализе: чем выше их значение, тем значительнее вклад данной переменной в дискриминацию объектов. Так как мы делили объекты на два класса, то выводится две функции. С помощью этих функций возможно прогнозирование: в данном случае отнесение какого-либо клиента с указанными характеристиками к благонадежным или неблагонадежным заемщикам. Характеристики заемщика подставляются в обе функции, значение какой функции больше, к тому классу и относится (теоретически) заемщик. Не следует забывать, что для того чтобы прогноз был надежным, показатели расстояния Махаланобиса и значимости должны соответствовать своим «условно эталонным» значениям.
5. Кроме того, для определения надежности дискриминирующих функций можно построить и проанализировать таблицу сопряженности (табл. 14). В нее требуется внести данные по результатам решения, отнесения объектов к классам и сравнить с тем, как было предопределено это отнесение до решения задачи. В этом нам помогут результаты, выводимые после дискриминирующих функций, в частности колонка *Класс*. Нам требуется посчитать, сколько объектов, определенных нами в первый класс, неправильно классифицированы, то есть отнесены программой во второй класс, и наоборот, сколько объектов, отнесенных нами во второй класс, отнесены программой в первый. И сколько и там и там классифицировано правильно (истинно положительные и отрицательные случаи).
6. Заполним следующую таблицу (табл. 17)

Таблица 17 - Таблица сопряженности

	Фактически		**Итого**
Модель	положительно	отрицательно	
положительно	12	1	13
отрицательно	5	9	14
Итого	17	10	

Комментарий к заполнению таблицы 17.

Для того чтобы получить число 12, так называемые истинно положительные случаи, следует посчитать количество объектов, отнесенных к первому классу программой, среди объектов первого класса, определенных нами до решения задачи. То есть, мы должны посчитать количество единиц среди

первых 17 объектов (именно столько мы отнесли заемщиков к первому классу).

Чтобы получить число 5 – положительные примеры, классифицированные как отрицательные, необходимо в той же группе (первые 17 клиентов), посчитать количество двоек, то есть, сколько объектов первого класса программа отнесла ко второму. Либо можно вычесть из общего числа объектов первого класса, то есть из 17, количество верно классифицированных объектов – 12, и получим число 5.

Для определения истинно отрицательных случаев, то есть количества клиентов, отнесенных и нами, и программой ко второму классу, считаем их количество во второй половине таблицы результатов решения, начиная с 18 объекта. Верно классифицированных объектов 9 из 10. Следовательно, неверно классифицирован один объект, так называемый ложно положительный случай.

7. Для более лучшего понимания полученных в таблице 17 данных рассчитаем доли истинно положительных, истинно отрицательных, ложно положительных, ложно отрицательных случаев в процентах. Для этого вычислим отношение чисел каждого столбца к сумме по этому столбцу. Например, разделив 12 на 17, мы узнаем количество верно классифицированных положительных примеров, их 71%, количество неверно классифицированных положительных объектов (благонадежных заемщиков, отнесенных программой к неблагонадежным) – 29% (ошибка первого рода).
8. Следуя той же логике расчетов, при анализе отрицательных случаев получим 90% верно классифицированных заемщиков, не вернувших кредит, и 10% неблагонадежных заемщиков, отнесенных программой к благонадежным (ошибка второго рода).
9. В целом по анализу и расчетам таблицы 17 можно сделать вывод, что ошибки первого и второго рода достаточно велики, особенно если учесть размер выборки, поэтому проведенная дискриминация объектов на классы неудачна.
10. Об этом также можно судить по значениям D^2 (расстояния Махаланобиса внутри каждого объекта). Все эти значения по объектам значительно больше общего показателя расстояния Махаланобиса, равного 3,066, что говорит о том, что разница средних внутри объектов и общим средним невелика, то есть объекты однородны, и делить их на классы не следует.
11. Также следует обратить внимание на показатель *Вероятность отнесения* каждого объекта к определенному классу. Естественно, что абсолютная вероятность равна 1, в нашем же случае вероятность отнесения многих объектов к первому или второму классам недостаточно высока.

В целом следует сделать вывод, что проведенная классификация заемщиков по классам с помощью дискриминантного анализа не дала положительные результаты, что подтверждают рассчитанные числовые характеристики. То есть данная группа заемщиков однородна, либо переменные, характеризующие клиентов (возраст, семейное положение, наличие недвижимости, доход, сумма кредита и т.д.) не являются теми переменными, которые

однозначно делят клиентов на два класса.

Далее рассмотрим другой пример, в котором классификация объектов четко видна.

Задание 2.

По данным таблицы 18, содержащей показатели, характеризующие деятельность 12 банков второго уровня Республики Казахстан, провести дискриминантный анализ, позволяющий разделить совокупность банков на 2 класса. В первый класс должны попасть банки, имеющие высокие характеристики, в другой – относительно невысокие.

Таблица 18 – Исходные данные[1]

млрд. тенге

Номер банка	Наименование банка	Активы	Ссудный портфель	Обязательства	Собственный капитал	Расходы, связанные с выплатой вознаграждения	Чистый доход, связанный с получением вознаграждения
1	АО «БТА Банк»	1 993,994	1 644,707	1 812,572	181,422	200,164	-66,608
2	АО «Народный Банк Казахстана»	2 023,522	1 224,416	1 755,757	267,765	87,581	93,251
3	АО «KASPI BANK»	352,905	288,704	317,049	35,856	26,646	23,898
4	АО «Банк ЦентрКредит»	1 211,057	719,928	1 134,266	76,790	79,982	12,465
5	ДБ АО «Сбербанк»	282,692	176,645	242,780	39,913	9,163	11,891
6	АО «Евразийский Банк»	358,761	227,546	333,260	25,502	22,039	7,173
7	АО «АТФ Банк»	982,965	849,475	919,034	63,931	56,517	24,235
8	АО «Цеснабанк»	227,211	149,288	203,640	23,571	14,217	7,112
9	АО «Альянс Банк»	489,546	545,529	455,598	33,948	37,798	5,363
10	АО «Жилстройсбербанк Казахстана»	101,188	67,702	76,138	25,050	1,097	4,163
11	АО «Метрокомбанк»	18,683	7,597	12,160	6,523	0,329	0,494
12	АО «Казкоммерцбанк»	2 430,236	2 344,296	2 151,743	278,494	142,521	134,572

[1] Источник данных – сайт Комитета по контролю и надзору финансового рынка и финансовых организаций Национального Банка Республики Казахстан (www.afn.kz)

Технология выполнения:

1. Прежде чем выполнять дискриминантный анализ, определим, к какому из двух классов относится каждый из банков. Банки под номерами 1, 2, 4, 7, 12 отнесем к первому классу, остальные - ко второму[1]. Назовем первый класс – крупные банки, второй – средние.
2. Внесем исходные данные в программу Stadia.
3. Выберем пункт меню *Статист-F9*, в открывшемся окне укажем пункт *Дискриминантный анализ*.
4. Получим следующие результаты:

```
ДИСКРИМИНАНТНЫЙ АНАЛИЗ.  Файл:

    Расстояние Махаланобиса=5,864E6, значимость=0
    Класс  <--- Коэффициенты дискриминантной функции:a[0],a[1],... --->
        1   -37,02     6139  0,02068    -6139    -6139   -1,163  -0,5077
        2  -0,6508   -388,80,0002165    388,8    388,8  -0,1341 -0,05536

    Объект   Класс      D^2   Значим Вероят.отнесения
        1        1    7,372   0,2878        1
        2        1    7,092   0,3123        1
        3        2    4,308    0,635        1
        4        1    4,159   0,6551        1
        5        2    2,454   0,8736        1
        6        2    4,394   0,6235        1
        7        1    1,552   0,9559        1
        8        2    0,323   0,9993        1
        9        2    6,558   0,3636        1
       10        2    2,308   0,8892        1
       11        2    1,428   0,9641        1
       12        1    3,891   0,6914        1
```

5. Во-первых, величина расстояния Махаланобиса достаточно высока (5864000), а значимость (лямбда Уилкса) равна нулю. Это свидетельствует о хорошей дискриминации объектов.
6. Во-вторых, вероятность отнесения каждого объекта к определенному классу равна 1, то есть каждый из объектов классифицирован правильно.
7. Так все исходные объекты (банки) отнесены именно к тому классу, к которому и предполагалось, считать таблицу сопряженности нет смысла, ошибки первого и второго рода равны нулю.
8. Далее сделаем прогноз с целью отнесения банка АО «Нурбанк» к одному из двух классов[2]. Характеристики банка следующие: активы – 259,187 млрд. тенге, ссудный портфель – 198,016 млрд. тенге, обязательства – 214,621 млрд. тенге, собственный капитал – 44,566 млрд. тенге, расходы, связанные с выплатой вознаграждения – 17,776 млрд. тенге, чистый доход, связанный с получением вознаграждения – 8,134 млрд. тенге. Подставим эти значения в каждую из двух дискриминирующих функций:

[1] Для определения класса, к которому следует отнести банк, можно построить гистограмму, отображающую параметры каждого банка, и визуально приблизительно определить, какие банки относятся к классу с высокими характеристиками (т.е. к первому классу) и какие - с относительно невысокими (ко второму классу).

[2] Сведения по банку АО «Нурбанк» взяты с сайта www.afn.kz (Комитет по контролю и надзору финансового рынка и финансовых организаций Национального Банка Республики Казахстан)

1 класс: -37,02 + 6139×259,187 + 0,02068×198,016 + (-6139) ×214,621 + (-6139) ×44,566 + (-1,163) ×17,776 + (-0,5077) ×8,314 = -20708518,42

2 класс: -0,6508 -388,8×259,187 + 0,0002165×198,016 + 388,8×214,621 + 388,8×44,566 + (-0,1341) ×17,886 + (-0,05536) ×8,134 = -2791228,44

9. Сравниваем значения обеих функций, значение второй функции больше, следовательно, АО «Нурбанк» следует отнести ко второму классу, т.е. к классу средних банков.

В целом можно сказать, что классификация банков проведена верно, расчетные показатели, полученные с использованием инструмента дискриминантного анализа, подтверждают это разделение на два класса. Полученными дискриминирующими функциями можно пользоваться для предсказания, к какому классу принадлежит определенный объект (банк), не входящий в первоначальную выборку, что и было продемонстрировано на примере АО «Нурбанк».

Задание 3.

Провести дискриминантный анализ по данным, представленным в таблице 19. В таблице представлены 32 объекта, характеризующиеся четырьмя показателями x_1, x_2, x_3, x_4. На основе полученных дискриминирующих функций определить принадлежность объектов 33, 34, 35, 36 к соответствующим классам.

Таблица 19 – Исходные данные [3]

№ п/п	Товарная продукция, тыс. усл. ден. ед. (x_1)	Фондоотдача, усл. ден. ед. (x_2)	Текущие затраты, тыс. усл. ден. ед. (x_3)	Единовременные затраты, тыс. усл. ден. ед. (x_4)	Класс
1	**2**	**3**	**4**	**5**	**6**
1	1.75	5.25	8	16	1
2	2.65	5.5	11	15	1
3	1.8	4.47	10	6	1
4	2.5	4.75	7	9	1
5	3.0	5.00	12	7	1
6	3.54	4.71	8	6	1
7	2.82	4.22	10	5	1
8	2.53	4.07	8	5	1
9	2.25	4.04	12	16	1
10	2.06	3.95	9	8	1
11	2.75	3.75	9	7	1
12	3.24	3.93	10	8	1
13	4.14	3.52	12	11	2
14	4.62	3.34	13	7	2

Продолжение табл. 19

1	2	3	4	5	6
15	4.27	3.11	11	8	2
16	3.78	2.92	15	6	2
17	4.25	2.76	14	5	2
18	4.53	2.60	14	5	2
19	5.28	5.00	13	7	3
20	5.21	4.54	10	5	3
21	6.01	5.00	12	9	3
22	5.25	4.23	15	8	3
23	6.00	4.37	14	8	3
24	5.76	3.92	15	7	3
25	3.5	2.00	8	4	4
26	2.77	1.54	4	4	4
27	3.51	1.40	3	6	4
28	4.00	1.75	7	2	4
29	3.03	1.00	8	2	4
30	3.25	0.88	6	3	4
31	3.69	0.49	5	2	4
32	4.20	1.00	7	2	4
33	2.30	6.11	10	14	
34	5.21	1.18	13	4	
35	4.75	3.25	11	8	
36	3.13	2.50	7	3	

Задание 4.

Провести дискриминантный анализ по представленным ниже данным (табл. 20), разделив совокупность объектов на 2 класса и самостоятельно определив, к какому классу может относиться каждый из объектов. Составить и рассчитать таблицу сопряженности и ошибки первого и второго рода. Интерпретировать полученный результат.

Таблица 20 – Исходные данные[1]

тыс. тенге

	Наименование организации	Активы	Обязательства	Собственный капитал	Нераспределенный доход (непокрытый убыток) отчетного периода
1	2	3	4	5	6
	Накопительные пенсионные фонды				
1.	АО "Накопительный пенсионный фонд "Ұлар Үміт"	7 071 621	701 677	6 369 944	-651 475
2.	АО "Открытый Накопительный Пенсионный Фонд "АМАНАТ КАЗАХСТАН"	1 337 185	66 271	1 270 914	-454 656

[1] Все данные взяты с сайта Комитета по контролю и надзору финансового рынка и финансовых организаций Национального Банка Республики Казахстан – www.afn.kz.

Продолжение табл. 20

1	2	3	4	5	6
3.	АО Накопительный пенсионный фонд "Атамекен" ДО АО "Нурбанк"	1 772 831	78 287	1 694 544	24 526
4.	АО Накопительный пенсионный фонд "ГНПФ"	18 394 752	511 713	17 883 039	3 285 772
5.	АО "Накопительный пенсионный фонд БТА Казахстан "дочерняя организация АО "БТА Банк"	11 605 223	1 280 287	10 324 936	-1 250 805
	Банки второго уровня				
6.	АО "КАЗКОММЕРЦБАНК"	2 430 236 299	2 151 742 785	278 493 514	104 475
7.	АО "Народный Банк Казахстана"	2 023 521 731	1 755 756 612	267 765 119	27 418 021
8.	АО "БТА Банк"	1 993 993 880	1 812 571 968	181 421 912	1 150 232 213
9.	АО "Банк ЦентрКредит"	1 211 056 597	1 134 266 337	76 790 260	- 29 306 934
10.	АО "АТФБанк"	982 965 005	919 033 776	63 931 229	- 37 674 485
	Организации, осуществляющие брокерскую и дилерскую деятельности				
11.	АО "VISOR Capital" (ВИЗОР Капитал)	2 374 313	572 401	1 801 912	-150 900
12.	АО "ФК "Альянс Капитал"	899 661	29 351	870 310	-165 540
13.	АО "Smart Group"	688 377	471 054	217 323	-35 922
14.	АО "ДБ Секьюритиз (Казахстан)"	475 986	32 129	443 857	27 033
15.	АО "Premier Capital"	372 993	28 261	344 732	88 544
	Ипотечные организации				
16.	АО «Алматинская Региональная Ипотечная Организация»	824 935	527	824 408	-52 160
17.	АО «Ипотечная организация «Астана-финанс»	12 778 698	5 524 960	7 253 738	-1 362 432
18.	АО «Ипотечная организация «Элитстрой Финанс»	939 729	58 986	880 743	26 339
19.	АО «Ипотечная Организация «Бирюза»	953 848	7 299	946 549	-1 914
20.	АО "Ипотечная Организация "Казахстанская Ипотечная Компания"	106 373 716	85 869 928	20 503 788	63 394
	Организации, управляющие инвестиционным портфелем, совмещающие деятельность с брокерской и (или) дилерской и отдельными видами банковских операций				
21.	АО "HomeBroker"	30 845 885	6 477 127	24 368 758	6 173 266
22.	АО «Финансовая компания «REAL-INVEST.kz»	5 304 051	2 142 263	3 161 788	365 781
	Организациях, осуществляющие управление инвестиционным портфелем				
23.	АО "Тройка Диалог Казахстан"	945 563	412 262	533 301	77 095
24.	АО "Gold Investment Group" ("Голд Инвестмент Групп")	940 058	3 470	936 588	3 339
25.	АО "Верный Капитал"	915 108	88 144	826 964	181 909
26.	АО "FIVE BROKERS CAPITAL"	632 969	5 810	627 159	327 276
27.	АО "Алматинский Финансовый Центр"	544 982	19 301	525 681	38 956

Задание 5.

По данным по вариантам, приведенным в приложении А настоящего практикума, провести дискриминантный анализ, разделив представленные коммерческие банки на два класса (к какому классу относится каждый банк, определить самостоятельно). Перед процедурой дискриминации провести корреляционный анализ взаимодействия показателей деятельности банков между собой с целью выявления мультиколлинеарности факторов, исключить такие факторы. Составить и рассчитать таблицу сопряженности и ошибки первого и второго рода. Интерпретировать полученные результаты.

Тема 5. Многомерное шкалирование

Цель: изучить основы многомерного шкалирования, приобрести умения и навыки проведения исследования с помощью методики многомерного шкалирования и интерпретации полученных результатов.

Многомерное шкалирование (МНШ) – методы, реализующие «погружение» анализируемых объектов, о которых известна лишь структура их попарных расстояний (мер сходства), в некоторое специально подобранное координатное пространство небольшой размерности.

Основные типы задач, решаемые методами МНШ:

- Поиск и интерпретация латентных переменных, объясняющих заданную структуру попарных расстояний. Этот тип задач многомерного шкалирования предусматривает не только построение вспомогательных шкал (координатных осей), в системе которых затем рассматриваются анализируемые объекты, но и содержательную интерпретацию этих шкал;
- Верификация геометрической конфигурации системы анализируемых объектов в координатном пространстве латентных переменных;
- Сжатие исходного массива данных с минимальными потерями в их информативности.

Многомерное шкалирование можно рассматривать как альтернативу факторному анализу. Целью последнего является поиск и интерпретация латентных переменных, дающих возможность исследователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков. В факторном анализе сходства/различия между объектами выражаются с помощью матрицы коэффициентов корреляции. В методе МНШ дополнительно к корреляционным матрицам в качестве исходных данных можно использовать произвольный тип матрицы сходства объектов. *На входе алгоритма МНШ* – матрица, элементы которой содержат сведения о попарном сходстве/различиях анализируемых объектов. *На выходе алгоритма МНШ* получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во «вспомогательных шкалах», связанных с латентными переменными), причем размерность нового пространства признаков существенно меньше размерности исходного.

Логику МНШ можно проиллюстрировать на следующем простом примере. Предположим, что имеется матрица попарных расстояний (т.е. сходства некоторых признаков) между городами определенной страны. Анализируя матрицу, стремятся расположить точки с координатами городов в двумерном пространстве (на плоскости), максимально сохранив реальные расстояния между ними. Полученное размещение точек на плоскости впоследствии можно использовать в качестве приближенной географической карты государства.

В общем случае метод МНШ позволяет таким образом расположить «объекты» (города в нашем примере) в пространстве некоторой небольшой размерности (в данном случае она равна двум), чтобы достаточно адекватно воспроизвести наблюдаемые расстояния между ними. В результате можно «измерить» эти расстояния в терминах найденных латентных переменных. Так, в нашем примере можно объяснить расстояния в терминах пары географических координат Север/Юг и Восток/Запад.

Ориентация осей координат. Как и в факторном анализе, ориентация осей может быть выбрана произвольной. Возвращаясь к нашему примеру, можно поворачивать карту страны произвольным образом, но расстояния между городами при этом не изменятся. Таким образом, окончательная ориентация осей на плоскости или в пространстве является, в большей степени результатом содержательного решения в конкретной предметной области (т.е. решением пользователя, который выберет такую ориентацию осей, которую легче всего интерпретировать). В примере можно было бы выбрать ориентацию осей, отличающуюся от пары Север/Юг и Восток/Запад, однако последняя удобнее, как «наиболее осмысленная» и естественная.

Основной тип данных в многомерном шкалировании – меры близости между двумя объектами, измеряющие насколько эти объекты похожи. Наиболее часто используются коэффициенты корреляции и совместные вероятности.

Обозначим меру близости двух объектов i и j как δ_{ij}. Если эта мера такова, что самые большие ее значения соответствуют парам наиболее похожих объектов, то δ_{ij} – мера сходства, если ее наибольшие значения соответствуют парам наименее похожих объектов, то δ_{ij} – мера различия (далее по умолчанию предполагается, что δ_{ij} – мера различия).

Необходимо, чтобы исходные данные представляли матрицу расстояний между объектами γ_{ij}, т.е. выполнен набор аксиом расстояний:

1) $\gamma_{ij} \geq 0$, $\gamma_{ij} = 0$, $i,j=1,\ldots,n$;
2) $\gamma_{ij} = \gamma_{ji}$, $i,j=1,\ldots,n$;
3) $\gamma_{ij} + \gamma_{jk} \geq \gamma_{jk}$, $i,j,k=1,\ldots,n$.

При этом часто матрица, оценивающая различия объектов, строится по результатам опросов экспертов на основе шкал сравнения или в категоризованной форме. В этом случае свойство 1) выполнено, а свойства 2) и 3) оказываются нарушенными. Для удовлетворения свойства 2) полагают $\gamma_{ij} = \gamma_{ji} = \frac{1}{2}(\delta_{ij} + \delta_{ji})$.

В *классической модели МНШ, предложенной У. Торгерсоном*, предполагается, что по этой матрице можно построить точки $x_1,\ldots,x_n$ в пространстве некоторого небольшого числа измерений p, причем так, чтобы каждый элемент матрицы γ_{ij} представлял собой евклидово расстояние

$$\gamma_{ij} = \sqrt{\sum_{r=1}^{p} (x_i^{(r)} - x_j^{(r)})^2} \qquad (5.1)$$

между объектом i и объектом j, т.е. в матричном виде:

$$\underset{n\times n}{\Gamma} = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} \end{bmatrix} \qquad (5.2)$$

Неизвестны координаты объектов:

$$X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)}),\ i = 1,...,n \qquad (5.3)$$

Требуется на основании известных расстояний вида (5.2) восстановить неизвестную размерность *p* пространства признаков и приписать каждому объекту координаты (5.3) так, чтобы вычисленные евклидовы расстояния совпали с заданными (5.2).

Таким образом, *цель МНШ в том, чтобы преобразовать информацию о конфигурации исходных многомерных данных, заданную матрицей расстояний, в геометрическую конфигурацию в многомерном пространстве*.

В неметрическом многомерном шкалировании предполагается, что различие между объектами измерено в рангах. Процедура неметрического многомерного шкалирования строится так, чтобы ранговый порядок попарных расстояний между точками минимально отличался от того, который задан.

Дж. Краскалом для определения меры соответствия предложен следующий критерий, названный им «*стресс*»:

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\gamma_{ij}^{*} - \gamma_{ij}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}^{2}}},$$

где γ_{ij}^{*} - расстояние между объектами в р-мерном пространстве;

γ_{ij} - расстояние между объектами в исходном пространстве данных.

Вообще говоря, чем больше размерность пространства, используемого для воспроизведения расстояний, тем лучше согласие воспроизведенной матрицы с исходной (меньше значение стресса). Если взять размерность пространства равной числу переменных, то возможно абсолютно точное воспроизведение исходной матрицы расстояний. Однако нашей *целью является упрощение решаемой задачи, с тем, чтобы объяснить матрицу сходства (расстояний) в терминах лишь нескольких важнейших факторов (латентных переменных или вспомогательных шкал)*.

Возвращаясь к нашему примеру с расстояниями между городами, если получена двумерная карта, намного проще представить себе расположение городов и планировать передвижение между ними, чем если бы имелась только матрица попарных расстояний.

Уменьшение числа факторов (или вспомогательных шкал) может приводить к ухудшению представления исходной матрицы. Для демонстрации этого предположения обозначим буквами A, B, C и D, E, F две тройки горо-

дов. Соответствующие им точки и попарные расстояния между ними показаны в двух матрицах.

	A	**B**	**C**
A	0		
B	90	0	
C	90	90	90

	D	**E**	**F**
D	0		
E	90	0	
F	180	90	0

Первой матрице соответствует случай, когда города удалены друг от друга в точности на 90 километров, а второй - когда города ***D*** и ***F*** удаляются на 180 километров. Можно ли три точки, соответствующие городам (объектам) расположить в одномерном пространстве (на прямой)? Действительно, три точки, соответствующие городам ***D***, ***E*** и ***F*** могут быть расположены на прямой линии:

D---*90 км*---***E***---*90 км*---***F***

D удален на 90 км от города ***E***, и ***E*** - на 90 км от ***F***, а город ***D*** удален на 90+90=180 км от ***F***. Если попытаться проделать тоже самое с городами ***A***, ***B*** и ***C***, то видно, что соответствующие им точки уже нельзя разместить на прямой с сохранением исходной структуры расстояний. Однако эти точки можно расположить на плоскости, например, в виде треугольника:

A

90 км 90 км

B 90 км **C**

Располагая эти три точки так, можно в точности воспроизвести все расстояния между ними. Без лишних деталей, этот пример показывает, как конкретная матрица расстояний (сходств) связана с числом искомых латентных переменных (размерностью результирующего пространства). Конечно, «реальные» данные никогда не являются такими «точными», и содержат случайный шум, т.е. случайную изменчивость, влияющую на различие между воспроизведенной и исходной матрицей.

Обычно, *для выбора размерности пространства, в котором будет воспроизводится наблюдаемая матрица, используют график зависимости стресса от размерности (график каменистой осыпи)*. Этот критерий впервые был предложен Кэттелом в контексте решения задачи снижения размерности в факторном анализе, Краскал и Виш применили этот график в методе МНШ.

Кэттел предложил найти такую абсциссу на графике (в методе факторного анализа по оси абсцисс идут собственные значения), в которой график стресса начинает визуально сглаживаться в направлении правой, пологой его части, и, таким образом, уменьшение стресса максимально замедляется. Образно говоря, линия на рисунке напоминает скалистый обрыв, а черные точки

на графике напоминают камни, которые ранее упали вниз. Таким образом, внизу наблюдается как бы каменистая осыпь из таких точек. Справа от выбранной точки на оси абсцисс, лежит только «факторная осыпь».

Вторым критерием для решения вопроса о размерности с целью интерпретации является «ясность» полученной конфигурации точек. Иногда, как в нашем примере с городами, результирующие координаты легко интерпретируются. В других случаях, точки на графике могут образовывать ту или иную разновидность «случайного облака», и не существует непосредственного способа для интерпретации латентных переменных. В последнем случае следует постараться немного увеличить число координатных осей и рассмотреть получаемые в результате конфигурации. Чаще всего, получаемые решения проще удается проинтерпретировать. Однако если точки на графике не следуют какому-либо образцу, а также если график стресса не показывает какого-либо явного «изгиба» (и не похож на «край обрыва»), то данные, скорее всего, являются случайным «шумом».

Преимущество метода МНШ в том, что можно анализировать произвольный тип матрицы расстояний или сходства. Эти сходства могут представлять собой оценки экспертов относительно сходства данных объектов, результаты измерения расстояний в некоторой метрике, процент согласия между судьями по поводу принимаемого решения, и мн.др. Например, методы МНШ весьма популярны в психологическом исследовании восприятия личности. В этом исследовании анализируются сходства между определенными чертами характера с целью выявления основополагающих личностных качеств. Также они популярны в маркетинговых исследованиях, где их используют для выявления числа и сущности латентных переменных (факторов), например, с целью изучения отношения людей к товарам известных торговых марок.

В общем случае, методы МНШ позволяют исследователю задать клиентам в анкете относительно ненавязчивые вопросы («насколько похож товар фирмы А на товар фирмы В») и найти латентные переменные для этих анкет незаметно для респондентов.

Для получения мер близости используются в основном четыре класса мер близости: прямые оценки различия, условные вероятности, совместные вероятности и индексы различия профилей. Подробно о способах получения мер близости изложено в [3, глава 3].

Многомерное шкалирование и факторный анализ

Даже, несмотря на то, что имеется много сходства в характере исследуемых вопросов, методы МНШ и факторного анализа имеют ряд существенных отличий.

Факторный анализ требует, чтобы исследуемые данные подчинялись многомерному нормальному распределению, а зависимости были линейными. Методы МНШ не накладывают таких ограничений.

Методы МНШ могут быть применимы, пока сохраняет смысл порядок следования рангов сходств.

В терминах различий получаемых результатов, факторный анализ стремится извлечь больше факторов (координатных осей или латентных переменных) по сравнению с МНШ; в результате чего МНШ часто приводит к проще интерпретируемым решениям.

Методы МНШ можно применять к любым типам расстояний или сходств, тогда как методы факторного анализа требуют, чтобы первоначально была вычислена матрица корреляций.

Методы МНШ могут быть основаны на прямом оценивании сходств между объектами, тогда как факторный анализ требует, чтобы объекты были оценены через их признаки по некоторому списку атрибутов.

Таким образом, схема решения задач МНШ состоит из следующих этапов:

1. Формулирование проблемы (задачи).
2. Подготовка исходных данных. Из статистических данных рассчитываются коэффициенты, показывающие меру различия между объектами (γ_{ij}). Коэффициенты могут быть разной природы, представлены в виде попарных разностей, коэффициентов корреляции, экспертных оценок.
3. Формирование из рассчитанных ранее коэффициентов квадратной матрицы, представляющей собой меру различия между исследуемыми объектами.
4. Выбор метода МНШ.
5. Проведение непосредственно процесса расчета с помощью одного из программных продуктов (Stadia, Statistica, SPSS и др.), предназначенных для подобного рода вычислений. При решении задачи формируются новые координатные оси, соответствующие сокращенному количеству признаков, характеризующих исходные объекты, и соответствующая им матрица.
6. Определение оптимального количества осей (выбор координатных осей).
7. Интерпретация полученного решения в выбранных координатных осях.
8. Оценивание надежности и достоверности решения.

Суммируя вышесказанное, можно сказать, что методы МНШ потенциально применимы к более широкому классу исследовательских задач.

Задание 1.

Имеются данные по благоустройству жилья по некоторым областям Республики Казахстан за 2009 год (табл. 21). Требуется путем проведения процедуры многомерного шкалирования сократить количество признаков, подобрав необходимое и наглядное координатное пространство, дав полученным осям значимые названия и интерпретировать полученный результат.

Таблица 21 – Данные по благоустройству жилья в 2009 по отдельным областям Казахстана[1]

в процентах

	A	B	C	D	E	F
1		Благоустройство жилья				
2		Признаки / Области	Водопровод	Центральное отопление	Газ	Горячее водоснабжение
3		Акмолинская	45,4	31,4	93,8	16
4		Актюбинская	64,6	51,4	92,7	42,5
5		Алматинская	35,6	12,9	95,5	11,8
6		Атырауская	63,6	37,8	97,2	37,5
7		Западно-Казахстанская	22,2	19,4	95,2	17,4
8		Жамбылская	62,8	17,8	99,3	16,6
9		Карагандинская	87,1	67,8	65,3	49,9
10		Костанайская	50,5	45,7	97,5	42,7
11		Кызылординская	25,8	16,6	96,8	1,8
12		Мангистауская	70,7	68,5	97,6	67,9

Технология выполнения:

1. Как уже говорилось в разделе теории к данной теме, отправной точкой при проведении процедуры многомерного шкалирования служат меры сходства/различия между объектами (областями в нашем случае). Поэтому необходимо преобразовать исходные данные (табл. 21) в квадратную матрицу попарных расстояний.
2. Для этого выполним следующие действия:

2.1 Введем данные из таблицы 21 на рабочий лист MS Excel, начиная с ячейки B2.

2.2 Далее на пустом месте этого же листа создадим таблицу следующего вида (рис. 46).

	A	B	C	D	E	F	G	H	I	J	K	L
20												
21												
22			Акмолинская	Актюбинская	Алматинская	Атырауская	Западно-Казахстанская	Жамбылская	Карагандинская	Костанайская	Кызылординская	Мангистауская
23	1	Акмолинская										
24	2	Актюбинская										
25	3	Алматинская										
26	4	Атырауская										
27	5	Западно-Казахстанская										
28	6	Жамбылская										
29	7	Карагандинская										
30	8	Костанайская										
31	9	Кызылординская										
32	10	Мангистауская										

Рисунок 46 – Расположение таблицы расчета значений попарных расстояний на листе MS Excel

[1]Данные взяты из статистического сборника «Уровень жизни населения в Казахстане», 2010. – с. 162

2.3 Теперь рассчитаем значения попарных расстояний между объектами (областями) по формуле евклидова расстояния (см. формулу 5.1). Например, для получения значения расстояния между Акмолинской и Актюбинской областями формула будет выглядеть так: $\sqrt{(45{,}4-64{,}6)^2+(31{,}4-51{,}4)^2+(93{,}8-92{,}7)^2+(16-42{,}5)^2}=38{,}37$

2.4 Такие расчеты следует провести по всем объектам. Для упрощения этой процедуры введем формулу в соответствующие ячейки, используя математические функции КОРЕНЬ и СУММКВРАЗН (сумма квадратов разностей) (нажав в строке формул f_x). Пример ввода формул представлен на рис. 47.

2.5 Обратите внимание, что на пересечении одинаковых областей расстояние равно нулю.

2.6 В данной матрице-таблице имеются пустые ячейки, для получения значений в них необходимо «зеркально» отобразить уже рассчитанные числа, поскольку объекты те же самые, но сочетаются по строкам.

2.7 Для проведения процедуры «зеркального отображения» сделаем следующее:

2.7.1 Выделим диапазон ячеек `C23:L32` и скопируем данные.

2.7.2 Установив курсор на пустую ячейку, например, `B35`, вызовем контекстное меню (нажатием правой кнопки мыши), выберем в нем пункт *Специальная вставка*, в открывшемся окне в разделе *Вставить* укажем опцию *Значения*, в нижней части окна отметим опцию *Транспонировать*. Нажмем *ОК*.

2.7.3 Теперь снова выделим весь вновь созданный диапазон ячеек (`B35:K44`), скопируем его, установим курсор в ячейку `C23`, вызовем контекстное меню, пункт *Специальная вставка*, в отрывшемся окне отметим опцию *Пропускать пустые ячейки*. Нажмем *ОК*.

2.7.4 В итоге получим требуемую матрицу попарных расстояний (рис. 48).

3. Откроем программу Stadia и введем в качестве исходных данных полученную в MS Excel матрицу.

	B	C	D	E
19				
20				
21				
22		Акмолинская	Актюбинская	Алматинская
23	Акмолинская	0		
24	Актюбинская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C4:F4))	0	
25	Алматинская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C5:F5))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C5:F5))	0
26	Атырауская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C6:F6))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C6:F6))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C6:F6))
27	Западно-Казахстанская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C7:F7))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C7:F7))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C7:F7))
28	Жамбылская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C8:F8))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C8:F8))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C8:F8))
29	Карагандинская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C9:F9))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C9:F9))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C9:F9))
30	Костанайская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C10:F10))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C10:F10))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C10:F10))
31	Кызылординская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C11:F11))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C11:F11))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C11:F11))
32	Мангистауская	=КОРЕНЬ(СУММКВРАЗН($C$3:$F$3;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$4:$F$4;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$5:$F$5;C12:F12))

	B	F	G	H
19				
20				
21				
22		Атырауская	Западно-Казахстанская	Жамбылская
23	Акмолинская			
24	Актюбинская			
25	Алматинская			
26	Атырауская	0		
27	Западно-Казахстанская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C7:F7))	0	
28	Жамбылская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C8:F8))	=КОРЕНЬ(СУММКВРАЗН($C$7:$F$7;C8:F8))	0
29	Карагандинская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C9:F9))	=КОРЕНЬ(СУММКВРАЗН($C$7:$F$7;C9:F9))	=КОРЕНЬ(СУММКВРАЗН($C$8:$F$8;C9:F9))
30	Костанайская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C10:F10))	=КОРЕНЬ(СУММКВРАЗН($C$7:$F$7;C10:F10))	=КОРЕНЬ(СУММКВРАЗН($C$8:$F$8;C10:F10))
31	Кызылординская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C11:F11))	=КОРЕНЬ(СУММКВРАЗН($C$7:$F$7;C11:F11))	=КОРЕНЬ(СУММКВРАЗН($C$8:$F$8;C11:F11))
32	Мангистауская	=КОРЕНЬ(СУММКВРАЗН($C$6:$F$6;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$7:$F$7;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$8:$F$8;C12:F12))

	B	I	J	K	L
19					
20					
21					
22		Карагандинская	Костанайская	Кызылординская	Мангистауская
23	Акмолинская				
24	Актюбинская				
25	Алматинская				
26	Атырауская				
27	Западно-Казахстанская				
28	Жамбылская				
29	Карагандинская	0			
30	Костанайская	=КОРЕНЬ(СУММКВРАЗН($C$9:$F$9;C10:F10))	0		
31	Кызылординская	=КОРЕНЬ(СУММКВРАЗН($C$9:$F$9;C11:F11))	=КОРЕНЬ(СУММКВРАЗН($C$10:$F$10;C11:F11))	0	
32	Мангистауская	=КОРЕНЬ(СУММКВРАЗН($C$9:$F$9;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$10:$F$10;C12:F12))	=КОРЕНЬ(СУММКВРАЗН($C$11:$F$11;C12:F12))	0

Рисунок 47 – Расчетные формулы для вычисления попарных расстояний

	B	C	D	E	F	G	H	I	J	K	L
22	Объекты	Акмолинская	Актюбинская	Алматинская	Атырауская	Западно-Казахстанская	Жамбылская	Карагандинская	Костанайская	Кызылординская	Мангистауская
23	Акмолинская	0,00	38,37	21,42	29,09	26,19	22,77	70,89	30,94	28,53	68,74
24	Актюбинская	38,37	0,00	57,22	15,21	58,80	42,97	39,76	15,95	66,26	31,60
25	Алматинская	21,42	57,22	0,00	45,47	15,91	28,31	89,61	47,50	14,54	86,46
26	Атырауская	29,09	15,21	45,47	0,00	49,60	29,01	51,22	16,16	56,15	43,79
27	Западно-Казахстанская	26,19	58,80	15,91	49,60	0,00	40,85	92,22	46,24	16,33	85,55
28	Жамбылская	22,77	42,97	28,31	29,01	40,85	0,00	73,18	40,18	39,95	72,58
29	Карагандинская	70,89	39,76	89,61	51,22	92,22	73,18	0,00	54,01	98,41	40,46
30	Костанайская	30,94	15,95	47,50	16,16	46,24	40,18	54,01	0,00	55,95	39,53
31	Кызылординская	28,53	66,26	14,54	56,15	16,33	39,95	98,41	55,95	0,00	95,29
32	Мангистауская	68,74	31,60	86,46	43,79	85,55	72,58	40,46	39,53	95,29	0,00

Рисунок 48 – Итоговая таблица матрицы попарных расстояний

4. Выберем в меню пункт *Статист=F9*. В открывшемся окне в разделе *Многомерные методы* выберем *Шкалирование*.
5. Появится следующее окно (рис. 49).

Рисунок 49 – Окно выбора метода шкалирования

6. Укажем тип шкалирования и выберем соответствующий метод: *Метрическое*, *1=Торгерсон*.
7. Далее будет выведено окно для указания проекции объектов на координатные оси (рис. 50)

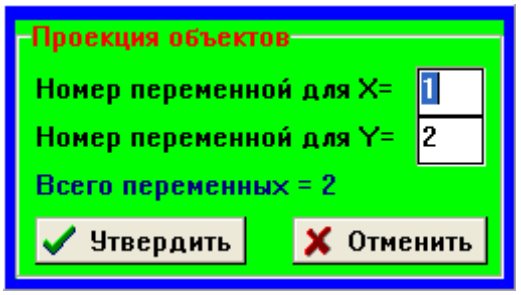


Рисунок 50 – Окно Проекции объектов

8. Нажмем *Утвердить* и при повторном появлении этого окна укажем *Отменить*. Если пространство больше чем двумерное, возможно построение графиков с разным сочетанием осей, например, 1 и 3, 2 и 3, 1 и 2. Номера этих осей (в терминах программы - переменных) указываются в окне *Проекция объектов*.
9. Получим следующие результаты (рис. 51):

```
МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ.  Файл:

              Собственные значения координатных осей шкалирова-
ния
Координата:       1
Собств.зн  1,065E4     1165     653,7
 Дисперс%    84,77    9,276     5,203
Накоплен%    84,77    94,04     99,24

        Метод: Торгерсон
            Координаты объектов в результате шкалирования
```

```
Координата:       1         2         3
       x1    15,71     3,492    -2,872
       x2   -21,99    -1,669     1,297
       x3    34,67     2,662    0,5315
       x4   -9,873     1,216     9,114
       x5    34,91    -10,44    -8,487
       x6    14,68     15,12     15,82
       x7    -52,3     19,03    -13,38
       x8    -10,3    -12,11   0,06322
       x9    43,78   -0,1224    -6,371
      x10   -49,28    -17,18     4,281

Стресс 1-4 = 0,02057; 0,4535, 0,0455, 1184
Корреляция исходных и конечных расстояний=0,9997
```

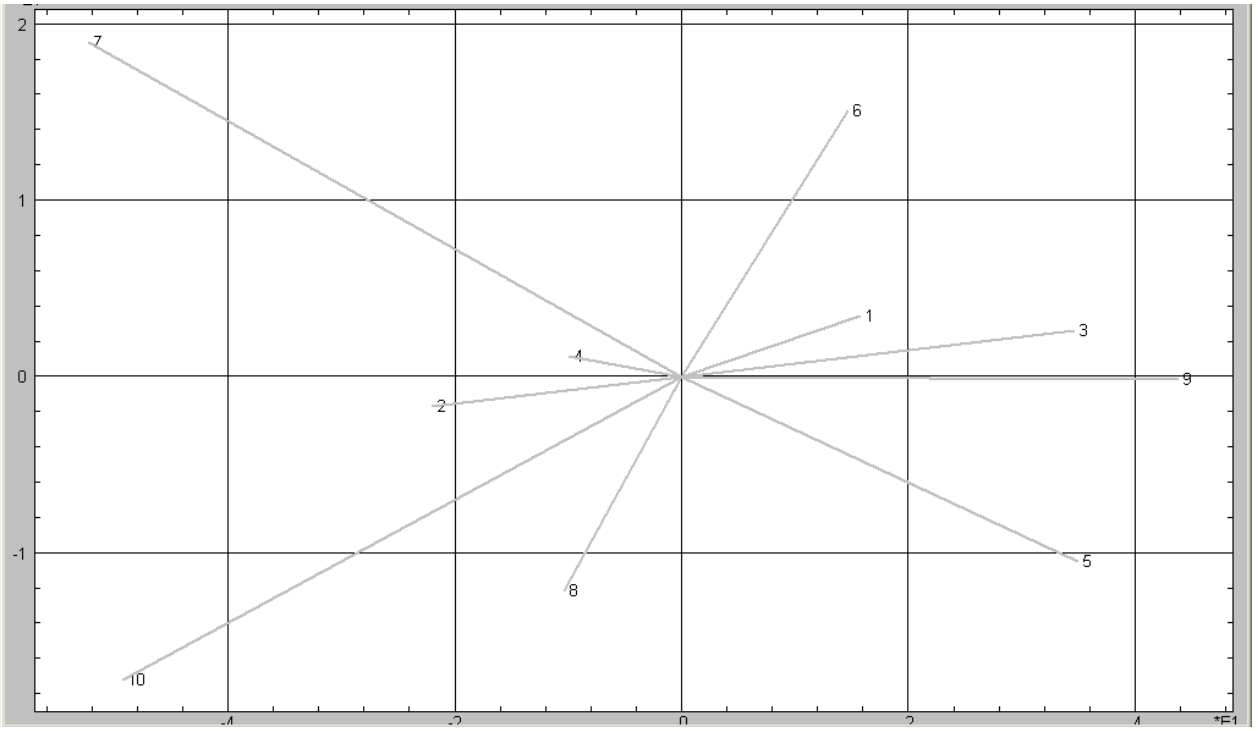


Рисунок 51 – График, выводимый программой

Рассмотрим полученные результаты.

1. Программа выделила три координатных оси. Нам следует определиться, сколько осей оставить: две или три. Для этого в программе MS Excel построим график «каменистой осыпи», пользуясь собственными значениями. По оси абсцисс отложим количество осей, по оси ординат – собственные значения (их величина приводится в выводимом программой решении). Получим следующий график (рис. 52).
2. Как видно из графика, оставить следует две оси: первую и вторую (см. тему «Факторный анализ», стр. 32-33).
3. Расположение объектов (областей) в данном координатном пространстве мы уже получили (см. рис. 51)

4. Для более наглядного представления объектов в данном двумерном пространстве нами построен такой же график в MS Excel с нанесением названий областей возле каждой точки (рис. 53).

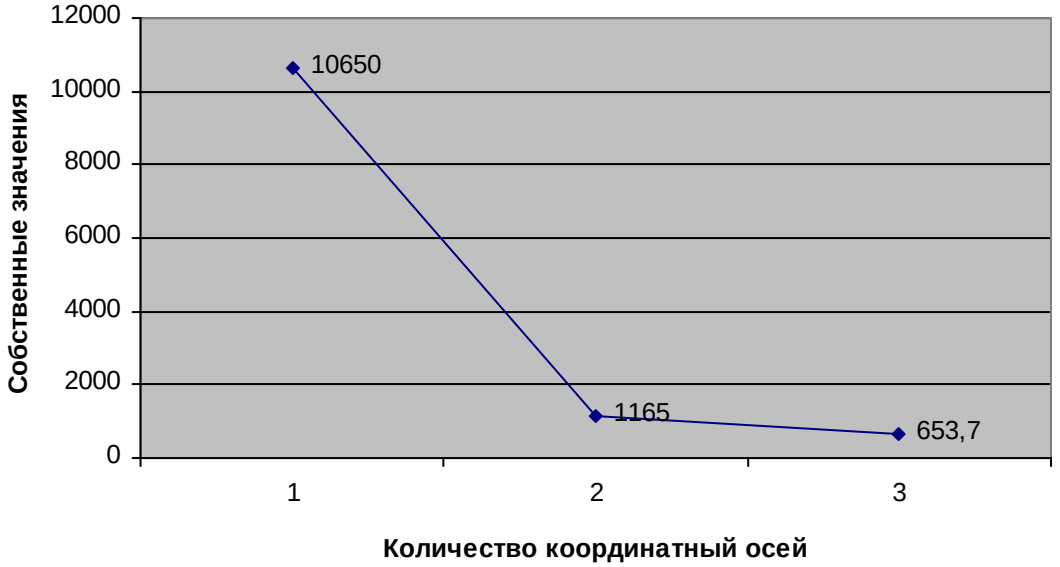


Рисунок 52 – График «каменистой осыпи» для определения оптимального количества осей

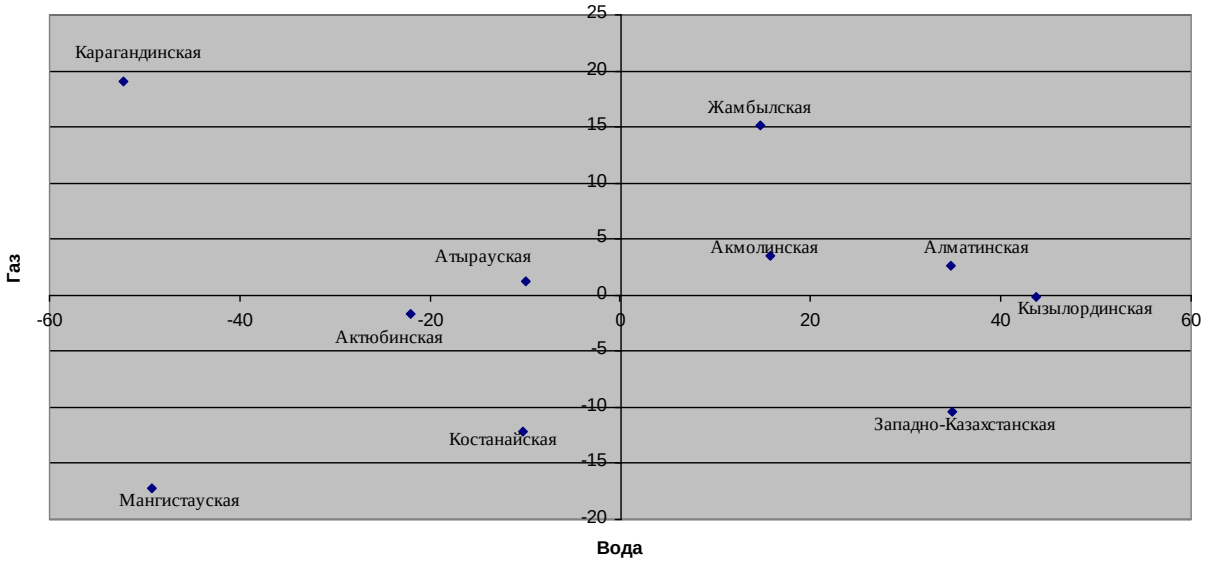


Рисунок 53 – График расположения областей в выделенном двумерном пространстве

5. Теперь нам нужно интерпретировать полученную картину и дать название осям. Исходя из рассматриваемых признаков, характеризующих благоустройство жилья (наличие водопровода, центрального отопления, газа, горячего водоснабжения), предположим, что газоснабжение можно выделить в отдельную характеристику (одну ось), водопровод и горячее водоснабжение – в другую ось. Небольшое затруднение визуально отнести к той или иной оси представляет центральное отопление, поскольку оно связано и с водой, и с газом.

6. Для того чтобы разрешить эту проблему, рассчитаем коэффициенты корреляции, определяющие, насколько взаимосвязан каждый из признаков с каждой осью. Для этого воспользуемся функцией КОРРЕЛ программы MS Excel (процедуру расчета парных коэффициентов корреляции см. в приложении Б).
7. Получатся следующие значения коэффициентов корреляции (табл. 22):

Таблица 22 – Значения коэффициентов корреляции

Признак / Ось	водопровод	центральное отопление	газ	горячее водоснабжение
1	-0,90831	-0,9721	0,511153	-0,95232
2	0,361885	-0,12674	-0,56276	-0,2841

8. Из значений, приведенных в таблице 22, можно сделать вывод, что ось 1 связана с признаками: водопровод, центральное отопление и горячее водоснабжение, ось 2 – с признаком: газ, так как соответствующие коэффициенты высокие.
9. Поэтому условно назовем ось 1 – ось абсцисс – *Вода*, ось 2 – ось ординат – *Газ*.
10. Так как коэффициенты корреляции, говорящие о сильной взаимосвязи осей и признаков, имеют отрицательную зависимость, правильнее будет рассматривать первый квадрант как третий, второй как четвертый и т.д. Или развернуть оси на 270° (рис. 54). При этом координаты объектов изменятся.
11. Для доведения полученных окончательных результатов проведенного многомерного шкалирования до логического завершения и их интерпретации сформируем карту расположения объектов в новой системе координат.
12. Чтобы определить новые координаты объектов (областей), можно выполнить вычисления, приведенные в приложении В.
13. Теперь если внимательно посмотреть на график на рисунке 54, то видно, что в первом квадранте находятся области, имеющие наибольший процент населения, обеспеченного условиями благоустроенного жилья. Соответственно, в третьем квадранте области, наименее обеспеченные.

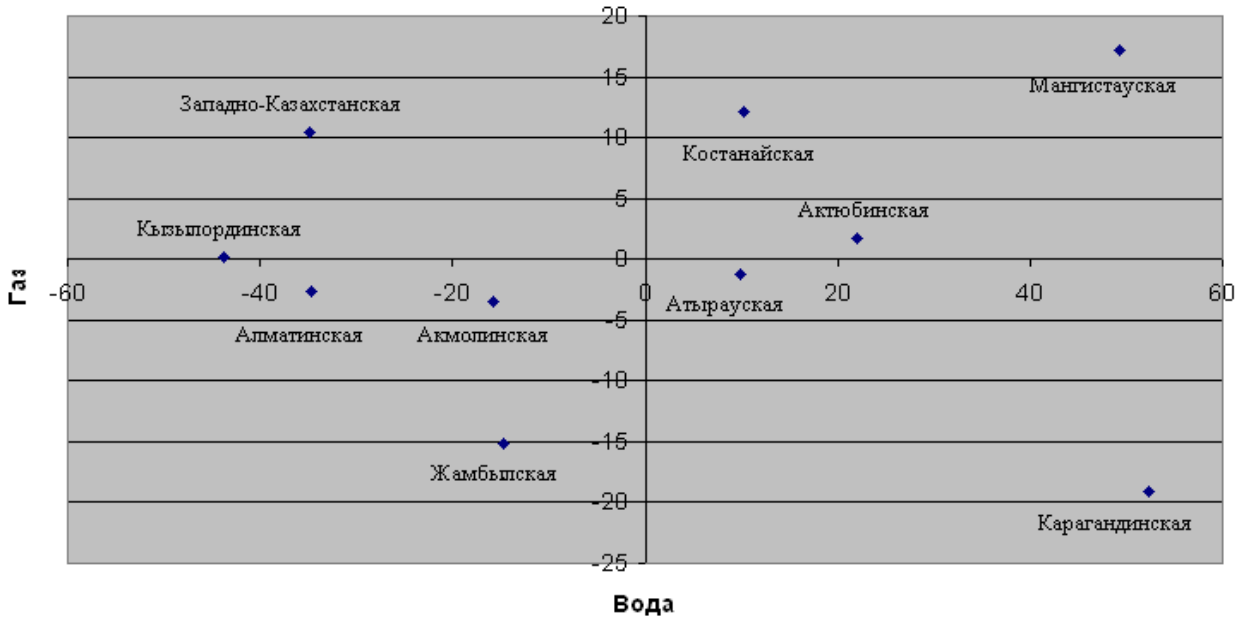


Рисунок 54 – Расположение объектов в новой системе координат

14. Для того чтобы окончательно принять полученное решение, оценим его на надежность и достоверность. Для этого требуется рассмотреть коэффициент корреляции, показывающий, насколько близки полученные результаты к исходным данным, и значение стресса - меру несоответствия модели или долю дисперсии оптимально шкалированных данных, которую не учитывает модель МНШ. Значения стресса варьируются в зависимости от метода МНШ и анализируемых данных. Для значений стресса, вычисляемых по стресс-формуле Краскела, даются следующие рекомендации (табл. 23).

Таблица 23 – Критерии соответствия модели исходным данным по значению стресса[1]

Стресс (%)	Критерий соответствия модели
20	Плохое
10	Удовлетворительное
5	Хорошее
2,5	Отличное
0	Превосходное (полное)

15. В нашем решении коэффициент корреляции равен 0,9997, значение стресса = 0,020 или 2%, что говорит о том, что построенная модель многомерного шкалирования хорошо соответствует исходным данным и можно пользоваться ее результатами.

[1] Данный для таблицы 23 взяты с сайта: http://www.mngt.ru/ - Маркетинговые исследования, раздел 3, п. 3.9 – Многомерное шкалирование и совместный анализ.

Итак, после проведения процедуры многомерного шкалирования по признакам, характеризующим благоустройство жилья в областях Казахстана, получили двумерное пространство, наглядно отображающее основные факторы (оси), по которым классифицируются объекты (области). Оси названы *Вода* и *Газ*, расположение областей республики в координатном пространстве этих осей дает четкое визуальное представление о том, какие области являются по этим параметрам наиболее обеспеченными, какие - наименее. Наиболее обеспеченными являются Мангистауская, Костанайская и Актюбинская области, наименее – Алматинская, Акмолинская и Жамбылская области. Следовательно, необходимо государству предусмотреть выделение финансовых средств из госбюджета и привлечь инвестиции из сферы бизнеса, в первую очередь, для повышения жилищного благоустройства в Алматинской, Акмолинской и Жамбылской областях.

Задание 2.

Определить, как можно классифицировать и наглядно отобразить в системе координат регионы Республики Казахстан по схожести в структуре денежных доходов населения в 2009 г. Исходные данные представлены в таблице 24.

Задание 3.

По имеющимся по вариантам данным (приложение А практикума) по отдельным характеристикам некоторых банков второго уровня Казахстана необходимо путем проведения процедуры многомерного шкалирования сократить количество признаков, подобрав необходимое и наглядное координатное пространство, дать полученным осям значимые названия и интерпретировать полученный результат.

Таблица 24 – Структура денежных доходов населения по регионам в 2009 году[1]

в процентах

№ п/п	Регионы Республики Казахстан	Доля в денежном доходе							
		Доходы от трудовой деятельности		Социальные трансферты				Доход от собственности	Материальная помощь от родственников, алименты и прочие доходы
		Доход от работы по найму	Доход от самостоятельной занятости и предпринимательской деятельности	пенсии	пособия	АСП и жилищная помощь	стипендии		
1.	Акмолинская	63,8	10,6	15,8	3,8	0,1	0,1	0,9	4,9
2.	Актюбинская	73,4	9,1	9,9	3,7	0,1	0,1	0,6	3,1
3.	Алматинская	70,8	9,7	12,7	3,4	0,1	0,4	0,3	2,6
4.	Атырауская	84,9	5,9	6,2	1,9	0,0	0,1	0,6	0,4
5.	Западно-Казахстанская	68,1	11,4	14,2	3,5	0,1	0,4	0,1	2,2
6.	Жамбылская	67,2	13,8	13,2	3,6	0,1	0,5	0,0	1,6
7.	Карагандинская	71,4	4,9	14,3	4,1	0,1	0,4	0,4	4,4
8.	Костанайская	65,9	8,4	17,7	3,0	0,3	0,3	0,5	3,9
9.	Кызылординская	61,3	16,5	8,6	4,3	0,2	0,2	0,1	8,8
10.	Мангистауская	85,8	4,6	5,7	2,1	0,2	0,2	03	1,1
11.	Южно-Казахстанская	64,4	18,9	9,8	4,6	0,0	0,3	0,2	1,8
12.	Павлодарская	72,9	9,0	12,5	2,1	0,0	0,5	0,1	2,9
13.	Северо-Казахстанская	63,9	11,0	14,0	4,2	0,1	0,4	1,1	5,3
14.	Восточно-Казахстанская	69,3	7,7	13,4	3,1	0,0	0,3	0,5	5,7
15.	г. Астана	72,9	8,5	9,4	2,3	0,0	0,2	3,4	3,3
16.	г. Алматы	76,7	5,8	11,2	2,0	0,0	0,2	0,9	3,2

[1] Данные взяты из статистического сборника «Уровень жизни населения в Казахстане». Астана, 2010. – 184 с.

Задание 4. (задача взята из [3]).

Приведенная в таблице матрица содержит данные о различии всех возможных пар восьми стран. Проведите по этим данным процедуру многомерного шкалирования с целью отражения объектов-стран на координатной плоскости (подберите соответствующее пространство) и дайте интерпретацию полученным результатам.

Таблица 25 - Данные о различии стран

Страна	Ангола	Аргентина	Австралия	Китай	Куба	Япония	США	Зимбабве
Ангола	0,00	1,41	1,00	1,00	1,41	1,41	1,73	0,71
Аргентина	1,41	0,00	1,00	1,73	1,41	1,41	1,00	1,41
Австралия	1,00	1,00	0,00	1,41	1,73	1,00	1,41	1,00
Китай	1,00	1,73	1,41	0,00	1,00	1,00	1,41	1,00
Куба	1,41	1,41	1,73	1,00	0,00	1,41	1,00	1,41
Япония	1,41	1,41	1,00	1,00	1,41	0,00	1,00	1,41
США	1,73	1,00	1,41	1,41	1,00	1,00	0,00	1,73
Зимбабве	0,71	1,41	1,00	1,00	1,41	1,41	1,73	0,00

Раздел 2. Пример использования многомерных методов статистического анализа данных при сегментировании рынка.

Цель: на примере программы Сегментирование и рынки фирмы КонСи показать применение многомерных методов анализа данных в реальной экономике, в частности, в сфере маркетинга.

Имеются данные, полученные путем опроса потребителей о предпочтении ими колбасных изделий мясокомбината, оцениваемых по различным параметрам, оценка респондентами определенных торговых марок, а также сведения о самих респондентах. Требуется на основании полученных данных провести сегментирование рынка потребителей колбасных изделий, позиционировать (т.е. разместить) торговые марки (бренды производителей) на построенных сегментах с целью выявления предпочтений потребителей в отношении колбасных изделий и предпочтений определенных брендов-производителей.

Сегментирование проводится по следующей укрупненной схеме:

1. Выдвижение гипотезы о количестве сегментов, на которые будет разбиваться совокупность респондентов.
2. Проведение процедуры кластеризации потребителей по определенным признакам, которые выбираются аналитиком из всего многообразия характеристик, которыми описываются предпочтения респондентов.
3. Проведение факторного анализа с целью сокращения количества ранее выбранных признаков для наглядного визуального отображения потребителей в координатном пространстве.
4. Проведение однофакторного дисперсионного анализа с целью выявления признаков, влияющих на разделение потребителей на классы (сегменты).
5. Построение карт позиционирования – карт восприятия брендов.
6. Принятие полученного результата либо проведение нового варианта сегментирования.

Технология выполнения:

1. Запустим программу *Сегментирование и рынки*, выбрав ее название либо в главном меню, либо ярлык на рабочем столе.
2. Откроем сетевую папку *Инструкции по примеру*, в которой содержится файл position_meat.chm – скомпилированный файл справки, содержащий пошаговые указания по выполнению работы в программе.
3. Исходные данные содержатся в базе данных K_sensory_3.mdb, расположенной в сетевой папке *Учебная база данных*.
4. Выполним процедуру сегментирования рынка потребителей и позиционирования брендов.

   Рассмотрим получаемые результаты.

Выдвинута гипотеза о разделении потребителей колбасных изделий на 3 сегмента. Этим сегментам даны условные названия: первый, второй, третий.

Далее выбраны признаки, по которым предполагается проводить кластеризацию потребителей:

- частота покупки вареных колбас;
- частота покупки копченых колбас;
- частота покупки сосисок;
- сумма, на которую в среднем делается покупка вареных колбас;
- сумма, на которую в среднем делается покупка копченых колбас;
- сумма, на которую в среднем делается покупка сосисок;
- среднее покупаемое количество вареной колбасы;
- среднее покупаемое количество копченой колбасы;
- среднее покупаемое количество сосисок;
- приемлемый ценовой диапазон за 1 кг вареной колбасы;
- приемлемый ценовой диапазон за 1 кг копченой колбасы;
- приемлемый ценовой диапазон за 1 кг сосисок.

После проведения кластерного анализа получены следующие размеры кластеров (рис. 55) и произведено отнесение каждого респондента к определенному кластеру (рис. 56).

Параметры сегментов				
	Сегмент	1 с	2с	3 с
	Число объектов	259	500	298

Рисунок 55 – Размер кластеров (сегментов)

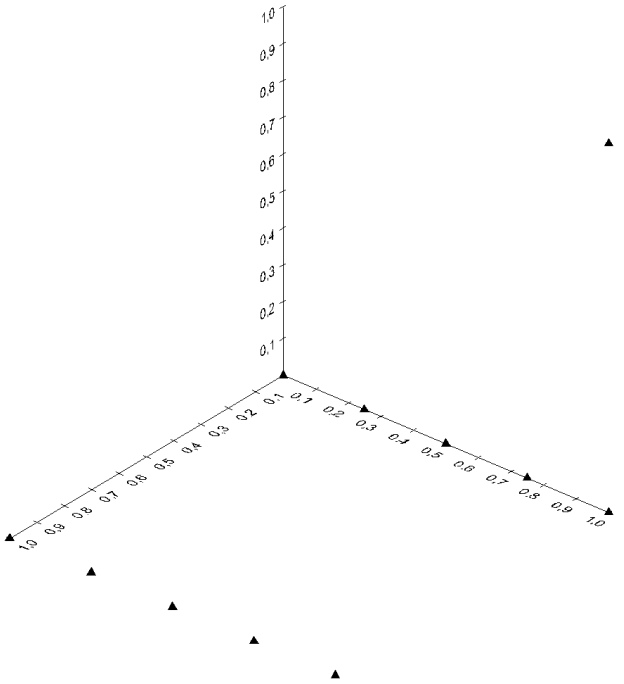


Рисунок 56 - График разделения потребителей на сегменты по выделенным сегментам

В качестве осей для графика на рисунке 56 программой по умолчанию выбраны первые три признака: пол, возраст, покупатели. Соответственно, точки респондентов могут располагаться в любой комбинации исходных признаков (признаки для осей можно выбирать, нажав кнопку - оси графика). Нам нужно расположить объекты (потребителей) в рамках тех признаков, по которым проводилась сегментация. Однако, здесь есть неудобство: признаков 12, следовательно, получится и 12 осей, анализировать которые очень неудобно, поэтому программой предусмотрен факторный анализ, который призван сократить количество исходных признаков (факторов).

В результате факторного анализа выделено три фактора, об этом свидетельствует таблица факторных нагрузок (рис. 57) и график собственных значений (рис. 58).

Матрица факторных нагрузок

Переменные	Фактор 1	Фактор 2	Фактор 3
Частота_Вареные	-0,5854	-0,5374	-0,3501
Частота_Копченые	-0,4209	0,6320	-0,3525
Частота_Сосиски	-0,6283	0,1138	0,5988
Сумма_Вареные	-0,6493	-0,6057	-0,3678
Сумма_Копченые	-0,5287	0,6365	-0,4405
Сумма_Сосисек	-0,7081	0,0627	0,6416
Количество_Вареные	-0,6498	-0,6037	-0,3717
Количество_Копченой	-0,5141	0,6572	-0,4550
Количество_Сосиски	-0,7008	0,0928	0,6446
Цена_Вареной	-0,6184	-0,6202	-0,3602
Цена_Копченой	-0,4872	0,6542	-0,4074
Цена_Сосисек	-0,6778	0,0645	0,6483

Рисунок 57 – Матрица факторных нагрузок

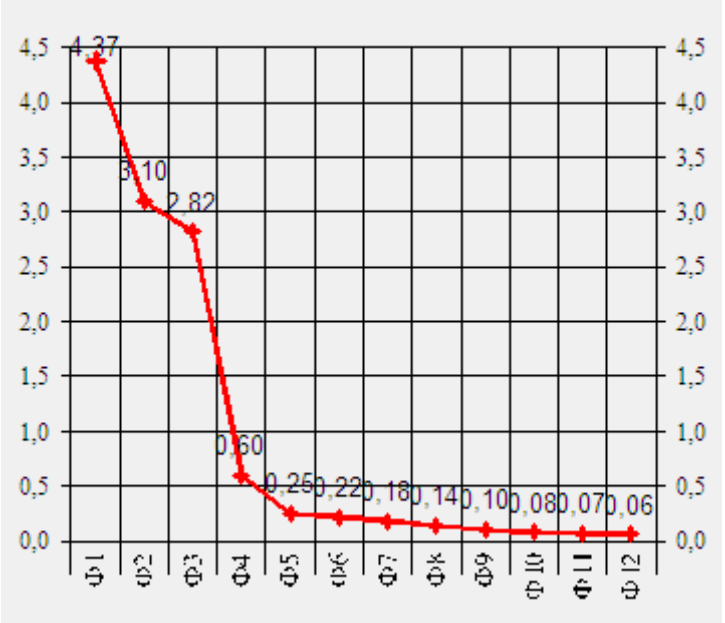


Рисунок 58 – График собственных значений (критерий каменистой осыпи)

И по величине собственных значений, и по графику следует оставить три фактора (количество факторов должно быть столько, сколько собственных значений больше 1, см. тему 2 «Факторный анализ»).

Согласно значениям коэффициентов корреляции из матрицы факторных нагрузок (рис. 57) определяется, как каждый из признаков взаимодействует с вновь выделенными факторами. Эти факторы и являются координатными осями, в которых наглядно отображаются респонденты (рис. 59).

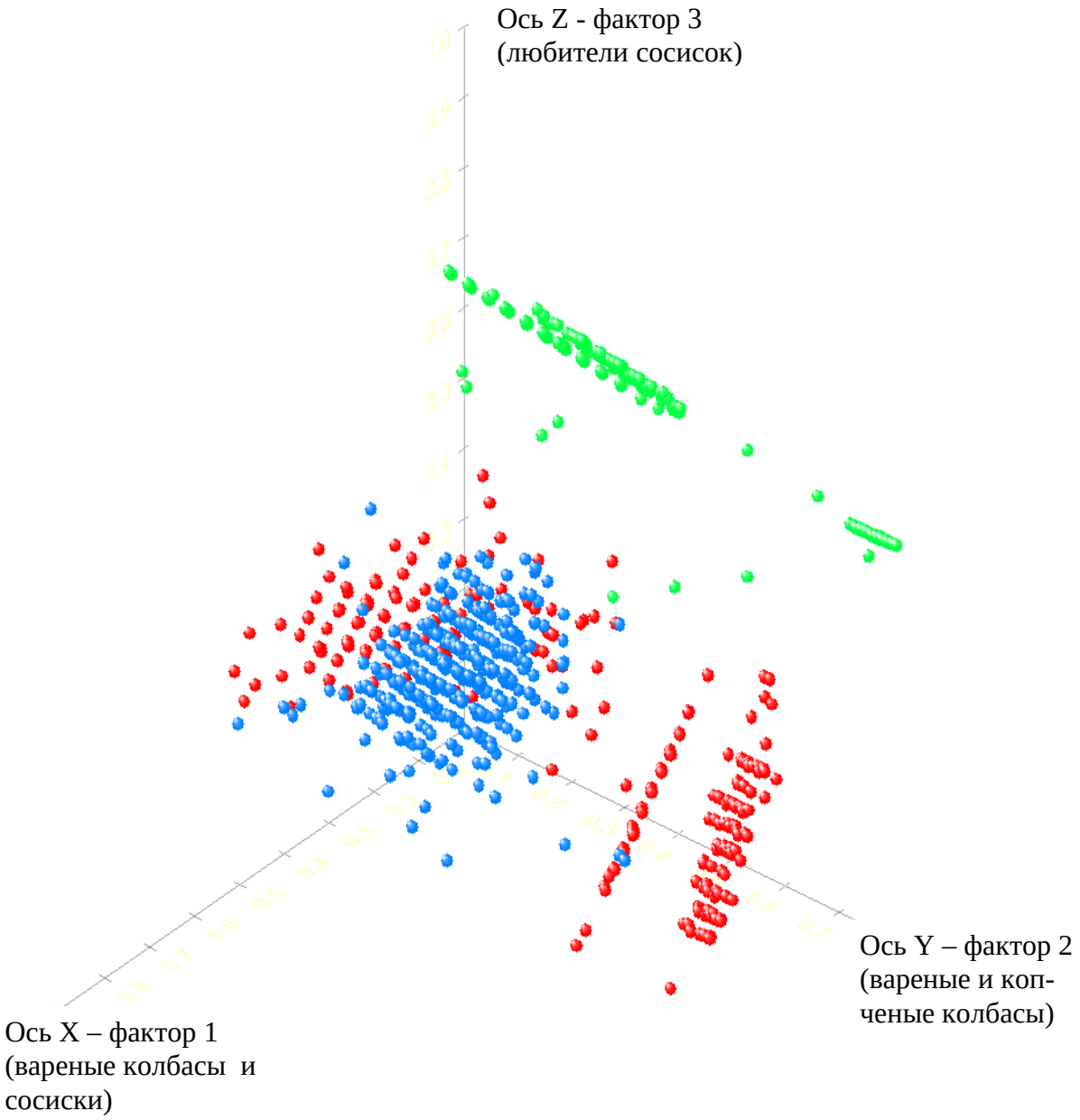
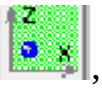


Рисунок 59 – График разделения потребителей на сегменты по выделенным факторам

Для того чтобы было более понятно, какие потребители (какие точки) расположены ближе к той или иной оси, оси можно представлять в разном

сочетании друг с другом в двумерном пространстве, нажимая кнопки  , , .

Итак, сегмент, выделенный красным цветом – это любители вареных и копченых колбас, синим цветом – вареных колбас и сосисок, зеленым - сосисок.

После определения сегментов по признакам предпочтения видов продукции предполагается разделение потребителей по оценкам ими качества колбасной продукции определенной марки (бренда). Колбасная продукция пяти брендов, названных условно бренд А, бренд В, бренд С, бренд D и бренд Е, оценивалась по следующим характеристикам:

– Качественная структура фарша;
– Доверие к упаковке внешний вид упаковки;
– Лучшее соотношение цена-качество;
– Гармоничность вкусовых свойств, специй, запаха;
– Известность марки продукции;
– Мясной цвет среза колбасного изделия;
– Свежий вид продукции мясокомбината на прилавке.

Для сокращения количества признаков снова был проведен факторный анализ, получено три значимых признака: вкусовые и визуальные предпочтения, известность марки и упаковка, цена-качество. В пространстве вновь выделенных осей были отображены респонденты, разбитые по сегментам и торговые марки производителей колбасных изделий (рис. 60)

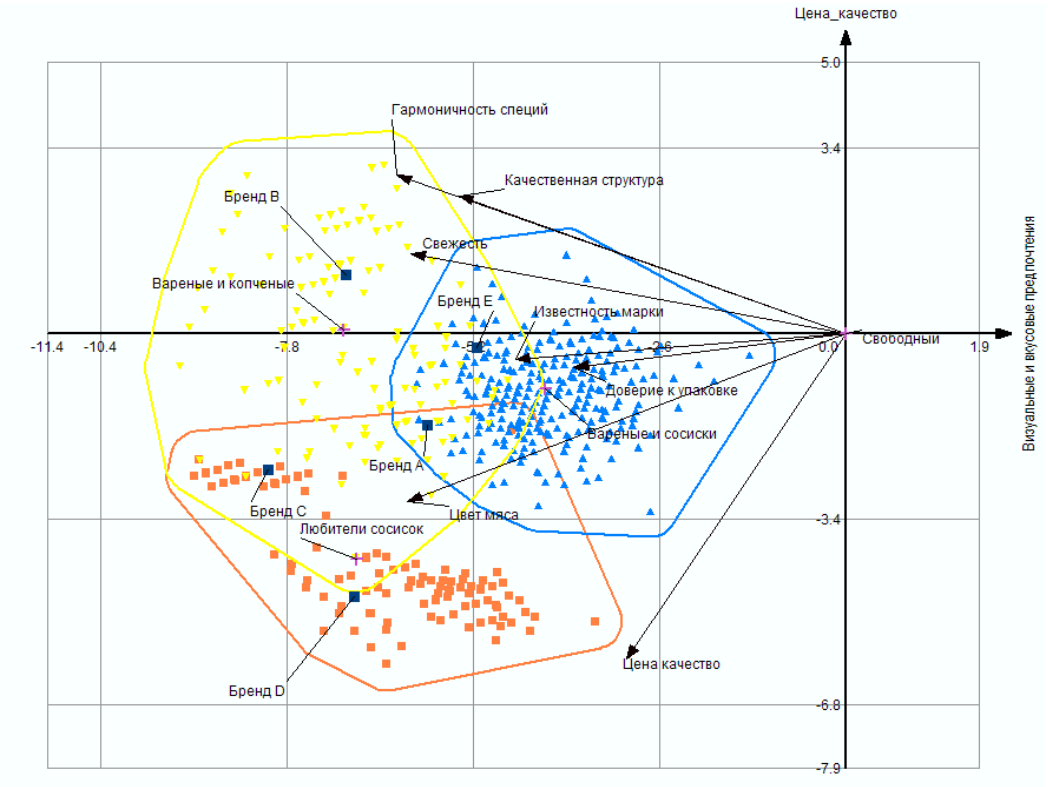


Рисунок 60 – Карта восприятия брендов (карта позиционирования)

Соответственно, чем ближе бренд расположен к центру сегмента, тем больше он удовлетворяет требованиям потребителей.

Далее программой определяется значимость каждой из переменных для выделения сегментов, то есть, проводится однофакторный дисперсионный анализ. Данные по результатам решения экспортируются в программу MS Excel с определенным выводом, в котором содержится краткая интерпретация решения. Так, например, такие параметры как ценовой диапазон колбасных изделий разных видов, количество приобретаемого товара, частота покупки оказывают влияние на разделение потребителей по сегментам, а такие характеристики респондентов как пол, возраст, образование – нет.

В целом по результатам проведенного исследования можно сказать, что:

- Проведенный кластерный и факторный анализ позволил с учетом отзывов покупателей выявить наиболее предпочитаемые виды колбасных изделий и брендов производителей. Данная информация имеет большую значимость для прогнозирования деятельности производителей колбасных изделий. Она может быть использована каждым производителем и любым управляющим региональным органом для размещения заказов по колбасным изделиям.
- Для серьезных экономических расчетов, анализа данных целесообразно применять методы многомерного анализа, особенно в сочетании друг с другом, использовать специализированные пакеты прикладных программ, выполняющих решение задач как отдельными методами, так и интегрированные пакеты, что и было продемонстрировано на примере программы фирмы КонСи Сегментирование и рынки.

Список источников

1. Берк, Кеннет, Кэйри, Патрик. Анализ данных с помощью Microsoft Excel.: Пер. с англ. – М.: Издательский дом «Вильямс», 2005. – 560 с.: ил. – Парал. тит. англ.
2. Вардиашвили Н.Н. Оптимизационное моделирование задач микроэкономики: в 2-х книгах, книга 2 . – Костанай, 2002. - 406 с. (гл. 6.1).
3. Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных. М.: Финансы и статистика, 1988. – 254 с.
4. Захарченко Н.И. Бизнес-статистика и прогнозирование в MS Excel. Самоучитель: М.: Издательский дом «Вильямс», 2004. – 208 с.: ил.
5. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.: ил.
6. Мирошник В.А. Методическое руководство по математической статистике (в помощь студентам экономического факультета, изучающим многомерные методы статистического анализа). Кустанай, 1997. – 39 с.
7. Ниворожкина Л.И., Арженовский С.В. Многомерные статистические методы в экономике: Учебник. – М.: Издательско-торговая корпорация «Дашков и К»; Ростов н/Д: Наука-Спектр, 2008. – 224 с.
8. Симчера В.М. Методы многомерного анализа статистических данных: учеб. Пособие. – М.: Финансы и статистика, 2008. – 400 с.: ил.
9. Толстова Ю.Н. Основы многомерного шкалирования: учебное пособие. – М.:КДУ, 2006. – 160 с., ил.
10. Уэйн Л. Винстон. Microsoft Excel: анализ данных и построение бизнес-моделей / Пер. с англ. – М.: Издательско-торговый дом «Русская редакция», 2005. – 576 с.: ил.
11. Электронный учебник StatSoft (по статистике). (c) Copyright StatSoft, Inc., 1984-2001.
12. Microsoft Excel корпорации Microsoft - программа для работы с электронными таблицами.
13. ППП Stadia 6.2. Пакет статистических расчетов. Создатель: Кулаичев А.П. Распространитель: НПО «Информатика и компьютеры» (ИнКо).

Приложение А.
Данные по вариантам для выполнения СРС[1]

№ п/п	Наименование банка	Активы, тыс. тенге	Ссудный портфель, тыс. тенге	Обязательства, тыс. тенге	Собственный капитал, тыс. тенге	Расходы, связанные с выплатой вознаграждения, тыс. тенге	Чистый доход, связанный с получением вознаграждения, тыс. тенге	Коэф. достаточности собственного капитала (k1-1)	Коэф. текущей ликвидности (k4)
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Вариант 1 (данные на 01.01.2011)									
1	АО «БТА Банк»	1 993 993 880	1 644 706 718	1 812 571 968	181 421 912	200 164 256	-66 607 932	0,138	1,437
2	АО «Народный Банк Казахстана»	2 023 521 731	1 224 415 801	1 755 756 612	267 765 119	87 581 139	93 251 113	0,109	1,101
3	АО «KASPI BANK»	352 905 200	288 704 195	317 048 955	35 856 245	26 646 440	23 897 522	0,085	1,051
4	АО «Банк ЦентрКредит»	1 211 056 597	719 928 174	1 134 266 337	76 790 260	79 982 377	12 464 966	0,073	1,469
5	ДБ АО «Сбербанк»	282 692 348	176 644 571	242 779 773	39 912 575	9 162 572	11 890 706	0,129	0,661
6	АО «Евразийский Банк»	358 761 229	227 546 034	333 259 525	25 501 704	22 039 022	7 172 894	0,071	1,402
7	АО «АТФБанк»	982 965 005	849 475 452	919 033 776	63 931 229	56 516 565	24 234 909	0,077	0,790
8	АО «Цеснабанк»	227 211 450	149 287 906	203 640 435	23 571 015	14 217 450	7 111 775	0,101	0,701
9	АО «Альянс Банк»	489 545 766	545 529 221	455 598 109	33 947 657	37 798 365	5 362 609	0,089	1,681
10	АО «Жилстройсбербанк Казахстана»	101 188 283	67 701 723	76 138 468	25 049 815	1 097 157	4 163 034	0,248	26,782
11	АО «МЕТРОКОМБАНК»	18 683 184	7 596 882	12 160 138	6 523 046	328 636	493 889	0,349	1,519
12	АО «КАЗКОММЕРЦБАНК»	2 430 236 299	2 344 296 478	2 151 742 785	278 493 514	142 521 037	134 571 640	0,123	0,675
Вариант 2 (данные на 01.02.2011)									
1	АО «КАЗКОММЕРЦБАНК»	2 361 139 774	2 321 894 392	2 081 806 545	279 333 229	141 155 599	130 789 688	0,126	0,587
2	АО «Народный Банк Казахстана»	2 161 900 497	1 214 909 748	1 891 287 084	270 613 413	86 695 998	93 470 630	0,114	1,091
3	АО «БТА Банк»	1 979 436 475	1 642 432 926	1 803 665 926	175 770 549	195 260 182	-57 093 046	0,136	2,129

[1] Данные для выполнения СРС взяты с сайта Комитета по контролю и надзору финансового рынка и финансовых организаций Национального Банка Республики Казахстан www.afn.kz.

1	2	3	4	5	6	7	8	9	10
4	АО «Банк ЦентрКредит»	1 133 902 740	720 984 357	1 056 759 810	77 142 930	79 133 829	11 109 388	0,078	1,383
5	АО «АТФБанк»	999 205 747	844 435 708	935 712 404	63 493 343	55 065 813	24 134 727	0,075	1,019
6	АО «Альянс Банк»	483 185 384	544 749 230	452 303 382	30 882 002	38 617 726	3 364 452	0,087	1,934
7	АО «Евразийский Банк»	369 249 518	236 944 350	343 548 291	25 701 227	21 834 398	5 640 417	0,071	1,740
8	АО «Нурбанк»	259 187 273	198 016 862	214 621 106	44 566 167	17 776 055	8 134 564	0,165	1,325
9	АО «KASPI BANK»	348 901 143	290 410 972	312 874 914	36 026 229	26 889 930	25 032 951	0,098	1,100
10	ДБ АО «Сбербанк»	283 646 421	181 741 505	242 669 661	40 976 760	9 362 233	12 511 468	0,138	0,650
11	АО «Ситибанк Казахстан»	287 686 118	36 501 797	265 381 162	22 304 956	987 613	3 012 163	0,085	0,853
12	АО «Цеснабанк»	232 664 316	148 778 103	208 992 877	23 671 439	14 345 238	7 541 448	0,101	0,846
Вариант 3 (данные на 01.03.2011)									
1	АО «КАЗКОММЕРЦБАНК»	2 363 188 645	2 304 175 165	2 082 379 389	280 809 256	139 327 260	128 569 282	0,126	0,649
2	АО «Народный Банк Казахстана»	2 151 630 993	1 213 694 524	1 879 058 711	272 572 282	85 395 442	94 250 364	0,115	1,173
3	АО «БТА Банк»	1 982 521 890	1 642 656 514	1 805 547 690	176 974 200	190 620 346	-52 426 208	0,136	2,175
4	АО «Банк ЦентрКредит»	1 151 885 346	743 082 268	1 075 052 514	76 832 832	77 826 035	9 932 648	0,076	1,352
5	АО «АТФБанк»	1 020 610 559	850 892 781	973 811 519	46 799 040	53 886 567	22 979 478	0,054	1,044
6	АО "Альянс Банк"	502 176 611	553 128 793	471 831 835	30 344 776	39 342 547	1 638 202	0,082	1,856
7	АО «Евразийский Банк»	400 666 273	238 354 737	374 344 987	26 321 286	21 523 964	8 872 963	0,065	1,706
8	АО «Нурбанк»	348 048 602	291 813 152	311 865 090	36 183 512	17 502 765	7 776 983	0,174	1,533
9	АО «KASPI BANK»	290 260 663	188 366 029	249 349 916	40 910 747	27 169 132	25 970 210	0,099	1,139
10	ДБ АО «Сбербанк»	253 518 557	38 268 575	231 142 260	22 376 297	9 474 183	13 128 528	0,135	0,572
11	АО «Ситибанк Казахстан»	246 696 179	197 196 763	202 754 594	43 941 585	872 284	3 097 357	0,089	0,743
12	АО «Цеснабанк»	240 499 500	165 754 503	216 756 342	23 743 158	14 531 447	7 862 492	0,098	0,864
Вариант 4 (данные на 01.01.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 351 990 834	2 312 247 489	2 082 564 189	269 426 645	163 062 754	214 203 768	0,128	0,588
2	АО «Народный Банк Казахстана»	1 971 441 914	2 524 360 149	3 445 335 309	1 473 893 395	103 423 170	83 712 473	0,111	0,842
3	АО «БТА Банк»	1 988 880 350	1 236 928 371	1 748 444 327	240 436 023	218 994 960	-42 382 862	-0,966	0,535

1	2	3	4	5	6	7	8	9	10
4	АО «Банк ЦентрКредит»	1 074 592 153	847 964 520	979 554 162	95 037 991	75 236 195	54 078 471	0,088	1,953
5	АО «АТФБанк»	1 152 094 706	669 233 315	1 063 893 918	88 200 788	74 695 345	42 217 420	0,102	1,985
6	АО «Альянс Банк»	529 987 266	504 958 095	978 440 170	448 452 904	84 960 376	3 508 622	-0,850	1,390
7	АО «Евразийский Банк»	301 771 119	251 212 112	270 511 081	31 260 038	19 576 057	5 751 210	0,078	2,243
8	АО «Нурбанк»	294 892 793	235 068 073	249 914 008	44 978 785	20 826 082	11 309 892	0,147	0,567
9	АО «KASPI BANK»	324 090 059	188 555 607	300 383 285	23 706 774	23 931 368	16 645 948	0,099	1,336
10	ДБ АО «Сбербанк»	179 143 285	262 116 866	262 599 986	83 456 701	5 800 976	8 043 623	0,155	1,713
11	АО «Ситибанк Казахстан»	182 741 196	38 354 132	164 062 655	18 678 541	1 393 622	2 566 093	0,066	1,021
12	АО «Цеснабанк»	221 333 718	96 463 558	184 671 113	36 662 605	11 562 313	7 749 478	0,092	0,732
Вариант 5 (данные на 01.02.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 375 572 093	2 324 130 852	2 103 736 417	271 835 676	162 474 236	210 993 068	0,127	0,680
2	АО «БТА Банк»	2 000 964 169	2 440 627 373	3 477 146 423	1 476 182 254	221 420 864	-62 218 122	-0,953	0,513
3	АО «Народный Банк Казахстана»	1 979 237 222	1 228 294 806	1 736 435 554	242 801 668	103 095 617	82 335 345	0,112	0,756
4	АО «Банк ЦентрКредит»	1 072 422 571	844 414 119	978 131 996	94 290 575	74 494 050	55 199 765	0,091	2,057
5	АО «АТФБанк»	1 181 418 068	666 129 420	1 087 860 188	93 557 880	74 583 111	40 415 633	0,103	1,901
6	АО «Альянс Банк»	537 776 703	499 267 314	974 604 019	436 827 316	78 422 565	3 629 073	-0,833	1,944
7	АО «KASPI BANK»	301 952 461	250 971 893	270 820 187	31 132 274	24 349 285	16 818 802	0,099	1,506
8	АО «Нурбанк»	310 420 734	239 623 184	264 980 193	45 440 541	20 682 321	10 845 631	0,142	0,772
9	АО «Евразийский Банк»	339 333 035	203 189 827	315 018 471	24 314 564	20 231 026	5 414 064	0,074	2,669
10	АО «Цеснабанк»	178 634 156	118 228 202	160 523 767	18 110 389	11 769 609	7 523 673	0,101	0,869
11	АО «Ситибанк Казахстан»	205 056 545	38 078 882	186 549 727	18 506 818	1 339 642	2 726 490	0,097	1,008
12	ДБ АО «Сбербанк»	228 118 073	99 423 409	191 138 954	36 979 119	6 212 957	8 026 411	0,160	1,587
Вариант 6 (данные на 01.03.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 391 598 641	2 308 700 444	2 117 475 727	274 122 914	158 545 201	207 569 392	0,126	0,698
2	АО «БТА Банк»	1 994 019 810	2 380 980 968	3 481 797 009	1 487 777 199	224 415 363	-82 074 586	-0,963	0,520
3	АО «Народный Банк Казахстана»	2 036 833 348	1 221 578 077	1 793 513 341	243 320 007	101 431 346	81 426 420	0,109	0,986

1	2	3	4	5	6	7	8	9	10
4	АО «АТФБанк»	1 032 526 354	832 130 515	940 472 280	92 054 074	73 562 585	37 962 282	0,104	1,701
5	АО «Банк ЦентрКредит»	1 232 163 285	678 660 074	1 125 574 988	106 588 297	75 047 129	54 413 667	0,097	2,218
6	АО «Альянс Банк»	468 426 880	587 507 476	964 001 356	495 574 476	73 358 975	2 763 053	-1,063	2,069
7	АО «KASPI BANK»	315 576 625	250 532 633	284 817 844	30 758 781	24 494 282	17 134 326	0,094	1,676
8	АО «Нурбанк»	311 548 729	239 677 804	265 872 915	45 675 814	20 295 153	10 762 191	0,142	0,840
9	АО «Евразийский Банк»	320 148 934	197 383 668	295 647 564	24 501 370	20 883 459	5 181 722	0,077	2,644
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	179 643 725	258 074 385	265 439 937	85 796 212	28 994 931	-5 159 619	-0,481	1,643
11	АО «Ситибанк Казахстан»	220 253 893	37 759 938	201 004 034	19 249 859	1 399 745	2 754 075	0,088	0,994
12	ДБ АО «Сбербанк»	221 586 868	107 223 613	184 101 290	37 485 578	6 593 774	8 159 646	0,165	1,623
Вариант 7 (данные на 01.04.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 411 832 678	2 330 716 207	2 138 030 930	273 801 748	156 269 723	168 744 830	0,124	0,779
2	АО «БТА Банк»	1 944 208 945	2 275 785 208	3 467 642 549	1 523 433 604	226 534 238	-126 750 129	-0,984	0,509
3	АО «Народный Банк Казахстана»	2 055 141 278	1 219 409 077	1 807 659 381	247 481 897	100 121 775	84 024 649	0,109	1,109
4	АО «АТФБанк»	1 055 966 144	831 657 456	953 855 091	102 111 053	72 197 709	37 815 505	0,111	1,821
5	АО «Банк ЦентрКредит»	1 250 318 971	671 321 573	1 143 516 919	106 802 052	76 093 817	33 499 230	0,096	2,290
6	АО «Альянс Банк»	457 031 489	577 060 973	426 445 087	30 586 402	72 794 288	-3 081 837	0,061	1,950
7	АО «KASPI BANK»	323 567 149	251 925 909	292 173 246	31 393 903	24 686 867	19 376 418	0,093	2,240
8	АО «Нурбанк»	326 443 214	237 559 095	281 184 985	45 258 229	19 990 792	11 464 868	0,136	1,011
9	АО «Евразийский Банк»	329 120 408	196 783 800	306 091 842	23 028 566	21 310 429	5 309 152	0,073	2,341
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	184 620 412	256 188 594	270 519 628	85 899 216	28 258 125	-3 874 520	-0,468	1,850
11	АО «Ситибанк Казахстан»	249 765 509	39 899 387	230 519 152	19 246 357	1 397 922	2 866 969	0,076	1,004
12	ДБ АО «Сбербанк»	223 724 920	113 754 514	186 431 482	37 293 438	6 909 707	8 429 873	0,163	1,090
Вариант 8 (данные на 01.05.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 387 886 725	2 337 351 658	2 113 674 663	274 212 062	152 043 453	166 584 374	0,125	0,724
2	АО «БТА Банк»	1 928 493 520	2 198 995 707	3 461 040 470	1 532 546 950	221 256 773	-133 281 055	-0,997	0,473

1	2	3	4	5	6	7	8	9	10
3	АО «Народный Банк Казахстана»	2 081 157 153	1 201 366 454	1 834 014 541	247 142 612	97 634 006	81 500 194	0,105	1,115
4	АО «АТФБанк»	1 047 539 286	828 389 130	945 886 602	101 652 684	69 308 132	33 624 982	0,112	1,803
5	АО «Банк ЦентрКредит»	1 259 203 616	684 765 566	1 153 042 080	106 161 536	78 300 494	30 222 797	0,084	2,293
6	АО «Альянс Банк»	443 373 914	574 183 566	413 915 895	29 458 019	60 308 223	-5 917 924	0,060	1,426
7	АО «KASPI BANK»	320 018 106	249 867 476	289 090 536	30 927 570	25 961 905	15 965 182	0,093	2,476
8	АО «Нурбанк»	332 746 465	238 061 409	286 091 832	46 654 633	19 914 871	12 255 010	0,133	1,050
9	АО «Евразийский Банк»	335 019 768	200 214 403	310 509 092	24 510 676	21 982 885	5 268 833	0,071	2,393
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	183 740 682	254 116 990	271 656 817	87 916 135	27 831 386	-4 066 038	-0,481	2,382
11	АО «Ситибанк Казахстан»	272 947 597	43 833 054	253 312 959	19 634 638	1 388 250	2 906 731	0,070	1,025
12	ДБ АО «Сбербанк»	231 180 709	116 618 185	192 749 054	38 431 655	7 733 025	8 902 316	0,158	0,829
Вариант 9 (данные на 01.06.2010)									
1	АО «КАЗКОММЕРЦБАНК»	2 389 626 096	2 339 994 262	2 112 828 231	276 797 865	152 043 453	166 584 374	0,125	0,710
2	АО «БТА Банк»	1 807 352 128	2 142 246 621	3 437 056 653	1 629 704 525	97 634 006	81 500 194	-1,071	0,483
3	АО «Народный Банк Казахстана»	2 168 610 419	1 207 745 120	1 923 330 035	245 280 384	221 256 773	-133 281 055	0,101	1,150
4	АО «АТФБанк»	1 083 205 627	824 277 157	981 399 295	101 806 332	78 300 494	30 222 797	0,108	2,014
5	АО «Банк ЦентрКредит»	1 257 779 633	678 640 985	1 151 404 867	106 374 766	69 308 132	33 624 982	0,084	2,257
6	АО «Альянс Банк»	450 693 757	571 713 405	422 720 850	27 972 907	60 308 223	-5 917 924	0,056	1,226
7	АО «KASPI BANK»	313 054 641	253 590 814	283 379 637	29 675 004	21 982 885	5 268 833	0,091	2,065
8	АО «Нурбанк»	323 574 430	233 592 971	277 405 149	46 169 281	19 914 871	12 255 010	0,137	1,070
9	АО «Евразийский Банк»	369 288 242	198 674 371	346 021 568	23 266 674	25 961 905	15 965 182	0,065	1,954
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	185 321 078	252 668 539	273 919 526	88 598 448	7 733 025	8 902 316	-0,480	2,784
11	АО «Ситибанк Казахстан»	243 632 003	48 300 697	223 783 186	19 848 817	1 388 250	2 906 731	0,082	0,989
12	ДБ АО «Сбербанк»	230 187 318	116 170 482	192 113 272	38 074 046	12 671 564	7 166 384	0,159	0,888

1	2	3	4	5	6	7	8	9	10
Вариант 10 (данные на 01.07.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 363 649 532	2 343 168 773	2 096 180 174	267 469 358	150 887 043	160 829 649	0,126	0,663
2	АО «БТА Банк»	1 869 604 990	1 961 142 794	3 499 078 029	1 629 473 039	216 023 069	-139 443 145	-1,047	0,445
3	АО «Народный Банк Казахстана»	2 114 515 153	1 204 258 387	1 861 190 253	253 324 900	96 726 640	85 904 477	0,104	1,158
4	АО «АТФБанк»	1 131 677 883	820 794 074	1 041 894 750	89 783 133	67 925 926	33 702 009	0,093	2,139
5	АО «Банк ЦентрКредит»	1 290 360 164	701 430 432	1 183 007 789	107 352 375	79 415 708	27 612 256	0,093	2,261
6	АО «Альянс Банк»	461 272 347	499 794 259	435 001 611	26 270 736	54 711 584	-3 265 147	0,051	1,478
7	АО «KASPI BANK»	318 498 157	257 730 584	289 440 094	29 058 063	26 018 047	16 695 578	0,087	1,701
8	АО «Нурбанк»	325 765 518	233 863 172	280 371 305	45 394 213	19 781 602	12 031 130	0,136	0,890
9	АО «Евразийский Банк»	365 012 097	193 746 381	341 618 590	23 393 507	22 105 161	5 541 892	0,066	1,792
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	186 072 412	252 015 716	167 901 647	18 170 765	29 867 352	-5 389 812	0,098	3,206
11	АО «Ситибанк Казахстан»	274 864 716	49 536 174	254 920 710	19 944 006	1 372 393	2 895 351	0,070	0,996
12	ДБ АО «Сбербанк»	252 080 011	124 830 625	214 600 092	37 479 919	8 103 596	9 153 498	0,145	0,837
Вариант 11 (данные на 01.08.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 383 951 843	2 334 207 160	2 116 050 272	267 901 571	149 202 849	155 794 630	0,125	0,698
2	АО «БТА Банк»	1 756 571 312	1 895 517 673	3 384 771 470	1 628 200 158	212 187 345	-133 704 926	-1,121	0,453
3	АО «Народный Банк Казахстана»	2 139 138 488	1 206 066 915	1 882 723 889	256 414 599	96 293 950	84 691 973	0,103	1,105
4	АО «АТФБанк»	1 076 779 838	809 923 190	991 259 307	85 520 531	66 495 081	33 377 668	0,094	2,022
5	АО «Банк ЦентрКредит»	1 313 602 020	707 271 006	1 207 120 970	106 481 050	80 234 293	25 013 376	0,090	2,162
6	АО «Альянс Банк»	462 230 904	530 711 091	422 724 500	39 506 404	50 198 671	-422 417	0,080	2,338
7	АО «KASPI BANK»	311 094 194	261 984 095	282 199 705	28 894 489	26 112 836	17 606 038	0,088	1,983
8	АО «Нурбанк»	322 182 828	227 116 261	276 844 342	45 338 486	19 526 097	12 326 070	0,138	0,830
9	АО «Евразийский Банк»	375 379 716	193 750 500	352 269 021	23 110 695	22 240 317	5 608 088	0,064	1,936

1	2	3	4	5	6	7	8	9	10
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	196 246 358	250 494 152	177 575 699	18 670 659	28 497 345	-4 044 392	0,095	3,221
11	АО «Ситибанк Казах-стан»	270 090 906	46 241 684	249 709 168	20 381 738	1 331 204	2 867 999	0,071	0,926
12	ДБ АО «Сбербанк»	245 247 329	131 217 992	207 856 721	37 390 608	8 392 064	9 452 102	0,149	0,729
Вариант 12 (данные на 01.09.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 380 933 759	2 319 973 110	2 113 356 214	267 577 545	148 979 390	146 993 363	0,125	0,776
2	АО «БТА Банк»	2 082 230 390	1 191 820 450	1 823 094 305	259 136 085	227 789 468	-143 940 954	0,108	0,370
3	АО «Народный Банк Казахстана»	1 957 332 019	1 937 158 704	1 851 718 949	105 613 070	94 933 800	86 053 489	0,106	1,135
4	АО «АТФБанк»	1 297 987 232	706 604 369	1 191 589 221	106 398 011	64 514 696	33 069 075	0,089	1,809
5	АО «Банк ЦентрКре-дит»	1 072 828 900	812 623 217	992 182 295	80 646 605	80 725 685	21 902 406	0,091	2,134
6	АО «Альянс Банк»	456 672 131	533 628 588	422 644 174	34 027 957	46 098 130	882 799	0,069	2,200
7	АО «KASPI BANK»	344 892 847	197 653 022	321 613 975	23 278 872	26 069 584	18 748 875	0,085	1,415
8	АО «Нурбанк»	325 050 871	221 858 956	279 853 731	45 197 140	19 006 040	12 810 150	0,136	0,757
9	АО «Евразийский Банк»	317 139 634	266 700 291	288 617 497	28 522 137	22 380 970	5 708 686	0,070	2,035
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	277 154 966	136 524 497	239 933 369	37 221 597	27 077 720	-2 068 544	0,084	3,401
11	АО «Ситибанк Казах-стан»	256 105 315	43 986 492	235 327 666	20 777 649	1 234 798	2 905 988	0,079	0,914
12	ДБ АО «Сбербанк»	202 336 516	129 122 849	183 901 040	18 435 476	8 657 015	9 791 512	0,132	0,857
Вариант 13 (данные на 01.10.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 440 600 125	2 344 299 043	2 171 975 560	268 624 565	147 292 614	144 252 360	0,122	0,803
2	АО «БТА Банк»	1 993 856 943	1 220 818 475	1 730 425 989	263 430 954	221 512 409	-114 042 206	0,104	1,359
3	АО «Народный Банк Казахстана»	1 949 747 216	1 926 003 260	1 853 034 310	96 712 906	93 088 470	86 691 696	0,110	1,036
4	АО «АТФБанк»	1 311 591 727	709 368 842	1 204 295 849	107 295 878	62 485 429	31 850 941	0,087	1,615
5	АО «Банк ЦентрКре-дит»	1 034 141 911	859 648 645	958 467 113	75 674 798	80 917 913	21 602 728	0,091	1,747
6	АО «Альянс Банк»	458 726 651	537 005 919	421 438 444	37 288 207	43 939 166	1 174 381	0,088	1,943
7	АО «KASPI BANK»	335 515 168	195 727 061	311 817 611	23 697 557	26 130 545	21 975 057	0,091	1,415
8	АО «Нурбанк»	318 986 895	225 137 265	273 052 413	45 934 482	18 442 820	13 359 560	0,138	0,803

1	2	3	4	5	6	7	8	9	10
9	АО «Евразийский Банк»	327 246 875	271 580 643	296 078 525	31 168 350	22 422 067	5 979 340	0,073	1,860
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	240 798 849	141 649 492	201 699 413	39 099 436	25 775 219	-468 390	0,091	3,584
11	АО «Ситибанк Казах-стан»	264 766 368	49 366 405	243 953 832	20 812 536	1 151 662	2 937 182	0,073	0,983
12	ДБ АО «Сбербанк»	210 167 721	133 853 720	191 621 533	18 546 188	8 739 533	10 348 912	0,152	0,708
Вариант 14 (данные на 01.11.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 514 498 715	2 319 516 628	2 245 515 648	268 983 067	144 739 096	141 933 510	0,119	0,754
2	АО «БТА Банк»	1 935 676 149	1 222 961 557	1 672 350 728	263 325 421	215 038 140	-114 823 618	0,092	1,118
3	АО «Народный Банк Казахстана»	1 943 667 071	1 868 127 403	1 870 154 004	73 513 067	91 337 229	88 256 590	0,114	1,141
4	АО «АТФБанк»	1 280 935 438	720 453 579	1 174 846 634	106 088 804	60 743 238	30 729 900	0,078	1,660
5	АО «Банк ЦентрКре-дит»	1 091 041 038	840 215 689	1 019 454 866	71 586 172	80 720 547	19 069 811	0,093	1,709
6	АО «Альянс Банк»	479 965 636	541 067 532	432 032 299	47 933 337	40 087 748	3 067 401	0,119	1,926
7	АО «KASPI BANK»	361 806 264	203 901 390	337 891 678	23 914 586	26 235 884	22 798 944	0,091	1,360
8	АО «Нурбанк»	306 670 549	204 558 107	261 310 013	45 360 536	18 402 499	12 587 452	0,144	0,715
9	АО «Евразийский Банк»	329 640 047	273 256 937	297 903 289	31 736 758	22 363 707	6 296 912	0,069	1,715
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	251 833 138	148 670 380	213 769 327	38 063 811	24 511 562	967 825	0,092	3,751
11	АО «Ситибанк Казах-стан»	293 265 307	50 482 865	271 642 609	21 622 698	1 073 199	3 005 822	0,067	0,916
12	ДБ АО «Сбербанк»	214 271 950	135 878 365	195 335 727	18 936 223	8 841 833	10 933 630	0,145	0,682
Вариант 15 (данные на 01.12.2010)									
1	АО «КАЗКОММЕРЦ-БАНК»	2 472 282 901	2 311 179 971	2 203 234 718	269 048 183	143 176 343	138 559 891	0,121	0,769
2	АО «БТА Банк»	1 968 668 869	1 216 900 650	1 704 369 073	264 299 796	201 160 481	-89 288 305	0,100	1,192
3	АО «Народный Банк Казахстана»	1 968 032 119	1 665 053 927	1 874 712 779	93 319 340	89 608 703	89 706 912	0,112	1,139
4	АО «АТФБанк»	1 238 373 907	719 305 684	1 132 258 706	106 115 201	58 557 887	29 858 761	0,075	1,487
5	АО «Банк ЦентрКре-дит»	1 020 818 609	848 590 684	956 012 434	64 806 175	80 423 566	17 227 747	0,096	1,568
6	АО «Альянс Банк»	476 902 647	544 879 446	439 159 602	37 743 045	35 339 564	6 719 466	0,099	1,794

1	2	3	4	5	6	7	8	9	10
7	АО «KASPI BANK»	342 031 645	209 321 282	317 166 665	24 864 980	26 492 058	23 681 628	0,089	1,166
8	АО «Нурбанк»	317 802 161	204 904 489	274 141 171	43 660 990	18 087 042	12 371 731	0,134	0,745
9	АО «Евразийский Банк»	338 587 966	277 584 691	304 333 362	34 254 604	22 161 884	6 790 401	0,077	1,840
10	ДО АО «БТА БАНК» - АО «ТЕМІРБАНК»	249 976 754	154 108 918	212 190 839	37 785 915	23 451 715	2 732 768	0,094	3,883
11	АО «Ситибанк Казах-стан»	208 571 987	40 048 314	186 733 435	21 838 552	1 055 626	3 016 376	0,092	0,863
12	ДБ АО «Сбербанк»	225 360 149	141 301 293	206 373 951	18 986 198	9 080 065	11 283 117	0,146	0,697

Приложение Б

Расчет коэффициентов корреляции осей с признаками

	A	B	C	D	E	F	G	H	I
1		Благоустройство жилья						Оси	
2		Признаки / Области	Водопровод	Центральное отопление	Газ	Горячее водоснабжение		1	2
3		Акмолинская	45,4	31,4	93,8	16		15,71	3,492
4		Актюбинская	64,6	51,4	92,7	42,5		-21,99	-1,669
5		Алматинская	35,6	12,9	95,5	11,8		34,67	2,662
6		Атырауская	63,6	37,8	97,2	37,5		-9,873	1,216
7		Западно-Казахстанская	22,2	19,4	95,2	17,4		34,91	-10,44
8		Жамбылская	62,8	17,8	99,3	16,6		14,68	15,12
9		Карагандинская	87,1	67,8	65,3	49,9		-52,3	19,03
10		Костанайская	50,5	45,7	97,5	42,7		-10,3	-12,11
11		Кызылординская	25,8	16,6	96,8	1,8		43,78	-0,1224
12		Мангистауская	70,7	68,5	97,6	67,9		-49,28	-17,18
13									
14		Признак / Ось	Водопровод	Центральное отопление	Газ	Горячее водоснабжение			
15		1	=КОРРЕЛ(C3:C12;$H$3:$H$12)	=КОРРЕЛ(D3:D12;$H$3:$H$12)	=КОРРЕЛ(E3:E12;$H$3:$H$12)	=КОРРЕЛ(F3:F12;$H$3:$H$12)			
16		2	=КОРРЕЛ($I$3:$I$12;C3:C12)	=КОРРЕЛ($I$3:$I$12;D3:D12)	=КОРРЕЛ($I$3:$I$12;E3:E12)	=КОРРЕЛ($I$3:$I$12;F3:F12)			

Приложение В

Пример вычислений для определения новых координат объектов

1. Определить углы между исходными и повернутыми координатными осями. Они будут следующими: между осями[1] 1 и 1’ - 180°, осями 1 и 2’ - 270°, между осями 2 и 1’ - 90°, между осями 2 и 2’ - 180°.
2. Определить косинусы этих углов и составить матрицу ортогонального преобразования:

$$T = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Примечание.

Для определения косинусов можно воспользоваться калькулятором Windows. Для этого нужно:

- Открыть в *Главном меню* пункт – *Программы* – *Стандартные* – *Калькулятор*.
- В меню *Вид* выбрать - *Инженерный*, опцию - *Градусы*.
- Ввести нужное число, например, 180 и нажать кнопку *cos*. Получаемый результат: -1.

3. Затем нужно исходную матрицу координат умножить на вычисленную матрицу[2], в результате чего получится новая система координат:

$$\begin{bmatrix} 15{,}71 & 3{,}492 \\ -21{,}99 & -1{,}669 \\ 34{,}67 & 2{,}662 \\ -9{,}873 & 1{,}216 \\ 34{,}91 & -10{,}44 \\ 14{,}68 & 15{,}12 \\ -52{,}3 & 19{,}03 \\ -10{,}3 & -12{,}11 \\ 43{,}78 & -0{,}1224 \\ -49{,}28 & -17{,}18 \end{bmatrix} \times \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} -15{,}71 & -3{,}492 \\ 21{,}99 & 1{,}669 \\ -34{,}67 & -2{,}662 \\ 9{,}873 & -1{,}216 \\ -34{,}91 & 10{,}44 \\ -14{,}68 & -15{,}12 \\ 52{,}3 & -19{,}03 \\ 10{,}3 & 12{,}11 \\ -43{,}78 & 0{,}1224 \\ 49{,}28 & 17{,}18 \end{bmatrix}$$

[1] Оси 1 и 1’ – исходная и повернутая на 270° первая ось, 2 и 2’ – исходная и повернутая вторая ось.

[2] Сайт on-line калькулятора для умножения матриц: http://www.webmath.ru/web/prog5_1.php.