МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра системного анализа и управления

Т.В. ГАИБОВА, Н.А.ШУМИЛИНА

СТАТИСТИЧЕСКИЕ МЕТОДЫ СИСТЕМНОГО АНАЛИЗА

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
К ЛАБОРАТОРНОМУ ПРАКТИКУМУ

Рекомендовано к изданию Редакционно-издательским советом
государственного образовательного учреждения высшего
профессионального образования
«Оренбургский государственный университет»

Оренбург 2005

УДК519.8 (07)
ББК22.18 я7
Г 14

Рецензент
кандидат технических наук С.Г. Сергеев

Гаибова Т.В., Шумилина Н.А.
Г 14 **Статистические методы системного анализа: Методические указания к лабораторному практикуму. - Оренбург: ГОУ ОГУ, 2005. - 18 с.**

Приведены методические указания к 3 лабораторным занятиям по применению статистических методов к анализу и синтезу сложных систем. Каждая работа включает теоретическое изложение материала, описание методики реализации и контрольные вопросы для самоподготовки.

Методические указания предназначены для выполнения лабораторных занятий по дисциплине "Системный анализ и принятие решений" для студентов специальности 553000 и практических занятий по дисциплине «Теория системного анализа и принятия решений» для студентов специальности 330100.

ББК22.18 я7

1 Лабораторная работа №1. Разработка регрессионных моделей объекта по результатам экспериментов

1.1 Теория вопроса

Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между некоторыми наблюдаемыми переменными и построения с помощью полученных экспериментальных данных математического описания исследуемого объекта (задача идентификации) /1/. Исследуемый объект или процесс в планировании эксперимента представляют "черным ящиком", на входе которого действуют управляющие x_i и возмущающие факторы. Выходы объекта называют откликами y_i. Устанавливая факторы на тех или иных уровнях, получают разные реализации $y_j = f_j(x_j)$, поэтому функцию f_j называют функцией отклика. Задача регрессионного анализа - выбор вида функции отклика и анализ свойств результата. Ее выбирают в виде отрезка полинома, характер которого зависит от необходимой точности определения функции:

$$y = f(X,\beta) = \beta_0 + \sum_{i=1}^{m}\beta_i x_i + \sum_{i=1}^{m}\beta_{ii}x_i^2 + \sum_{i\neq n}\beta_{in}x_i x_n + \ldots, \qquad (1)$$

где все β - коэффициенты регрессии, причем

$$\beta_1 = \partial f \big/ \partial x_1 ; \ldots \beta_{11} = (1/2)\partial^2 f \big/ \partial x_1^2 ; \ldots \beta_{12} = \partial^2 f \big/ \partial x_1 \partial x_2 ; \ldots \qquad (2)$$

По результатам опытов можно найти только выборочные значения функции и точечные оценки $\bar{y}, b_0, b_1, \ldots$ В итоге получают регрессионную модель

$$\bar{y} = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m} b_{ii} x_i^2 + \ldots . \qquad (3)$$

Решение этой задачи можно разбить на следующие этапы.

1 Выдвинуть гипотезу о виде зависимости $y = f(X,\beta)$.

2 По экспериментальным данным $\{x_{iu}\}$ и $\{y_u\}$ найти оценки $b_0, b_1 \ldots$ коэффициентов регрессии $\beta_0, \beta_1 \ldots$; $u = \overline{1,n}$; n - число наблюдений или опытов.

3 Определить значимость оценок коэффициентов регрессии (отличие от нуля).

4 Проверить адекватность построенной модели объекту.

5 Проверить работоспособность модели.

Вид функции отклика должен быть по возможности простым, но в то же время хорошо выражать реальную зависимость. Ее первоначальный выбор базируется на материалах решения аналогичных задач и интуиции. Если проверка адекватности модели покажет, что она не соответствует объекту, то нужно перейти к более сложной модели.

Проверка значимости коэффициентов регрессии. В результате проверки устанавливается статистическая значимость или незначимость отличия оценок коэффициентов регрессии от нуля., т.е. проверяется, обусловлено ли отличие b_j от нуля влиянием помех или это отличие не случайно и влияние j-го фактора существенно $(b_j \neq 0)$.

Для проверки используется статистика:

$$t_j = b_j \Big/ \sigma\{b_j\} \quad , \tag{4}$$

где $\sigma\{b_j\}$ - среднеквадратическое отклонение оценки b_j. Найденное значение t_j сравнивается с табличным значением t-распределения с числом степеней свободы $\nu_{b_j} = n(k-1)$. Если $t_j > t_{табл}$, то считается, что b_j отличается от нуля не случайно, коэффициент b_j статистически значим и должен быть сохранен в уравнении регрессии. Если же $t_j < t_{табл}$, то коэффициент b_j статистически незначим и может быть исключен из уравнения регрессии.

Проверка адекватности модели. Идея проверки адекватности заключается в сравнении дисперсии предсказания на основе исследуемой регрессионной модели с дисперсией шума.

Дисперсия предсказания (остаточная дисперсия) определяется по формуле

$$\sigma_R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n-d}, \tag{5}$$

где d - число значимых коэффициентов в регрессионной модели.

Дисперсия шума:

$$\sigma_Y^2 = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{(n-1)}. \tag{6}$$

Рассчитывается статистика:

$$\hat{F} = \frac{\sigma_R^2}{\sigma_Y^2}. \tag{7}$$

Расчетное значение $\hat{F}$ сравниваем с табличным значением F-критерия Фишера со степенями свободы $\nu_Y = n-1$ и $\nu_R = n-k-1$.

При $\hat{F} > F_{табл}$ делается вывод с уровнем значимости α о неадекватности модели. Это значит, что необходимо усовершенствовать модель (усложнить, ввести нелинейность и т.п.).

Анализ работоспособности модели. Модель считается работоспособной и годится для предсказания поведения объекта (процесса) если она в 2 раза уменьшает ошибку предсказания по сравнению с предсказанием по среднему. При этом коэффициент детерминации, который показывает, какая доля вариации отклика объекта обусловлена коэффициентами регрессии, будет $R^2 \geq 0{,}75$.

Коэффициент детерминации вычисляется по формуле

$$R^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \qquad (8)$$

1.2 Цель занятия

Закрепить теоретические знания и приобрести практические навыки в построении регрессионных моделей объекта по экспериментальным данным с использованием программы Statistica.

1.3 Задание на занятие

По результатам наблюдений за функционированием объектов получены экспериментальные данные в виде характеризующих параметров (x,y) и показатели качества объектов (z). Построить регрессионные модели объектов по заданным экспериментальным данным.

Решение общей задачи разбивается на несколько этапов:

а) предварительная обработка данных с целью стандартизации результатов наблюдения;

б) оценка параметров регрессионных моделей;

в) проверка значимости коэффициентов регрессии;

г) оценка точности регрессионных моделей;

д) выводы о возможности применения составленных регрессионных моделей.

1.4 Методические указания по выполнению работы

1.4.1 Приобретение практических навыков применения регрессионного анализа для построения моделей объекта по экспериментальным данным и заданному виду функции отклика.

Каждый студент обрабатывает свой вариант экспериментальных данных, выданный преподавателем в соответствии с порядковым номером. Для вычисления необходимо использовать систему *Statistica*, модули *Множественная регрессия* и *Нелинейное оценивание*.

Обработка данных ведется применительно к трем видам уравнений регрессии:

- линейная регрессия $z = a*x + b$ - модуль *Множественная регрессия*;

- параболическая регрессия второго порядка - $z = c*x + d*x + e$ - модуль *Нелинейное оценивание*;

- множественная регрессия $z=f*x+g*y+p$ - модуль *Множественная регрессия.*

Для минимизации отклонений между экспериментальными и расчетными значениями рекомендуется использовать метод наименьших

квадратов. Оценивать значимость коэффициентов регрессии, адекватность и работоспособность уравнений регрессии и формировать выводы о возможности их использования необходимо на основе коэффициента детерминации модели, значений стандартных ошибок, а также на значениях критериев Стьюдента и Фишера.

1.4.2 Идентификация зависимости инвестиционного лага проекта развития промышленного предприятия τ от величины проектных затрат G.

Каждый студент обрабатывает свой вариант экспериментальных данных в соответствии с порядковым номером (примерный вариант исходных данных представлен в таблице 1). Для вычисления необходимо использовать систему *Statistica*, модуль *Нелинейное оценивание*. Вид аппроксимирующей функции определяется студентом самостоятельно. Для минимизации отклонений между экспериментальными и расчетными значениями рекомендуется использовать метод наименьших квадратов. Оценивать значимость полученных коэффициентов аппроксимирующей зависимости, адекватность и работоспособность уравнений и формировать выводы о возможности их использования необходимо на основе коэффициента детерминации модели, значений стандартных ошибок, а также на значениях критериев Стьюдента и Фишера.

1.5 Содержание отчета

Отчет должен содержать (для каждого уравнения регрессии):

- заданные выборки экспериментальных данных в исходном и стандартизированном видах;
- значения коэффициентов уравнений регрессии;
- значения коэффициентов детерминации моделей и стандартные ошибки;
- значения критериев Стьюдента и Фишера и уровень значимости;
- отклонения расчетных значений от фактических значений функции;
- выводы по результатам обработки экспериментальных данных.

1.6 Контрольные вопросы

Виды зависимостей случайных величин.

Постановка задачи построения регрессионной модели.

Для чего осуществляется стандартизация матрицы наблюдений?

Как определяется коэффициент корреляции двух случайных величин?

Какова область значений коэффициента корреляции?

В чем сущность метода наименьших квадратов для расчета коэффициентов регрессионных моделей?

Что показывает коэффициент детерминации модели и какова область его значений?

Как определяется значимость коэффициентов регрессии?

Как определяется адекватность полученной модели?

Таблица 1 – Пример варианта исходных данных для идентификации зависимости инвестиционного лага τ от величины проектных затрат G /2/

№ проекта	Проектные затраты G, отн.единицы	Инвест.лаг τ, мес.
1	14,5	8
2	34,4	6
3	1,3	6
4	5,3	4
5	41,0	4
6	12,7	6
…	…	…
20	7,4	6
21	97,2	12
22	3,3	6

2 Лабораторная работа №2. Применение метода главных компонент для формирования обобщенных критериев в задачах многокритериальной оптимизации

2.1 Теория вопроса

Для решения задач многокритериальной оптимизации на практике часто используется метод свертки критериев в один обобщенный показатель (метод взвешенных сумм). Используемые для определения весовых коэффициентов экспертные методы зачастую дают очень субъективные результаты. Поэтому для построения объективных обобщенных показателей на основании достоверной информации о структуре множества критериальных переменных необходимо применять обоснованные научные методы, например, корреляционный и факторный анализ.

Очень важно установление взаимосвязей критериальных переменных, так как в случаях, когда две или более переменные очень сильно связаны между собой, можно без ущерба для качества принятия решений исключить одну или несколько переменных из рассмотрения. Это приводит к задаче отыскания небольшого набора существенных (наиболее информативных) критериев, которые позволяют осуществлять построение алгоритма принятия решений при сокращенном числе переменных без значительной потери информации.

Возможны два способа выбора существенных переменных /3/.

К первому способу относятся методы, позволяющие сократить размерность пространства без видоизменения переменных. Среди них можно выделить так называемую группировку взаимно коррелированных переменных, когда, например, матрица связей преобразуется в блочно-диагональный вид, а затем из каждого блока (группы переменных) выбирается одна переменная, образующая в сочетании с другими представителями других групп совокупность существенных переменных. Распространенные методы выделения существенных переменных основаны на построении линейной регрессионной модели. В этом случае поиск переменных осуществляется последовательным исключением (или введением) из модели переменных, а их отнесение к разряду существенных определяется в соответствии с изменениями множественного коэффициента корреляции.

Ко второму способу относятся методы, в которых снижение размерности пространства происходит одновременно с его преобразованием. Это – факторный анализ, метод главных компонент, канонический анализ. Характерной особенностью этого подхода является то, что происходит выбор и оценка значимости не отдельных переменных, а информативных по совокупности групп переменных.

Метод главных компонент – это один из широко использующихся методов многомерной математической статистики. Он применяется для решения следующих задач:

- причинный анализ взаимосвязей показателей и определение их стохастической связи с главными компонентами;
- построение обобщенных технико-экономических показателей;
- ранжирование объектов или наблюдений по главным компонентам;
- классификация объектов наблюдений;
- сжатие исходной информации;
- построение уравнений регрессии по обобщенным технико-экономическим показателям.

В общем виде задача снижения размерности признакового пространства и построения интегрального показателя может быть сформулирована следующим образом.

Пусть N - число исследуемых объектов;

n - число признаков (измеряемых характеристик объектов - критериев);

матрица Y порядка $n \times N$ - совокупность всех N наблюдаемых значений всех параметров n после нормализации.

Необходимо описать набор критериев n числом главных компонент $m << n$, обеспечивающих долю дисперсии $\gamma \geq 0{,}95$ и сформировать интегральный показатель оптимальности на основе матрицы A весовых коэффициентов, учитывающих тесноту связи между исходными критериями и главными компонентами.

$Y = A \cdot F$, где матрица F включает совокупность всех N полученных значений всех n главных компонент.

Или в развернутом матричном виде:

$$\begin{pmatrix} y_{11} & y_{12} & \dots & y_{1N} \\ y_{21} & y_{22} & \dots & y_{2N} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nN} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nN} \end{pmatrix} \cdot \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1N} \\ f_{21} & f_{22} & \dots & f_{2N} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nN} \end{pmatrix} \quad (9)$$

Задача сводится к определению матрицы A.

Для исследования начальными основными данными являются коэффициенты корреляции.

Связь между главными компонентами и коэффициентами корреляции для объекта i:

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \dots + a_{jn} f_{ni}; \quad i = 1,2,\dots,N, \quad (10)$$

где y_{ji} - нормированное значение j-го критерия для i-го объекта; f_{1i} - значение первой главной компоненты для i-го объекта.

Для устранения неоднородности рассматриваемых критериев может быть использован следующий вид нормализации:

$$y_{ji} = \frac{x_{ji} - x_{cpj}}{\sigma_j},\ \text{j=1…n} \qquad (11)$$

где x_{ji} -значение случайной величины X_j при i-м измерении ;

x_{cpj} - среднее значение случайной величины x_j по результатам N измерений;

σ_j - среднее квадратическое отклонение X_j .

Среднее значение случайной величины X_j определяется по формуле

$$x_{cpj} = \frac{1}{N}\sum_{i=1}^{N} x_{ji}\ , \qquad (12)$$

а среднее квадратическое отклонение

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N}(x_{ji} - x_{cpj})^2}{N-1}}\ . \qquad (13)$$

Коэффициент корреляции r_{jk} характеризует связь между двумя случайными величинами X_j и X_k в случае линейной корреляции между ними. Для любых признаков и случайных величин

$$r_{jk} = \frac{1}{N}\sum_{i=1}^{N} y_{ji} y_{ki}\ , \qquad (14)$$

где y_{ki} - нормированное значение случайной величины X_k для i-го измерения (объекта).

Вариабельность, зависящая от особенностей объекта, является причиной разброса значений критериев от объекта к объекту относительно математического ожидания. Полная дисперсия статистического признака выражается через дисперсию главных компонент:

$$\sigma_o^2 = \frac{1}{N}\sum_{i=1}^{N} y_{ji}^2 = \frac{1}{N}[a_{ji}^2\sum_{i=1}^{N} f_{1i}^2 + a_{j2}^2\sum_{i=1}^{N} f_{2i}^2 + ... + a_{jn}^2\sum_{i=1}^{N} f_{ni}^2 + 2(a_{j1}a_{j2}\sum_{i=1}^{N} f_{1i}f_{2i} + \\ + a_{j1}a_{j3}\sum_{i=1}^{N} f_{1i}f_{3i} + ... + a_{j(n-1)}a_{jn}\sum_{i=1}^{N} f_{(n-1)i}f_{ni})] \qquad (15)$$

Так как дисперсии нормированных величин равны единице, а главные компоненты ортогональны, то выражение упрощается :

$$\sigma_j^2 = a_{j1}^2 + a_{j2}^2 + ... + a_{jn}^2 = 1 \qquad (16)$$

Слева записана полная дисперсия, а справа – доли полной дисперсии, относящиеся к соответствующим главным компонентам. Дисперсия является характеристикой изменчивости случайной величины, ее отклонений от среднего значения. Полный вклад r-го фактора в дисперсию всех n признаков определяет ту долю общей дисперсии, которую данная главная компонента объясняет. Этот вклад вычисляется по формуле:

$$V_r = \sum_{j=1}^{n} a_{jr}^2\ , \qquad (17)$$

где j - индекс признака (показателя коммерческой эффективности); r - индекс главной компоненты.

Несмотря на то, что вместо n признаков получается такое же количество главных компонент, вклад большей части главных компонент в объясняемую дисперсию оказывается небольшим. Исключают из рассмотрения те главные компоненты, вклад которых мал. При помощи m первых (наиболее весомых) главных компонент можно объяснить основную часть суммарной дисперсии. Таким образом:

- метод главных компонент позволяет описать большой набор признаков (показателей коммерческой эффективности) n небольшим числом главных компонент m, $m << n$;
- различия между объектами зависят от доли изменчивости, связанной с данной главной компонентой;
- связи между признаками и факторами (главными компонентами) – линейные;
- для каждого признака эффект воздействия факторов суммируется.

На основе выявленных наиболее весомых главных компонент, как критериальных характеристик, предлагается проводить компонентный анализ для выявления индивидуальных значений этих главных компонент для рассматриваемых объектов и ранжирования объектов по весу этих значений, что позволит выбрать наиболее оптимальные объекты. Эта процедура может использоваться, например, при проведении экспертизы инвестиционных проектов и отбора их для финансирования по критериям коммерческой эффективности и для оптимизации программ развития промышленности региона в целом.

2.2 Цель занятия

Закрепить теоретические знания и приобрести практические навыки в обработке результатов метода корреляционных плеяд и метода главных компонент с использованием программы Statistica при построении обобщенного критерия оптимальности для многокритериальных задач принятия решений.

2.3 Задание на занятие

Построить обобщенный показатель коммерческой эффективности инвестиционных проектов развития промышленных предприятий из семи частных показателей коммерческой эффективности:

- y1 – доход на капитал;
- y2 – срок окупаемости;
- y3 – будущая стоимость проекта;
- y4 – чистый дисконтированный доход (NPV);
- y5 – внутренняя норма рентабельности (IRR);
- y6 – индекс доходности (PI);

- y7 – индекс возвратности (PBP)

на основании системного анализа критериев с применением метода корреляционных плеяд и метода главных компонент /2/.

Решение общей задачи разбивается на несколько этапов:

а) стандартизировать (нормировать) исходные экспериментальные данные;

б) построить корреляционную матрицу частных критериев коммерческой эффективности;

в) на базе матрицы корреляций, изменяя пороговое значение степени связей частных критериев *G* (0,3; 0,5; 0,7; 0,9) проследить динамику образования корреляционных плеяд – не пересекающихся и достаточно удаленных друг от друга подмножеств – с целью отыскания наиболее взаимосвязанных и информативных переменных (рисунок 1);

г) провести факторный анализ для исходного критериального пространства и для наиболее информативных критериев, отобранных в результате применения метода корреляционных плеяд. Выделить главные компоненты. Определить доли объясняемой ими дисперсии;

д) проанализировать весовые коэффициенты главных компонент и интерпретировать их;

е) построить обобщенный показатель коммерческой эффективности на основе результатов проведенного анализа.

Таблица 2 - Матрица корреляций критериев коммерческой эффективности инвестиционных проектов

	y1	y2	y3	y4	y5	y6	y7
y1	1						
y2	0,16	1					
y3	0,76	0	1				
y4	0,79	0,31	0,81	1			
y5	0,73	0,13	0,95	0,82	1		
y6	0,77	0,12	0,93	0,96	0,88	1	
y7	0,13	0,97	-0,09	0,2	0,09	0,1	1

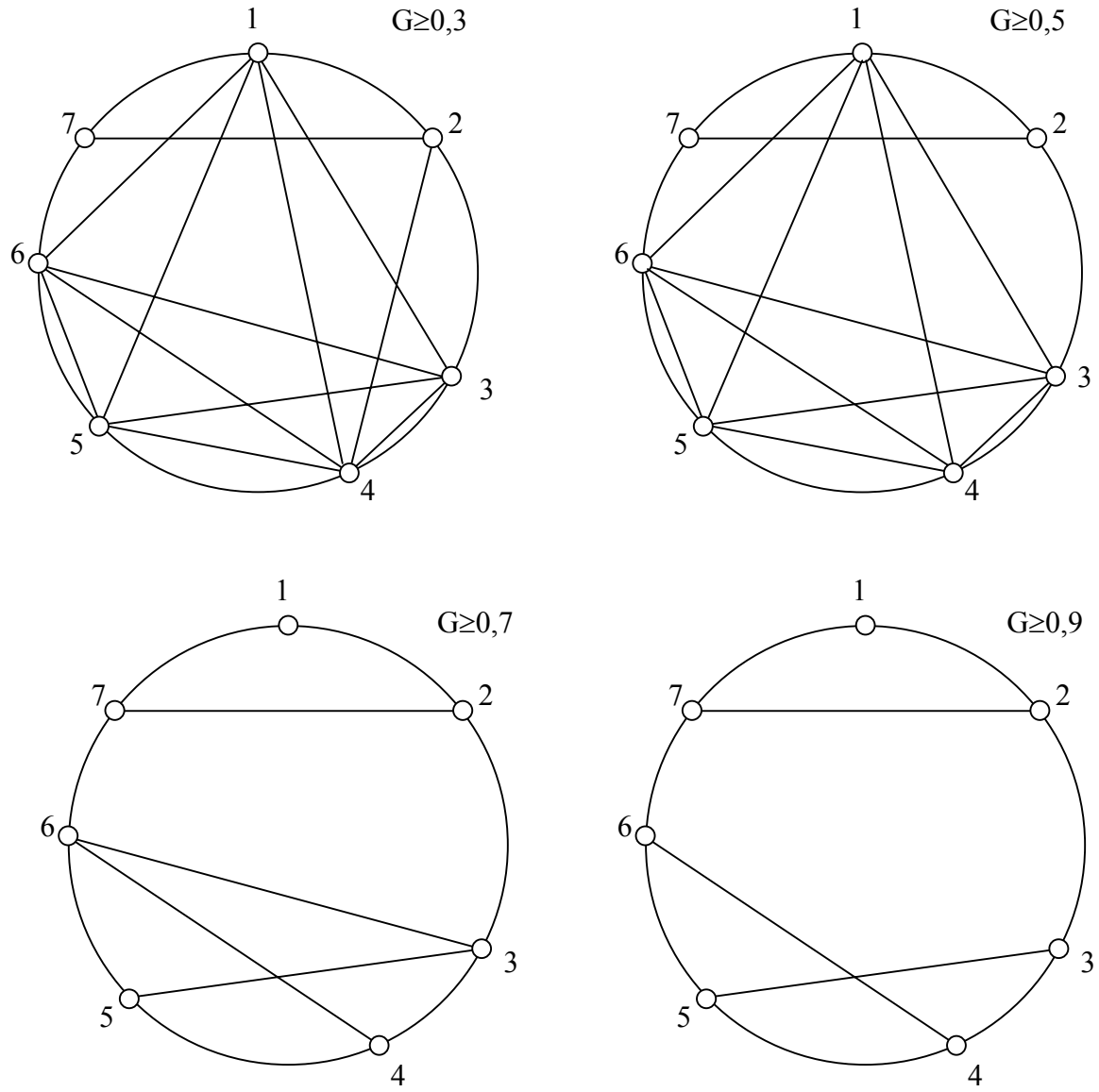

Рисунок 1 – Пример динамики образования корреляционных плеяд для пространства из семи частных критериев на основе корреляционной матрицы, представленной в таблице 2

2.4 Методические указания по выполнению работы

Каждый студент обрабатывает свой вариант экспериментальных данных, выданный преподавателем (пример исходных данных приведен в табл.3). Для вычисления необходимо использовать систему *Statistica*, модуль *Факторный анализ*.

2.5 Содержание отчета

Отчет должен содержать:

- заданные экспериментальные данные;

- исходные данные после нормировки;
- матрицу корреляций частных критериев;
- корреляционные плеяды для четырех пороговых значений степени связей частных критериев между собой;
- весовые коэффициенты выделенных главных компонент с указанием доли объясняемой ими дисперсии;
- построенный обобщенный показатель коммерческой эффективности;
- распределение инвестиционных проектов по сформированному обобщенному показателю.

2.6 Контрольные вопросы

Постановка задачи снижения размерности признакового пространства

Сущность метода главных компонент

Возможности применения метода главных компонент

Основные идеи метода корреляционных плеяд

Таблица 3 - Пример варианта исходных данных – показатели коммерческой эффективности инвестиционных проектов

№ проекта	Доход на капитал %	Срок окуп, год	Будущая ст.-сть проекта, отн.ед.	IRR, %	NPV, отн.ед	PI	PBP, год
1	46,6	3,5	5565,9	24	497,4	1,04	5,1
2	37,3	3,02	14293	26	8927	1,51	3,5
…							
21	24,3	1,9	82465,2	44,3	23452	2,2	2,2
22	12,5	6,59	4800,2	11	5100,2	0,73	20

3 Лабораторная работа №3. Применение кластерного анализа для задач классификации при отсутствии априорной информации

3.1 Теория вопроса

Кластерный анализ включает в себя набор различных алгоритмов классификации для организации наблюдаемых данных в наглядные структуры. Задачу классификации можно сформулировать следующим образом /3/.

Имеется некоторое конечное множество объектов произвольной природы, представленных совокупностью соответствующих векторов. Необходимо классифицировать эти объекты, т.е. разбить их множество на заданное или произвольное количество групп (кластеров, классов, таксонов) таким образом, чтобы в каждую группу оказались включенными объекты, близкие между собой в том или ином смысле. Априорная информация о классификации объектов при этом отсутствует. Таким образом, необходимо разбить множество векторов X на k попарно непересекающихся классов $X_1,....,X_k$ так, чтобы $\bigcup_{i=1}^{k} X_i = X$, причем $1 \le k \le M$, где M - число векторов. Для оценки расстояния между двумя векторами $x, y \in R^n$ могут быть использованы следующие меры:

1) евклидово расстояние:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad ; \tag{18}$$

2) квадрат евклидова расстояния:

$$\sum_{i=1}^{n}(x_i - y_i)^2 \quad ; \tag{19}$$

3) расстояние городских кварталов (манхэттенское расстояние)

$$\sum_{i=1}^{n}|x_i - y_i| \quad ; \tag{20}$$

3) расстояние Чебышева:

$$\max|x_i - y_i|; \tag{21}$$

4) степенное расстояние:

$$\sum_{i=1}^{n}(|x_i - y_i|^p)^{\frac{1}{r}} \tag{22}$$

где p, r - параметры, задаваемые пользователем;

5) процент несогласия $(число \ \ x_i \ne y_i)/i$.

3.2 Цель занятия

Закрепить теоретические знания и приобрести практические навыки в обработке результатов кластерного анализа данных с

использованием системы Statistica с целью выявления плохо определенных зависимостей и недостаточно изученных закономерностей.

3.3 Задание на занятие

Разбить рассматриваемую совокупность инвестиционных проектов на отдельные кластеры по типу производства и степени его совершенства на основании исследования параметров производственных функций.

Решение общей задачи разбивается на несколько этапов:

а) провести кластерный анализ исходных данных методом k средних. В качестве меры расстояния между кластерами использовать евклидово расстояние;

б) проанализировать описательные статистики полученных кластеров, расстояния элементов до центров кластеров;

в) провести кластерный анализ исходных данных методом иерархической кластеризации по правилам простого и полного связывания;

г) проанализировать построенную древовидную диаграмму и сравнить результаты иерархической кластеризации с результатами кластеризации методом k-средних.

3.4 Методические указания по выполнению работы

Каждый студент обрабатывает свой вариант экспериментальных данных, выданный преподавателем (примерный вариант исходных данных приведен в таблице 4). Для вычисления необходимо использовать систему *Statistica*, модуль *Кластерный анализ*.

3.5 Содержание отчета

Отчет должен содержать:

- заданные экспериментальные данные;
- описательные статистики кластеров, полученных методом k-средних;
- элементы каждого кластера, полученного методом k-средних, и их расстояние до центра кластера;
- результаты иерархической кластеризации в виде древовидной диаграммы по правилам простого и полного связывания;
- выводы по результатам обработки экспериментальных данных.

3.6 Контрольные вопросы

Постановка задачи классификации объектов при отсутствии априорной информации.

Основные идеи методов кластерного анализа.

Виды мер расстояний между объектами кластеризации.
Правило простого связывания.
Правило полного связывания.

Таблица 4 – Пример варианта исходных данных для кластеризации -показатели развития объектов инвестирования

№ проекта	Осн. фонды, отн.ед.	Кол-во работающих, отн. ед.	Годовая выручка (при полном освоении мощностей), отн.ед.
1	9,5	2,82	29,44
2	28,7	2,9	59,58
3	0,7	1,36	3,75
4	5,2	1,18	2,55
5	34,6	5,44	65,38
…			
20	5,3	2,1	8,2
21	80,1	9,3	98,83
22	3	1,23	5,73

4 Литература, рекомендуемая при подготовке к выполнению лабораторных работ

4.1 Гмурман В.Е. Теория вероятностей и математическая статистика. Учеб. пособие для ВУЗов. - М.: Высшая школа, 1999. - 479 с.

4.2 Горковенко В.А. Методы статистического анализа в задачах управления производством: Учебное пособие. -Л.: изд. ЛПИ,1988. - 85 с.

4.3 Многомерные статистические методы и основы эконометрики: Учебно-практическое пособие. / Сост. А.М. Дубров и др. – М.: МЭСИ, 1998. – 92 с.

4.4 Казанцев В.С. Задачи классификации и их программное обеспечение. – М.:Наука, 1990. – 136 с.

Список использованных источников

1 Горковенко В.А. Методы статистического анализа в задачах управления производством: Учебное пособие. -Л.: изд. ЛПИ,1988. - 85 с.

2 Гаибова Т.В. Многокритериальная оптимизация инвестиционных проектов развития промышленных предприятий: Дисс… канд. техн. наук: Спец. 05.13.01: Системный анализ, управление и обработка информации (промышленность)/ Т.В. Гаибова; Рук. Р.Т. Абдрашитов. – Самара: СамГТУ, 2004. – 136 с.: ил.

3 Казанцев В.С. Задачи классификации и их программное обеспечение. – М.:Наука, 1990. – 136 с.