Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

И. Ф. КУРБЫКО А. С. ЛЕВИЗОВ С. В. ЛЕВИЗОВ

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

Учебное пособие

Владимир 2018

УДК 519.22/25
ББК 22.172
К93

Рецензенты:

Доктор физико-математических наук, профессор
профессор кафедры алгебры и геометрии
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых
Н. И. Дубровин

Кандидат экономических наук, доцент
зам. директора Владимирского филиала автономной некоммерческой
образовательной организации высшего образования
Центросоюза Российской Федерации
«Российский университет кооперации»
М. А. Шумилина

Издается по решению редакционно-издательского совета ВлГУ

Курбыко, И. Ф. Методы прикладной статистики : учеб. по-
К93 собие / И. Ф. Курбыко, А. С. Левизов, С. В. Левизов ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2018. – 184 с. – ISBN 978-5-9984-0845-8.

Представлены в обобщающем виде основные этапы комплексного анализа данных на основе многомерных методов математической статистики. Изложены теоретические аспекты прогнозирования временных рядов, наиболее часто используемых в экономической практике. Материал по каждой теме проиллюстрирован формулами и примерами подробного решения задач. В качестве аналитического инструмента моделирования используется статистический пакет STADIA.

Адресовано студентам, обучающимся по направлениям подготовки 01.03.02 «Прикладная математика и информатика», 38.03.05 «Бизнес-информатика», 38.03.01 «Экономика», 02.04.01 «Математика и компьютерные науки», аспирантам и научным работникам, интересующимся задачами прикладной статистики.

Рекомендовано для формирования профессиональных компетенций в соответствии с ФГОС ВО.

Ил. 44. Табл. 43. Библиогр.: 27 назв.

УДК 519.22/25
ББК 22.172

ISBN 978-5-9984-0845-8

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Методами и правилами обработки и анализа статистических данных (т. е. сведений о числе объектов, обладающих определенными признаками в какой-либо совокупности) занимается математическая статистика. Сами методы и правила строятся вне зависимости от того, какие статистические данные обрабатываются, но обращение с ними предусматривает обязательное понимание сущности явления, изучаемого с помощью этих правил.

Математическая статистика применима к экономике потому, что экономические данные зачастую представляют собой статистические сведения, т. е. информацию об однородных совокупностях объектов и явлений. Такими однородными совокупностями могут выступать выпускаемые товары и услуги, персонал предприятий и отраслей, данные о финансовых результатах и т. д.

Центральное понятие математической статистики – случайная величина – это «всякая наблюдаемая величина, изменяющаяся при повторении общего комплекса условий, в которых она возникает» [18, с. 332]. В зависимости от случая она может принимать те или иные значения с определенными вероятностями. В случае когда перечень значений этой величины неудобен для изучения (поскольку их много), математическая статистика позволяет получить нужные сведения о случайной величине по совокупности наблюдений над нею (выборке), т. е. зная существенно меньшее количество ее значений. Объясняется это тем, что статистические данные подчиняются таким законам распределения (или приводятся к ним искусственно), которые характеризуются лишь несколькими параметрами. Зная их, можно получить достаточно полное представление о значениях случайной величины.

Л. И. Лопатников выделяет пять основных типов задач, которые решаются в экономике средствами математической статистики [18, с. 184]:

1. Оценка статистических данных.

2. Сравнение этих данных с определенным стандартом и между собой.

3. Формирование групп данных и исследование связей между статистическими данными и их группами.

4. Нахождение наилучшего варианта измерения изучаемых данных.

5. Анализ временных рядов.

Особенно важно для экономики то, что математическая статистика позволяет на основании анализа течения событий в прошлом, т. е. исследований выбранных на определенные даты сведений о характерных чертах системы, предсказать вероятное развитие явления в будущем (при условии неизменности внутренних и внешних условий).

В управлении хозяйственными процессами используют разнообразные математико-статистические методы. С ними связаны теория массового обслуживания, позволяющая наиболее эффективно организовать ряд процессов производства и обслуживания населения; теория расписания, предназначенная для выработки оптимальной последовательности производственных, транспортных и иных операций; теория решений, теория управления запасами, а также теория планирования эксперимента и выборочного контроля качества продукции, сетевые методы планирования и управления.

На основе математико-статистической обработки данных строят математические модели экономических процессов, составляют экономические и технико-экономические прогнозы. Широкое распространение математико-статистических методов в различных сферах жизни общества опирается на развитие электронно-вычислительной техники. В условиях современных развитых компьютерных технологий обработка информации на основе многомерных методов математической статистики применяется в самых разнообразных видах человеческой деятельности.

Реализация многомерных методов математической статистики требует системного подхода и оптимального конструирования этапов получения и обработки информации. В частности, построение адекватных математических моделей сложных технических структур [23], изучение связей в многофакторных экономических системах [8 – 9] связаны с тщательным проведением множественного корреляционно-регрессионного анализа. Значительную роль играет комплексный анализ экспериментальных данных во многих областях науки и практики (см., в частности, [9; 14; 20; 26] и имеющуюся там библиографию).

Тематика и содержание настоящего пособия направлены на выработку навыков самостоятельной научной работы учащихся вузов, понимание практической значимости теории математической статистики. Материал подготовлен на основе опыта чтения спецкурсов для студентов направлений «Математика», «Математические методы в экономике», «Социология», «Экономика». Изложение теории сопровождается решением большого количества задач различной степени трудности.

В главе 1 изложены теоретические аспекты некоторых методов математической статистики. Здесь предложен алгоритм статистического анализа и моделирования связей в многофакторной системе показателей на основе многомерных методов математической статистики. Подробно описаны основные шаги каждого этапа исследования, приведена их краткая характеристика и обоснование. По ходу изложения алгоритма формулируются основные понятия и определения, которыми руководствуются при исследовании.

В главе 2 представлены задачи многомерного корреляционно-регрессионного анализа и проиллюстрированы результаты реализации алгоритма исследования на примере получения аналитической группировки регионов России по уровню инновационного потенциала и построения математических моделей зависимости объёмов результативного показателя инноваций от факторных переменных инновационной деятельности регионов.

Главы 3 – 4 иллюстрируют методику кластерного и факторного анализа. Здесь приведено в деталях решение задачи кластеризации объектов на основе факторного анализа, что представляет, на взгляд авторов, научный и практический интерес.

Главы 5 – 6 посвящены изучению динамики социально-экономических процессов на основе статистического анализа временных рядов, который представляет собой самостоятельную, весьма обширную и одну из наиболее интенсивно развивающихся областей прикладной статистики. Статистическая обработка данных проводилась с помощью пакета прикладных программ по математической статистике STADIA [12] и MS Excel.

Студентам старших курсов целесообразно использовать пособие при подготовке выпускных квалификационных работ, связанных с вопросами прикладной статистики. Авторы полагают, что изучение материала пособия будет способствовать развитию математической культуры учащихся и принесет ощутимую пользу при дальнейшем их обучении в аспирантуре. В пособии используется терминология работы авторов [13]. Книга также призвана содействовать научным исследованиям магистрантов и аспирантов, получению новых результатов и дальнейшему их приложению к практическим задачам. Пособие ориентировано на подготовленного читателя в области курса теории вероятностей и математической статистики.

Глава 1. ВВЕДЕНИЕ В ПРИКЛАДНУЮ СТАТИСТИКУ

1.1. Представление и обработка статистической информации

Каждое статистическое исследование начинается со сбора и систематизации первоначальных сведений об отобранных единицах наблюдения. Эти сведения представляются в виде статистических показателей, рассчитанных в соответствии с правилами научно-организованного статистического наблюдения. *Статистический показатель* – это «количественная характеристика свойства изучаемого явления, относящаяся к конкретным условиям места и времени» [26, с. 19]. Чтобы охарактеризовать изучаемое явление с разных сторон, дать его комплексную оценку, на основании экспериментальных данных строят систему показателей, связанных определенной целью исследования. Таким образом, получение показателей, определяющих функционирование изучаемой сферы или её отдельного элемента, есть важный подготовительный этап научного исследования.

С развитием статистики множество применяемых показателей неизбежно расширяется. Ниже приведена классификация экономических показателей. Их можно определить как «обобщающие параметры, которые количественно оценивают существенные стороны функционирования или развития какого-либо предприятия или хозяйства» [5, с. 10]. Данные по показателям сводят воедино, при необходимости группируют, рассчитывают обобщающие показатели по всем единицам в целом или для отдельных групп, вычисляют средние значения, получают расчетные относительные показатели, позволяющие в дальнейшем сравнивать и анализировать объекты наблюдения. Итоговые значения статистических переменных обычно представляют в виде таблицы, придерживаясь определенных правил оформления [5, с. 72 – 79].

Согласно табличному представлению информации удобно сформировать массив исходных данных в виде матрицы для дальнейшей его обработки с помощью прикладных компьютерных программ. Все это составляет подготовительный этап творческой работы исследователя, аналог которой показан в работах [15 – 17]. От правильности выбора, достоверности показателей, качества их организации в систему зависит все дальнейшее математико-статистическое исследование, в частности выявление закономерности в развитии изучаемого явления, поиск адекватной математической модели, составление прогноза.

Классификация экономических показателей

Признак классификации	**Класс показателей**
Цель работы	Плановые Отчетные
Содержание	Количественные Качественные
Форма	Абсолютные Относительные
Число объектов	Средние Общие Групповые Индивидуальные
Натуральность показателей	Натуральные Стоимостные
Постоянство показателей	Статические Динамические
Полнота совокупности данных	Генеральные Выборочные
Сложность показателей	Простые Сложные

Приведенная классификация несколько условна, так как каждый из классов показателей по ходу исследования может быть подразделен на подклассы или отдельные виды.

На подготовительном этапе исследования особое место при анализе сложных систем занимает графическое представление информации. В данном случае под графиком мы понимаем схематическое изображение сведений. Графики прежде всего служат наглядным представлением экспериментальных данных, иллюстрируют структуру изучаемого явления, позволяют сравнивать одноименные показатели объектов исследования и анализировать динамику их изменения. Графики помогают определить характер взаимосвязи статистических показателей и остановиться на выборе функциональной зависимости при поиске структуры математической модели. В конечном счете график искомой математической модели, выполненный согласно ее уравнению в системе координат с определенными масштабными ориентирами, дает представление о характере изучаемого явления и позволяет его прогнозировать.

1.2. Сущность экономико-математического моделирования

Изучение связей и зависимостей нужно доводить до количественного выражения, т. е. в итоге они должны быть представлены в математической форме. Критерий завершенности исследования статистико-экономической связи – конкретное математическое выражение, моделирующее ее. В последнее время широкое распространение в научных исследованиях получило математическое моделирование.

Под *моделированием* понимают «воспроизведение или имитацию поведения какой-либо реально существующей системы на специально построенном ее аналоге или модели» [5, с. 18]. Моделирование основано на принципе аналогии и дает возможность исследовать объект, труднодоступный для изучения, не непосредственно, а с помощью рассмотрения другого, сходного с ним и более доступного объекта, в качестве которого и выступает построенная модель. *Модель* – это «логическое или математическое описание компонентов и функций, отображающих существенные свойства моделируемого объекта или процесса (обычно рассматриваемых как системы или элементы системы)» [18, с. 204]. Исходя из свойств модели, можно характеризовать особенности изучаемого объекта, но не все, а только те, которые аналогичны и в модели, и в объекте и при этом важны для исследования (эти свойства называют существенными).

Типы подобия между моделируемым объектом и его моделью [18, с. 204]:

- *физическое* – когда объект и модель имеют одинаковую или сходную физическую природу;
- *структурное* – при сходстве между структурой объекта и структурой модели;
- *функциональное* – с точки зрения выполнения объектом и моделью сходных функций при соответствующих воздействиях;
- *динамическое* – между последовательно изменяющимися состояниями объекта и модели;
- *вероятностное* – между процессом вероятностного характера в объекте и модели;
- *геометрическое* – между пространственными характеристиками объекта и модели.

В соответствии с вышеприведенной классификацией различают и виды моделей: материальные (или вещественные), знаковые (графические, аналитические). Даже словесное описание можно рассматривать в качестве модели объекта.

В управлении хозяйственными процессами важное значение имеет применение экономико-математических моделей. *Экономико-математическая модель* – это «концентрированное выражение в математической форме наиболее существенных взаимосвязей и закономерностей процесса формирования экономической системы» [5, с. 18].

Необходимость построения и применения математических моделей в экономике объясняется следующими причинами [6, с. 6 – 7]:

1. Процесс построения математической модели дает возможность систематизировать сложные взаимосвязанные факторы, выявить существенные и несущественные для изучения процесса параметры и связи.

2. Математическое описание позволяет для каждого исследуемого фактора найти соответствующий символ (переменную) и установить взаимосвязь и взаимовлияние одних переменных на другие.

3. В отличие от вербальной (или словесной), математическая модель позволяет описать процесс компактно, в виде набора математических соотношений, без учета несущественных деталей, и определить строгие правила поведения переменных, характеризующих процесс.

4. В ходе построения математической модели информация упорядочивается, определяется ее необходимый объем и степень значимости.

5. Математическая модель формирует основу для количественного анализа поведения объекта через проведение численного эксперимента, что дает возможность выяснить вероятные альтернативные сценарии поведения объекта и количественно оценить последствия, к которым приведет их реализация. Это в особенности важно при изучении систем, для которых невозможен натурный эксперимент; к таким системам относятся также социально-экономические явления и процессы. При этом значительно увеличивается число сценариев, которые удается исследовать.

6. Математическая модель позволяет качественно изучить поведение реального объекта. Построив модель, ее можно исследовать, пользуясь правилами математики, абстрагируясь при этом от реального объекта. Это абстрагирование может дать новые, нетривиальные знания о моделируемом объекте.

7. Математическое моделирование как инструмент оценки влияния на объект латентных (скрытых) факторов.

8. Создание математической модели служит основой для эффективного применения современных информационных технологий. Потребность в них объясняется сложностью расчета моделей, необходимостью сбора и анализа больших массивов данных, длительностью проведения вычислительных процессов.

Для объективной характеристики того или иного изучаемого явления в природе или обществе необходимо прежде всего раскрыть его связь с другими явлениями, выяснить происхождение, дать оценку настоящего положения и на этом основании спрогнозировать его будущее развитие. В этой связи предлагается следующий алгоритм статистического анализа и моделирования связей в многофакторной системе экспериментальных данных.

Этап I. Формирование системы статистических показателей

Шаг 1. Определение объектов исследования.

Шаг 2. Построение матрицы исходных данных.

Шаг 3. Расчет многомерного среднего показателя системы.

Шаг 4. Табличное и графическое представление информации.

Этап II. Построение аналитической группировки

Шаг 1. Расчет описательной статистики распределения многомерного среднего показателя *P*, %.

Шаг 2. Построение интервального вариационного ряда и гистограммы распределения *P*, %.

Шаг 3. Формирование результативного признака *Y* системы.

Шаг 4. Классификация объектов по многомерному среднему показателю *P*, %, и оценка их уровня развития.

Шаг 5. Оценка положения интересующего объекта в системе построенной группировки.

Этап III. Анализ влияния многомерного среднего на вариацию результативного показателя на основе моделирования функциональной зависимости Y = Y(P)

Шаг 1. Проведение дисперсионного анализа и обоснование аналитической группировки.

Шаг 2. Построение корреляционного поля (диаграммы рассеивания) точек (*P*, *Y*).

Шаг 3. Вычисление выборочных коэффициентов корреляции, корреляционных отношений и проверка статистической гипотезы значимости связи между P и Y.

Шаг 4. Расчет линии регрессии на основе аналитической группировки.

Шаг 5. Выбор наиболее качественной модели $Y = Y(P)$, адекватной экспериментальным данным.

Шаг 6. Анализ остатков по уравнению регрессии для интересующих объектов.

Шаг 7. Точечный прогноз по модели $Y = Y(P)$.

Этап IV. Оценка изолированного влияния факторных переменных {X_j} (j = 1, …, k) на результативный показатель Y системы

Шаг 1. Расчет статистических характеристик выборочного распределения $\{X_j\}$ ($j = 1, \ldots, k$).

Шаг 2. Оценка силы прямой линейной корреляции между факторами $\{X_j\}$ и результативным показателем Y.

Шаг 3. Построение моделей регрессии $Y = Y(X_j)$ ($j = 1, \ldots, k$).

Шаг 4. Графическое представление результатов моделирования.

Этап V. Построение моделей множественной регрессии

Шаг 1. Расчёт множественного коэффициента корреляции.

Шаг 2. Регрессионный анализ зависимости $Y = Y(X_1, \ldots, X_k)$.

Шаг 3. Выделение факторных переменных, значимо влияющих на вариацию результативного признака, и построение (на основе пошаговой регрессии) более сокращенной модели $Y = Y(X_1, \ldots, X_s)(s < k)$.

Шаг 4. Анализ остатков по уравнениям регрессии и точечный прогноз для интересующих объектов исследования.

Шаг 5. Практическая интерпретация результатов моделирования.

Пример реализации данного алгоритма исследования приводится в работе [17]. В ней строится система относительных показателей инновационной деятельности российских регионов и вычисляется многомерный средний показатель (обобщающий в процентах), комплексно характеризующий уровень инновационного потенциала регионов:

$$P_i = \left(\sum_{j=1}^{k} \left(X_{ij} \,/\, \bar{X}_j \right) 100\% \right) / k; \;\; i = 1, 2, \ldots, n,$$

где n – число регионов, участвующих в исследовании; k – количество факторных переменных; $\{X_{ij}\}$ – двумерный массив абсолютных значений фак-

торных переменных, где строки отвечают за объекты исследования – регионы, столбцы – за показатели инновационного развития; $\bar{X}_j = \left(\sum_{S=1}^{n} X_{Sj} \right) / n (j = 1, 2, \ldots, k)$ – выборочное среднее значение j-го показателя. Предполагается, что различия в значениях объемов результативного показателя инноваций Y обусловлены вариацией факторного признака P_i – многомерного среднего показателя инновационного потенциала. Так формируется n-мерный вектор $P = \{P_i\}$ – многомерный средний показатель изучаемой многофакторной системы экспериментальных данных. По ходу изложения материала пособия приводится краткая характеристика и обоснование каждого этапа работы с данными, также формулируются основные понятия и определения, которыми рекомендуется пользоваться при исследовании.

Отметим, что в процессе построения моделей следует избегать ряда типичных недостатков: включения в модель несущественных переменных, либо не включения существенных; недостаточно точной оценки параметров модели; неправильного определения функциональной зависимости и т. п. Для построения более точной и подробной модели может потребоваться ее усложнение, что не всегда оправдывается возросшими трудностями расчетов. С другой стороны, при упрощении модели иногда имеет место снижение ее достоверности.

1.3. Описательная статистика выборочных распределений

При статистическом изучении случайной величины строят вариационный ряд её выборочных значений или, по-другому, ряд распределения. Согласно [8, с. 81] «вариацией значений какого-либо признака в совокупности называется различие его значений у разных единиц данной совокупности в один и тот же период или момент времени». Вариационный ряд имеет следующие формы: ранжированный ряд, дискретный ряд, интервальный ряд. *Ранжированный ряд* есть перечень единиц совокупности в порядке возрастания или убывания их значений: (x_1, x_2, …, x_n). *Дискретный ряд* распределения задают в виде таблицы, где участвуют выборочные числовые значения признака x_i и f_i – частота появления i-го значения. *Интервальный вариационный ряд* также представляет собой таблицу, состоящую, как правило, из двух граф (или строк) – интервалов исследуемого признака X и числа выборочных значений признака, попадающих в данный интервал, т. е.

частот. Количество групп k в вариационном ряду определяют, следуя формуле Стержесса [8, с. 85]: $k \approx 1{,}44 \cdot \ln(n) + 1$, где n – численность единиц наблюдения. При этом длина интервалов в ряду равна $l = (x_{max} - x_{min}) / k$. Интервальный ряд графически изображается в виде гистограммы – множества прямоугольников (столбиков), основания которых соответствуют интервалам варьирующего признака, а высоты пропорциональны частотам. Встречаются случаи, когда применение формулы Стержесса приводит к появлению «пустых» интервальных групп. Тогда для группировки могут использоваться неравные по длине интервалы, причем их границы задаются самим исследователем исходя из целей изучения варьирующего признака.

Ниже определяются основные числовые характеристики выборочных распределений, что составляет описательную статистику вариационного ряда. *Выборочное среднее значение* $\bar{x}$ вычисляют как сумму всех значений наблюдаемого признака X, деленную на их число – n. Число n называют *объемом выборки*. Для измерения величины вариации используют следующие абсолютные показатели: размах вариации; дисперсию; среднее квадратичное отклонение; среднее линейное отклонение. *Размах вариации* $R = X_{max} - X_{min}$ характеризует лишь максимальное различие значений признака. Показателем силы вариации выступает *среднее линейное отклонение*, или *средний модуль отклонений*: $a = \sum_{i=1}^{n} |x_i - \bar{x}| / n$. Но средний модуль отклонений нельзя поставить в соответствие с каким-либо вероятностным законом распределения, параметром которого является *среднее квадратичное отклонение* $\sigma = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / n}$; для интервального ряда $\sigma = \sqrt{\left(\sum_{j=1}^{k} (x_j - \bar{x}_j)^2 f_j\right) / \left(\sum_{j=1}^{k} f_j\right)}$, где k – количество групп в интервальном ряду, $f_1, \ldots, f_k$ – частоты значений $x_1, \ldots, x_k$. Отношение ($\sigma : a$) зависит от присутствия в выборке резких, выделяющихся отклонений. Чем больше величина ($\sigma : a$), тем сильнее «засоренность» выборки. Для нормального закона распределения [8, с. 94] значение ($\sigma : a$) $\approx 1{,}2$. *Выборочной дисперсией* называют число $D_в = \sigma^2 = \overline{x^2} - \bar{x}^2$. Понятие дисперсии используют многие методы статистики. При этом большое значение имеет правило сложения дисперсий [8, гл. 6]. В случае малых выборок ($n \leq 30$) вычисляют несмещенную оценку дисперсии – исправленную выборочную дисперсию: $D_{исп} = \frac{n}{n-1} D_в$.

При больши́х n $D_{\text{исп}}$ и $D_{\text{в}}$ практически совпадают. Дисперсия характеризует меру разброса выборочных значений около среднего значения признака. В отличие от дисперсии среднее квадратичное отклонение имеет те же единицы измерения, что и исходные значения (x_1, x_2, …, x_n). Величина σ – показатель степени однородности данных признака X. При малых выборках за среднее квадратичное отклонение (стандартное отклонение) принимают число

$$s = \sqrt{D_{\text{исп}}} = \sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \Big/ \left(n-1\right)}.$$

Медиана Me есть число, которое делит выборку пополам: количество наблюдаемых значений, меньших *Me*, равно количеству значений, больших *Me*. Для изучения характера вариации используют показатели *асимметрии* A_s и *эксцесса* E_x: $A_s = \left(\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3 \Big/ n\right) \Big/ \sigma^3$;

$E_x = \left(\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4 \Big/ n\right) \Big/ \sigma^4 - 3$ [8, с. 97 – 98].

Для вариационного ряда с нормальным распределением $A_s = 0$ и $E_x = 0$, поэтому в некоторых прикладных компьютерных программах по каждому из коэффициентов асимметрии и эксцесса определяют уровни значимости P нулевой гипотезы об отсутствии различий выборочного распределения от нормального, проверку выполняют по нормально распределенным статистикам [12, с. 129 – 131]. Если оказывается, что P больше 0,05, нет оснований отвергать нулевую гипотезу. Такой способ проверки нормальности распределения по оценкам A_s и E_x удобно применять в случае малых выборок. Однако большей мощностью для выборок длины $n \leq 50$ обладает критерий Шапиро – Уилка, который базируется на отношении оптимальной линейной несмещённой оценки дисперсии к её обычной оценке. Статистика критерия имеет вид [11, с. 238]

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{k} a_{n-i+1}\left(x_{n-i+1} - x_i\right)\right]^2;\ \ s^2 = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2;\ \ k = \left[\frac{n}{2}\right].$$

Числитель представляет собой квадрат оценки среднеквадратического отклонения Ллойда. Коэффициенты a_{n-i+1} приведены в специальной таблице [11, с. 237 – 238]. Критические значения статистики $W(\alpha)$ приведены в [11, с. 240]. Если $W < W(\alpha)$, то нулевая гипотеза нормальности распределения отклоняется на уровне значимости α. Для выборок большого объема ($n > 50$) обычно пользуются специаль-

ными критериями, в частности *D*-критерием Колмогорова, критериями «хи-квадрат, χ^2» и «омега-квадрат, ω^2», [12, с. 135 – 138].

После проведения выборки рассчитывают также возможные ошибки выборочных показателей (ошибки репрезентативности), которые используют для оценки результатов выборки или получения основных характеристик генеральной совокупности. *Ошибка репрезентативности* – это разность между значением выборочного показателя и генеральным параметром μ (математическим ожиданием) [8, гл. 6]: $\Delta = \overline{x} - \mu = ts_{\overline{x}}$, где $s_{\overline{x}} = s/\sqrt{n}$. При малых выборках значение t находят по таблице распределения Стьюдента с $df = (n - 1)$ степенями свободы. Вероятность, которая принимается при расчете ошибки выборочной характеристики, называется *доверительной вероятностью*. Величина Δ указывает на допустимый размер погрешности оцениваемого показателя с определенной доверительной вероятностью.

Ниже приводится пример расчёта (на основе статистического пакета STADIA) описательной статистики показателя $X = (x_1, x_2, \ldots, x_{18})$ площади территорий регионов ЦФО РФ согласно данным Росстата за 2015 год [25, с. 18], приведённым в табл. 1.

Таблица 1. Исходные данные

Регион	**Площадь территории *X*, тыс. км^2**
1. Белгородская область	27,1
2. Брянская область	34,9
3. Владимирская область	29,1
4. Воронежская область	52,2
5. Ивановская область	21,4
6. Калужская область	29,8
7. Костромская область	60,2
8. Курская область	30,0
9. Липецкая область	24,0
10. Московская область	44,3
11. Орловская область	24,7
12. Рязанская область	39,6
13. Смоленская область	49,8
14. Тамбовская область	34,5
15. Тверская область	84,2
16. Тульская область	25,7
17. Ярославская область	36,2
18. г. Москва	2,6

ФАЙЛ. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Переменная *X*

Размер	<Диапазон>	Среднее	Ошибка	Дисперсия	Ст.откл.	Сумма
18	2,6 – 84,2	36,13	4,183	315	17,75	650,3
Медиана	**<Квартили>**	**Асимметрия**	**Значимость**	**Эксцесс**	**Значимость**	
32,25	25,45 – 45,67	0,93	0,0296	1,49	0,0095	

Таким образом, по данной выборке объёма n = 18 среднее значение площади регионов ЦФО составляет $\bar{x}$ = 36,13 тыс. км^2. Размах 95 % доверительного интервала среднего составил Δ = 8,721 тыс. км^2. Абсолютные показатели вариации следующие: размах вариации R = 81,6 тыс. км^2; дисперсии $D_в = \sigma^2$ = 297,5 и $D_{исп}$ = 315,0; стандартное отклонение s = 17,75 тыс. км^2; среднее линейное отклонение a = 12,62 тыс. км^2.

На этой основе вычислены относительные показатели вариации: коэффициент вариации $V_\sigma = \frac{\sigma}{\bar{x}} 100\ \%$ = 47,7 %; коэффициент осцилляции $V_\sigma = \frac{R}{\bar{x}} 100\ \%$ = 225,8 %; относительное линейное отклонение (линейный коэффициент вариации) $V_a = \frac{a}{\bar{x}} 100\ \%$ = 34,9 %.

По ранжированному вариационному ряду определены структурные средние. Медиана Me = 32,25 тыс. км^2 показывает, что половина регионов ЦФО имеет площадь территории меньше, чем 32,25 тыс. км^2, а остальные регионы – больше, чем медиана. Нижний квартиль Q_1 = 25,45 тыс. км^2 и верхний квартиль Q_4 = 45,67 тыс. км^2 показывают, что четверть регионов имеют площадь меньшую, чем 25,45 тыс. км^2, а другая четверть – большую, чем 45,67 тыс. км^2.

Показатели формы распределения имеют положительные значения: асимметрия A_s = 0,93 и эксцесс E_x = 1,49, что указывает на правостороннюю асимметрию и островершинный вид кривой распределения. Значение эксцесса значимо отлично от нуля, на этом основании отвергаем гипотезу о нормальном характере распределения исследуемой величины *X*.

Ниже по данной выборке построен интервальный вариационный ряд, где указаны левая граница интервалов в тыс. км^2 и единицах стандартного отклонения. На рис. 1 приведена соответствующая гистограмма распределения *X* с наложенной кривой нормального распределения.

Данные программы:

X-лев.	*X*-станд	Частота	Част., %	Накопл.	Накопл., %
2,6	–1,889	1	5,56	1	5,556
18,92	–0,970	10	55,56	11	61,11
35,24	–0,050	4	22,22	15	83,33
51,56	0,870	2	11,11	17	94,44
67,88	1,789	1	5,56	18	100
84,20	2,709				

Рис. 1. Гистограмма распределения *X*

1.4. Краткая характеристика многомерных методов прикладной статистики

Для решения типовых задач математико-статистической обработки данных созданы и используются разнообразные прикладные компьютерные программы, которые в руках умелого пользователя становятся мощным инструментом исследования. При этом знание математической статистики совершенно необходимо. В то же время не стоит полагать, что очень легко овладеть статистическими методами настолько, чтобы свободно пользоваться ими на практике, даже принимая все необходимые меры предосторожности во избежание ошибок. Математическая статистика – это наука, которая становится еще и искусством, когда ее применяют не в абстрактных условиях чистой математики, а в системной практике реальной действительности.

Корреляционно-регрессионный анализ базируется на математической теории корреляции. Сам термин «корреляция» (от позднелатин-

ского correlatio – соотношение) означает вероятностную (статистическую) зависимость между величинами, не имеющую строго функционального характера. В отличие от функциональной, корреляционная зависимость в сложной многофакторной системе возникает тогда, когда один из исследуемых показателей зависит не только от данного второго, но и от других случайных факторов или когда среди условий исследуемой сферы, от которых зависят эти показатели, имеются общие для них обоих условия.

Основная задача корреляционного анализа состоит в выявлении взаимосвязи между случайными переменными на основе вычисления выборочных коэффициентов корреляции или корреляционных отношений и проверки статистической гипотезы значимости связи. Измерение тесноты связи между тремя и более переменными изучается путем вычисления и проверки значимости частных и множественных коэффициентов корреляции и, соответственно, коэффициентов детерминации. Кроме того, с помощью корреляционного анализа в многофакторных системах решается задача отбора факторов, оказывающих наиболее существенное влияние на результативный признак. Дальнейшее исследование состоит в поиске конкретного вида зависимости между изучаемыми величинами, т. е. в построении модели регрессии в виде математической функции от независимых переменных: $Y = f(X_1, X_2, \ldots, X_n)$.

Термин «регрессия» (от лат. regression – движение назад) был введен английским ученым Ф. Гальтоном. В математической статистике под регрессией понимают зависимость среднего значения какой-либо случайной величины от некоторой другой величины или от нескольких. Если, например, при каждом значении $X = x_i$ наблюдается n_i значений $y_{i1}, \ldots, y_{in_i}$ случайной величины Y, то зависимость средних арифметических $\overline{y_i} = \left(y_{i1} + \ldots + y_{in_i}\right)/n_i$ этих значений от x_i и является регрессией в статистическом понимании этого термина. В прикладной статистике ставится задача приближенной оценки регрессии. Решением этой задачи может служить выбор из всех функций $f(x)$ заданного класса такой функции, при которой достигается минимум математического ожидания $M\{(Y - f(x))^2\}$. Такая функция $y(x)$ называется *средней квадратической регрессией*. Уравнение $y = y(x)$ называется *уравнением регрессии*, а соответствующий график – *линией регрессии*, здесь функция $y(x)$ и есть регрессия случайной величины Y по X.

При практическом изучении связей в многофакторной системе оба метода, корреляционный и регрессионный анализ, как правило, применяются комплексно, т. е. объединяются в один вид исследования – корреляционно-регрессионный анализ. При этом следует обратить внимание на ряд предпосылок корреляционно-регрессионного анализа, выполнение которых необходимо для того, чтобы свойства полученных математических моделей были научно обоснованы и в конечном счете обеспечили прикладную полезность решения реальных задач.

Важную роль в статистическом анализе играют аналитические группировки, служащие для изучения связей между признаками. При этом один из признаков принимают за результативный, а другие – за факторные. Результативный признак меняется под воздействием вариации факторных признаков. Связь между признаками называется прямой, если с увеличением значений факторного признака возрастают значения результативного признака. В противном случае связь называется обратной. При построении аналитической группировки в качестве группировочного признака выбирают факторный признак, а значения результативного показателя вычисляют как среднее арифметическое значений данной группы. Таким образом, с помощью аналитической группировки выявляют наличие связи между признаками и ее направление.

В отличие от анализа однофакторных аналитических группировок, при исследовании сложных систем построение групп осуществляют на основе множества факторов и в дальнейшем изучают связи между признаками в однородных группах объектов исследования.

Кроме методов многомерных аналитических группировок для исследования различных форм ассоциаций (сходства, близости) объектов по ряду показателей применяют методы кластерного анализа [12, гл. 11]. Здесь все факторы выступают как равноправные, без деления на зависимые и независимые. Результаты кластеризации (разбиения объектов на группы) можно использовать в дальнейшем для изучения связей между признаками в отдельных группах сходных объектов. Объединение объектов в кластеры осуществляется на основе вычисления расстояния d между парами объектов в многомерном пространстве, размерность которого равна количеству факторных переменных. В зависимости от типа исходных данных и выбранной стратегии объединения вводится расстояние по определенным фор-

мулам. В частности, евклидовы расстояния $d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2}$ между i-м и j-м объектами в пространстве m факторных переменных применимы для факторов, измеренных в одних и тех же единицах. Нормированное евклидово расстояние $d_{ij} = \sqrt{\sum_{k=1}^{m} \frac{\left(x_{ik} - x_{jk}\right)^2}{s_k^2}}$ вводится в случае факторных переменных, значительно различающихся по величине и измеренных в разных единицах для отдельных факторов. Поэтому здесь для нормировки используют средние квадратические отклонения исходных значений факторов s_k (k = 1, 2, …, m).

Объединение изучаемых объектов в заданное количество групп (или кластеров), выполненное на основе одного или множества качественных факторных переменных, полезно обосновать с помощью дисперсионного анализа. В частности, задача однофакторного дисперсионного анализа состоит в оценке отношения средней межфакторной дисперсии, характеризующей непосредственно влияние факторной переменной на отклики, к средней внутрифакторной дисперсии (или остаточной дисперсии), учитывающей случайный разброс откликов в пределах каждой группы относительно своей групповой средней. Таким образом, проверяется статистическая значимость различий между средними значениями результативного признака-отклика Y в выделенных группах. Исходные данные представляются в виде матрицы $\{y_{ij}\}$ (i = 1, …, n_j; j = 1, …, k), где столбцы отвечают за значение результативного признака по каждой группе объектов, соответствующей определенному уровню факторного признака P. Испытуемая гипотеза H_0 формулируется так: средние значения для групп откликов, измеренных при различных значениях факторного признака, не имеют существенных различий между собой. Значимость таких различий проверяется на основе параметрического F-критерия [12, с. 177 – 179]: $F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}$. Здесь $S^2_{\text{факт}} = \frac{D_{\text{факт}}}{k-1}$, $S^2_{\text{ост}} = \frac{D_{\text{ост}}}{n-k}$, $S^2_{\text{общ}} = \frac{D_{\text{общ}}}{n-1}$ – три оценки генеральной дисперсии, пропорциональные степеням свободы, вычисленные на основе

$D_{\text{общ}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{ij} - \overline{y}\right)^2$ – полной вариации;

$D_{\text{факт}} = \sum_{j=1}^{k} \left(\bar{y}_j - \bar{y} \right)^2 n_j$ – межгрупповой факторной вариации;

$D_{\text{ост}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y}_j \right)^2$ – внутригрупповой (остаточной) вариации,

где y_{ij} – значение признака Y i-го региона в j-й группе; n_j – число объектов в группе; $\bar{y}_j$ – среднее значение признака y в j-й группе; $\bar{y}$ – общее среднее значение; k – количество групп; n – общее число объектов, равное объёму выборки значений результативного признака Y.

Согласно вышеприведенным формулам строится табл. 2.

Таблица 2. Дисперсионная таблица

Компоненты вариации	**Сумма квадратов отклонений**	**Число степеней свободы**	**Средний квадрат отклонений**
Между группами	$D_{\text{факт}}$	$k-1$	$S_1^2 = D_{\text{факт}} / (k-1)$
Внутри групп	$D_{\text{ост}}$	$n-k$	$S_2^2 = D_{\text{ост}} / (n-k)$
Полная вариация	$D_{\text{общ}}$	$n-1$	$S^2 = D_{\text{общ}} / (n-1)$

Вычисляется расчетное F-отношение $F_{\text{факт}} = S_1^2 / S_2^2$ и сравнивается с табличным значением с уровнем значимости α. Если $F_{\text{факт}} > F_{\text{табл}}(\alpha, k-1, n-k)$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1: есть влияние фактора на отклик, тогда мы считаем, что влияние признака-фактора P на Y статистически значимо. Так, в примере ниже (см. п. 2.6) предполагается, что различия в значениях объемов инновационных товаров, работ и услуг Y обусловлено различиями факторного признака P – многомерного среднего показателя инновационной деятельности регионов. Статистическая значимость таких различий устанавливается на основе дисперсионного анализа.

В случае принятия гипотезы H_1 представляет интерес более детальный анализ на наличие различий между конкретными уровнями факторного признака P или группами уровней с помощью парных сравнений Шеффе. При этом для всех пар групп вычисляется разность средних значений результативного показателя Y, размах доверительного интервала разности, уровень значимости нулевой гипотезы об отсутствии различий между средними значениями [12, с. 178 – 179]. В дальнейшем можно провести более концентрированный анализ на наличие и возрастание факторного эффекта с помощью рангового

(непараметрического) критерия Джонкхиера [12, с. 181 – 182]. Здесь альтернативная гипотеза формулируется так: различия сдвига существуют и медианы выборки упорядочены по возрастанию. Таким образом, выявляется значимое влияние уровня *P* на объемы результативного показателя *Y*.

Факторный анализ [12, с. 315 – 350] – наиболее значимый и часто используемый из многомерных методов, а методы кластеризации и дискриминантного анализа дают наглядные результаты, если их применяют к данным, уже прошедшим процедуру факторизации. Переменные, значения которых можно измерять в эксперименте, имеют для исследуемого объекта или явления нередко достаточно условный характер, лишь опосредованно отражая его внутреннюю структуру, движущие силы (механизмы) или действующие на объект факторы. Здесь исследователь ограничен субъективными мнениями опрашиваемых или набором физических явлений, или показателей, доступных для экспериментальных измерений. На первом этапе исследования основная задача факторного анализа – нахождение сокращенной системы существенных факторов в пространстве регистрируемых переменных.

Важная последующая задача – исследование распределений объектов в их проекциях на плоскости главных факторов (в новых факторных координатах объектов), где особенности группировок объектов значительно более явно проявляются по сравнению с их проекциями на плоскости исходных переменных. Здесь наряду с визуальным анализом естественным продолжением становится применение методов кластерного, дискриминантного анализа, а также статистического анализа временных рядов.

Временной ряд – это последовательность упорядоченных по времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления. Важнейшая классическая задача при исследовании временных рядов – выявление и статистическая оценка основной тенденции изучаемого процесса, прогнозирование его будущего развития. Среди наиболее распространенных методов исследования временных рядов выделяют корреляционно-регрессионный и спектральный анализы, построение моделей авторегрессии и скользящей средней [7 – 9; 11 – 12]; получение адаптивных моделей в целях краткосрочного прогнозирования [1; 16]; построение тренд-сезонных

моделей [22]. В прикладной статистике выделяют следующие основные этапы анализа временных рядов [12, с. 222]:

- графическое представление и описание поведения временного ряда;
- расчёт показателей динамики временных рядов;
- исследование структуры временного ряда и выделение его закономерных составляющих (тренда, сезонных и циклических составляющих);
- построение математической модели процесса, представленного временным рядом;
- исследование случайной составляющей временного ряда и оценка качества математической модели;
- прогнозирование будущего развития случайного процесса на основе имеющегося временного ряда;
- преобразование временного ряда средствами сглаживания и фильтрации и исследование преобразованных рядов;
- изучение взаимосвязанных динамических рядов.

Контрольные вопросы

1. Какие основные этапы статистического изучения многофакторных систем экспериментальных данных вы можете выделить?
2. Какие многомерные методы математической статистики могут быть использованы на данных этапах исследования?
3. Какие числовые характеристики включает описательная статистика выборочных распределений?
4. Как построить интервальный вариационный ряд и гистограмму распределения?
5. Для чего служат аналитические группировки?
6. Каковы цели и задачи корреляционно-регрессионного анализа?
7. Для чего используют методы кластерного анализа?
8. Каково предназначение факторного анализа?
9. Какова цель дисперсионного анализа? Как строится расчетная дисперсионная таблица?
10. Укажите основные этапы статистического анализа временных рядов.

Глава 2. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. О всеобщей связи явлений

Статистическая корреляция в системе показателей экспериментальных данных представляет интерес, так как она указывает на наличие закономерной связи между рассматриваемыми факторами исследуемой сферы.

В мире все находится в диалектической связи. Явления природы и общества существуют не обособленно друг от друга, а пребывают во взаимных связях и зависимостях. *Связь* – это «соотношение причинности, обусловленности или общности между явлениями» [5, с. 5]. Существует всеобщая связь явлений, которую можно рассматривать в качестве наиболее общей закономерности окружающей действительности, представляющей собой результат универсального взаимодействия всех вещей и объектов. Посредством всеобщей связи вещей проявляется детерминированность или обусловленность явлений какими-либо материальными процессами, обнаруживаются единство и развитие материального мира. Под *зависимостью* понимается «отношение одного явления к другому как следствия к причине» [5, с. 5]. Ввиду того, что имеет место всеобщая связь явлений, окружающий нас мир представляет собой не хаотичное нагромождение вещей, а закономерный процесс движения и развития материи.

С точки зрения философии источник движения и развития внутренне присущ любой материальной системе как совокупности взаимосвязанных вещей. Развитие есть следствие единства и борьбы противоположностей, результат взаимного перехода количественных и качественных изменений, итог отрицания старого и отжившего и замены его новым и более прогрессивным.

Говоря о всеобщем характере связи явлений, нужно отметить, что эту связь не следует принимать упрощенно. Нельзя полагать, что это постоянная связь и взаимодействие какой-либо частицы Вселенной с любой другой, сколь угодно удаленной в пространстве и времени от первой. Бесконечно далекие и мгновенные связи практически невозможны либо крайне несущественны. Они исключаются как конечной скоростью распространения материальных взаимодействий, так и ограниченностью времени существования конкретного объекта, так как происходит постоянное превращение одних форм материи в другие.

Наряду с постоянной связью каждого объекта с ближним и далеким окружением существует также и относительная их независимость. Последнее характерно для сильно разделенных во времени и пространстве объектов, особенно если их существование относится к различным временным периодам, т. е. когда какие-либо объекты физически не могут взаимодействовать между собой.

Поскольку всеобщая связь явлений носит закономерный характер, во всех материальных системах существуют определенный объективный порядок и естественная последовательность событий. Следствие общего действия множества законов, отражающее совокупность связей и отношений, – *закономерность*. Если закон однозначно характеризует определенную связь, то понятие «закономерность» по своему интервалу намного шире понятия «закон». *Закон* можно понимать как «существенную и устойчивую связь явлений, которая обусловливает их упорядоченное изменение» [5, с. 6]. Зная закон, можно практически досконально точно предвидеть, спрогнозировать характер протекания исследуемого процесса или его результата. Для объективной характеристики того или иного изучаемого явления необходимо прежде всего раскрыть его связи с другими явлениями, выяснить происхождение, дать оценку его настоящему положению и развитию в будущем. Различные виды связей представлены в табл. 3.

Таблица 3. Виды связей

Классификационный признак	**Вид связей**
Происхождение	Естественные Искусственные
Обязательность существования	Необходимые Случайные
Опосредование связи	Прямые Опосредованные
Сложность связи	Простые Сложные
По времени	Постоянные Прерывные
Сила (теснота) связи	Сильные Слабые
Существенность влияния	Существенные Несущественные

Окончание табл. 3

Классификационный признак	**Вид связей**
Направление связи	Однонаправленные Двунаправленные
Соединение связей	Последовательные Параллельные
Удаленность в пространстве	Далекие Близкие
Отношение к системе	Внутренние Внешние
Своевременность связи	Своевременные Опаздывающие (опережающие)
Жесткость	Жесткие Гибкие
Форма	Вещественные Энергетические Информационные
Характер	Функциональные Корреляционные
Аналитическое выражение	Линейные Нелинейные

Так как каждое явление окружающего мира обусловлено множеством каких-либо других явлений, то в природе нет абсолютно изолированных объектов. Конечно, исключение может составлять временная искусственная изоляция. Но и в этом случае речь можно вести, по-видимому, не об абсолютной, а скорее об относительной изоляции. Если явление отделяется от его естественных связей, оно, как правило, превращается в нечто иррациональное, необъяснимое и бесполезное.

Как и все естественные явления, общественные также находятся в динамических связях. Научная и практическая необходимость познания этих связей требует исследования и влияния одного отдельно взятого явления на другое и совместного влияния множества явлений. Изучение влияния связей и взаимосвязей явлений – это одна из основных задач научных исследований, в том числе экономических. Поскольку все явления взаимосвязаны, взаимодействуют и взаимно обусловливают друг друга, это должно быть учтено и в методах их исследования. Исследование должно предполагать изучение объекта

со всех сторон, в его связях и отношениях с другими вещами с учетом окружающих условий, времени и места.

Одной из форм классификации связей между различными явлениями и их признаками (см. табл. 4) выступает их разделение на два вида: *функциональную,* или жестко детерминированную, с одной стороны, и *статистическую,* или стохастически детерминированную – с другой. Строго охарактеризовать отличия этих типов связи можно в том случае, если они получают математическую формулировку. «Если с изменением значения одной из переменных вторая изменяется строго определенным образом, т. е. значению одной переменной обязательно соответствует одно или несколько точно заданных значений другой переменной, связь между ними является функциональной» [8, с. 190].

Функциональная связь двух величин возникает только при условии, что вторая из них зависит от первой и ни от чего больше. В реальности таких связей не существует; они представляют собой лишь абстракции, полезные и необходимые для анализа явлений, но упрощающие реальную действительность. Функциональная зависимость конкретной величины y от многих факторов $x_1, x_2, \ldots, x_k$ возможна исключительно тогда, когда величина y зависит только от указанного набора факторов $x_1, x_2, \ldots, x_k$ и ни от чего больше. Однако же все явления и процессы неограниченно реального мира связаны между собой, и отсутствует какое-либо конечное число переменных k, абсолютно полно определяющих зависимую величину y. Это свидетельствует о том, что множественная функциональная зависимость переменных тоже представляет собой лишь упрощающую реальность абстракцию.

Тем не менее физика, химия, биология, экономика и другие науки с успехом пользуются представлением связей как функциональных не только в аналитических целях, но нередко и в целях прогнозирования. Это возможно благодаря тому, что в несложных системах интересующая нас переменная величина в основном зависит от немногих других переменных или только от одной. Иначе говоря, связь в подобной простой системе хотя и не абсолютно функциональна, но практически очень близка к таковой.

Стохастически детерминированная связь не подвержена ограничениям и условиям, свойственным функциональной связи: «Если с изменением значения одной из переменных вторая может в определен-

ных переделах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические (массовые) характеристики изменяются по определенному закону – связь является статистической» [8, с. 191]. Иными словами, в случае статистической связи разным значениям одной переменной соответствуют разные распределения значений другой.

Любые связи, которые можно измерить и выразить численно, подходят под определение «статистические связи», в том числе и функциональные. Функциональные связи можно представить как частный случай статистических, когда значениям одной переменной соответствуют «распределения» значений второй, состоящие из одного или нескольких значений и обладающие вероятностью, равной единице. Безусловно, качественное отличие действительно вероятностных распределений и отдельных значений, имеющих вероятность единицы (т. е. достоверных), настолько велико, что хотя функциональные связи и подпадают в широком плане под определение статистической связи, есть все основания говорить о двух видах связей.

2.2. Системный подход в исследовании связей

Исследование связей в экономике будет результативным лишь при системном подходе в исследованиях. Под *системным подходом* понимают «методологическое направление в науке, основная задача которого состоит в разработке методов исследования и конструирования сложноорганизованных объектов или систем» [5, с. 16]. Системный подход подразумевает совокупность методов познания, методов исследования и конструирования, способов описания и объяснения природы изучаемых естественных или искусственно создаваемых объектов.

Применяя системный подход, любую сложную проблему важно рассматривать как нечто целое, как систему во взаимодействии всех ее структурных элементов. *Система* – это «множество элементов, находящихся в отношениях и связях друг с другом, которое образует определенную целостность, единство» [18, с. 323]. Каждый элемент системы можно расчленить на ряд составляющих элементов более низкого порядка: любая система имеет иерархическую структуру. Си-

стемы низшего уровня служат подсистемами систем более высокого уровня, те же, в свою очередь, относятся к подсистемам в системах еще более высокого порядка и т. д.

Основной признак системы заключается в том, что ее составные элементы образуют во взаимодействии единое целое с качественно новыми свойствами. Особое значение имеют такие виды связей, как вещественные, энергетические и информационные. Степень расчлененности систем в процессе анализа относительна и определяется задачей исследования.

Использование системного подхода подразумевает учет всех изменений, происходящих в системе под воздействием трансформации ее элементов. При исследовании обращают внимание и на системное окружение, или среду, т. е. ту объективную реальность, в пределах которой развивается система. Изменение состояния среды не может не влиять на развитие системы. Окончательный результат оценивается по эффективности всей системы целом. Изучая связи в экономике, можно выделить множество различных систем. Конечно же, не все возможные системы равноценны. Ряд из них недостаточен, а другой, напротив, избыточен при проведении исследования. Проблему выбора конкретной системы решает индивидуально сам исследователь.

В процессе изучения статистико-экономических связей используют разнообразные системы: простые и сложные, детерминированные и стохастические, статические и динамические, естественные и искусственные, открытые и закрытые. Системы, применяемые для изучения статистико-экономических связей, имеют ряд свойств. Основные из них – объективность, расчлененность, связность, конструктивность, организованность, целостность, динамичность, целесообразность, адаптивность, экономичность. Разумеется, учитываются те свойства, которые необходимы исходя из задачи исследования.

Чтобы принять оптимальное решение по управлению системой, следует провести анализ ее развития, определить цели ее функционирования, цели ее подсистем и альтернативы достижения этих целей. *Системный анализ* – это «методология аналитических исследований сложных вопросов и проблем на основе формирования целенаправленных структурированных систем» [5, с. 18]. Основная задача си-

стемного анализа – превращение сложной проблемы в ряд простых решаемых вопросов.

Как и во многих других, в экономических системах имеются такие элементы, как входы, ограничения, процессы, прямые и обратные связи, цели, выходы. Количественная определенность взаимодействия явлений и процессов дает возможность их математической оптимизации. Применение системного подхода позволяет комплексно решить поставленную задачу, предусмотреть достижение намеченной цели при наиболее эффективном использовании находящихся в распоряжении ресурсов. Для достижения максимального эффекта системный подход требует улучшения всего хозяйственного механизма.

2.3. Статистическое изучение корреляции

Определение наличия корреляционной зависимости между случайными величинами, установление ее направления и количественная оценка тесноты связи – основные задачи корреляционного анализа. Сам термин «корреляция» означает вероятностную (статистическую) зависимость между величинами, не имеющую строго функционального характера.

Рассмотрим подробнее методику корреляционного анализа. На первых шагах исследования строится *корреляционное поле* (или диаграмма рассеивания), что символизирует наблюдение и представляет собой совокупность точек на координатной плоскости, соответствующих двумерному выборочному распределению (X, Y). При большом объеме наблюдений данные обычно группируют в форме корреляционной таблицы, в клетках которой на пересечениях строк и столбцов указывают число случаев (часто́ты), в которых определенные уровни независимого фактора X сочетаются со значениями результативного показателя Y. Визуальный анализ таблицы позволяет исследователю выявить аномальные явления и при необходимости исключить их из дальнейшего рассмотрения.

По характеру расположения точек диаграммы рассеивания получают первоначальное представление о форме, силе и направлении связи между наблюдаемыми величинами. По форме корреляционная связь между варьирующими признаками X и Y может быть линейной или нелинейной, по направлению – прямой или обратной (рис. 2 – 5).

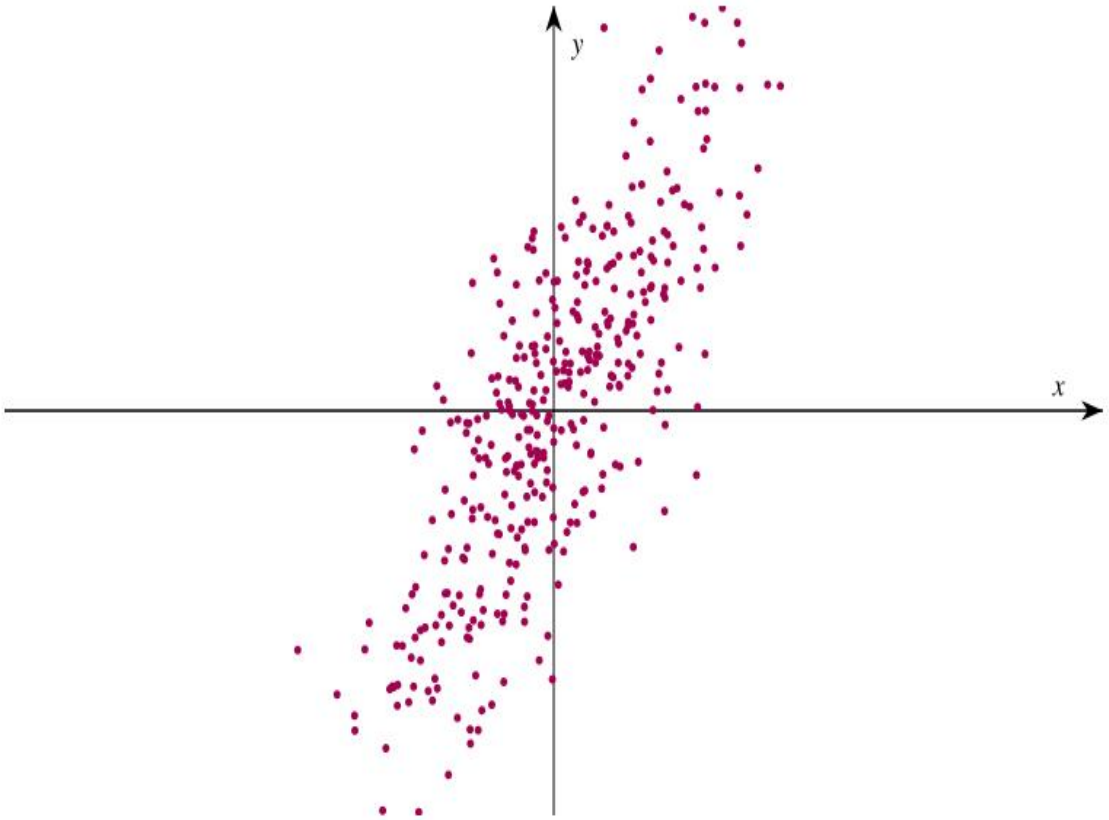

Рис. 2. Наличие прямой линейной взаимосвязи между *X* и *Y*

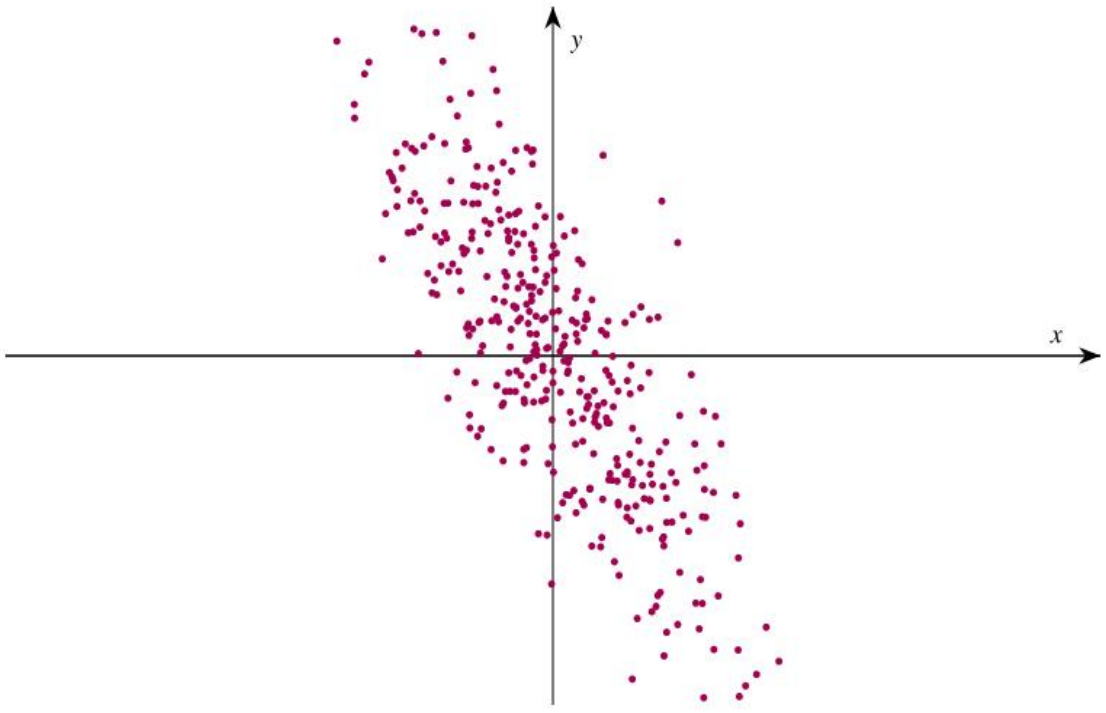

Рис. 3. Наличие обратной линейной взаимосвязи *X* и *Y*

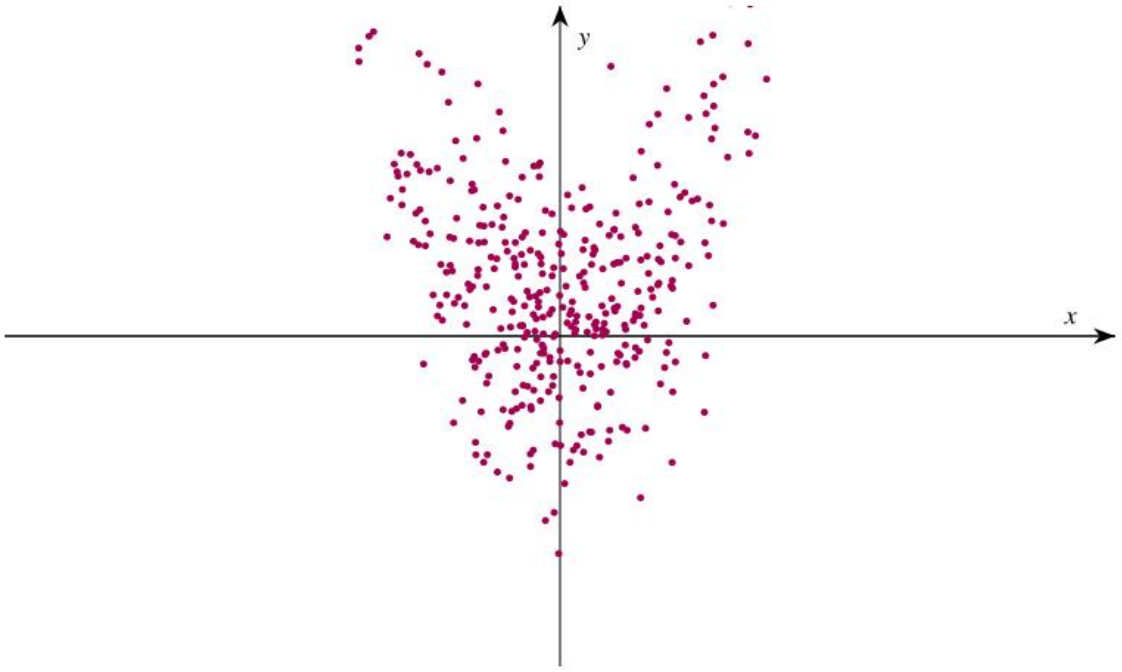

Рис. 4. Наличие нелинейной (параболической) связи между *X* и *Y*

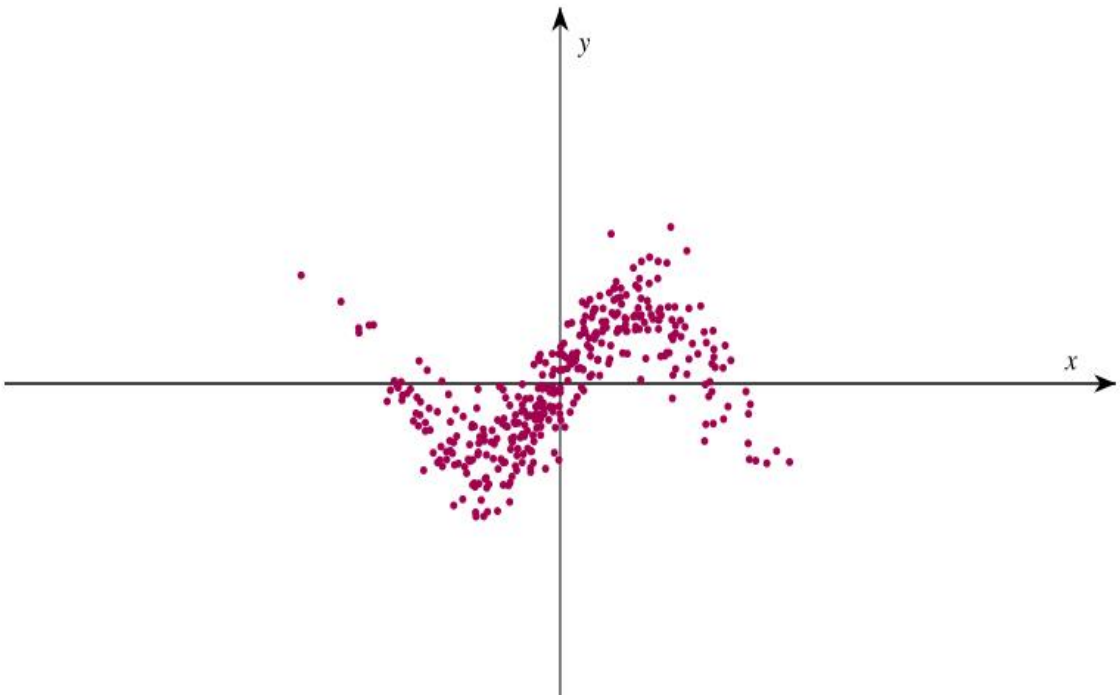

Рис. 5. Наличие нелинейной (синусоидальной) связи между *X* и *Y*

Более точную информацию о тесноте и направлении корреляционной связи дают *коэффициент корреляции* и *корреляционное отношение*. Меру линейной зависимости между двумя случайными вели-

чинами Y и X представляет параметрический парный коэффициент корреляции Пирсона r_{yx}, который вычисляют по формуле [22, с. 169]

$$r_{yx} = \frac{C_{yx}}{\sigma_y \sigma_x}, \qquad (1)$$

где $C_{yx} = \overline{yx} - \overline{yx} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y - \bar{y}\right) \Big/ n$. Число C_{yx} называют *ковариацией*, что представляет собой среднее произведение отклонений двух величин X и Y от их соответствующих средних $\bar{x}$, $\bar{y}$. Коэффициент корреляции r_{yx} принимает значение только на отрезке [–1; 1]. Его значение, близкое к +1 или −1, говорит о сильной прямой или обратной корреляции. В этом случае все точки корреляционного поля группируются около некоторой прямой. В то же время высокая корреляция необязательно означает наличие причинной связи между X и Y, поскольку, например, обе эти изучаемые величины могут зависеть от третьей величины Z. Значение r_{yx}, близкое к нулю, указывает на отсутствие линейной связи между X и Y, но не исключает возможности существования нелинейной связи между этими переменными, что требует дальнейшего исследования.

С увеличением числа наблюдений доверие к коэффициенту корреляции будет возрастать. Значимость коэффициента r_{yx} проверяется по t-критерию Стьюдента $t_{\text{расч}} = \sqrt{\dfrac{r_{yx}^2}{1 - r_{yx}^2}(n-2)}$ [22, с. 171] или с помощью таблицы Фишера – Йейтса [26, с. 185]. В таблице Фишера – Йейтса находят значение $r_{\text{кр}}$ при заданном уровне значимости α и $\upsilon = (n - 2)$ − количестве степеней свободы (в случае парной корреляции). Если $| r_{yx} | > r_{\text{кр}}$ (α, υ), коэффициент r_{yx} считается значимым. Заметим, что коэффициент корреляции лишь оценивает связь, исходя из выборочных значений факторов. Если эти значения хорошо отражают состояние изучаемой системы величин, то вычисленный коэффициент корреляции будет указывать на реальную связь.

Для *качественной оценки коэффициента корреляции* применяются различные шкалы, в частности *шкала Чеддока* [22, с. 170]. В зависимости от значения коэффициента корреляции связь оценивают так: 0,1 – 0,3 – слабая; 0,3 – 0,5 – заметная; 0,5 – 0,7 – умеренная; 0,7 – 0,9 – высокая; 0,9 – 1,0 – весьма высокая.

Пример расчёта коэффициента линейной корреляции Пирсона r_{yx} по формуле (1) приведён ниже в табл. 4. Предмет исследования – выборка (X, Y) показателей инновационной деятельности 74 российских регионов [24]. Здесь X – внутренние затраты на научные разработки, Y – объем инновационных товаров, работ и услуг, выраженные в процентах по отношению к своему среднему значению. Расчёты выполнены с помощью пакета прикладных программ по математической статистике MS Excel.

Таблица 4. Расчёт коэффициента линейной корреляции Пирсона

Номер региона	**Регион**	X	Y	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	Белгородская область	35,5812	106,2042	23217,85	152,9659	1884,5529
2	Брянская область	10,0317	50,07332	31656,76	1915,196	7786,456
3	Владимирская область	108,197	105,2623	6361,419	130,5537	911,32129
4	Воронежская область	102,016	75,36451	7385,6	341,2058	1587,4538
5	Ивановская область	4,86607	5,243277	33521,63	7848,718	16220,414
6	Калужская область	284,235	89,88488	9269,83	15,61345	380,43929
7	Костромская область	4,0512	26,85607	33820,68	4486,347	12317,927
8	Курская область	132,087	75,74519	3121,294	327,287	1010,722
9	Липецкая область	6,93327	375,0959	32768,94	79106,97	50914,157
10	Московская область	491,015	283,8143	91844,98	36091,66	57574,627
11	Орловская область	5,27596	7,35541	33371,7	7478,939	15798,258
12	Рязанская область	51,2698	47,44976	18682,9	2151,708	6340,3585
13	Смоленская область	12,3886	69,55755	30823,62	589,4561	4262,5309
14	Тамбовская область	97,3393	40,1833	8211,241	2878,641	4861,8118
15	Тверская область	135,123	25,07629	2791,224	4727,934	3632,7294
16	Тульская область	84,949	179,5221	10610,27	7342,054	8826,163
17	Ярославская область	134,642	137,3117	2842,308	1890,112	2317,8177

Продолжение табл. 4

Номер региона	**Регион**	X	Y	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
18	г. Москва	596,011	261,7843	166509,8	28206,55	68532,228
19	Республика Карелия	0,93779	2,24271	34975,51	8389,38	17129,59
20	Республика Коми	58,8723	161,4196	16662,39	4567,51	8723,8559
21	Амурская область	2,9358	40,1767	34232,18	2879,349	9928,0614
22	Вологодская область	4,96185	96,59344	33486,57	7,601972	504,5433
23	Калининградская область	24,3002	2,670691	26782,96	8311,163	14919,703
24	Ленинградская область	101,841	60,27604	7415,582	1126,289	2889,998
25	Мурманская область	7,38898	13,93754	32604,16	6383,806	14427,011
26	Новгородская область	78,1944	62,03174	12047,44	1011,528	3490,8908
27	Псковская область	2,9981	5,373457	34209,13	7825,669	16361,825
28	г. Санкт-Петербург	604,276	268,5674	173323,2	30530,97	72744,242
29	Республика Адыгея	3,90141	53,95264	33875,8	1590,704	7340,7339
30	Республика Калмыкия	2,33486	0,071559	34454,91	8791,821	17404,638
31	Краснодарский край	16,1866	2,989753	29504,45	8253,09	15604,578
32	Астраханская область	4,56905	22,40963	33630,48	5101,765	13098,658
33	Волгоградская область	75,6023	16,20311	12623,17	6026,908	8722,3092
34	Ростовская область	88,3621	94,42928	9918,792	0,351667	59,060216
35	Республика Дагестан	3,13926	0,06912	34156,93	8792,278	17329,664
36	Кабардино-Балкарская Республика	5,15674	9,689294	33415,28	7080,714	15381,937
37	Карачаево-Черкесская Республика	1,10931	3,27806	34911,39	8200,789	16920,429

Продолжение табл. 4

Номер региона	**Регион**	X	Y	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
38	Республика Северная Осетия	5,61097	0,098938	33249,42	8786,687	17092,462
39	Еврейская автономная область	1,89292	0,016677	34619,17	8802,116	17456,288
40	Ставропольский край	5,81264	59,21599	33175,91	1198,564	6305,8273
41	Республика Башкортостан	70,066	135,5952	13897,85	1743,81	4922,9277
42	Республика Марий Эл	2,07403	16,14592	34551,81	6035,79	14441,172
43	Республика Мордовия	31,8922	247,0008	24355,64	23459,37	23903,309
44	Республика Татарстан	106,299	569,9881	6667,718	226720,6	38880,7
45	Удмуртская Республика	21,8457	67,99315	27592,38	667,8669	4292,7891
46	Чувашская Республика	55,2208	91,37108	17618,43	6,077166	327,21567
47	Пермский край	171,318	469,3274	276,794	140993,6	6247,094
48	Кировская область	28,6779	49,65959	25369,27	1951,579	7036,3435
49	Нижегородская область	470,174	329,2795	79647,37	55433,49	66446,46
50	Оренбургская область	2,46438	26,64744	34406,84	4514,338	12462,911
51	Пензенская область	153,736	46,22648	1170,985	2266,692	1629,1907
52	Самарская область	277,344	511,2051	7990,314	174196,8	37308
53	Саратовская область	23,066	35,62744	27188,46	3388,268	9598,0101
54	Ульяновская область	150,227	156,7846	1423,43	3962,493	2374,9379
55	Курганская область	9,3211	29,56167	31910,14	4131,225	11481,636
56	Свердловская область	192,386	152,6864	19,62816	3463,343	260,72789
57	Тюменская область	87,6651	23,01701	10058,11	5015,368	7102,4715

Окончание табл. 4

Номер региона	**Регион**	X	Y	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
58	Челябинская область	160,421	144,6971	758,1088	2586,829	1400,392
59	Республика Алтай	0,49805	0,140226	35140,18	8778,949	17563,994
60	Республика Бурятия	4,63805	44,12191	33605,18	2471,518	9113,4953
61	Магаданская область	0,25619	226,2033	35230,92	17521,04	24845,168
62	Республика Хакасия	1,69091	0,393855	34694,38	8731,485	17404,985
63	Алтайский край	20,5507	26,44396	28024,28	4541,723	11281,778
64	Забайкальский край	4,08377	47,64122	33808,7	2133,982	8493,9492
65	Красноярский край	122,884	121,5438	4234,206	767,7067	1802,9499
66	Иркутская область	16,8773	13,98423	29267,64	6376,348	13660,917
67	Кемеровская область	5,97527	7,998775	33116,69	7368,075	15620,701
68	Новосибирская область	68,8112	80,41321	14195,29	180,1787	1599,2774
69	Омская область	64,2933	77,49596	15292,25	267,0056	2020,6727
70	Томская область	243,889	70,38173	3128,64	550,1157	1311,9124
71	Республика Саха (Якутия)	10,2619	62,4686	31574,92	983,9309	5573,8259
72	Камчатский край	1,99593	9,133138	34580,85	7174,621	15751,332
73	Приморский край	5,03573	7,390675	33459,53	7472,841	15812,582
74	Хабаровский край	10,261	107,9657	31575,21	199,6415	2510,7217
Среднее значение		**80,981**	**93,83827**	**28188,09**	**14397,27**	**13425,417**
Среднее квадратическое отклонение				**167,8931**	**119,9886**	–
Коэффициент корреляции						**0,67**

Вычисленное значение r_{yx} = 0,67 указывает на умеренную линейную связь между переменными X и Y.

Процедура проверки значимости коэффициента корреляции Пирсона r_{yx} опирается на допущение о нормальном распределении исходных данных. Для экспериментальных данных, распределение которых

существенно отличается от нормального, а также для выборок малого объема целесообразно использовать ранговую корреляцию, базирующуюся только на случайном характере исходных данных.

Непараметрическим аналогом классического коэффициента корреляции Пирсона является коэффициент ранговой корреляции Спирмена [26, с. 243]:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n},$$

где n – количество наблюдений, d_i^2 – квадрат разности рангов для i-го наблюдения. При этом рангом называют номер возрастающих (или убывающих) значений факторного или зависимого (результативного) признака. Коэффициент корреляции ρ имеет ту же интерпретацию взаимосвязи показателей, что и линейный коэффициент корреляции r_{yx}, но дает менее точную оценку взаимосвязи показателей. Однако простота расчета коэффициента Спирмена ρ делает его применимым на практике в тех случаях, когда требуется приблизительно оценить тесноту связи. Существенной считается связь, если |ρ| > 0,5 [26, с. 245]. Находит применение и другой показатель тесноты связи – ранговый коэффициент корреляции Кендэла [26, с. 247]:

$$\tau = \frac{2(P+Q)}{n(n-1)},$$

где P – величина, исчисленная для каждого ранга зависимого признака как число последующих рангов, бóльших взятого; Q – отрицательное количество последующих рангов, меньших каждого взятого ранга зависимого показателя; n – число наблюдений. Метод Кендэла отличается от метода Спирмена тем, что более детально анализирует тенденцию связи переменных, но принимает во внимание только факты близости или различия пар наблюдений без учета степени различий. Коэффициент корреляции знаков Фехнера вычисляется по формуле

$$kf = \frac{u - v}{u + v},$$

где u – число пар значений (X, Y), у которых знак отклонения от среднего значения совпадает; v – число пар, у которых знаки отклонения не совпадают. Расчёт ранговых коэффициентов корреляции позволяет в сложных условиях сократить объем работы.

Задача. Используя данные по двум показателям *X* и *Y*, измерить корреляцию между ними на основе расчёта:

а) коэффициента корреляции знаков Фехнера (табл. 5);

б) рангового коэффициента корреляции Спирмена (табл. 6);

в) рангового коэффициента корреляции Кендэла (табл. 7).

Исходные данные:

X	439	515	603	632	640	704	735	738	760	830	888	942	985	2093
Y	321	298	277	461	351	425	576	524	588	497	584	624	573	863

Таблица 5. Расчет коэффициента Фехнера

Номер	***X***	***Y***	**Знак отклонения от среднего по *X***	**Знак отклонения от среднего по *Y***
1	439	321	–	–
2	515	298	–	–
3	603	277	–	–
4	632	461	–	–
5	640	351	–	–
6	704	425	–	–
7	735	576	–	+
8	738	524	–	+
9	760	588	–	+
10	830	497	+	–
11	888	584	+	+
12	942	624	+	+
13	985	573	+	+
14	2093	863	+	+
Итого	**11504**	**6962**		
Средние	**821,714**	**497,286**		

Имеем: $u = 10$; $v = 4$; $kf = 0{,}429$. Таким образом, положительное значение коэффициента Фехнера указывает на наличие прямой связи между признаками *X* и *Y*, т. е. с ростом переменной *X* возрастает в среднем показатель *Y*.

Таблица 6. Расчет коэффициента ранговой корреляции Спирмена

Номер	***X***	***Y***	**Ранг по *X*, *R*(*x*)**	**Ранг по *Y*, *R*(*y*)**	**Разность рангов, d_i**	**d_i^2**
1	439	321	1	3	–2	4
2	515	298	2	2	0	0
3	603	277	3	1	2	4
4	632	461	4	6	–2	4
5	640	351	5	4	1	1
6	704	425	6	5	1	1
7	735	576	7	9	–2	4
8	738	524	8	8	0	0
9	760	588	9	12	–3	9
10	830	497	10	7	3	9
11	888	584	11	11	0	0
12	942	624	12	13	–1	1
13	985	573	13	10	3	9
14	2093	863	14	14	0	0

$\sum d_i^2 = 46$; $n = 14$; $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - 6 \cdot 46 / 2730 = 1 - 0{,}101 = 0{,}899.$

Вывод: коэффициент корреляции рангов Спирмена указывает на весьма тесную, прямую связь между показателями *X* и *Y*.

Таблица 7. Расчёт коэффициента ранговой корреляции рангов Кендэла

Номер наблю-дения	***X***	***Y***	**Ранг по *X***	**Ранг по *Y***	**Число P_i наблюдений по *X*, следующих за текущим с больши-ми значениями рангов**	**Число Q_i наблюдений по *Y*, следующих за текущим с меньши-ми значениями рангов**
1	439	321	1	3	11	2
2	515	298	2	2	11	1
3	603	277	3	1	11	0
4	632	461	4	6	8	2
5	640	351	5	4	9	0
6	704	425	6	5	8	0
7	735	576	7	9	5	2
8	738	524	8	8	6	1
9	760	588	9	12	2	3
10	830	497	10	7	4	0
11	888	584	11	11	2	1
12	942	624	12	13	1	1
13	985	573	13	10	1	0
14	2093	863	14	14	0	0
Итого					$\sum P_i = 79$	$\sum Q_i = 13$

$$\tau = \frac{2(P+Q)}{n(n-1)} = 2(79 - 13) / 14(14 - 1) = 120/182 = 0{,}659 > 0{,}5.$$

Вывод: коэффициент корреляции Кендэла указывает на наличие существенной прямой связи между показателями *X* и *Y*.

При расчете показателя тесноты связи по аналитической группировке используют формулу [8, с. 196]

$$\eta^2 = \left(\sum_{j=1}^{k}\left(\overline{y}_j - \overline{y}\right)^2 f_j\right) \Big/ \left(\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right),$$

где $\overline{y_j}$ – средние групповые значения показателя *Y*; $n = \sum_{j=1}^{k} f_j$ – объём выборки; f_j – частота в *j*-й группе; $\overline{y}$ – общая средняя результативного признака $\tilde{Y}$. В данном случае η называют эмпирическим корреляционным отношением, его уровень находится в зависимости от числа групп. Исходя из результатов корреляционного анализа проводят регрессионный анализ. По уравнению регрессии вычисляют теоретическое корреляционное отношение $\eta_{\text{теор}}$, квадрат которого равен [8, с. 197]

$$\eta^2_{\text{теор}} = \left(\sum_{i=1}^{n}\left(y_{\text{рег}.i} - \overline{y}\right)^2\right) \Big/ \left(\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right).$$

Корреляционное отношение в общем случае определяется формулой [5, с. 49]

$$\eta = \sqrt{\delta^2 / \sigma^2},$$

где δ^2 – дисперсия, обусловленная факторными признаками; σ^2 – общая дисперсия результативного признака. Значение η^2 называют *коэффициентом детерминации*. Для оценки качественной характеристики тесноты связи используют следующую шкалу [8, с. 226], отраженную в табл. 8:

Таблица 8. Оценка тесноты связи по коэффициенту детерминации

Значение коэффициента детерминации	**Характеристика связи**
Меньше 0,3	Весьма слабая
0,3 – 0,5	Слабая
0,5 – 0,8	Средняя
0,8 и выше	Сильная

Таким образом, коэффициент детерминации измеряет долю вариации результативного признака, который можно объяснить полученным уравнением регрессии, т. е. вариацией факторов, включенных в мо-

дель регрессии. В случае линейной зависимости между двумя признаками X и Y значение $\eta_{\text{теор}} = r_{yx}$. В случае множественной линейной корреляции значение $\eta_{\text{теор}} = R$, а его квадрат $\eta^2_{\text{теор}} = d^2$ – универсальный показатель тесноты связи.

На основе аналитической группировки расчет прямой линии регрессии $Y = a_0 + a_1 P$ проводится по формулам [8, с. 201]

$$a_1 = \sum_{j=1}^{k} \left(\overline{p}_j - \overline{p}\right)\left(\overline{y}_j - \overline{y}\right) f_j \,/\, \sum_{j=1}^{k} \left(\overline{p}_j - \overline{p}\right)^2 f_j;$$

$$a_0 = \overline{y} - a_1 \overline{p},$$

где $\overline{p_j}$ и $\overline{y_j}$ – групповые средние факторной переменной P и результативного признака Y; $\overline{p}$ – среднее значение P по всем наблюдениям; $\overline{y}$ – среднее значение Y.

Отметим, что расчет параметров корреляции на основе аналитической группировки приближенный, поскольку реальные значения факторов заменяются здесь их средними значениями. Поэтому мы предлагаем следующим шагом провести более детальное исследование зависимости $Y = Y(P)$ на основе парной выборки (P, Y) значений по всем n наблюдениям. Следует рассчитать ряд моделей $\tilde{Y} = Y(P)$, адекватных экспериментальным данным, т. е. для которых критерий Фишера [22, с. 196]

$$F = \left(\sum_{i=1}^{n} \left(\tilde{y}_i - \overline{y}\right)^2\right) / \left(\sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right)^2 / \left(n-2\right)\right),$$

где $\tilde{y}_i = Y(p_i)$ $(i = 1, \ldots, n)$ – значения результативного признака по уравнению регрессии; $y_i (i = 1, \ldots, n)$ – выборочные значения признака Y, выше критической границы $F_p(1, n-2)$ с $(1, n-2)$ степенями свободы, соответствующей уровню значимости p. Затем можно отобрать тренд по максимуму коэффициента детерминации d и критерия Фишера F либо по минимуму стандартной ошибки отклонения s регрессионных значений результативного признака Y от его экспериментальных значений; здесь $s^2 = \sum_{i=1}^{n} \left(\tilde{y}_i - y_i\right)^2 / \left(n-2\right)$.

Анализ остатков по изучаемым объектам позволит сделать вывод о том, какие объекты заметно отстают от выравненных значений по уравнению регрессии и какие в состоянии обеспечить высокий результативный показатель системы при определенном уровне многомерного среднего. Наиболее качественная модель даст возможность построить точечный прогноз результативного признака Y по значени-

ям многомерного среднего P, что осуществляется непосредственной подстановкой фиксированного P в уравнение модели. Однако уравнение модели $Y = Y(P)$ не позволяет оценить индивидуальное влияние каждой факторной переменной в отдельности на результативный признак системы. Поэтому на следующем этапе работы используются многомерные методы математической статистики: метод множественной линейной регрессии в сочетании с корреляционным анализом и пошаговой регрессией.

2.4. Основные задачи многомерного корреляционного анализа

Многомерный корреляционный анализ предназначен для решения следующих задач: оценки тесноты связи (на основе матрицы коэффициентов парной корреляции) одной случайной величины, например результативного показателя системы, с множеством остальных факторных переменных; измерения тесноты связи между двумя переменными при исключении влияния на них остальных факторов исследуемой системы. Решение этих задач предполагает расчёт соответственно коэффициентов множественной и частной корреляции. Вначале на основе таблицы многомерной выборки значений переменных Y и X (табл. 9) вычисляют коэффициенты парной корреляции r_{yx_j} и $r_{x_i x_j}$ для определения силы линейной связи различных пар переменных, входящих в систему.

Таблица 9. Значения переменных Y и X

Переменная / Номер наблюдения	Y	X_1	X_2	…	X_m
1	y_1	x_{11}	x_{21}	…	x_{m1}
…	…	…	…	…	…
n	y_n	x_{1n}	x_{2n}	…	x_{mn}

Матрица **R** коэффициентов парной корреляции имеет вид

$$\mathbf{R} = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_1x_2} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_1x_m} & r_{x_2x_m} & \dots & 1 \end{pmatrix}$$

Матрицу **R** называют также корреляционной, она симметрична относительно главной диагонали, на которой расположены единицы. Чтобы полностью оценить зависимость между переменной Y и остальными переменными (X_1, X_2, …, X_m), вычисляют выборочный коэффициент множественной корреляции по формуле [22, с. 173]

$$R_{y,1,2,\ldots,m} = \sqrt{1 - \frac{|R|}{R_{11}}},$$

где $|R|$ – определитель корреляционной матрицы R; R_{11} – алгебраическое дополнение элемента r_{11} матрицы R. Квадрат коэффициента множественной корреляции $D = R^2_{y,1,2,\ldots,m}$ называется выборочным множественным коэффициентом детерминации; D – доля вариации (случайного разброса) величины Y – объясняет вариацию остальных переменных (X_1, X_2, …, X_m). Коэффициенты множественной корреляции и детерминации принимают неотрицательные значения в интервале (0; 1). Чем ближе коэффициент D к единице, тем сильнее взаимосвязь случайных переменных. По величине коэффициентов $R_{y,1,2,\ldots,m}$ и D нельзя сделать вывод о направлении связи. Значимость коэффициента детерминации устанавливается сравнением расчетного значения F-критерия Фишера: $F_{\text{расч}} = \dfrac{R^2 / m}{(1 - R^2)/(n - m + 1)}$ с табличным значением $F_{\text{табл}} = F(\alpha, m, n - m + 1)$ [22, с. 173]. Коэффициент D значимо отличается от нуля, если выполняется неравенство $F_{\text{расч}} > F_{\text{табл}}$. Табличное значение F-критерия соответствует заданному уровню значимости α и степеням свободы m и $(n - m + 1)$. Ниже будет показано, что по уравнению регрессии также можно рассчитать множественный коэффициент детерминации и оценить его значимость. Отметим, что величина $D = R^2_{y,1,2,\ldots,m}$ может только увеличиться, если в модель регрессии включить новые переменные, и уменьшиться при исключении из анализа какой-либо из имеющихся факторных переменных.

Если изучаемые случайные переменные связаны попарно корреляцией, то на величине коэффициентов корреляции частично сказывается влияние других величин. В связи с этим возникает необходимость исследования частной корреляции между величинами при *исключении* влияния других случайных переменных (возможно, одной или нескольких). Исходя из этого вычисляют частные коэффициенты

корреляции. Выборочный частный коэффициент корреляции определяется по формуле [22, с. 174]

$$r_{jk(1,\,2,\,\ldots,\,m)} = -\frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}},$$

где R_{jk}, R_{jj}, R_{kk} – алгебраические дополнения к соответствующим элементам корреляционной матрицы **R**. Частный коэффициент корреляции может принимать значения в пределах от –1 до +1. При $m = 3$ получаем коэффициент $r_{12(3)}$ корреляции между x_1 и x_2 при фиксированном x_3: $r_{12(3)} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$.

2.5. Построение моделей парной регрессии

Задача IV этапа исследования – выявление характера и структуры взаимосвязи между элементами системы и оценка их влияния на результативный показатель, построение многофакторной математической модели исследуемой системы показателей, получение статистических оценок неизвестных параметров, входящих в уравнения регрессии $Y = Y(X_j)$ ($j = 1, \ldots, k$), проверка статистических гипотез о регрессии. Решение поставленных задач и интерпретация конечных результатов исследований требуют глубоких знаний многомерного статистического анализа и умение правильно оценивать полученную информацию.

Под *регрессией* понимают зависимость среднего значения какой-либо случайной величины Y от некоторой другой величины X или от нескольких величин. При этом изменение значений x независимой переменной X приводит к закономерному изменению условного математического ожидания $M\{Y|X = x\} = y(x)$ величины Y.

Уравнение $y = y(x)$ называется уравнением регрессии, а соответствующий график – линией регрессии, здесь функция $y(x)$ и есть регрессия случайной величины Y по X.

При практической реализации шагов 1 – 4 оба метода, корреляционный и регрессионный анализ, обычно применяются комплексно. *Первый шаг* данного этапа исследования содержит подготовительные действия для дальнейшего детального проведения корреляционно-

регрессионного анализа. Здесь вычисляют основные статистические характеристики выборочных распределений показателей: выборочное среднее, выборочную дисперсию, стандартное отклонение, медиану, асимметрию, эксцесс, ошибки репрезентативности (формулы см. выше). Также предполагается посмотреть наличие некоторых предпосылок применения корреляционно-регрессионного анализа, в частности проверить гипотезу нормального распределения. Таким образом, на первом шаге данного этапа необходимо представить себе, с какой числовой информацией (с точки зрения математической статистики) имеем дело, и исключить некоторые исходные данные из рассмотрения либо преобразовать к виду, удобному для дальнейшей работы.

На *втором шаге* изучается теснота связи не только между каждым фактором X_j (j = 1, …, k) системы и зависимым признаком Y – результативным показателем, но и между каждой парой факторов, что является предметом многофакторного корреляционного анализа. При этом вычисляют парные коэффициенты корреляции $r_{x_j y}$, $r_{x_j xi}$(i, j = 1, …, k). Строят корреляционную матрицу R, элементы которой есть парные коэффициенты корреляции между соответствующими факторами. Заметим, что матрица R симметрична относительно главной диагонали. На основе корреляционной матрицы в дальнейшем исследовании определяются частные и множественные коэффициенты корреляции и детерминации. После вычисления коэффициентов корреляции следует проверить их значимость. При этом задается уровень значимости α, равный вероятности принятия ошибочного решения. В расчетах обычно берут одно из следующих значений уровня значимости: 0,05; 0,02; 0,01; 0,001. Возможность ошибки следует из того факта, что для установления тесноты взаимосвязи взята выборка лишь ограниченного числа данных из некой генеральной совокупности объектов.

В случае парной регрессии для выбора аналитической зависимости $y(x)$ (в нашем случае на шаге 3 при моделировании зависимостей $Y=Y(X_j)$ (j = 1, 2, …, k) результативного признака Y от факторов X_j) целесообразно вначале построить корреляционное поле, визуальный анализ которого позволяет выбрать для расчета одну или несколько моделей регрессии. В предлагаемых ниже примерах модели получены процедурой простой регрессии с использованием пакета прикладных программ STADIA.

Следуя [12, с. 287], укажем основные типы регрессионных моделей, из числа которых отбираются наиболее качественные, удовлетворяющие задачам исследования:

1. Линейная: $y(x) = a + bx$.
2. Парабола: $y(x) = a + bx + cx^2$.
3. Полиномиальная: $y(x) = \sum_{i=0}^{m} a_i x^i$.
4. Логарифмическая: $y(x) = a + b \ln(x)$.
5. Степенная: $y(x) = ax^b$.
6. Показательная: $y(x) = ab^x$.
7. Экспонента: $y(x) = \exp(a + bx)$; $y(x) = \exp(a + b/x)$; $y(x) = a + b \cdot \exp(cx)$; $y(x) = \exp(a + bx + cx^2)$; $y(x) = \exp(a + b\sqrt{x})$; $y(x) = a + b \cdot \exp(cx)$.
8. Гипербола: $y(x) = a + b/x$; $y(x) = 1/(a + bx)$; $y(x) = a + b \cdot \exp(cx)$; $y(x) = 1/(a + b \ln x)$; $y(x) = 1/(a + b\sqrt{x})$; $y(x) = a + 1/(b + cx)$.
9. Оптимума: $y(x) = 1/(a + bx + cx^2)$; $y(x) = x/(a + bx + cx^2)$.
10. Логистическая: $y(x) = a + b/(1 + \exp(c + dx))$.
11. Линейная с синусом: $y(x) = a + bx + c \cdot \sin(d + \exp(x))$.

Коэффициенты линейной модели вычисляются согласно методу наименьших квадратов (МНК) по формулам

$$a = \left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right) \Big/ n,$$
$$b = \left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i\right) \Big/ \left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right). \quad (2)$$

Коэффициенты параболической модели также рассчитываются по МНК, т. е. таким образом, чтобы сумма квадратов отклонений наблюдаемых значений y_i от $y_{\text{рег}.i}$, вычисленных по уравнению регрессии, была минимальной: $S = \sum_{i=1}^{n} \left(y_i - y_{\text{рег}.i}\right)^2 \to \min$. Этот критерий и лежит в основе метода наименьших квадратов, согласно которому регрессионные коэффициенты (параметры модели) вычисляют такими, чтобы они обеспечивали минимальное значение суммы S. В этом случае на графике линия регрессии проходит там, где точки корреляционного поля наибольшим образом сконцентрированы, т. е. на минимальном удалении от них. В частности, в случае линейной модели для поиска регрессионных коэффициентов a и b строится система нормальных уравнений, представляющих необходимое условие экстремума функции $S = S(a, b)$:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases} \quad \text{или} \quad \begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}. \qquad (3)$$

Коэффициенты a и b – решение системы (3) – вычисляют по формулам (2). Между коэффициентами a и b и парным коэффициентом корреляции r_{yx} существует связь: $b = r_{yx}\dfrac{\sigma_y}{\sigma_x}$. При этом $a = \overline{y} - r_{yx}\dfrac{\sigma_y}{\sigma_x}\overline{x}$. Так что уравнение линейной регрессии выражается формулой

$$y(x) = \overline{y} + r_{yx}\frac{\sigma_y}{\sigma_x}(x - \overline{x}),$$

где $\overline{y}$ и $\overline{x}$ – среднее арифметическое значений исходных данных y и x соответственно. Вычисленные по уравнению регрессии значения $y_{рег}(x) = y(x)$ мы называем теоретическими, или выравненными значениями результативной переменной Y.

Ряд нелинейных моделей в процессе вычисления с помощью специальных преобразований [12, с. 288] приводят к линейным. Однако если при решении конкретной задачи можно ограничиться построением линейной модели, зачастую выбирают ее. Это объясняется тем, что расчет линейных моделей достаточно простой и сами модели легче интерпретируемы. Рекомендации по применению моделей можно найти в [12, с. 285 – 286]. Остаточная дисперсия вычисляется по формуле $s^2_{ост} = \sum_{i=1}^{n}(y_i - y_{рег.i})^2/(n - 2)$ [12, с. 279]. Рассматриваемая дисперсия $s^2_{ост}$ – это характеристика уклонений (рассеивания) относительно линии регрессии, т. е. она вызвана влиянием на Y факторов, отличных от X.

С помощью процедур пакета STADIA по F-критерию проверяют гипотезу адекватности модели экспериментальным данным. Кроме того, строят график модели на фоне корреляционного поля с указанием зоны доверительного интервала. Все это позволяет выбрать наиболее качественную модель, адекватную экспериментальным данным. Как правило, это модель с минимальной остаточной дисперсией либо максимальным значением $F_{набл}$-критерия и коэффициента множественной корреляции R.

Таким способом в работе [14] при исследовании зависимости объёмов туристских потоков Y от каждого фактора инфраструктуры в отдельности был рассчитан ряд адекватных моделей типа 1 – 11, затем отобраны модели, наиболее качественные по своим статистическим характеристикам $F_{\text{набл}}$, $s^2_{\text{ост}}$, R. Отметим, что для сильно зашумленных данных в некоторых случаях пользуются и неадекватной моделью, чтобы хорошо локализовать область максимума; в конечном счете все зависит от целей исследования.

На последнем шаге этого этапа полезно построить графики моделей, лучших по статистическим характеристикам, и выявить объекты, отстающие от линии регрессии. Это позволяет дать оценку положения объектов исследования относительно линии регрессии, выявить объекты, для которых изучаемые факторы обеспечивают достаточно высокий либо низкий результативный показатель системы Y. Кроме того, по уравнениям регрессии и графикам можно выполнить приближенный точечный прогноз показателя Y за пределами диапазона значений факторных переменных и также оценить доверительный интервал прогнозных значений, пользуясь процедурами программы STADIA.

Следуя [5, гл. 14], остановимся на аналитических свойствах параметров некоторых однофакторных моделей регрессии. В случае линейной модели $y = a + bx$ изменение результативного признака Y, обусловленное изменением факторной переменной X, есть величина $\Delta y = b\Delta x$. Если, например, факторная переменная возрастет на 1 единицу своего измерения, тогда результативная переменная изменится на b единиц. При моделировании параболой $y = a + bx + cx^2$ экстремальное значение результативного признака достигается при значении факторной переменной $x_0 = -b/2c$. В случае модели гиперболы $y = a + \dfrac{b}{x}$ повышение факторной переменной на 1 единицу измерения по отношению к своему среднему значению $\bar{x}$ влечет изменение результативной переменной на $\Delta y = \dfrac{b}{\bar{x}+1} - \dfrac{b}{\bar{x}} = -\dfrac{b}{\bar{x}(\bar{x}+1)}$ единиц. Во всех трех случаях выше свободный член a показывает усредненное влияние неучтенных факторов и частичное влияние фактора X на Y в той мере, в какой X коррелирует с другими факторами. При моделировании степенной функцией $y = ax^b$ (a, $b > 0$) увеличение факторной переменной

X на 1 % влечет увеличение результативной примерно на b %. При моделировании с помощью показательной функции $y = ab^x$ (a, $b > 0$) рост факторного признака на 1 единицу измерения увеличивает зависимый показатель Y в b раз. Таким образом, число b – это коэффициент роста, а в процентах выражает темп роста. Здесь свободный член a есть начальное значение фактора Y (при $x = 0$), которое может расти в заданном темпе. Частный случай показательной функции – экспонента $y = \exp(a + bx)$, которая при $b > 0$ показывает, что повышение факторной переменной на Δx единиц измерения приводит к увеличению результативной переменной в $\exp(b\Delta x)$ раз. Устойчивость параметров регрессионных моделей по отношению к изменению факторного и зависимого признаков изучается в [5, гл. 12]. В частности, показано, что в случае линейной модели большей устойчивостью характеризуется коэффициент при x по сравнению со свободным членом a.

Отметим, что конкретная интерпретация регрессионных коэффициентов может вырьироваться во времени и пространстве. Изменение количества наблюдений и видов зависимостей влечет изменение параметров регрессии. Без точного понимания аналитического смысла регрессионных коэффициентов невозможно уяснить их настоящего экономико-статистического содержания.

2.6. Построение моделей множественной регрессии

Вначале мы определяем тесноту линейной зависимости результативного показателя Y от факторных признаков X_1, X_2, …, X_k с помощью *множественного коэффициента корреляции* R, в качестве выборочной оценки которого используем выражение [22, с. 173].

Если величина $R < 0{,}3$, тогда теснота связи считается слабой, если $R > 0{,}7$ – сильной. Число $d = R^2$ называют *множественным коэффициентом детерминации*. Значимость множественного коэффициента детерминации проверяем на основании F-критерия. Число d показывает, какая доля дисперсии результативного признака объясняется влиянием факторных признаков. При небольшом числе наблюдений, если $(n - k) / k \leq 20$, вычисляют *скорректированный коэффициент детерминации* [22, с. 204]: $d_{\text{скор}} = 1 - (1 - R^2)(n - 1) / (n - k - 1)$. Отметим, что всегда $d_{\text{скор}} < d$.

Тесноту связи в зависимостях, отличных от линейной, рекомендуем измерять с помощью корреляционного отношения η, интерпретация которого зависит от вида исследуемой связи.

Заметим, что результаты корреляционного анализа не могут дать полного понимания роли факторных переменных в формировании результативного показателя, нельзя также представить корреляцию факторов как связь их уровней, что создает серьезные ограничения в исследовании многофакторных систем экспериментальных данных. Интерпретация результатов исследования корреляционных связей требует тщательного проведения регрессионного анализа.

Так при исследовании сложных систем в экономике (в том числе и в региональной сфере) строят многофакторные регрессионные модели, описывающие зависимость показателя *Y* от нескольких факторных переменных. Ниже приведены основные типы моделей множественной регрессии.

1. Линейная:

$y(x) = a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k$.

2. Степенная:

$y(x) = a_0 x_1^{a_1} x_2^{a_2} \cdot \ldots \cdot x_k^{a_k}$.

3. Показательная:

$y(x) = \exp(a_0 + a_1x_1 + a_2x_2 + \ldots + a_kx_k)$.

4. Параболическая:

$y(x) = a_0 + a_1x_1^2 + a_2x_2^2 + \ldots + a_kx_k^2$.

5. Гиперболическая:

$$y(x) = a_0 + \frac{a_1}{x_1} + \frac{a_2}{x_2} + \ldots + \frac{a_k}{x_k}.$$

Здесь a_0, a_1, a_2, …, a_k – коэффициенты регрессии, (x_1, x_2, …, x_k) – значения факторных переменных X_1, X_2, …, X_k. Остановимся подробнее на модели 1 множественной линейной регрессии. В матричной форме линейная многофакторная модель 1 (с учетом $\vec{\varepsilon}$ случайных ошибок (остатков)) имеет вид $\vec{Y} = X\vec{a} + \vec{\varepsilon}$. При этом предполагается, что компоненты случайного вектора $\vec{\varepsilon}$ имеют нормальный закон распределения и математическое ожидание, равное нулю. Здесь используют обозначения

$$\vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}; \quad \vec{a} = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_n \end{pmatrix}; \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

где $\vec{Y}$ – вектор-столбец значений зависимого признака Y; $\vec{a}$ – вектор-столбец регрессионных коэффициентов; X – матрица, столбцы которой, начиная со второго, соответствуют значениям факторных переменных X_1, X_2, …, X_k, номер строки соответствует номеру объекта наблюдения. Согласно МНК строят систему нормальных уравнений. Решение этой системы – вектор оценок регрессионных коэффициентов – представляют в виде [22, с. 211]

$$\vec{a} = \left(X^T X \right)^{-1} X^T \vec{Y},$$

где X^T – матрица, полученная транспонированием матрицы X; $(X^TX)^{-1}$ – матрица, обратная к матрице X^TX. После вычисления по данной формуле значений регрессионных коэффициентов получают уравнение регрессии линейной модели. По уравнению регрессии также можно рассчитать множественный коэффициент корреляции по формуле

$$R = \sqrt{1 - \left(\sum_{i=1}^{n} \left(y_{\text{рег}.i} - y_i \right)^2 \right) \Big/ \left(\sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2 \right)}, \qquad (4)$$

где $y_{\text{рег}.i}$ – индивидуальные значения признака Y по уравнению линейной регрессии; y_i – индивидуальные выборочные значения Y; $\overline{y}$ – выборочное среднее значение показателя Y; n – объем выборки. Значимость регрессионной модели проверяют по F-критерию Фишера [12, с. 298]. Для этого вычисляют

$$F_{\text{набл}} = \frac{\frac{1}{k} Q_R}{\frac{1}{n-k-1} Q_{\text{ост}}}, \qquad (5)$$

где $Q_R = \sum_{i=1}^{n} \left(y_{\text{рег}.i} - \overline{y} \right)^2$; $Q_{\text{ост}} = \sum_{i=1}^{n} \left(y_{\text{рег}.i} - y_i \right)^2$. Значение $F_{\text{набл}}$ определяет качество регрессионной модели: чем больше число $F_{\text{набл}}$, тем точнее модель описывает зависимость результативного признака Y от факторных переменных.

Если $F_{\text{набл}} > F_{\text{кр}}(\alpha, \upsilon_1, \upsilon_2)$, уравнение считается значимым и регрессионную модель признают адекватной экспериментальным данным при уровне значимости α с $\upsilon_1 = k$ и $\upsilon_2 = n - k - 1$ степенями свободы. Статистическую значимость отдельных коэффициентов регрессии можно проверить по *t*-критерию Стьюдента. Для этого вычисляют $t_{\text{набл}} = a_i \Big/ \sqrt{\sigma^2_{a_i}}$, где $\sigma^2_{a_i}$ – дисперсия соответствующего регрессионного коэффициента a_i данной модели. Значение $\sigma^2_{a_i}$ ($i = 0, 1, \ldots, k$) находится на главной диагонали ковариационной матрицы [8, с. 210] $V = \sigma^2_{\text{ост}} \left(X^T X \right)^{-1}$, где $\sigma^2_{\text{ост}} = \dfrac{Q_{\text{ост}}}{n-k-1}$ – несмещенная оценка остаточной дисперсии. По таблице *t*-распределения Стьюдента находим $t_{\text{крит}}(\alpha, \upsilon)$ для заданного уровня значимости α и числа степеней свободы $\upsilon = n - k - 1$. Если $|t_{\text{набл}}| > t_{\text{крит}}$, соответствующий коэффициент регрессии признают значимым.

При наличии незначимых коэффициентов регрессии в линейной многофакторной модели в дальнейшем исследовании используют методы пошаговой регрессии [12, с. 303 – 309]. Пошаговая регрессия позволяет из исходных факторных переменных $X_1, X_2, \ldots, X_k$ последовательно отбирать факторы, оказывающие наиболее существенное влияние на результативный признак *Y*.

Рассмотрим два вида процедуры отбора факторных переменных: включение переменных, исключение переменных. При методе *последовательного включения* вначале в модель вводят ту факторную переменную, которая имеет наибольший коэффициент корреляции с зависимой переменной *Y*. На последующих шагах в модель включают ту переменную X_i, которая имеет наибольший частный коэффициент корреляции $\sqrt{dR^2} = \sqrt{R^2 - R_i^2}$, где R_i^2 – коэффициент детерминации, который вычисляется, когда все независимые переменные, за исключением *i*-ой, присутствуют в модели. Значимое изменение R^2 указывает на то, что фактор X_i вносит существенный вклад в вариацию *Y*, и поэтому X_i следует включить в уравнение модели регрессии.

Процесс прекращается, когда ни одна из оставшихся переменных не обеспечивает заданное минимальное значение статистики Фишера, или *F*-значение включения, либо когда значение толерантности

меньше заданного уровня. Число $T = 1 - R_i^2$ есть *толерантность* [12, с. 307], которая показывает меру изменчивости регрессионных коэффициентов, не объясняемую другими переменными. Малое значение T означает высокую степень коррелированности i-й независимой переменной с другими независимыми переменными и в то же время большую стандартную ошибку в оцениваемом регрессионном коэффициенте a_i.

Метод *последовательного исключения* состоит в удалении на очередном шаге из полного набора факторных переменных той переменной, которая имеет наименьший частный коэффициент корреляции. Процесс исключения факторов останавливается на том шаге, при котором удаление очередной переменной из модели не приводит к значимому изменению коэффициента множественной корреляции R. В конечном итоге с помощью процедуры отбора переменных, даже в условиях мультиколлинеарности факторных признаков, мы формируем модель множественной линейной регрессии, адекватной для описания исходных данных, в которой регрессионные коэффициенты $\{a_i\}$ эмпирически значимы.

Регрессионные коэффициенты моделей линейной регрессии характеризуются определенной устойчивостью по отношению к изменению исходных данных. Модель множественной линейной регрессии позволяет нам оценивать влияние факторных переменных на вариацию результативного показателя Y, что достигается вычислением коэффициентов эластичности модели, Δ-коэффициентов и β-коэффициентов соответственно по формулам [22, с. 225 – 226]

$$Э_i = a_i \frac{\overline{x}_i}{\overline{y}}; \;\; \beta_i = a_i \frac{S_i}{S_y}; \;\; \Delta_i = r_{yx_i} \frac{\beta_i}{R^2},$$

где $\overline{x}_i$ и $\overline{y}$ – выборочные средние значения факторной переменной; σ_i – стандартное отклонение факторной переменной X_i ($i = 1, 2, \ldots, k$), σ_y – отклонение зависимой переменной Y. Числа β_i^2 ($i = 1, 2, \ldots, k$) показывают самостоятельный взнос каждого показателя X_i в вариацию результативного показателя Y. Таким образом, β-коэффициенты модели – это своего рода индикаторы относительной важности факторных переменных. Величина $\eta = R^2 - (\beta_1^2 + \ldots + \beta_k^2)$ составляет так называемый *системный эффект модели*, обусловленный взаимосвязями фак-

торных переменных. Отметим, что свободный член a_0 в уравнении регрессии отражает усредненное влияние факторов, не привлеченных к исследованию, и частичное влияние учтенных факторных признаков, которое зависит от тесноты их связи с неучтенными факторами.

К сожалению, методы корреляционно-регрессионного анализа можно применять не ко всем статистическим данным. Использование этих методов для получения статистико-математических моделей многофакторных систем основано на ряде предпосылок [13, с. 37 – 39], выполнение которых необходимо для того, чтобы свойства полученных математических моделей были научно обоснованы и в конечном счете обеспечили прикладную полезность решения реальных задач. Предпосылки множественного регрессионного анализа формулируют обоснованность полученных результатов и свойств моделей.

Возможности применения в прикладных задачах моделирования на основе корреляционно-регрессионного анализа весьма широки, хотя не стоит и переоценивать его значение. Математическое моделирование следует использовать как вспомогательное средство, в то время как окончательное решение всегда должно оставаться за специалистом.

2.7. Статистический анализ многофакторной системы показателей инновационного развития российских регионов

Первый этап работы начался со сбора и систематизации сведений об инновационной деятельности регионов. Эти сведения были представлены в виде многофакторной системы (X_1, X_2, …, X_9) статистико-экономических показателей на основе информации Федеральной службы государственной статистики за 2013 г. [24, с. 698 – 738]. Таким образом, предметом исследования [17] стали следующие показатели: X_1 – выпуск из аспирантуры и докторантуры (количество человек); X_2 – численность исследователей с учеными степенями (количество кандидатов наук и докторов наук); X_3 – численность персонала, занятого научными исследованиями и разработками (количество человек); X_4 – внутренние затраты на фундаментальные исследования (млн руб.); X_5 – внутренние затраты на прикладные исследования (млн руб.); X_6 – внутренние затраты на разработки (млн руб.); X_7 – число используемых передовых производственных технологий; X_8 – чис-

ло выданных патентов на изобретения и полезные модели; X_9 – объем инновационных товаров, работ и услуг (млн руб.). Из объектов исследования были исключены следующие регионы: Республика Ингушетия, Республика Тыва, Чеченская Республика, Чукотский и Ненецкий автономные округа (ввиду недостаточной информации по выбранной нами системе показателей), а также Архангельская и Сахалинская области (по причине весьма высоких значений отдельных показателей, несопоставимых со значениями по другим регионам). Например, по Сахалинской области абсолютное значение показателя X_9 составляет 321 867,5 млн руб., что примерно в 1,5 раза превышает значение X_9 по г. Санкт-Петербургу и в 1,4 раза – по Московской области, которые входят в число очевидных лидеров по остальным показателям. Удивляет динамика изменения значений X_9 по Сахалинской области [24, с. 735]: в 2011 г. по сравнению с предыдущим периодом этот показатель возрос более чем в 3000 раз, а доля инновационных товаров, работ, услуг в общем объеме производства по данному региону, начиная с 2011 г. стала составлять более 50 %, чего не наблюдается ни в одном другом регионе РФ. Статистика по Архангельской области показывает рост показателя X_9 в 2012 г. в 39 раз по сравнению с предыдущим годом. Исходные данные по Ханты-Мансийскому и Ямало-Ненецкому автономным округам включены в данные по Тюменской области. Таким образом, в качестве объектов исследования были взяты 74 региона РФ. В целях сравнения этих регионов по уровню инновационного развития абсолютное значение каждого показателя было представлено в расчете на 10 тыс. человек населения, занятого в экономике. Так была сформирована матрица ID (74, 9) исходных данных для последующей статистической обработки. Поскольку абсолютные значения показателей разнокачественны и выражаются в разных единицах измерения, исходные данные были переведены в процентные значения по отношению к своему выборочному среднему $\overline{X_j}$ (j = 1, …, 9), вычисленному по каждому столбцу. Так была построена матрица процентов PD (74, 9), вектор-столбцы которой PD_j (j = 1, …, 8) именуются в дальнейшем факторными переменными системы. Компоненты вектора PD_9 представляют показатель X_9, который принят за результативный показатель системы и в дальнейшем обозначается через Y.

Поясним практический смысл некоторых показателей. Согласно методологии Института проблем развития науки, фундаментальными

называют экспериментальные и теоретические исследования, направленные на получение новых знаний без какой-либо конкретной цели, связанной с использованием этих знаний. Прикладными исследованиями могут называться исследования, направленные на получение новых знаний с целью решения конкретных практических задач [10, с. 235]. К внутренним затратам на исследования и разработки относятся выраженные в денежной форме фактические затраты на выполнение исследований и разработок на территории страны (включающие в себя финансирование из-за рубежа, но исключающие выплаты, которые были произведены за рубежом). Один из факторов, сдерживающих инновационное развитие в РФ, – недостаточность финансирования. В «Стратегии инновационного развития РФ» указано, что к 2020 г. внутренние затраты на исследования и разработки должны увеличиться до 2,5 – 3 % от ВВП. Несмотря на положительную динамику внутренних затрат на исследования и разработки в РФ и возрастающую конкуренцию в сфере НИОКР, этот показатель пока остается ниже уровня других развитых стран [20].

В начале второго этапа на основе факторных переменных по каждому региону вычислен показатель $P_i(\%) = \sum_{j=1}^{8} PD_{ij} / 8$ (i = 1, …, 74) и помещен в десятый столбец матрицы процентов. Вектор $P(\%) = \{P_1, \ldots, P_{74}\}$ именуется в дальнейшем многомерным средним показателем, комплексно характеризующим инновационное развитие российских регионов. Затем был построен ранжированный вариационный ряд $\{P_i\}$ и, соответственно, отсортированы строки матрицы процентов так, что все регионы выстраиваются в порядке возрастания многомерного среднего и получают ранг согласно номеру строки. Чем выше ранг региона, тем больше его многомерный средний. Описательная статистика дискретного вариационного ряда $\{P_i\}$ следующая: $P_{\max}$ = 551,69 %; $P_{\min}$ = 23,76 %; размах вариации R = 527,93 %; выборочное среднее $\bar{P}$ = 100 %; стандартное отклонение s = 88,67 %; медиана Me = 72,57 %; коэффициенты асимметрии и эксцесса: A_s = 2,96 и E_x = 10,05; размах 95 %-го доверительного полуинтервала среднего $d\bar{P}$ = 20,76 %. Описательная статистика результативного показателя Y: $Y_{\max}$ = 617,33 %; $Y_{\min}$ = 0,02 %; размах вариации R = 617,31 %; выборочное среднее $\bar{Y}$ = 100 %; стандартное отклонение s = 130,50 %; медиана Me = 52,03 %; коэффициенты A_s = 2,17 и E_x = 4,66; $d\bar{Y}$ = 29,85%. Расчет описательной статистики P и Y дал возможность представить, с какой числовой ин-

формацией мы имеем дело: в частности, значения стандартных отклонений этих распределений свидетельствуют о большом разбросе значений P и Y около своих средних значений. Ранговый коэффициент корреляции Спирмена между P и Y невысокий и составляет 0,54. Высокие значения коэффициентов вариации результативного показателя Y и многомерного среднего P указывают на то, что каждая совокупность значений (Y_1, Y_2, …, Y_{74}) и (P_1, P_2, …, P_{74}) неоднородна и для дальнейшего статистического анализа возникает необходимость разбить множество объектов-регионов на группы. Проверка по известным критериям показала, что распределения показателей P и Y значимо отличаются от нормального, что не позволяет оценивать надежность выборочного коэффициента корреляции между P и Y и параметры регрессионных моделей, описывающих зависимость Y от факторов (X_1, X_2, …, X_8). Это обстоятельство также приводит к необходимости разбиения объектов исследования на группы с целью моделирования зависимостей между факторными переменными системы. Кроме того, представляет интерес выделение групп регионов, схожих по величине инновационного потенциала. Число групп в данном исследовании равно четырем. Такое количество взято исходя из гистограммы распределения многомерного среднего. Вначале была сформирована группа C со средним уровнем инновационного развития из регионов, для которых многомерный P(%) находился в пределах доверительного интервала среднего ($\overline{P} - d\overline{P}$, $\overline{P} + d\overline{P}$), т. е. в диапазоне 79 – 121 %. В состав группы C вошли 19 регионов: Республика Мордовия, Курская область, Республика Башкортостан и так далее, замыкает эту группу Воронежская область. Далее были отобраны регионы с многомерным средним ниже 79 %. Из этих регионов были сформированы две группы: группа A с регионами, для которых $23 < P \leq 50$, и группа B с регионами, для которых $50 < P \leq 79$. В группу A включены 15 регионов с очень низким значением P, т. е. эти регионы могут считаться субъектами со слабым уровнем инновационного развития: Республика Хакасия, Республика Калмыкия, Республика Дагестан и т. д. Соответственно, в группу B вошли 27 регионов с невысоким уровнем инновационного развития: Астраханская, Кировская, Амурская области, замыкает группу Республика Карелия. В группу D с высоким уровнем инновационного развития были включены регионы с многомерным средним выше 121 % (13 регионов): Нижегородская, Новосибирская,

Томская области и т. д. Самый высокий уровень инновационного развития в рамках нашей системы показателей имеет г. Москва – 551,69 %. Таким образом, построена классификация 74 российских регионов по уровню инновационного развития, представленная в табл. 11 – 12. Отметим, что построенная классификация согласуется с рейтингом инновационной активности российских регионов, составленным Национальной ассоциацией инноваций и развития информационных технологий. Графики ранжированного вариационного ряда значений многомерного среднего $P(\%)$ и соответствующих значений результативного показателя $Y(\%)$ представлены на рис. 6.

Таблица 10. Классификация регионов по многомерному среднему показателю

Группа	**Оценка многомерного среднего *P***	**Количество регионов**	$\bar{P}$**, %**	$\bar{Y}$**, %**
A	$23 < P \leq 50$	15	39,00	28,01
B	$50 < P \leq 79$	27	63,43	60,21
C	$79 < P \leq 121$	19	98,98	132,98
D	$121 < P \leq 552$	13	247,83	217,51

Таблица 11. Группы регионов и парная выборка (P, Y)

Номер региона	**Регион**	***P*, %**	***Y*, %**	***Y*, млн руб./ 10 тыс. занятых**
Группа А				
1	Оренбургская область	23,76	28,86	82,90
2	Республика Хакасия	27,65	0,43	1,23
3	Забайкальский край	30,50	51,60	148,21
4	Республика Калмыкия	34,66	0,08	0,22
5	Брянская область	35,21	42,32	121,55
6	Кемеровская область	35,23	8,66	24,88
7	Вологодская область	38,54	104,62	300,50
8	Республика Алтай	39,53	0,15	0,44
9	Республика Дагестан	40,52	0,08	0,22
10	Костромская область	44,28	25,88	74,35
11	Курганская область	45,24	32,02	91,97
12	Ставропольский край	45,73	64,13	184,22
13	Смоленская область	46,55	38,00	109,15
14	Псковская область	48,52	5,82	16,72
15	Республика Марий Эл	49,05	17,49	50,23

Продолжение табл. 11

Номер региона	**Регион**	***P*, %**	***Y*, %**	***Y*, млн руб./ 10 тыс. занятых**
Группа B				
16	Липецкая область	50,72	351,47	1009,58
17	Астраханская область	51,57	24,27	69,72
18	Кировская область	52,22	53,78	154,49
19	Амурская область	52,64	43,51	124,99
20	Карачаево-Черкесская Республика	52,93	3,55	10,20
21	Алтайский край	53,13	28,64	82,27
22	Республика Адыгея	55,33	58,43	167,85
23	Краснодарский край	55,56	3,24	9,30
24	Еврейская автономная область	56,47	0,02	0,05
25	Белгородская область	58,86	105,64	303,43
26	Калининградская область	60,47	2,89	8,31
27	Чувашская республика	62,68	98,96	284,25
28	Республика Северная Осетия	62,74	0,107	0,31
29	Волгоградская область	62,93	17,55	50,41
30	Удмуртская Республика	65,60	73,64	211,53
31	Кабардино-Балкарская Республика	66,19	10,49	30,14
32	Рязанская область	66,54	41,24	118,45
33	Новгородская область	67,65	67,18	192,98
34	Тульская область	68,42	161,96	465,20
35	Хабаровский край	68,43	116,93	335,88
36	Тюменская область	70,73	24,93	71,61
37	Тамбовская область	72,20	18,12	52,06
38	Республика Бурятия	72,95	47,79	137,26
39	Омская область	73,85	83,93	241,09
40	Республика Коми	75,53	174,83	502,18
41	Орловская область	77,37	10,16	29,17
42	Республика Карелия	78,90	2,43	6,98
Группа C				
43	Республика Мордовия	80,46	267,52	768,42
44	Курская область	82,87	52,45	150,65
45	Республика Башкортостан	86,36	146,86	421,84
46	Саратовская область	86,53	38,59	110,84
47	Ростовская область	86,57	102,27	293,77
48	Пензенская область	86,81	50,07	143,81
49	Красноярский край	91,88	131,64	378,12
50	Ленинградская область	94,74	65,28	187,52
51	Иркутская область	95,36	15,15	43,50
52	Челябинская область	98,18	156,72	450,15
53	Владимирская область	102,13	123,91	355,93

Окончание табл. 11

Номер региона	**Регион**	***P*, %**	***Y*, %**	***Y*, млн руб./ 10 тыс. занятых**
54	Тверская область	102,20	110,01	315,99
55	Республика Саха (Якутия)	103,50	67,66	194,34
56	Ивановская область	106,31	3,29	9,46
57	Пермский край	108,11	508,31	1460,07
58	Республика Татарстан	113,74	617,33	1773,23
59	Камчатский край	116,11	9,89	28,41
60	Мурманская область	118,45	15,10	43,36
61	Воронежская область	120,33	44,53	127,92
Группа D				
62	Приморский край	121,92	8,01	22,99
63	Ярославская область	134,23	125,23	359,70
64	Самарская область	144,29	553,67	1590,36
65	Свердловская область	145,50	165,37	475,01
66	Ульяновская область	180,71	169,81	487,76
67	Магаданская область	183,52	244,99	703,72
68	Калужская область	214,46	113,10	324,86
69	Нижегородская область	241,18	356,63	1024,39
70	Новосибирская область	243,76	87,09	250,16
71	Томская область	301,05	76,23	218,96
72	Московская область	343,89	277,25	796,39
73	г. Санкт-Петербург	415,60	290,87	835,51
74	г. Москва	551,69	359,38	1032,28

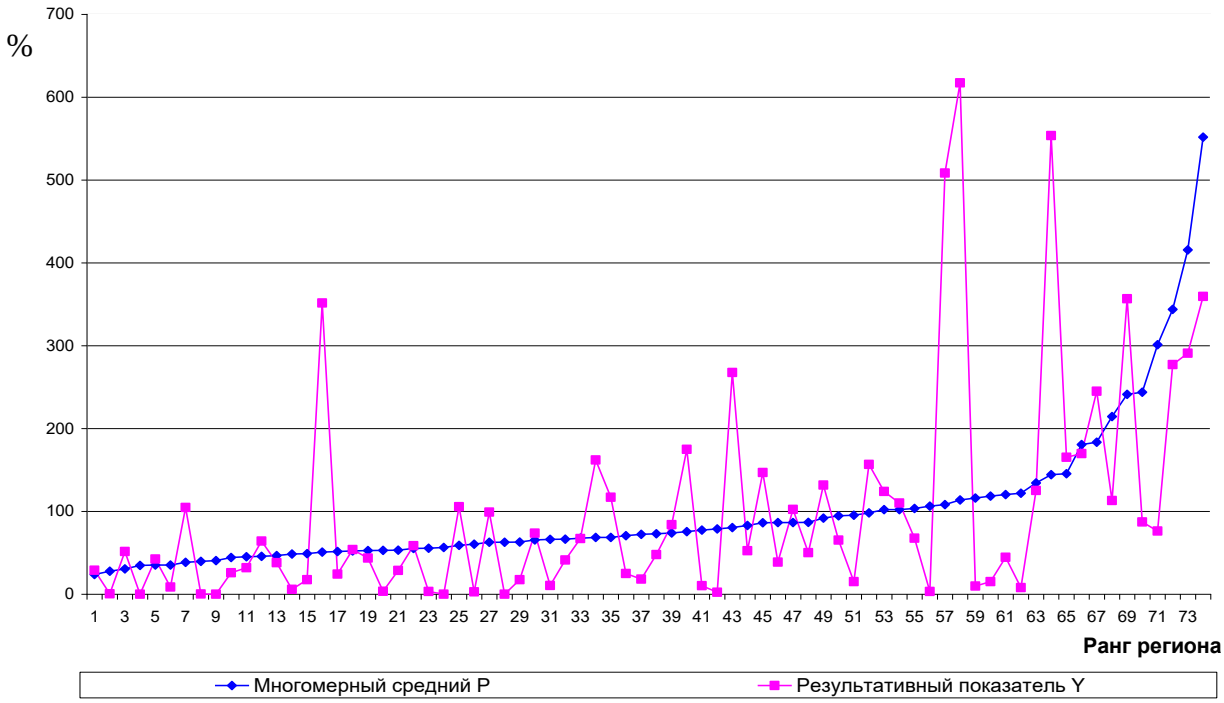

Рис. 6. График двумерной выборки (*P*, *Y*) %

На третьем этапе исходя из непараметрических критериев однофакторного анализа Крускала – Уоллиса и Джонкхиера [12, с. 181 – 183] на уровне значимости 0,05 была принята гипотеза: различие сдвига существует, и медианы выборок групповых значений переменной *Y* упорядочены по возрастанию, т. е. группы *A, B, C, D*, образованные по факторному признаку *P*, в целом различны по результативному показателю *Y* системы инноваций. Это подтверждает статистическую значимость построенной группировки. Таким образом, обосновано влияние многомерного среднего *P* на показатель *Y*. В этой связи по выборке средних значений ($\overline{P}$, $\overline{Y}$) % в сформированных группах было рассчитано уравнение линейной модели регрессии $\tilde{Y} = 12{,}73 + 0{,}8632P$, адекватной экспериментальным данным по *F*-критерию. Полученное уравнение показывает, что увеличение многомерного среднего показателя инновационного развития региона на 1 % влечет за собой рост объема инновационных товаров, работ и услуг в среднем на 0,86 %. При этом отметим, что расчет параметров регрессии на основе группировки приближенный, поскольку реальные значения факторов заменены здесь их средними значениями.

Анализ парной выборки значений (*P*, *Y*) % показал, что большинство регионов с высоким значением многомерного среднего имеют и высокий результативный показатель инноваций. Встречаются и аномальные случаи, когда регионы с большим значением *Y* имеют невысокий многомерный средний на фоне показателей других регионов. К таким регионам относятся: Липецкая область (*Y* = 351 %; *P* = 51 %), Тульская область (*Y* = 162 %; *P* = 68 %). Некоторые регионы со средним уровнем показателя *P* могут обеспечивать достаточно высокий показатель *Y*: Республика Татарстан (*Y* = 617 %; *P* = 114 %), Пермский край (*Y* = 508 %; *P* = 108 % – немногим выше среднего), что указывает на достаточно эффективное использование инновационных факторов в целях экономического развития регионов. Часть регионов с многомерным средним показателем, бóльшим 100 %, имеют значение *Y*, бóльшее 100 %. Количество таких регионов равно 14 (в основном регионы группы *D*). Состав группы *D* показывает, что высокий уровень инновационного развития этих регионов связан прежде всего с высоким уровнем их научного потенциала. К аналогичным выводам приходят и другие современные исследователи на основе методов дис-

персионного анализа и экспертных оценок. Отметим также, что Еврейская автономная область, Республика Дагестан, Республика Калмыкия, Республика Северная Осетия, Республика Алтай имеют очень низкие значения результативного показателя (менее 0,5 %), в то время как значение многомерного среднего этих регионов находится в диапазоне $34 < P < 63$. Последние значения этих показателей указывают на незначительное использование инновационных факторов в этих регионах.

На следующем этапе с целью выявления группировок среди регионов со средним уровнем инновационного развития, близких по всей совокупности факторов инновационной деятельности (включая результативный показатель системы), был проведен кластерный анализ. В многомерном пространстве переменных (X_1, X_2, …, X_9) нормированное эвклидово расстояние между объектами-регионами с номерами i и j вычислялось по формуле: $d(i,j)=\sqrt{\sum_{k=1}^{9}\left(\left(x_{ik}-x_{jk}\right)^2/s_k^2\right)}$ [12, с. 357]. Здесь $x_{ik}(k = 1, …, 9)$ и $x_{jk}(k = 1, …, 9)$ – координаты (значения факторов) для i-го и j-го регионов соответственно; номера регионов принимают значения $i = 43, …, 61$; $j = 43, …, 61$ (согласно табл. 11). Расчеты показали, что наименее удаленными друг от друга оказались следующие пары регионов: Пензенская и Челябинская области, Владимирская и Тверская области, Камчатский край и Мурманская область, Курская и Ростовская области, для которых расстояние соответственно равно: $d(48,52) = 1,34$; $d(53,54) = 1,46$; $d(59,60) = 1,57$; $d(44,47) = 1,58$. Дендрограммы по стратегиям Уорда и дальнего соседа обнаружили формирование четырех-пяти кластеров. Далее в целях сравнения группировок с различным количеством кластеров была применена дивизивная стратегия (с эвклидовой метрикой), показавшая, что при переходе от пяти к шести кластерам среднее внутрикластерное расстояние уменьшается незначительно. По этой причине за основу кластеризации мы приняли разбиение регионов группы *C* на следующие пять кластеров: (45, 53, 54, 58, 43); (46, 56); (44, 47, 49, 61); (48, 52, 57); (50, 51, 55, 59, 60), где геометрическими центрами кластеров предстают следующие регионы: Республика Башкортостан, Сара-

товская, Ростовская, Челябинская, Мурманская области. На основе дискриминантного анализа было показано, что разбиение на данные кластеры статистически достоверно.

Отметим, что для некоторых регионов характерна высокая степень неоднородности значений показателей: высокий процент по одному показателю сочетается с очень низким значением по другому. Так, например, для Ивановской области оценка показателя X_6 составляет 5 %, а для показателя X_8 – 455 %; в то же время результативный показатель Y равен 3 %. В целях моделирования влияния факторных переменных на вариацию результативного показателя системы на пятом этапе часть регионов пришлось исключить из исследования. Из группы *A* были исключены три региона (7, 14, 15); из группы *B* – восемь регионов (16, 23, 24, 26, 28, 31, 41, 42); из группы *C* – два региона (43, 56); из группы *D* – один регион (64). Поскольку результаты статистического анализа по парной выборке (P, Y) не могут дать полного понимания роли отдельных факторов X_1, X_2, …, X_8 в формировании вариации результативного показателя Y, пятый этап работы был посвящен корреляционно-регрессионному анализу показателей (X_1, X_2, …, X_8, Y).

В рамках группы *A* расчет матрицы парных линейных коэффициентов корреляции показал, что результативный показатель системы Y связан статистически значимой корреляцией (по *t*-критерию на уровне значимости 0,05) с факторными переменными X_5, X_6. Значимыми положительными коэффициентами корреляции связаны между собой факторные переменные (X_1, X_5); (X_2, X_3), для которых коэффициенты корреляции следующие: $r(X_1, X_5) = 0{,}65$; $r(X_2, X_3) = 0{,}63$.

Для регионов группы *B* значимая корреляция с Y выявлена для фактора X_5. Кроме того, статистически значимы коэффициенты корреляции факторов (X_2, X_4); (X_3, X_6). Здесь $r(X_2, X_4) = 0{,}83$; $r(X_3, X_6) = 0{,}57$.

В группе *C* с результативным показателем имеют значимую связь факторы X_6, X_8. Отметим, что в рамках группы *C* также значимо коррелируют между собой факторы (X_1, X_8); (X_2, X_4); (X_3, X_6); (X_4, X_5), для которых $r(X_1, X_8) = 0{,}72$; $r(X_2, X_4) = 0{,}80$; $r(X_3, X_6) = 0{,}67$; $r(X_4, X_5) = 0{,}58$. Таким образом, самая высокая корреляция проявилась для факторов X_2 и X_4.

Для регионов группы *D* значимо коррелируют с показателем Y факторы X_3, X_5, X_6, X_7. Кроме того, обнаружены следующие стати-

стически значимые коэффициенты корреляции (между факторными переменными): $r(X_1, X_2) = 0{,}85$; $r(X_1, X_8) = 0{,}85$; $r(X_2, X_3) = 0{,}72$; $r(X_2, X_4) = 0{,}70$; $r(X_2, X_8) = 0{,}83$; $r(X_3, X_5) = 0{,}64$; $r(X_3, X_6) = 0{,}95$; $r(X_3, X_8) = 0{,}74$; $r(X_5, X_8) = 0{,}66$; $r(X_6, X_8) = 0{,}69$. Так в группе с высоким уровнем инновационного развития наблюдается наиболее выраженная взаимная корреляционная связь между факторными переменными.

Далее в рамках групп *A*, *B*, *C*, *D* на основе процедуры пошаговой регрессии были отобраны факторные переменные, вносящие наибольший самостоятельный вклад в вариацию результативного признака *Y*, и построены линейные модели регрессии, описывающие зависимость объемов инновационных товаров, работ и услуг от соответствующих факторных переменных. Результаты моделирования и статистические характеристики моделей (*R* – множественный коэффициент корреляции и расчетный *F*-критерий) представлены в табл. 12.

Таблица 12. Модели регрессии

Группа	**Уравнение модели**	**Статистические характеристики**	
		R	***F***
A	$\tilde{Y} = -10{,}23 + 1{,}001X_5 + 2{,}239X_6$	0,78	7,09
B	$\tilde{Y} = 17{,}91 + 1{,}285X_5$	0,58	8,67
C	$\tilde{Y} = -55{,}45 + 1{,}763X_8$	0,53	5,92
D	$\tilde{Y} = -77{,}42 + 0{,}1961X_5 + 0{,}2174X_6 + 0{,}8963X_7$	0,93	16,34

Для регионов группы *A* около 61 % вариации результативного показателя учтено в линейной модели и обусловлено влиянием факторов X_5 и X_6, при этом одновременный рост этих факторных переменных на 1 % приводит к росту показателя *Y* в среднем на 3 %; в группе *B* около 34 % вариации *Y* модель объясняет влиянием фактора X_5, увеличение X_5 на 1 % приводит к росту *Y* в среднем на 1,3 %; в группе *C* около 28 % вариации *Y* объясняется влиянием фактора X_8, рост переменной X_8 на 1 % влечет за собой возрастание *Y* на 1,8 %; в группе *D* модель объясняет 86 % вариации *Y* за счет влияния факторов (X_5, X_6, X_7), причем одновременный рост факторных переменных на 1 % влечет за собой увеличение результативного показателя примерно на 1,3 %.

Анализ остатков по уравнениям регрессии показал наибольшее отставание от регрессионных значений результативного показателя для регионов: в группе *A* – Республики Калмыкии и Кемеровской области; в группе *B* – Тюменской, Астраханской, Волгоградской областей; в группе *C* – Курской и Воронежской областей; в группе *D* – Томской и Калужской областей. Значительно выше регрессионных значений показатель *Y* достигает для регионов: в группе *A* – Забайкальского края; в группе *B* – Республики Коми, Чувашской Республики и Хабаровского края; в группе *C* – Республики Татарстан и Пермского края; в группе *D* – Новосибирской, Нижегородской областей и г. Москвы. Отставание *Y* от регрессионных значений говорит о том, что включенные в модели факторы инновационной инфраструктуры в недостаточной степени используются для достижения хорошего результата. В то же время опережение регрессионных значений говорит о рациональном использовании имеющихся ресурсов для достижения высоких значений объема инновационных товаров, работ и услуг.

Из результатов расчетов следует:

- на результативный показатель системы существенное влияние оказывают объемы затрат на фундаментальные и прикладные исследования;

- лидерами по факторам инновационной деятельности вышли регионы с сильным научным потенциалом;

- в рамках группы с высоким уровнем инновационного развития корреляционные связи между наблюдаемыми переменными оказались сильнее выражены, чем в других группах;

- разрыв между регионами по показателям инновационной деятельности весьма значительный (отдельные значения показателей отличаются в десятки, а то и в сотни раз);

- содержательный экономический анализ группировок объектов, однородных по инновационному потенциалу, может быть плодотворным материалом для разработки программ инновационного развития территорий.

В дальнейшем предполагается классифицировать регионы на основе инструментов факторного и кластерного анализа. Комбинация

универсальных многомерных методов математической статистики может эффективно работать в целях количественной и качественной оценки инновационного развития российских регионов.

Контрольные вопросы

1. Что означают понятия «корреляция», «регрессия», «уравнение регрессии»? Каковы цели и задачи корреляционно-регрессионного анализа?

2. Укажите формулы для расчета выборочного коэффициента линейной корреляции Пирсона, коэффициента множественной корреляции, коэффициента детерминации. Каков диапазон изменения этих коэффициентов и их практический смысл?

3. Что представляет собой корреляционная матрица?

4. Укажите формулы для расчета коэффициентов множественной корреляции и детерминации. Каков диапазон изменения этих коэффициентов и их практический смысл?

5. Какова сущность метода наименьших квадратов (МНК)? На основе МНК выведите формулы для расчета параметров линейной и параболической моделей парной регрессии.

6. Каковы предпосылки МНК?

7. Каковы наиболее употребительные модели парной регрессии? Дайте рекомендации, для моделирования каких зависимостей их целесообразно использовать.

8. Какие виды нелинейных зависимостей поддаются линеаризации подходящей заменой переменных?

9. Какие вы знаете модели множественной регрессии? Приведите примеры моделей линейных и нелинейных относительно независимых переменных; также линейных и нелинейных по параметрам.

10. Как осуществляется проверка адекватности модели регрессии на основе F-критерия? Как выполняется оценка статистической значимости параметров регрессии? Какова практическая интерпретация значимых коэффициентов линейной модели регрессии? Что характеризуют Δ-коэффициенты модели множественной регрессии?

Глава 3. КЛАСТЕРНЫЙ АНАЛИЗ

3.1. Назначение кластерного анализа

«Кластерный анализ» – это общее название множества вычислительных процедур, используемых при создании классификации. В результате работы с процедурами образуются кластеры, или группы очень похожих объектов. Более точно, *кластерный метод* – это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Отметим, что кластерный анализ используют также в случаях, если необходимо построить группировку не только объектов, но и переменных. В зависимости от типа исходных данных для кластеризации осуществляется выбор расстояния или метрики.

Количественное оценивание сходства объектов отталкивается от понятия метрики. При этом объекты представляются точками координатного пространства, причем сходства и различия между точками определяются в соответствии с метрическими расстояниями между ними. Размерность метрического пространства совпадает с числом переменных, использованных для описания объектов. Известны четыре стандартных критерия, которым должна удовлетворять мера сходства, чтобы быть метрикой:

1) *симметрия*: расстояние между любыми двумя объектами x и y удовлетворяет условию: $\rho(x, y) = \rho(y, x) \geq 0$;

2) *неравенство треугольника*: для любых трёх объектов (x, y, z) расстояние удовлетворяет условию: $\rho(x, y) \leq \rho(x, z) + \rho(y, z)$;

3) *различимость нетождественных объектов*: если $x \neq y$, то $\rho(x, y) \neq 0$;

4) *неразличимость идентичных объектов*: для двух идентичных объектов x и x' $\rho(x, x') = 0$.

Два объекта идентичны, если описывающие их переменные принимают одинаковые значения. В этом случае расстояние между ними равно нулю. Меры расстояния обычно не ограничены сверху и зависят от выбора шкалы (масштаба) измерений.

Одни из наиболее известных расстояний – эвклидово и нормированное эвклидово расстояние. Нормированное эвклидово расстоя-

ние между I-м и J-м объектами (I = 1, 2, …, n; J = 1, 2, …, n) вычисляется по формуле

$$\rho(I,J)=\sqrt{\sum_{k=1}^{p}\frac{\left(x_{ik}-x_{jk}\right)^2}{s_k^2}},$$

где (x_{i1}, x_{i2}, …, x_{ip}) и (x_{j1}, x_{j2}, …, x_{jp}) – факторные координаты объектов; (s_1^2, s_2^2, ..., s_p^2) – несмещенные оценки дисперсий соответствующих переменных. Нормировку вводят в случае факторных переменных, значительно различающихся по величине и измеренных в разных единицах для отдельных факторов. Эвклидово расстояние рассчитывается по формуле $\rho(I,J)=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2}$ и применяется в случае переменных одинаковой размерности. Определяют и другие расстояния, в зависимости от постановки задачи и типа переменных [12, с. 353 – 357]. Так, хорошо известной мерой является «манхеттеновское» расстояние, или «расстояние городских кварталов», которое обычно применяется для номинальных или качественных переменных и определяется следующим образом:

$$\rho(I,J)=\frac{1}{n}\sum_{k=1}^{p}\left|\left(x_{ik}-x_{jk}\right)\right|,\quad i=1,...,n;\quad j=1,...,n.$$

Отметим, что имеет значение выбор переменных для проведения кластерного анализа. В исследовании очень важно, чтобы результаты кластеризации наилучшим образом отражали сходства объектов. Здесь окончательный выбор и интерпретация результатов обработки данных должна стоять за специалистом в данной области исследования. Главная цель кластерного анализа – нахождение групп схожих объектов в выборке данных. Эти группы называют *кластерами*.

Кластеры обладают следующими свойствами: плотностью, дисперсией, размерами, формами и отделимостью. *Плотность* – это свойство, которое определяет кластер как скопление точек в пространстве данных, относительно плотное по сравнению с другими областями пространства, содержащими либо мало точек, либо не содержащими других точек. В исследовании *дисперсия* представляет собой характеристику степени рассеяния точек в пространстве относительно центра кластера. В этой связи кластер называют «плотным», если все его точки находятся вблизи его же геометрического центра

(т. е. центра тяжести), и «неплотным», если они весьма разбросаны вокруг центра. Понятие *размера* кластера тесно связано с определением дисперсии. Если кластер можно идентифицировать, то можно и измерить его «радиус», что полезно в случаях, если рассматриваемые кластеры представляют собой гиперсферы (т. е. имеют круглую форму) в многомерном пространстве исследуемых переменных. Под *формой* кластера понимают расположение его точек в пространстве. Кластеры можно представить в форме эллипсоидов, гиперсфер или удлинённых кластеров. Отделимость характеризует степень перекрытия кластеров и расстояние друг от друга в пространстве. Так, кластеры могут быть относительно близки друг к другу и не иметь четких границ или же разделены широкими участками пространства.

3.2. Методы кластерного анализа

Среди основных методов кластерного анализа, разработанных в научной литературе, можно выделить следующие: иерархические агломеративные методы; дивизивные методы; методы поиска модальных значений плотности; факторные аналитические методы; методы сгущений; методы, использующие теорию графов. Выбор метода зависит от области исследования. В то же время применение разных методов к одним и тем же исходным данным может привести к очень различным результатам кластеризации. В конкретных отраслях науки могут оказаться особенно полезными определённые методы. Например, иерархические агломеративные методы часто применяют в биологии, факторные аналитические методы пользуются популярностью в психологии. Важно представлять, какой метод и мера сходства могут соответствовать ожидаемой кластеризации. В социологии и экономике успешно применяют иерархические агломеративные, дивизивные и итеративные методы.

Следуя [12, с. 355], укажем объединяющие стратегии кластеризации объектов, которые используют алгоритмы иерархических агломеративных методов:

- стратегия ближайшего соседа очень сильно сжимает пространство исходных переменных и рекомендуется для получения минимального дерева классификации. В этом случае расстояние между двумя

кластерами S_l и S_m определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах:

$$\rho_{\min}(S_l, S_m) = \min_{x_i \in S_l, x_j \in S_m} \rho(x_i, x_j);$$

- стратегия дальнего соседа растягивает пространство так, что удаленные друг от друга объекты становятся ещё более далёкими, при этом расстояние между кластерами определяется наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. наиболее удалёнными соседями):

$$\rho_{\max}(S_l, S_m) = \max_{x_i \in S_l, x_j \in S_m} \rho(x_i, x_j);$$

- стратегия группового соседа сохраняет метрику пространства. Расстояние между кластером S_{i+j} и всеми другими кластерами S_k есть

$$\rho(S_{i+j}, S_k) = \frac{\rho(S_i, S_k)n_i + \rho(S_j, S_k)n_j}{n_i + n_j}, \quad n_i + n_j = n_{i+j};$$

- гибкая стратегия сжимает или растягивает пространство, в зависимости от выбора β-коэффициента расстояние есть

$$\rho(S_{i+j}, S_k) = \frac{(1-\beta)(\rho(S_i, S_k) + \rho(S_j, S_k))}{2 - \beta \cdot \rho(S_i, S_j)};$$

- метод Уорда минимизирует внутрикластерный разброс объектов и глубоко разделяет кластеры, другими словами, минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге, однако при этом стремится создавать кластеры малого размера. В целом данный метод очень эффективен.

Таким образом, по ходу иерархических агломеративных процедур строится последовательность возрастающих по числу объектов кластеров и рассчитывается расстояние, на уровне которого происходит образование нового кластера. Это правило позволяет нанизывать объекты вместе для формирования кластеров, при этом результирующие кластеры имеют тенденцию быть представленными длинными цепочками. В агломеративных процедурах начальным является разбиение, состоящее из одноэлементных классов, а конечным – из одного кластера. К недостаткам иерархических процедур следует отнести громоздкость их вычислительной реализации, большинство этих алгоритмов исходит из матрицы расстояний (сходства) между объектами.

Программа STADIA реализует алгоритмы иерархической классификации и предусматривает графическое представление результатов кластеризации в виде дендрограммы. *Дендрограмма* – это дерево объединений кластеров с порядковыми номерами объектов по горизонтальной оси и шкалой расстояний по вертикальной оси.

На основе разделяющей (дивизивной) стратегии группируют объекты в заданное количество кластеров. В западных статистических пакетах прикладных программ аналоги дивизивной стратегии называют методом *K*-средних или анализом центров. Этот метод кластеризации существенно отличается от аггломеративных методов. Для реализации данного метода нужно заранее иметь представление о конечном количестве кластеров. В начале решения задачи необходимо определить количество *K* кластеров так, чтобы они были настолько различны, насколько это возможно. В общем случае метод *K*-средних строит ровно *K* различных кластеров, расположенных на возможно бóльших расстояниях друг от друга. Принцип его работы заключается в перемещении объектов в кластер с ближайшим центром тяжести. Иллюстрация работы алгоритма при $K = 2$ показана на рис. 7.

Таким образом, алгоритм состоит из следующих шагов:

1) выбирают *K* точек, которые считаются центрами или центроидами кластеров, и объекты распределяют по кластерам так, что в результате каждый назначен определенному кластеру;

2) вычисляют центры кластеров, которые затем и далее являются покоординатными средними кластеров;

3) объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий: кластерные центры стабилизировались, т. е. все наблюдения принадлежат кластеру, которому принадлежали до текущей операции; число итераций достигло максимально возможного значения. *Общая идея алгоритма*: наблюдения сопоставляются с кластерами так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. *Достоинства* алгоритма *K*-средних: простота и быстрота использования; понятность и прозрачность. *Недостатки* алгоритма *K*-средних: алгоритм слишком чувствителен к выбросам, которые могут искажать среднее.

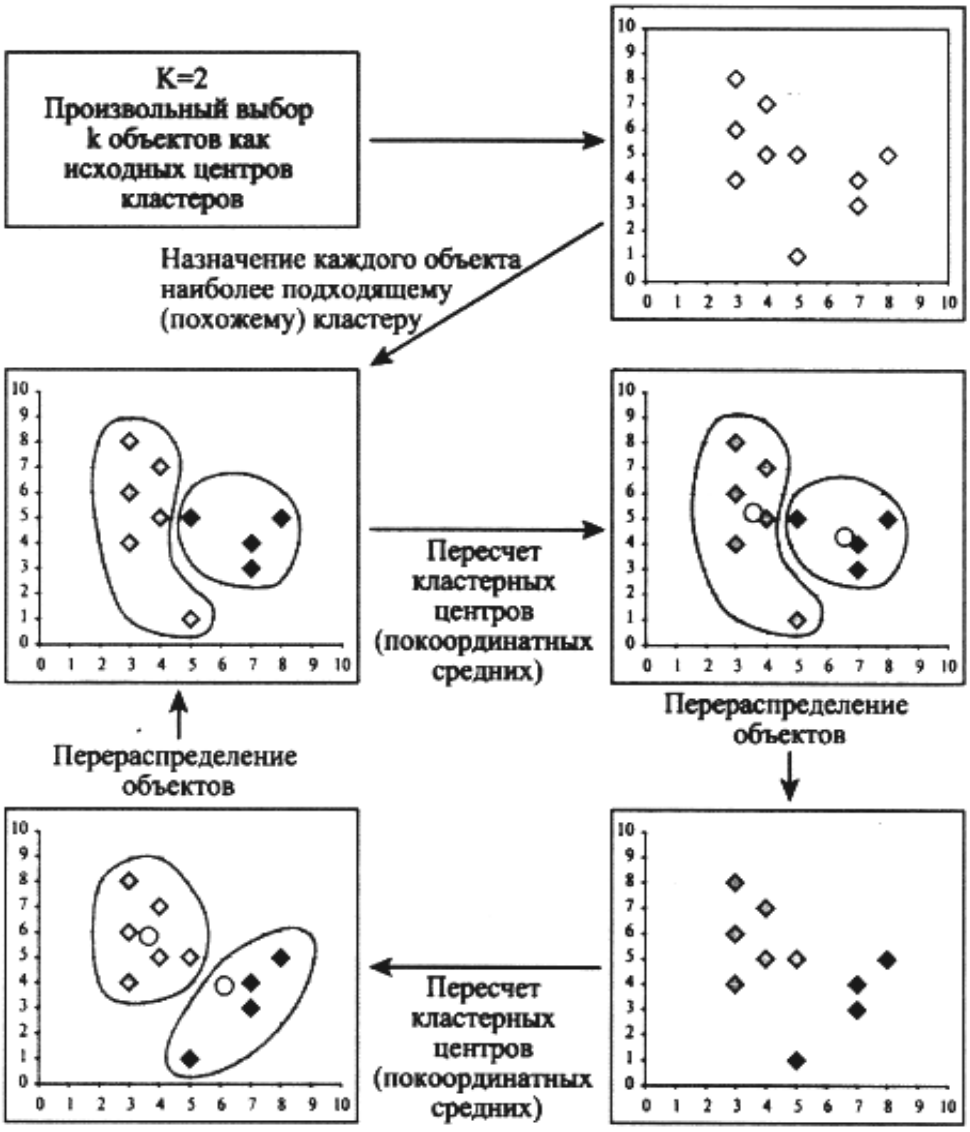

Рис. 7. Схема алгоритма *K*-средних

3.3. Пример решения задачи классификации объектов методом кластерного анализа

Постановка задачи. Имеется 15 предприятий, по каждому из которых известны данные об объёмах продаж (млн руб.) пяти видов товаров. Требуется на основе агломеративной стратегии, используя нормированную эвклидову метрику, образовать последовательность кластеров возрастающей общности предприятий по объёмам продаж. Также произвести кластеризацию объектов исследования (предприятий) на основе дивизивной стратегии с целью выявления имеющихся группировок. Исходные данные (условные) представлены в табл. 13.

Таблица 13. Исходные данные задачи кластеризации объектов

Номер показателя / Номер объекта	*X*1	*X*2	*X*3	*X*4	*X*5
1	751	584	363	197	453
2	2317	2141	2588	1375	384
3	2300	1104	931	1023	1203
4	1904	1454	1272	1128	1028
5	2028	1868	2412	2758	4235
6	423	240	1368	409	373
7	2090	1858	588	1223	1121
8	517	522	1345	801	787
9	2097	1705	1345	2229	1925
10	1187	626	1015	1773	1788
11	1556	965	867	876	812
12	1724	1738	2396	2303	2415
13	1925	1960	1979	1421	1502
14	965	803	1371	1294	1282
15	820	622	1167	1174	1037

Решение данной задачи представим фрагментами протокола выполнения процедур кластерного анализа по программе STADIA.

1. Рассчитываем матрицу расстояний.

Файл. Таблица расстояний

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(2)	5,294													
(3)	3,027	3,133												
(4)	3,032	2,477	1,024											
(5)	7,025	3,813	4,987	4,686										
(6)	1,745	5,027	3,555	3,310	6,807									
(7)	3,359	3,179	1,419	1,296	4,997	4,167								
(8)	1,783	4,443	3,050	2,701	6,003	0,889	3,560							
(9)	4,600	2,522	2,328	2,000	3,135	4,790	2,127	3,996						
(10)	2,930	3,968	2,304	2,208	4,461	2,957	2,774	2,174	2,424					
(11)	1,866	3,633	1,276	1,236	5,526	2,416	1,844	1,922	2,895	1,914				
(12)	5,439	2,146	3,541	2,964	2,167	5,080	3,591	4,295	1,816	3,117	3,875			
(13)	4,346	1,172	2,327	1,535	3,673	4,258	2,240	3,575	1,708	2,950	2,684	1,849		
(14)	2,403	3,578	2,276	1,851	4,865	2,081	2,749	1,243	2,786	1,156	1,479	3,181	2,607	
(15)	1,932	4,085	2,475	2,173	5,417	1,652	2,967	0,847	3,261	1,368	1,450	3,758	3,119	0,586

2. По методу Уорда объединяем объекты в кластеры возрастающей общности и указываем величину расстояния объединения. Стро-

им дендрограмму (рис. 8). Здесь по горизонтали указаны номера объектов, по вертикали – расстояние объединения.

Файл. Метод Уорда.
К л а с т е р ы.
(список объектов) -> расстояние
(15,14) --> 0,5859
(8,6) --> 0,8894
(4,3) --> 1,024
(13,2) --> 1,172
(11,4,3) --> 1,325
(15,10,14) --> 1,423
(11,7,4,3) --> 1,687
(12,9) --> 1,816
(8,1,6) --> 1,971
(13,12,9,2) --> 2,513
(15,8,1,6,10,14) --> 3,179
(13,5,12,9,2) --> 3,85
(15,11,7,4,3,8,1,6,10,14) --> 4,965
(15,13,5,12,9,2,11,7,4,3,8,1,6,10,14) --> 8,76

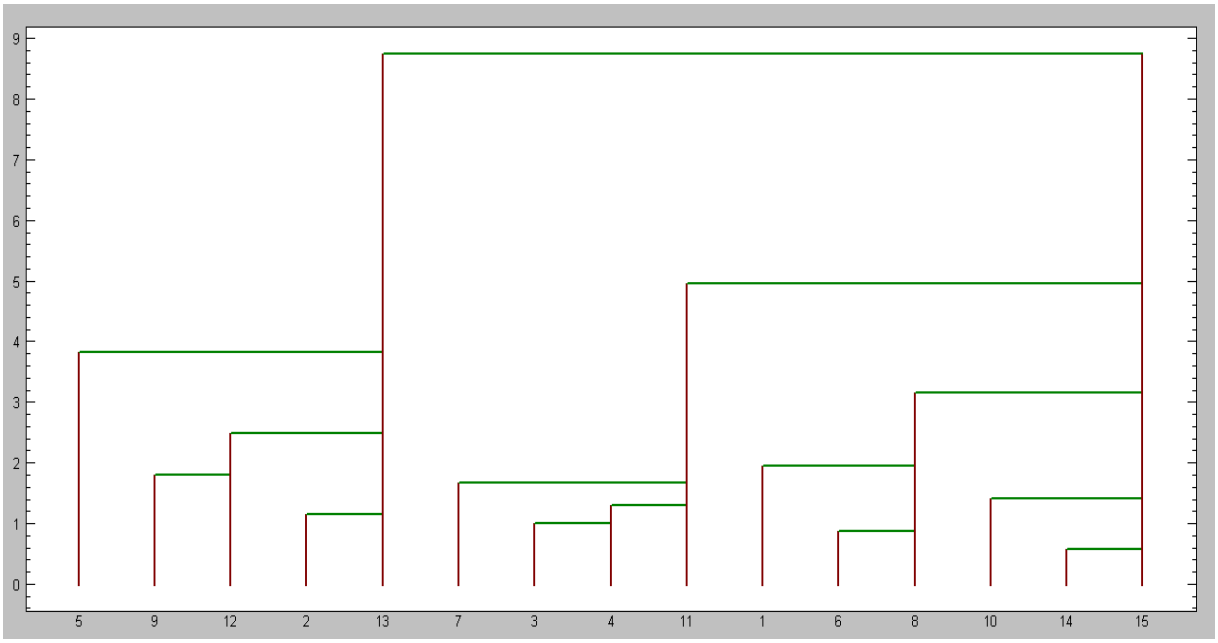

Рис. 8. Дендрограмма по стратегии Уорда

3. Используем стратегию ближайшего соседа и иллюстрируем результаты дендрограммой, представленной на рис. 9.

Файл. Норм. эвклид. + Ближ. сосед
К л а с т е р ы:
(список объектов) -> расстояние
(15,14) --> 0,5859
(15,8,14) --> 0,8465
(15,6,8,14) --> 0,8894
(4,3) --> 1,024
(15,10,6,8,14) --> 1,156
(13,2) --> 1,172
(11,4,3) --> 1,236
(11,7,4,3) --> 1,296
(15,11,7,4,3,10,6,8,14) --> 1,45
(15,13,2,11,7,4,3,10,6,8,14) --> 1,535
(15,9,13,2,11,7,4,3,10,6,8,14) --> 1,708
(15,1,9,13,2,11,7,4,3,10,6,8,14) --> 1,745
(15,12,1,9,13,2,11,7,4,3,10,6,8,14) --> 1,816
(15,5,12,1,9,13,2,11,7,4,3,10,6,8,14) --> 2,167

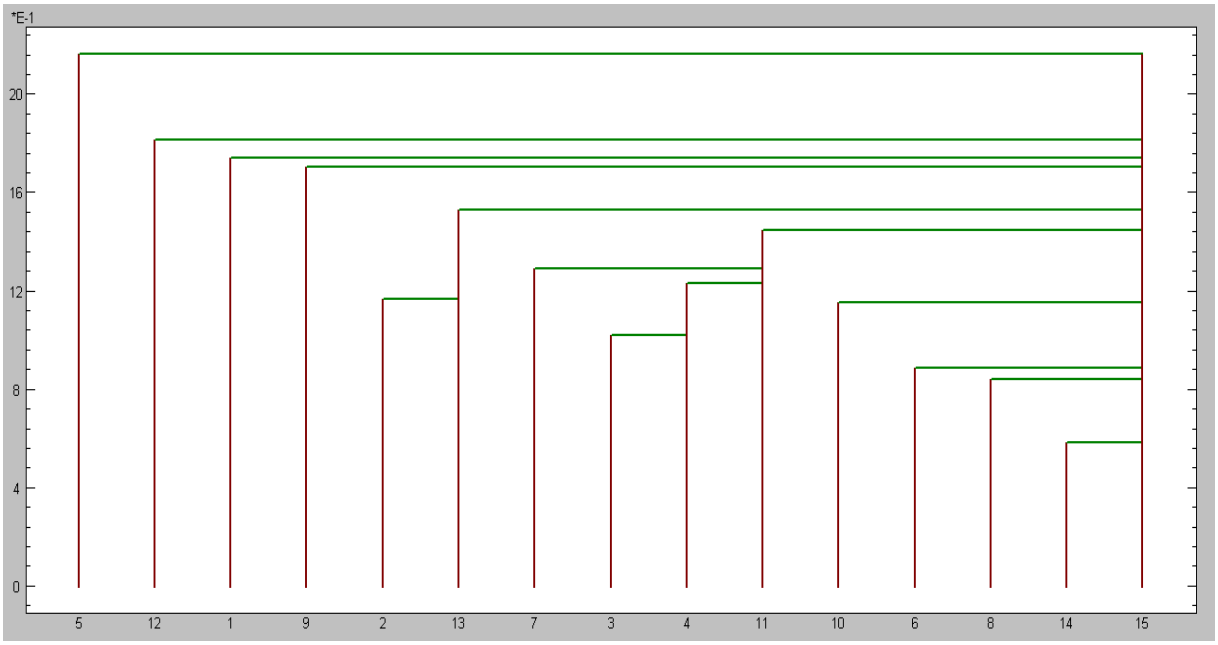

Рис. 9. Дендрограмма по стратегии ближайшего соседа

4. Используем стратегию группового соседа и иллюстрируем результаты дендрограммой, представленной на рис. 10.

Файл: Норм. эвклид. + Групп. сосед
К л а с т е р ы:
(список объектов) -> расстояние

(15,14) --> 0,5859
(8,6) --> 0,8894
(4,3) --> 1,024
(13,2) --> 1,172
(11,4,3) --> 1,256
(15,10,14) --> 1,262
(11,7,4,3) --> 1,52
(8,1,6) --> 1,764
(12,9) --> 1,816
(15,8,1,6,10,14) --> 2,024
(13,12,9,2) --> 2,056
(15,11,7,4,3,8,1,6,10,14) --> 2,608
(13,5,12,9,2) --> 3,197
(15,13,5,12,9,2,11,7,4,3,8,1,6,10,14) --> 3,866

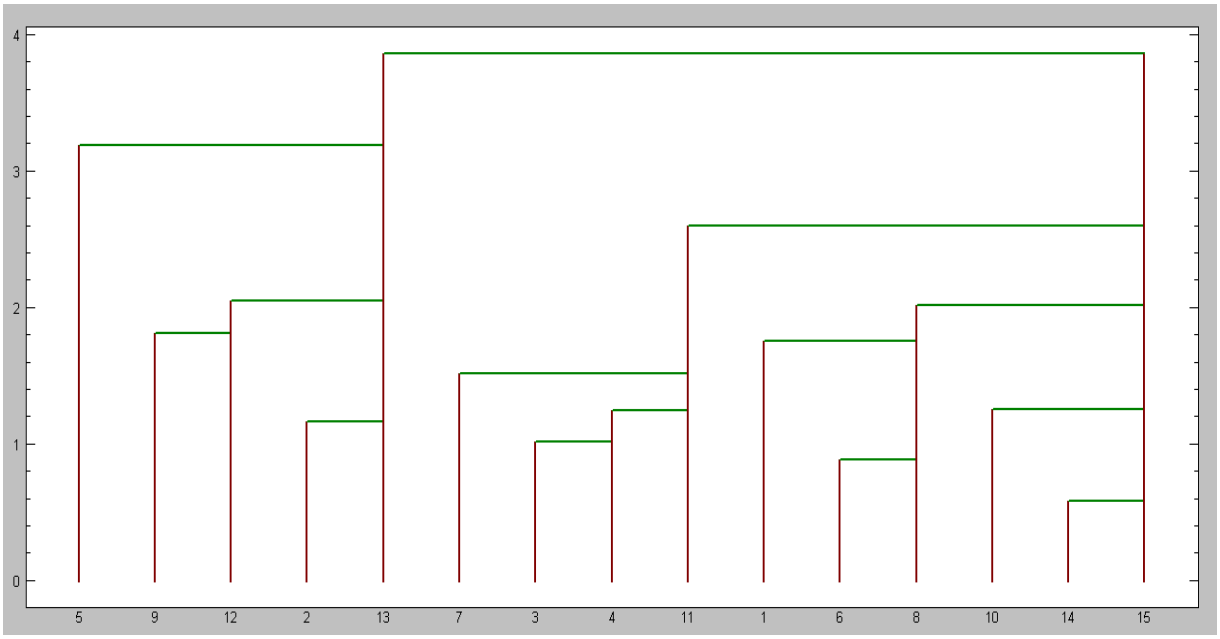

Рис. 10. Дендрограмма по стратегии группового соседа

5. На основе дивизивной стратегии с использованием нормированной эвклидовой метрики получаем разбиение объектов на три кластера предприятий, сходных по реализации продукции: I (5, 9, 12*, 2); II (13, 7, 3, 4*, 11); III (1, 6, 8, 10, 14, 15*). Здесь указаны номера предприятий, символом «*» обозначены геометрические центры кластеров в пятимерном пространстве переменных, характеризующих объёмы продаж. Полученные группировки объектов можно заметить на дендрограмме (см. рис. 10). На рис. 11 показаны результаты дивизивной кластеризации в проекции на плоскость (X_1, X_2).

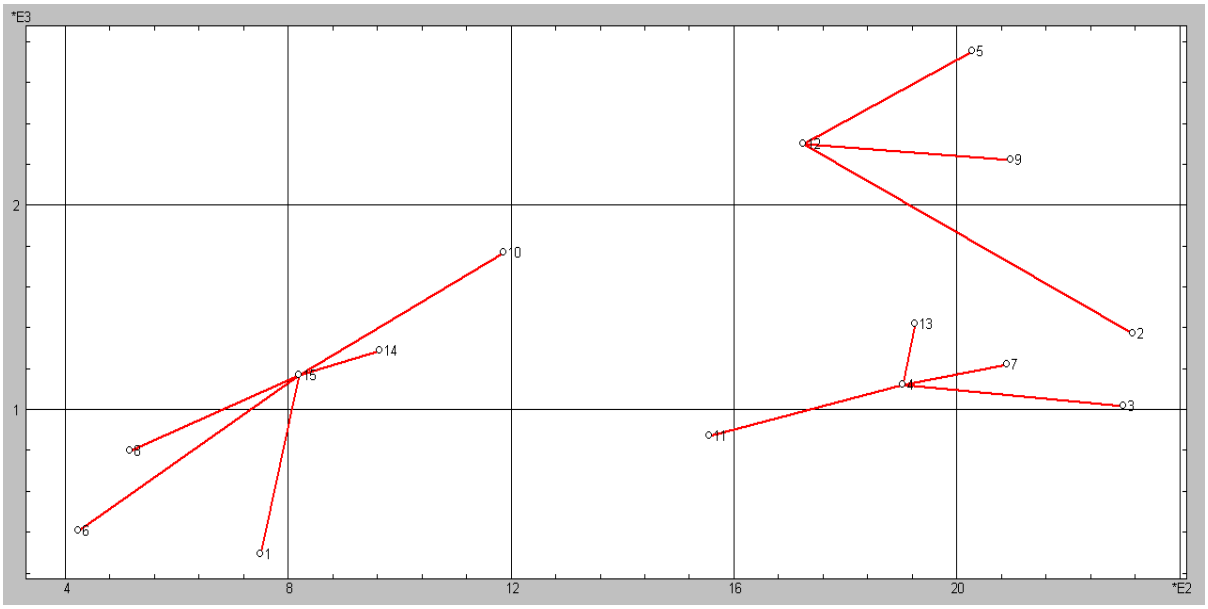

Рис. 11. Проекции объектов на плоскость (X_1, X_2)

Контрольные вопросы

1. Каковы цели и задачи кластерного анализа?
2. В какой форме представляются исходные данные для выполнения процедур кластерного анализа с помощью программы STADIA?
3. На чём основано количественное оценивание сходства объектов?
4. От чего зависит выбор расстояния между объектами исследования?
5. Укажите четыре стандартных критерия, которым должна удовлетворять мера сходства, чтобы быть метрикой.
6. Приведите формулы и поясните все обозначения для расчета эвклидового и нормированного эвклидового расстояний между объектами.
7. Что представляют собой кластеры? Каковы их основные свойства?
8. Какие стратегии кластеризации объектов вы знаете?
9. Чем отличается аггломеративная стратегия от дивизивной?
10. Что представляет собой дерево кластеризации объектов по аггломеративной стратегии?
11. Что представляет собой и в каких целях используется дендрограмма?
12. Поясните алгоритм построения кластеров методом «К-средних».
13. Каковы достоинства и недостатки алгоритма «К-средних»?
14. Каким методам кластеризации объектов отдают предпочтение при исследованиях в биологии, психологии, социологии, экономике?

Глава 4. ФАКТОРНЫЙ АНАЛИЗ

4.1. Теоретические аспекты факторного анализа

При изучении сложных многофакторных систем экспериментальных данных возникает необходимость построения сокращённой системы факторных переменных, отвечающих за подавляющую часть вариации исследуемого явления или признака. В таких случаях инструментом исследования выступает факторный анализ, основная задача которого – преобразование исходных данных в систему новых показателей (главных компонент).

Главные компоненты не коррелированы и упорядочены по величине их дисперсий. При этом исходные переменные (X_1, X_2, …, X_m) представляют собой линейную комбинацию главных компонент (F_1, F_2, …, F_m) так, что модель главных компонент представляется в виде $X_s = \sum_{J=1}^{m} u_{sj} F_j$, где u_{sj} ($s = 1, \ldots, m$; $j = 1, \ldots, m$) – коэффициенты перехода от системы переменных X к системе компонент F. В то же время коэффициенты u_{sj} составляют матрицу собственных векторов U корреляционной матрицы R исходных показателей X, т. е. $RU = \lambda U$, где $\lambda = (\lambda_1, \ldots, \lambda_m)$ – вектор собственных значений матрицы R. Затем строится матрица A факторных нагрузок такая, что $R = AA^T$, элементы которой есть $a_{sj} = u_{sj}\sqrt{\lambda_j}$. Причём сумма $\sum_{i=1}^{m} \lambda_i = m$ и собственное значение, равное $\lambda_i = \sum_{s=1}^{m} a_{si}^2$, характеризует вклад i-й главной компоненты в суммарную дисперсию всех признаков. Удельный вклад i-й компоненты составляет $(\lambda_i / m)\,100\,\%$. Обычно для анализа используют k первых компонент, суммарный вклад которых превышает 60 – 70 % общей дисперсии, их называют значимыми, или общими факторами.

Выделение главных компонент отвечает наиболее распространенной, так называемой *канонической модели* факторного анализа, в которой исходные переменные линейно выражаются через факторы, а факторы не коррелированы между собой и со случайными (специфическими) остатками модели, а случайные остатки, в свою очередь, взаимно не коррелированы и нормально распределены. При этом число объектов или наблюдений должно быть не меньше числа переменных, ещё лучше, чтобы их было в 2 – 3 раза больше. Невыполнение

указанного условия может привести к неадекватной модели главных компонент. Для решения задач снижения размерности и классификации объектов используют k первых главных компонент, число которых должно быть значительно меньше числа исходных переменных. При наличии результативного показателя (признака) Y может быть построено уравнение регрессии на главных компонентах.

Следуя [12, с. 317 – 318], укажем последовательные этапы выполнения факторного анализа:

- вычисление *главных* факторных *компонент* как новой координатной системы, расположенной по осям *эллипсоида рассеяния* анализируемых данных;

- волевой (содержательный) выбор в качестве *факторов* тех из главных компонент, которые отвечают за бóльшую часть дисперсии (рассеяния) анализируемых данных;

- коррекция выделенных факторов специальными методами факторного анализа с целью достижения большей адекватности факторной модели исходным данным по ряду критериев;

- вращение выделенных факторов с целью обеспечения их лучшей проецируемости на исходные переменные для облегчения последующей предметной интерпретации;

- содержательная интерпретация факторов в предметных (физических) терминах, что является творческой задачей, выходящей за рамки формального метода, однако она может принести много полезного для дальнейшего понимания объекта исследования.

- продолжением исследования может быть применение методов кластерного, дискриминантного, корреляционно-регрессионного видов анализа в пространстве факторных координат объектов.

4.2. Алгоритм выделения главных факторов

В основе факторного анализа лежит методика выделения главных компонент, которая включает решение следующих задач.

1. Строят матрицу нормированных данных $Z = (z_{ij})$, где $Z_{ij} = \dfrac{x_{ij} - \overline{x}_j}{s_{X_j}}$,

x_{ij} – компоненты матрицы исходных данных, в ней столбцы отвечают

за исходные переменные, строки соответствуют определённым объектам исследования.

2. Рассчитывают матрицу парных коэффициентов корреляции: $R = (1/n)Z^TZ$ – это симметричная матрица, на главной диагонали которой находятся единицы.

3. Находят вектор собственных значений $\lambda = (\lambda_1, \ldots, \lambda_m)$ и соответствующие собственные векторы матрицы R. Собственные значения выстраиваются в порядке возрастания, и определяется матрица нормированных собственных векторов U. Элементы матрицы U равны коэффициентам перехода от системы исходных координат X к системе координат главных факторов F. Так как дисперсия каждой стандартизированной переменной равна единице, то сумма $\sum_{i=1}^{m}\lambda_i = m$, где m есть число всех исходных переменных. Поэтому λ_i/m есть доля дисперсии, отвечающая i-й компоненте. Собственное значение λ_i компоненты с номером i представляет часть общей дисперсии стандартизированных исходных данных.

4. Вычисляют матрицу

$$A = U\begin{pmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \sqrt{\lambda_m} \end{pmatrix}.$$

Элементы матрицы A есть $a_{sj} = u_{sj}\sqrt{\lambda_j}$, они называются факторными нагрузками. Сумма (по всем строкам s) квадратов нагрузок для конкретного фактора F (так как по теореме Пифагора сумма квадратов коэффициентов поворота факторных осей равна единице) равна собственному значению этого фактора. Поэтому по величине факторной нагрузки можно определить, насколько геометрически близка переменная к фактору. Сумма произведений нагрузок двух строк, соответствующих двум переменным, равна коэффициенту корреляции между ними. При этом корреляционная матрица представляется в виде $R = AA^T$.

5. Строят матрицу F, столбцы которой и есть новые переменные – главные компоненты (F_1, F_2, …, F_m), так что нормированные исходные данные $Z_s = \sum_{J=1}^{m} u_{sj} F_j$ ($s = 1, \ldots, m; j = 1, \ldots, m$).

Новые координаты объектов в системе главных факторов служат важным материалом для последующих статистических исследований: выделения основных группировок объектов средствами кластерного анализа; статистической верификации наиболее оптимальной кластеризации объектов методом дискриминантного анализа; статистических оценок парных различий выделенных кластеров; нахождения регрессионных зависимостей для распределений объектов в пространстве главных факторов.

В дальнейшем на основе результатов факторного анализа возможно решение задачи классификации объектов, исходя из k первых главных компонент, а также построение на главных компонентах уравнения регрессии $Y = Y(F_1, F_2, \ldots, F_k)$. Ниже представлено решение задачи кластеризации объектов в пространстве координат главных факторов.

4.3. Решение задачи кластеризации объектов на основе факторного анализа

Предложенный выше алгоритм факторного анализа проиллюстрируем на основе процедур статистического пакета STADIA для решения следующей задачи.

Задача. По многомерной выборке социально-экономических показателей регионов ЦФО Российской Федерации [25] требуется построить классификацию регионов. Исходные данные представлены ниже в табл. 14.

Предметом исследования стали следующие показатели: X_1 – среднегодовая численность занятых в экономике (тыс. чел.); X_2 – среднедушевые денежные доходы в месяц (руб.); X_3 – среднедушевые денежные расходы в месяц (руб.); X_4 – среднемесячная номинальная начисленная заработная плата работников организаций (руб.); X_9 – оборот розничной торговли (млрд руб.); X_{10} – инвестиции в основной капитал (млн руб.). В целях сравнения регионов (объектов исследования) показатели представлены в расчёте на 10 тыс. чел. населения региона.

Таблица 14. Социально-экономические показатели регионов

№ п/п	Регион	X_1	X_2	X_3	X_4	X_9	X_{10}
1	Белгородская область	1547,9	25372	18035	23895	253,67	120,39
2	Брянская область	1233,0	22039	16980	20911	196,19	66,83
3	Владимирская область	1405,6	20569	15230	22581	182,62	75,67
4	Воронежская область	2360,9	25505	19328	24001	422,90	243,26
5	Ивановская область	1036,9	20409	15099	20592	143,72	37,34
6	Калужская область	1010,5	24984	17726	28248	162,84	99,79
7	Костромская область	654,4	19320	13027	20867	75,34	27,51
8	Курская область	1117,4	23188	16132	23099	164,20	71,74
9	Липецкая область	1157,9	25263	18617	23133	198,28	110,10
10	Орловская область	765,2	19981	14580	20885	100,89	44,93
11	Рязанская область	1135,4	21988	15096	24280	159,37	58,21
12	Смоленская область	964,8	21788	16022	22279	144,96	56,75
13	Тамбовская область	1062,4	22377	16868	20757	167,90	112,71
14	Тверская область	1315,1	20602	16137	23866	202,16	74,49
15	Тульская область	1513,6	23040	16605	25873	231,98	95,44
16	Ярославская область	1271,6	23876	16456	25434	190,28	76,49
17	Московская область	7231,1	34948	24751	38598	1582,37	594,50

Основные этапы решения задачи выполняем и анализируем исходя из протокола расчётов и графической иллюстрации результатов программы факторного анализа.

1. Файл: факторный анализ.

Переменная	Среднее
X_1	1575,5
X_2	23250
X_3	17194
X_4	24076
X_9	269,39
X_{10}	112,66

2. Корреляционная матрица

	X_1	X_2	X_3	X_4	X_9	X_{10}
X_1	1	0,888	0,741	0,886	0,998	0,975
X_2	0,888	1	0,808	0,895	0,894	0,921
X_3	0,741	0,808	1	0,637	0,747	0,765
X_4	0,886	0,895	0, 637	1	0,888	0,870
X_9	0,998	0,894	0,747	0,888	1	0,976
X_{10}	0,975	0,921	0,765	0,870	0,976	1

Критическое значение: 0,4758; число значимых коэффициентов: 15.

3. Собственные значения и процент объясняемой дисперсии факторов:

Фактор:	1	2	3
Собств. знач.	5,3133	0,4050	0,1853
Дисперсия, %	88,555	6,7506	3,0886
Накоплен дисп.,%	88,555	95,305	98,394

4. Собственные векторы (коэффициенты поворота факторных осей):

а) фактор:

		1	2	3
переменная:	X_1	0,42329	–0,18319	0,39905
	X_2	0,41570	0,08276	–0,47471
	X_3	0,35832	0,87482	–0,04285
	X_4	0,39912	–0,39512	–0,61529
	X_9	0,42439	–0,17113	0,37707
	X_{10}	0,42453	–0,09423	0,30464

б) факторные нагрузки до вращения:

фактор:		1	2	3
переменная:	X_1	0,9757	–0,1166	0,1718
	X_2	0,9582	–0,2044	0,0009
	X_3	0,8250	0,5568	0,0015
	X_4	0,9200	–0,2515	–0,2649
	X_9	0,9782	–0,1089	0,1623
	X_{10}	0,9786	0,0304	0,1311

5. Переменная <--- Факторные координаты объектов --->:

фактор:		1	2	3
номер объекта:	1.	438,47	396,18	–378,56
	2.	–12,731	2377,3	845,35
	3.	–1015,7	–519,72	866,05
	4.	850,07	760,50	–281,79
	5.	–1454,2	–225,37	1318,8
	6.	914,71	–366,39	–1472,9
	7.	–1966,4	–996,86	1420,9
	8.	–420,86	–176,72	176,54
	9.	304,92	749,42	–252,16
	10.	–1605	–444,91	1288,0
	11.	–580,41	–766,1	131,99
	12.	–840,64	–116,97	610,77
	13.	–825,85	426,36	891,09
	14.	–687,3	–398,99	516,34
	15.	147,8	–489,09	–409,84
	16.	144,1	–422,87	–505,21
	17.	6609	214,28	–4765,3

Графическая иллюстрация результатов факторного анализа приведена на рис. 12 – 16.

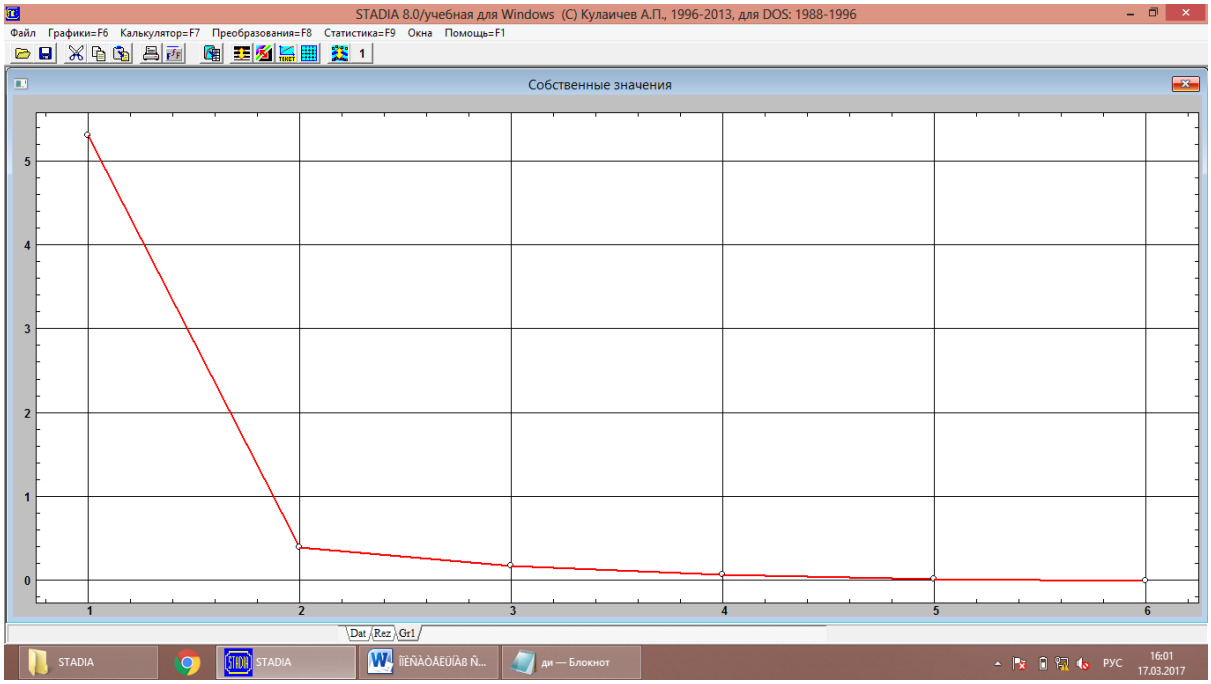

Рис. 12. График собственных значений

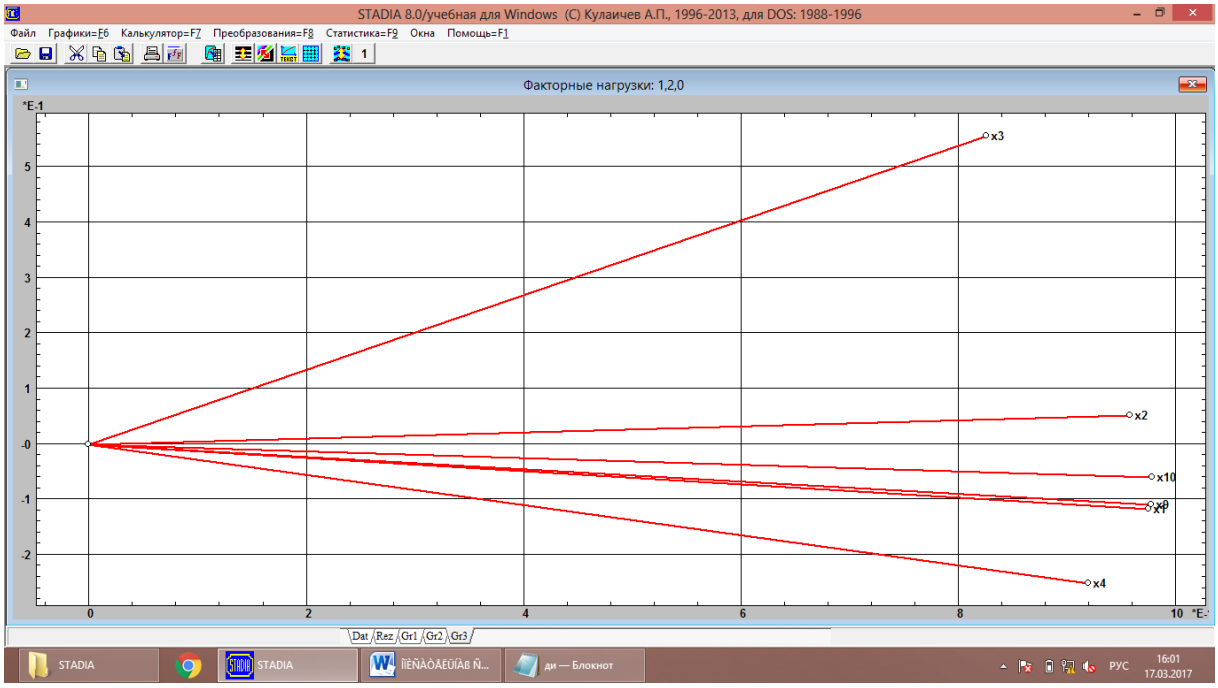

Рис. 13. Векторы факторных нагрузок в двумерном пространстве

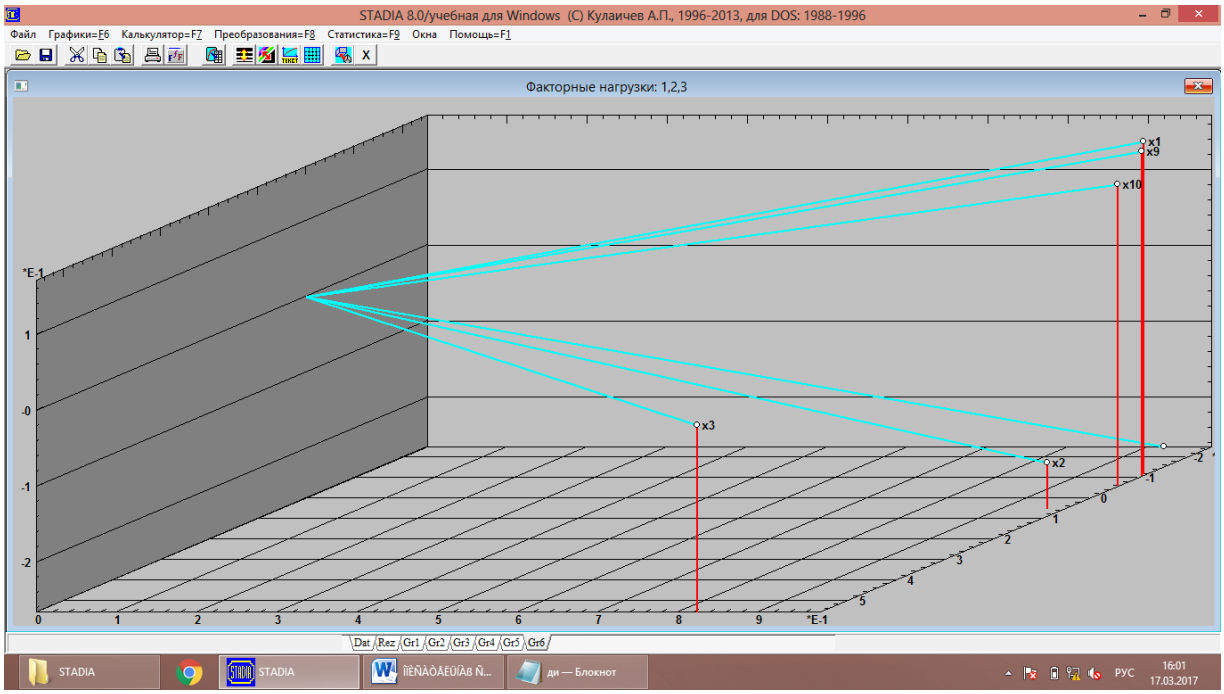

Рис. 14. Векторы факторных нагрузок
в трёхмерном пространстве

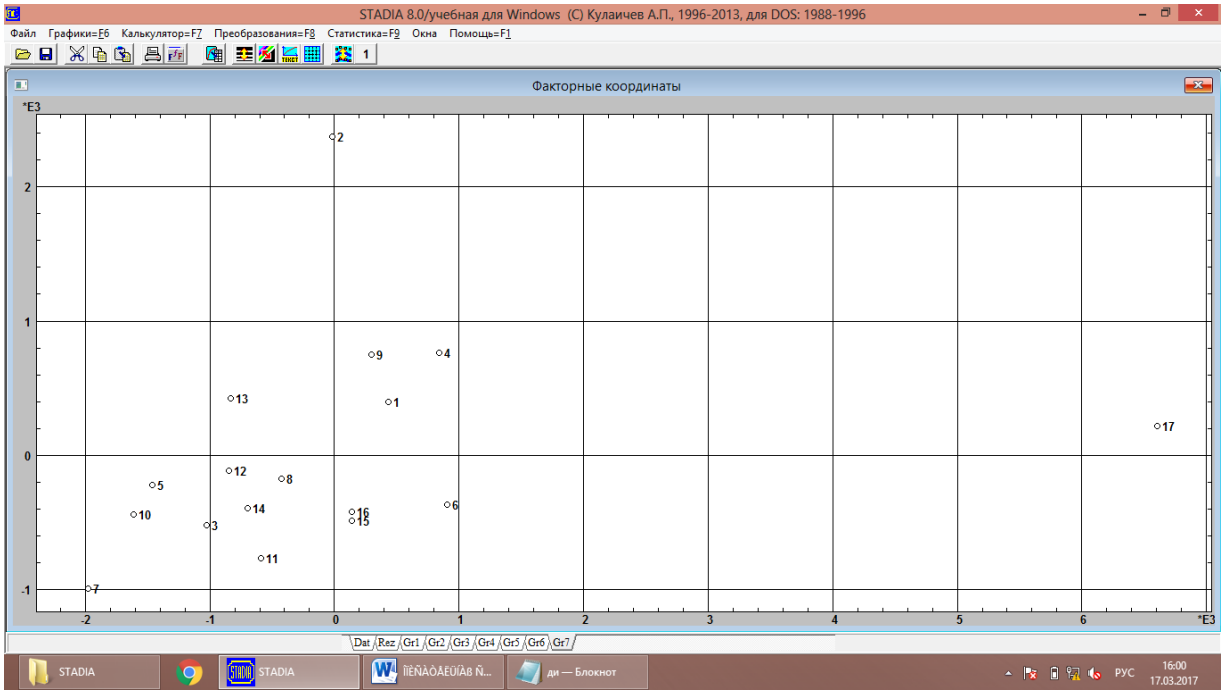

Рис. 15. Проекции объектов на плоскость ($F1$, $F2$)

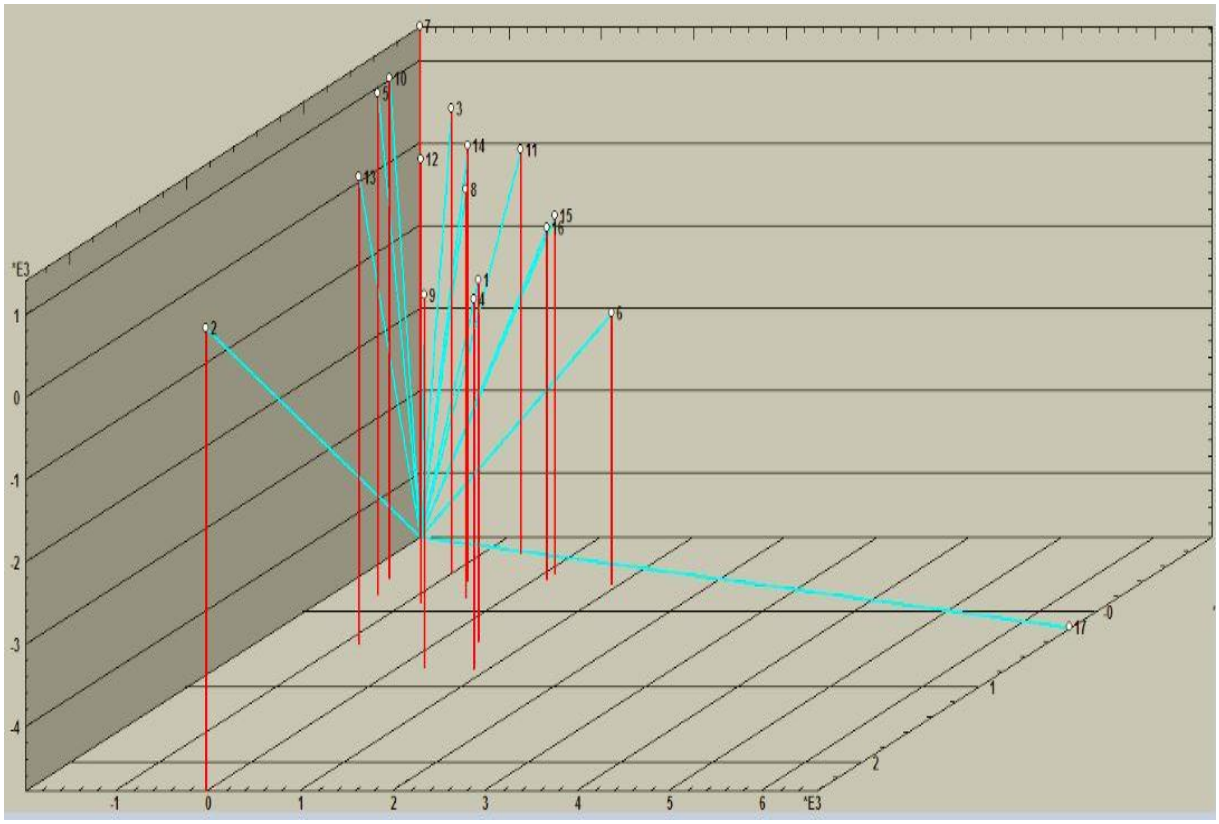

Рис. 16. Объекты-регионы в трёхмерном пространстве
главных компонент (*F*1, *F*2, *F*3)

Анализируя результаты расчётов, приходим к следующим выводам:

1. Как показывает корреляционная матрица, все исходные переменные (X_1, X_2, X_3, X_4, X_9, X_{10}) связаны значимой корреляционной связью между собой, здесь рассчитывались коэффициенты линейной корреляции Пирсона. Анализируя таблицу, график собственных значений и процент объясняемой дисперсии, приходим к выводу, что в качестве главных компонент (новых факторных переменных, которые получены в результате линейных преобразований исходных показателей) достаточно взять *первые 3 главных фактора*, которые объясняют около 98,38 % дисперсии исходных показателей. На суммарную дисперсию остальных компонент приходится менее 2 %, эти компоненты в выдаче результатов не участвуют. В рассматриваемой модели главные компоненты не коррелированны и устроены так, что дисперсия *i*-й главной компоненты равна собственному значению λ_i корреляционной матрицы, представленной выше.

2. Факторные нагрузки характеризуют тесноту линейной связи попарно между исходными показателями и главными компонентами. Чем выше факторная нагрузка, тем более проявляется исходный показатель в главной компоненте. В данном случае самые высокие факторные нагрузки у переменных X_1 и X_2 составляют соответственно 0,9782 и 0,9757, что указывает на высокую линейную корреляцию между показателями X_1, X_2 и первой главной компонентой. Анализируя таблицу факторных нагрузок, приходим к выводу, что все исходные показатели преимущественно проецируются на первый главный фактор (условно назовём его фактором оценки социально-экономического развития регионов). На второй фактор в основном проецируется показатель X_3 и далее X_4. На третий фактор наиболее выраженно проецируется показатель X_4 и далее X_9. Графики векторов факторных нагрузок в проекции на плоскость главных компонент ($F1$, $F2$) и в пространстве ($F1$, $F2$, $F3$) показывают вектор наименьшей длины по переменной X_3.

3. В трехмерном пространстве трёх главных компонент наиболее близкие векторы визуально соответствуют показателям X_1 и X_9 (по численности занятых в экономике и обороту розничной торговли).

4. Отметим, что выявленные три главных фактора не коррелированны между собой, что показывает распечатка программы STADIA.

5. Рассмотрим проекции объектов-регионов на плоскость главных компонент ($F1$, $F2$) и в трёхмерном пространстве ($F1$, $F2$, $F3$). На рисунках выше номера регионов представлены точками на соответствующей координатной плоскости и в пространстве. Визуальный анализ проекций указывает на три – четыре группировки наиболее близко расположенных друг от друга точек. Точка, отвечающая региону № 17 (Московской области) наиболее удалена от остальных, что действительно отражает реальное положение этого региона на фоне остальных. Владимирская область (регион № 3) ближе всего расположена к Тверской (региону № 14).

Таким образом, на основе факторного анализа выявлены три главных фактора (в пространстве регистрируемых показателей), которые можно рассматривать в какой-то мере как движущие силы в сфере экономического развития.

На следующем этапе применяем кластерный анализ объектов-регионов в трёхмерном пространстве их факторных координат. Чтобы установить близость объектов друг к другу, вначале вычисляем попарные расстояния между объектами. Затем по стратегии Уорда выстраиваем возрастающие цепочки кластеров всё большей общности. Ниже приводится протокол расчёта в программе STADIA.

КЛАСТЕРНЫЙ АНАЛИЗ. Файл:
Норм.Эвклид.+Уорда
Таблица расстояний

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**
(2)	2,07															
(3)	1,77	2,67														
(4)	1,29	2,40	2,61													
(5)	2,09	2,71	0,63	3,01												
(6)	1,16	2,62	2,08	2,01	2,49											
(7)	2,80	3,53	1,19	3,72	0,87	3,00										
(8)	1,09	2,37	0,85	2,19	1,10	1,47	1,73									
(9)	0,43	1,80	1,83	1,51	2,05	1,28	2,79	1,11								
(10)	2,29	2,93	0,74	3,21	0,33	2,58	0,61	1,25	2,25							
(11)	1,55	2,78	0,62	2,57	1,01	1,62	1,42	0,59	1,64	1,06						
(12)	1,67	2,41	0,77	2,75	0,70	1,94	1,36	0,69	1,60	0,88	0,72					
(13)	1,28	2,06	0,97	2,13	1,02	1,97	1,79	0,73	1,18	1,22	1,14	0,97				
(14)	1,58	2,43	0,45	2,42	0,95	1,76	1,53	0,78	1,63	1,06	0,59	0,79	1,00			
(15)	0,98	2,47	1,19	1,89	1,68	0,98	2,24	0,79	1,20	1,81	0,85	1,26	1,31	0,89		
(16)	0,90	2,49	1,25	2,01	1,64	0,91	2,19	0,62	1,06	1,77	0,80	1,14	1,25	1,02	0,37	
(17)	8,34	9,08	9,60	7,32	10,11	8,42	10,71	9,27	8,62	10,30	9,46	9,81	9,40	9,37	8,68	8,85
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**

К л а с т е р ы:
(список объектов) -> расстояние
(10, 5) --> 0,32869
(16, 15) --> 0,36728
(9, 1) --> 0,42641
(14, 3) --> 0,45409
(11, 8) --> 0,58665
(12, 11, 8) --> 0,73824
(10, 7, 5) --> 0,84506
(14, 12, 11, 8, 3) --> 0,91552
(16, 6, 15) --> 1,0701
(14, 13, 12, 11, 8, 3) --> 1,1178

(16, 9, 1, 6, 15) --> 1,5205
(14, 10, 7, 5, 13, 12, 11, 8, 3) --> 2,0159
(16, 4, 9, 1, 6, 15) --> 2,1383
(16, 2, 4, 9, 1, 6, 15) --> 2,8388
(16, 14, 10, 7, 5, 13, 12, 11, 8, 3, 2, 4, 9, 1, 6, 15) --> 4,623
(17, 16, 14, 10, 7, 5, 13, 12, 11, 8, 3, 2, 4, 9, 1, 6, 15) --> 12,576

Здесь рассчитывалось нормированное эвклидово расстояние между *I*-м и *J*-м объектами по формуле: $d(I,J)=\sqrt{\sum_{k=1}^{3}\frac{\left(x_{ik}-x_{jk}\right)^2}{s_k^2}}$, где ($x_{i1}$, x_{i2}, x_{i3}) и (x_{j1}, x_{j2}, x_{j3}) – факторные координаты объектов; s_1^2, s_2^2, s_3^2 – дисперсии главных факторов. Отметим, что самыми близкими объектами оказались регионы № 10 и № 5, находящиеся на расстоянии *d*(10, 5) = 0,329. Владимирская область (регион № 3) ближе всего расположена к Тверской (региону № 14), расстояние между ними составляет *d* (14, 3) = 0,454, далее к ним присоединяются регионы № 12, 11 и 8. На последнем шаге к цепочке всех объектов присоединяется регион № 17 – Московская область (рис. 17).

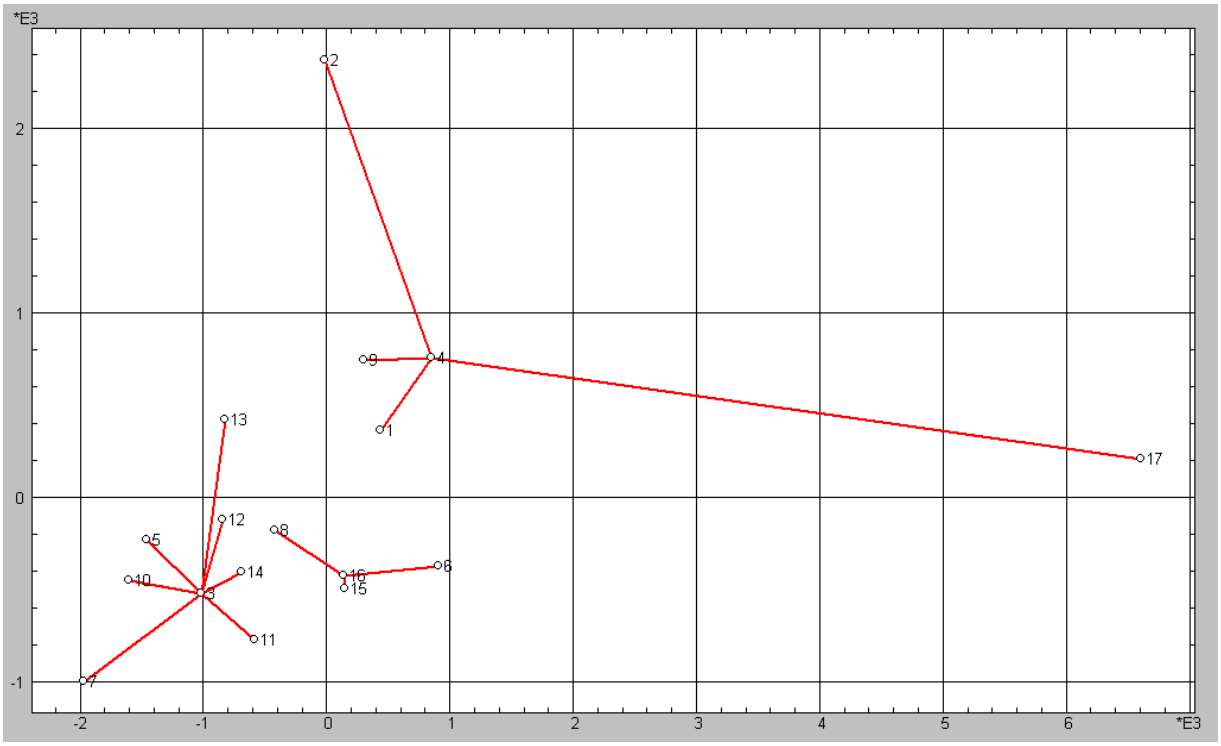

Рис. 17. Три кластера в проекции на плоскость (*F*1, *F*2)

На основе факторных координат объектов с применением дивизивной стратегии на следующем этапе была построена классификация регионов ЦФО РФ по состоянию социально-экономического развития, представленная в табл. 15. Результаты разбиения всех объектов на три кластера демонстрирует рис. 17. Здесь нетрудно заметить удалённость объекта № 17 от всех регионов, так что его потребовалось определить в четвёртый кластер. В табл. 15 символом «*» указаны геометрические центры кластеров в трёхмерном пространстве главных компонент.

Таблица 15. Группы регионов по социально-экономическому развитию

Номер региона	**Регион**	**Координаты объектов в трехмерном пространстве главных факторов (*F*1, *F*2, *F*3)**	**Результативный показатель (ВРП) *Y*, млн руб.**
Кластер 1			
6	Калужская область	(914,71; –366,39; –1472,9)	293,434
8	Курская область	(–420,86; –176,72; 176,54)	272,238
15	Тульская область	(147,8; –489,09; –409,84)	347,06
16*	Ярославская область	(144,1; –422,87; –505,21)	360,732
Кластер 2			
3*	Владимирская область	(–1015,7; –519,72; 866,05)	307,486
5	Ивановская область	(–1454,2; –225,37; 1318,8)	157,735
7	Костромская область	(–1966,4; –996,86; 1420,9)	143,108
10	Орловская область	(–1605; –444,91; 1288,0)	164,526
11	Рязанская область	(–580,41; –766,1; 131,99)	278,732
12	Смоленская область	(–840,64; –116,97; 610,77)	225,595
13	Тамбовская область	(–825,85; 426,36; 891,09)	235,86
14	Тверская область	(–687,3; –398,99; 516,34)	291,408

Окончание табл. 15

Номер региона	**Регион**	**Координаты объектов в трехмерном пространстве главных факторов (*F*1, *F*2, *F*3)**	**Результативный показатель (ВРП) *Y*, млн руб.**
Кластер 3			
1	Белгородская область	(438,47; 396,18; –378,56)	569,414
2	Брянская область	(–12,731; 2377,3; 845,35)	223,324
4*	Воронежская область	(850,07; 760,5; –281,79)	606,668
9	Липецкая область	(304,92; 749,42; –252,16)	314,79
Кластер 4			
17	Московская область	(6609; 214,28; –4765,3)	2551,284

Контрольные вопросы

1. Какова основная задача факторного анализа?
2. Что называют канонической моделью факторного анализа?
3. Невыполнение какого условия может привести к неадекватной модели главных компонент?
4. Укажите последовательные этапы выполнения факторного анализа и поясните их предназначение.
5. Что называют главными компонентами и за что они отвечают в процедурах факторного анализа?
6. Поясните основные процедуры алгоритма выделения главных компонент?
7. Для чего нужен расчёт собственных значений корреляционной матрицы стандартизированных исходных данных?
8. На чём основан выбор наиболее значимых главных компонент?
9. Для чего нужны проекции объектов на плоскость пары главных компонент?
10. Что характеризуют факторные нагрузки переменных?
11. Как могут быть использованы результаты факторного анализа для последующих статистических исследований?

Глава 5. ВВЕДЕНИЕ В АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ

5.1. Общие сведения о временных рядах

Научная методология исследования случайных процессов в природе и обществе базируется на анализе их исторического развития. В прикладной математической статистике разработаны специальные методы [7; 8; 12; 22], предназначенные для изучения развития и изменения явлений во времени, т. е. в динамике. В этой связи статистика вводит понятие *динамического ряда*. Согласно [22, с. 284], динамическим рядом называют «совокупность наблюдений некоторого явления (показателя), упорядоченных в зависимости от последовательности значений другого явления (признака)». В статистике также вводят понятие *временного ряда*, который представляет собой ряд последовательных значений некоторого статистического показателя Y, расположенных в хронологическом порядке и характеризующих развитие явления, как правило, в равноотстоящих точках времени t: $\{y_t, t \in T\}$. Значение y_t называют *уровнем* временного ряда Y.

Во временном ряде содержится информация об особенностях и закономерностях протекания изучаемого процесса, а статистический анализ позволяет выявить закономерности и использовать их для оценки характеристик процесса в будущем, т. е. для прогнозирования. При построении временного ряда следует иметь в виду, что его уровни должны подчиняться требованиям сопоставимости, т. е. должны рассчитываться по единой методологии с одинаковыми единицами измерения по всем объектам и характеризовать явление, относящееся к одной и той же территории, сопоставимому периоду времени.

Если уровни ряда даны по состоянию на определённую дату, такой ряд называют *моментным*. Если уровни ряда обозначают значение показателя за определённый временной промежуток (месяц, квартал, год и т. д.), такой ряд называют *интервальным*. В зависимости от вида статистического показателя ряды динамики подразделяют на ряды абсолютных, относительных и средних величин. По расстоянию между уровнями временные ряды подразделяются на ряды с равноотстоящими и неравноотстоящими значениями времени. В равноотсто-

ящих рядах даты регистрации периода следуют друг за другом, в неравноотстоящих равные интервалы не соблюдаются. На рис. 18 приведена классификация временных рядов.

Рис. 18. Классификация временных рядов

Одна из главных задач изучения временных рядов состоит в выявлении закономерностей (тенденций) в развитии процесса или явления. Конечно, для этого требуется иметь достаточное количество информации, т. е. располагать данными за как можно более длительный период времени.

Если изучаются временные ряды сложных экономических систем, тогда возникает необходимость провести сопоставление нескольких рядов динамики и сравнить интенсивность развития явлений во времени одновременно по нескольким показателям или по отношению к разным объектам. В этом случае ряды абсолютных показателей не дают такого наглядного представления о процессах, как ряды относительных величин, приведённые к одному основанию.

С этой целью уровни всех рядов выражают в коэффициентах или процентах по отношению к определенным значениям сравниваемых величин за один и тот же период времени (как правило, к начальным уровням – базе сравнения), что позволяет выявить наиболее интенсивно изменяющиеся статистические показатели. Существуют определенные требования к базе сравнения. Данные базы должны быть взяты в момент времени или за период с «типичным» значением показателя. Нельзя брать в качестве базы сравнения периоды с экстремальными для остальных моментов времени условиями, иначе получим неверное представление о темпах развития явления. Так, если база сравнения содержит высокий уровень показателей, то темпы роста будут занижены, если низкий – завышены.

Математическое моделирование экономических процессов, представленных одномерными временными рядами, сводится к решению следующих задач:

1. Табличное и графическое представление временных рядов.

2. Предварительный анализ данных.

3. Исследование структуры временного ряда (тренда, сезонных и циклических составляющих).

4. Преобразование временного ряда средствами сглаживания или фильтрации.

5. Построение математической модели процесса, представленного временным рядом.

6. Исследование случайной составляющей временного ряда и оценка качества математической модели.

7. Прогнозирование будущего развития экономического процесса.

В пособии приведены примеры решения этих задач на фоне статистического анализа временных рядов некоторых социально-экономических показателей.

Во всем многообразии приведенных выше задач в качестве основных выступают: определение тенденции развития, построение математической модели изучаемого экономического явления и прогнозирование его в будущем. Факторы, влияющие на формирование значений показателя развития экономического явления, можно подразделить на следующие типы:

- факторы, действующие постоянно и формирующие основную тенденцию развития явления;

- факторы, действующие периодически, вызывающие сезонные и циклические колебания;

- факторы, способствующие случайным колебаниям уровней динамического ряда.

При этом под тенденцией развития мы понимаем сформировавшееся определенное направление развития экономического процесса, проявляющееся в том, что фактические значения динамического ряда группируются вокруг некоего тренда – функции времени $\tilde{y}(t)$, описывающей характер соответствующего статистического показателя.

В отдельных случаях визуальный анализ графика расположения уровней динамического ряда позволяет определенно сказать о тенденции возрастания или убывания в ряду значений либо о колеблемости точек вокруг некоторой гладкой кривой. Однако не всегда по графику можно проследить определенную тенденцию в ряду, в особенности если уровни ряда хаотично возрастают или убывают на большом временном отрезке. К методам определения основной тенденции развития динамического ряда относят следующие:

- метод укрупнения интервалов;

- метод скользящей средней;

- аналитическое выравнивание динамического ряда.

Метод укрупнения интервалов заключается в построении динамического ряда с более крупными временными промежутками, где обнаруживается тенденция к возрастанию (или убыванию) уровней ряда. Однако применение этого метода неэффективно, если уровни исходного временного ряда представлены через длительные интервалы времени, например по годам. Тенденцию роста или убывания в динамическом ряду можно выявить по характеру изменения скользящих средних, т. е. средних арифметических значений новых m-членных укрупненных интервалов. Первый из них включает первые m значений ряда, второй содержит также m уровней, куда входят $(m - 1)$ значений первого интервала, начиная со второго его элемента, и $(m + 1) - e$ значение ря-

да и так далее вплоть до включения последнего уровня ряда. По подвижным скользящим средним делают вывод о наличии тенденций в ряду.

Более глубокий анализ временного ряда с целью дальнейшего прогнозирования его значений заключается в построении математической модели – уравнения тренда, описывающего характер развития изучаемого явления во времени, позволяющего уловить основные нюансы и направления его изменения в будущем. Представление временного ряда с помощью уравнения тренда называют аналитическим выравниванием исходного ряда. Современные пакеты прикладных программ по статистике позволяют подбирать процедурой простой регрессии в качестве модели тренда аналитическое выражение произвольного вида, задаваемого самим пользователем. Наиболее употребительны следующие регрессионные модели:

- линейная: $\tilde{y}(t) = a_0 + a_1 t$;
- параболическая: $\tilde{y}(t) = a_0 + a_1 t + a_2 t^2$;
- показательная: $\tilde{y}(t) = a_0 a_1^t$;
- логарифмическая: $\tilde{y}(t) = a_0 + a_1 \ln t$;
- степенная: $\tilde{y}(t) = a_0 t^{a_1}$;
- гиперболическая: $\tilde{y}(t) = a_0 + \frac{a_1}{t}$.

Линейный тренд используют, если уровни ряда меняются в арифметической прогрессии, в этом случае цепные абсолютные приросты уровней приблизительно одинаковы. Линейная функция хорошо отражает тенденцию изменений экономического явления под воздействием совокупности разнообразных факторов, равнодействующая которых зачастую выражается в примерно постоянной скорости изменения, равной параметру a_1 линейного тренда.

Пример 1. Приведем расчет линейного тренда во временном ряду показателя обеспеченности жильем населения на конец года в РФ с 1991 по 2008 гг. (Данные см. http://www.gks.ru).

Определим цепные абсолютные приросты для выявления динамики изменения временного ряда (табл. 16 и рис. 19).

Таблица 16. Показатели обеспеченности жильем
в Российской Федерации на конец года

***t* – условное время**	**Год**	**Обеспеченность жильем на конец года, м²/1 чел.**	**Абсолютное изменение**	**Разность величин абсолютного изменения**
1	1991	16,5	–	–
2	1992	16,8	0,3	0
3	1993	17,4	0,6	0,3
4	1994	17,7	0,3	–0,3
5	1995	18,1	0,4	0,1
6	1996	18,3	0,2	–0,2
7	1997	18,6	0,3	0,1
8	1998	18,8	0,2	–0,1
9	1999	19,1	0,3	0,1
10	2000	19,3	0,2	–0,1
11	2001	19,7	0,4	0,2
12	2002	20,0	0,3	–0,1
13	2003	20,2	0,2	–0,1
14	2004	20,5	0,3	0,1
15	2005	20,9	0,4	0,1
16	2006	21,3	0,4	0
17	2007	21,5	0,2	–0,2
18	2008	21,7	0,2	0
В среднем		19,24	0,31	–0,006

На графике изобразим фактическую обеспеченность жильем с целью графического выявления тренда.

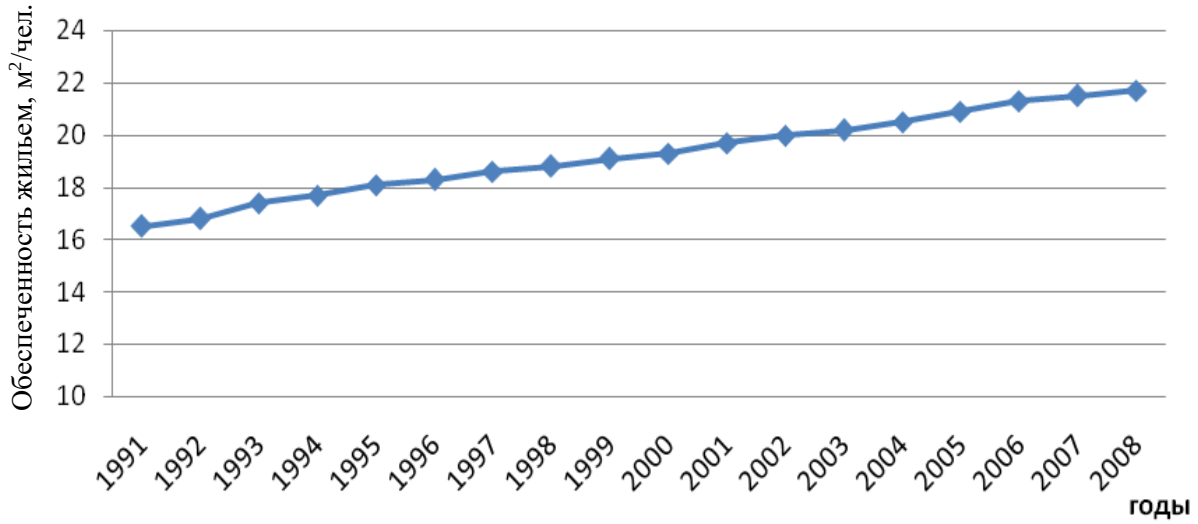

Рис. 19. Динамика обеспеченности населения жильем на конец года

Из табл. 16 и рис. 19 видно, что обеспеченность жильем имеет линейную тенденцию к росту со средним абсолютным изменением 0,31 м^2. Разность абсолютных изменений за последние периоды приближается к нулю. Кроме того, визуальный анализ корреляционного поля также позволяет предположить, что можно использовать в качестве тренда линейную функцию.

Выполним оценку тесноты линейной связи между $X = t$ и Y по коэффициенту линейной корреляции Пирсона (1): $r_{XY} = 0{,}9976$.

Значимость коэффициента проверяем по t-критерию Стьюдента:

$t_{\text{расч}} = 63{,}54$

$t_{\text{крит}}(0{,}05 : 16) = 2{,}12$.

$|t_{\text{расч}}| > t_{\text{крит}} \Rightarrow r_{XY}$ статистически значим.

$|r_{XY}| > 0{,}7$, значит, связь достаточно сильная.

Уравнение модели линейной регрессии

$\tilde{Y} = 16{,}41 + 0{,}30t$.

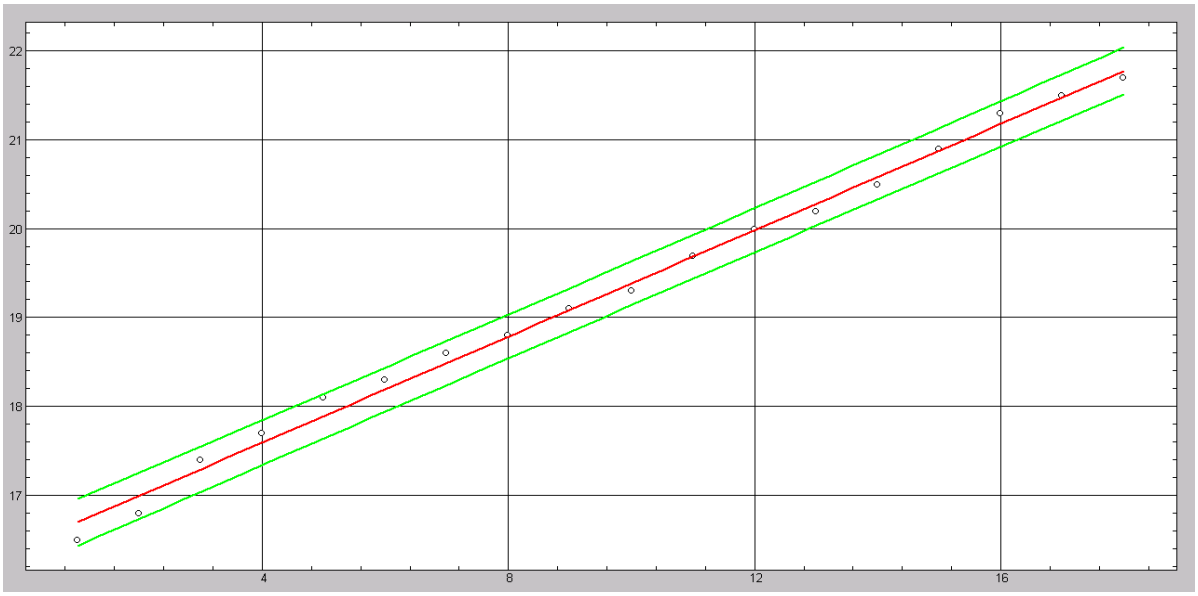

Рис 20. График линейной модели на фоне корреляционного поля

Оценим значимость модели регрессии по F-критерию Фишера:

$$F_{\text{расч}} = \frac{\text{дисперсия регрессии}}{\text{дисперсия остатков}} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$$

$F_{\text{расч}} = 3316$; $F_{\text{крит}}(0{,}05, 1, n-2) = 4{,}41$.

Так как $F_{\text{расч}} \geq F_{\text{крит}} \Rightarrow$ модель адекватна экспериментальным данным, т. е. статистически значимо уравнение регрессии. Коэффициент детерминации $R^2 \approx 0{,}99 \Rightarrow$ модель линейной регрессии на 99 % объясняет вариацию результативного показателя жилищного фонда (Y) с течением фактора времени t. Из уравнения модели следует, что с каждым годом жилищный фонд возрастает на 0,3 м^2 в среднем на одного жителя РФ.

Параболическую формулу тренда используют при ускоренном или замедленном изменении уровней ряда, т. е. когда постоянны цепные абсолютные приросты цепных приростов – вторые разности уровней. Величина $2a_2$ дает значение постоянного ускорения, координаты вершины параболы указывают на экстремальные точки в развитии процесса.

В случае показательной формы тренда параметр a_1 показывает темп изменения в рядах. Если $a_1 > 1$, показательная функция выражает тенденцию ускоренного и все более ускоряющегося возрастания уровней. При $a_1 > 1$ тренд означает тенденцию постоянно замедляющегося снижения уровней временного ряда. Выравнивание по показательной функции следует использовать, если уровни ряда меняются в геометрической прогрессии, т. е. цепные коэффициенты роста относительно постоянны.

Логарифмическая модель пригодна для отображения тенденции замедляющегося роста уровней при отсутствии предельного возможного значения. При достаточно большом t логарифмическая кривая мало отличается от прямой линии.

Модель в виде степенной функции пригодна для выражения изменений во времени с разной мерой пропорциональности. Обязательным является условие прохождения кривой тренда через начало координат.

Выравнивание по гиперболе используют, если уровни динамического ряда снижаются, постепенно замедляя скорость, но никогда не смогут достичь нуля. При $a_1 > 0$ гиперболическая модель выражает тенденцию замедляющегося снижения значений ряда, стремящихся к пределу a_0. При $a_1 < 0$ гипербола выражает тенденцию замедляющегося роста уровней, стремящихся в пределе к a_0. Таким образом, выравнивание по гиперболе подходит для выражения тенденции экономических процессов, ограниченных некоторым предельным значением уровня.

При наличии периодических колебаний в динамическом ряду используют выравнивание по Фурье-моделям вида [26, с. 362 – 369]

$$\tilde{y}(t) = a_0 + \sum_{k=1}^{m} \left(a_k \cos kt + b_k \sin kt \right).$$

Хорошие результаты даёт применение моделей авторегрессии (*AR*) к процессам, в которых прослеживается наличие одной или нескольких доминирующих корреляций. В *AR* модели текущее значение динамического ряда $y_t (t = 1, \ldots, n)$ с нулевым средним значением выражается как линейная комбинация значений y_{t-i} и белого шума $W(t)$, содержащего информацию, приходящую из настоящего: $W(t) = a_0 y_t + a_1 y_{t-1} + \ldots + a_p y_{t-p}$ [12, с. 254]. Чем ниже дисперсия белого шума, тем меньше новой информации дает очередное наблюдение. В качестве дополнения к моделям авторегрессии и для детального описания шумовой составляющей используют модель скользящего среднего (*MA*) [12, с. 216]

$$y_t = W(t) + b_1 W(t-1) + \ldots + b_q W(t-q),$$

где $(b_1, \ldots, b_k)$ – неизвестные параметры. Для описания и прогнозирования нестационарных процессов, статистические свойства которых могут изменяться с течением времени, полезны модели ARIMA [12, с. 255 – 259]. Они сочетают в себе комбинацию моделей авторегрессии и скользящего среднего. Итерационный процесс построения параметров таких моделей основан на сложном алгоритме, по количеству процедур превосходящем аналитическое выравнивание средствами регрессионного и спектрального анализа при построении Фурье-моделей.

При аналитическом выравнивании динамических рядов, как правило, рассчитывают несколько математических моделей в виде уравнений тренда и в итоге отбирают модель с наименьшей стандартной ошибкой отклонений (остатков) фактических уровней ряда от выравненных значений по уравнению тренда. Качество модели можно оценить также, исследуя остатки на автокорреляцию с помощью метода Дарбина – Уотсона:

$$d = \sum_{t=1}^{n-1} \left(\varepsilon_{t+1} - \varepsilon_t \right)^2 \Big/ \left(\sum_{i=1}^{n} \varepsilon_t^2 \right),$$

где $\varepsilon_t = y_t - \tilde{y}_t$ – отклонение фактического уровня ряда от его выравненного значения. Расчетное значение d сравнивается с критическим [22, с. 310] и на определенном уровне значимости принимается решение о наличии автокорреляции в ряду остатков. Если автокорреляция

отсутствует, приходят к следующему выводу: модель адекватно отражает тенденцию развития экономического явления. Наряду с исследованием автокорреляции остатков анализ динамических рядов предполагает изучение характера колеблемости ряда, т. е. отклонений уровней ряда от выравненных значений. Типы колебаний довольно разнообразны, среди них выделяют:

- пилообразную, или маятниковую колеблемость;
- циклическую долгопериодическую колеблемость;
- случайно распределенную во времени колеблемость.

С учетом степеней свободы основные абсолютные показатели колеблемости вычисляют по формулам:

$\bar{a}(t) = \frac{1}{n-p}\sum_{i=1}^{n}|y_i - \tilde{y}_i|$ – среднее линейное отклонение;

$s(t) = \left(\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2\right)^{1/2}$ – среднее квадратическое отклонение,

где y_i – фактический уровень ряда; $\tilde{y}_i$ – выравненный уровень; n – число уровней в ряду; p – количество параметров тренда. Относительная мера колеблемости – коэффициент колеблемости (аналог коэффициента вариации) – $V_y(t) = S(t)/\bar{y}$. Чем меньше значение $V_y(t)$, тем слабее колеблемость уровней ряда.

При исследовании экономических процессов нередко обнаруживают, что изменение во времени одного статистического показателя приводит к изменению другого. В этом случае интерес представляет изучение взаимосвязи соответствующих динамических рядов. Для того чтобы точнее определить силу корреляционной связи, полезно предварительно проверить ряды на существование доминирующих корреляций и их лагов (временных задержек) внутри каждого ряда. С этой целью вычисляют значения автокорреляционной функции $r_{yy}(j)$ для последовательных лагов j [12, с. 225]:

$$r_{yy}(j) = \left(\sum_{t=1}^{n-j}(y_t - \bar{y})(y_{t+j} - \bar{y})\right) \Big/ \left((n-1)s_y^2\right);\ j = 0, \ldots, (n-8),$$

где $\bar{y}$ – средний уровень исходного ряда; $s_y^2 = \left(\sum_{t=1}^{n}(y_t - \bar{y})^2\right) \Big/ (n-1)$.

Значение $r_{yy}(j)$ называют также коэффициентами автокорреляции. Автокорреляционная функция показывает, в какой степени динамика изменения заданного фрагмента ряда воспроизводится в сдви-

нутых во времени его же отрезках. Таким образом, по величине значимых коэффициентов автокорреляции судят о зависимости уровней динамического ряда от их значений в предыдущее время. Если оба исследуемых временных ряда содержат автокорреляцию, тогда ее исключают преобразованием исходных рядов и только затем рассчитывают коэффициент корреляции между остаточными величинами данных рядов. Преобразование рядов обычно осуществляют по формуле $d_y = y - \tilde{y}_t = \varepsilon_t$. Высокое значение коэффициента корреляции между преобразованными рядами может служить индикатором причинно-следственных связей между этими рядами, и тогда становится актуальным построение модели регрессии $\tilde{y}_{x,y} = \tilde{y}(x, t)$, где уровни одного из исходных временных рядов выступают в качестве факторной переменной X; вторая факторная переменная – время t, позволяющая учесть в модели влияние автокорреляции. Для анализа взаимосвязанных динамических рядов полезно также вычислить кросскорреляционную функцию [12, с. 188]. Кросскорреляционная функция позволяет определить, в какой степени динамика изменения заданного участка одного динамического ряда воспроизводится в сдвинутых во времени участках второго ряда. Так получают картину, отражающую влияние одного экономического явления на другое, и определяют задержки (величину лага) в передаче взаимодействия, что имеет значение для выработки стратегии в практических действиях.

5.2. Компонентный состав временных рядов

В практике исследования динамики явлений и прогнозирования принято считать, что значения уровней временных рядов могут содержать следующие компоненты (составные части, или структурообразующие элементы):

- тренд;
- сезонную компоненту;
- циклическую компоненту;
- случайную составляющую.

Под *трендом* понимают изменение, определяющее общее направление развития, основную тенденцию временного ряда. Это систематическая составляющая долговременного действия.

Наряду с долговременными тенденциями во временных рядах часто имеют место более или менее регулярные колебания – периодические составляющие рядов динамики. Если период колебаний не превышает одного года, то их называют *сезонными*. Чаще всего причиной их возникновения считаются природно-климатические условия. Иногда причины сезонных колебаний имеют социальный характер, например увеличение закупок в предпраздничный период или платежей в конце квартала.

При большем периоде колебания считают, что во временных рядах имеет место *циклическая* составляющая. В экономических временных рядах редко предоставляется возможность для выделения и дальнейшего анализа циклической компоненты, так как ряды динамики экономических показателей часто оказываются слишком короткими для проведения такого исследования. Хорошо изучены циклические составляющие в рядах динамики, относящихся к естественным наукам. Например, по многолетним наблюдениям установлена цикличность солнечной активности (с периодом колебаний в одиннадцать лет).

Если из временного ряда удалить тренд и периодические составляющие, то останется нерегулярная компонента. Экономисты разделяют факторы, под действием которых формируется нерегулярная компонента, на два вида: факторы резкого, внезапного действия; текущие факторы. Факторы первого вида (например, стихийные бедствия, эпидемии, война, кризис и т. д.), как правило, вызывают более значительные отклонения. Иногда такие отклонения называют катастрофическими колебаниями. Факторы второго вида вызывают случайные колебания, являющиеся результатом действия большого числа побочных причин. Влияние каждого из текущих факторов незначительно, но ощущается их суммарное воздействие.

Если временной ряд представляется в виде суммы соответствующих компонент $y_t = u_t + s_t + v_t + \varepsilon_t$, то полученная модель носит название *аддитивной.* Представление временного ряда в виде произведения компонент соответствует *мультипликативной* модели $y_t = u_t s_t v_t \varepsilon_t$, где y_t – уровни временного ряда; u_t – трендовая составляющая; s_t – сезонная компонента; v_t – циклическая компонента; ε_t – случайная компонента.

Можно выделить еще модели *смешанного типа*, предусматривающие как сложение, так и умножение соответствующих компонент.

Примером модели этого типа может служить выражение, в котором соответствующие значения случайной составляющей прибавляются к произведению трендовой компоненты и периодических составляющих: $y_t = u_t s_t v_t + \varepsilon_t$.

В процессе формирования значений уровней каждого временного ряда не обязательно участие одновременно всех компонент. В изменении значений одного показателя может отсутствовать трендовая компонента, другого – периодические составляющие, динамика третьего показателя может описываться лишь случайной компонентой.

Решение любой задачи по анализу и прогнозированию временных рядов начинается с построения графика исследуемого показателя, так как на этом этапе можно исследовать компонентный состав временных рядов, а также сделать первые шаги к выбору модели для описания их динамики.

Пример 2. Имеются данные Федеральной службы государственной статистики об объеме платных услуг населению (млн руб.) по Владимирской области за 2005 – 2011 гг., представленные в табл. 17 (см. http//vladimirstat.gks.ru).

Таблица 17. Объем платных услуг населению

Месяц	**Год**						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Январь	970,1	1333,4	1685,6	2094,1	2519,8	2924,4	3459,3
Февраль	1025,2	1434,6	1674,2	2052,6	2596,2	2805,9	3559,4
Март	1197,2	1460,4	1716,8	2141,7	2582,6	2890,8	3612,5
Апрель	1093,9	1423,2	1689,4	2180,9	2498,7	2731	3535,6
Май	1049,6	1329,3	1590,2	1950	2230,8	2628,5	3529,8
Июнь	1081,8	1314,3	1694,9	2032,7	2225,7	2798,9	3663,6
Июль	1113,8	1350,4	1677,7	2022,8	2353,3	2752,9	3700,7
Август	1133,4	1333,9	1630,6	2082,5	2162,5	2598,2	2991,4
Сентябрь	1140,2	1348,3	1632,8	2051,1	2140,2	2663,6	2919,6
Октябрь	1235,5	1485,6	1827,6	2307,7	2450,6	2780,8	3252,1
Ноябрь	1397,4	1628,3	1974,9	2352,4	2485,5	3191,2	3398,1
Декабрь	1533,9	1717,6	2121,2	2520,4	2707,5	3587,8	3643,1

На основании месячных данных об объеме платных услуг населению построим график временного ряда (рис. 21). Для временного ряда, представленного на рис. 21, характерно наличие ярко выраженного возрастающего тренда и сезонных колебаний. В данном случае с

учетом характера изменения периодических колебаний для описания данного ряда целесообразно выбрать модель с аддитивной сезонностью.

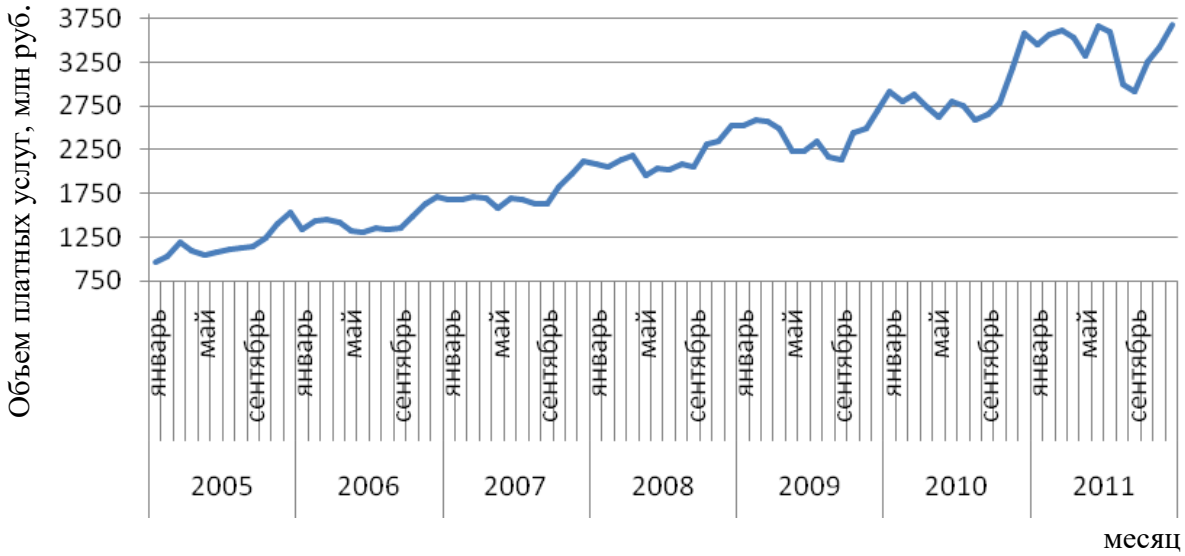

Рис. 21. Месячные данные объемов платных услуг населению

Пример 3. Имеются следующие поквартальные данные по инвестициям в основной капитал (млн руб.) по Владимирской обл. за 2003 – 2011 гг., см. табл. 18 (http//vladimirstat.gks.ru).

Таблица 18. Инвестиции в основной капитал, млн руб.

Квартал	**Год**								
	2003	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
I	1117	1486	1726,2	2410,7	3814,7	4969,9	5199,7	4850,8	5631
II	2391,6	2945,6	2948,6	3972	7077,7	8567,3	8441,7	9071,2	11149,6
III	2037	2193,1	4116,5	4856,5	9487,5	10486,9	11541,6	11464,4	15457,3
IV	4781,1	5866,4	8535,6	11014,1	12444,7	22796,9	17766,7	22347,3	25749,1

Динамика уровней данного временного ряда представлена на рис. 22. На графике прослеживаются устойчивые сезонные колебания, наслаивающиеся на монотонно возрастающий тренд. Отчетливо видны ежегодно повторяющиеся пики активности, приходящиеся на последний квартал года. Причем амплитуда сезонных колебаний возрастает с ростом объемов инвестиций, что приводит к выводу о мультипликативном характере сезонности. Таким образом, на стадии проведения графического анализа можно исследовать компонентный состав временных рядов, а также сделать выбор модели для описания их динамики и последующего прогнозирования.

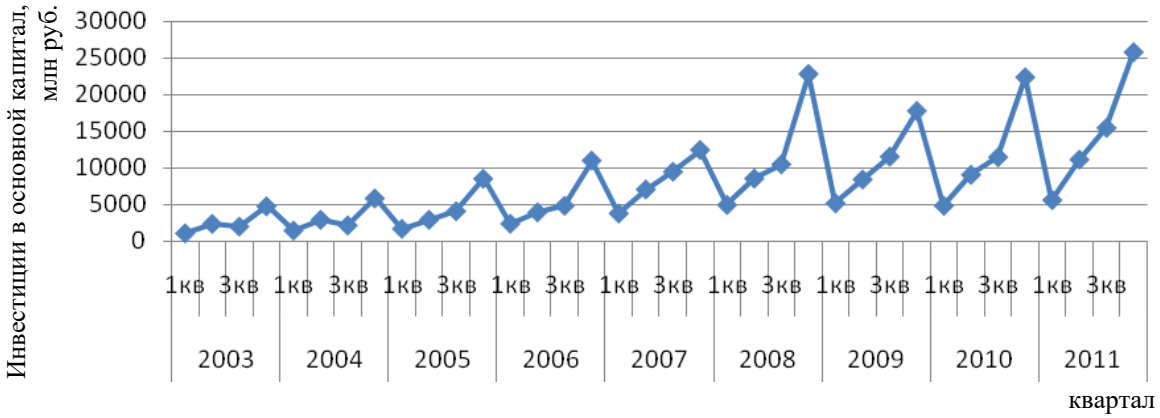

Рис. 22. Поквартальная динамика инвестиций в основной капитал

5.3. Сглаживание временных рядов с помощью скользящих средних

Важнейшая классическая задача при исследовании экономических временных рядов – выявление и статистическая оценка основной тенденции развития изучаемого процесса и отклонений от неё. В некоторых случаях судить о наличии тенденции в динамическом ряду можно на основе его визуального анализа, когда чётко видно, что при переходе от одного момента времени к другому уровни ряда возрастают или убывают. Однако в других случаях подобный визуальный анализ данных не позволяет обнаружить тенденцию к росту или падению значений показателя: они могут, как кажется, хаотично возрастать и так же хаотично убывать. Нельзя сразу сказать, что в таком временном ряду тенденции нет вовсе, к подобным данным следует применить специальные методы.

Один из способов выявления наличия тенденции в динамическом ряду основан на сглаживании ряда с помощью так называемых скользящих (подвижных) средних. Процедуры расчёта и анализа скользящих средних опираются на известную теорему Вейерштрасса, согласно которой любая гладкая функция при самых общих допущениях может быть локально (т. е. в ограниченном интервале изменения ее аргумента t) представлена алгебраическим полиномом подходящей степени.

Скользящими (подвижными) средними называют вычисленные средние арифметические значения новых l-членных укрупнённых интервалов. Правила построения этих интервалов следующие. Первый из них включает первые l уровней ряда динамики, второй образуется

путём исключения первого члена укрупненного интервала и замены его последующим (l + 1) очередным элементом ряда динамики и так далее до включения в интервал последнего уровня ряда. На рис. 23 представлена укрупнённая схема сглаживания временного ряда с помощью простой скользящей средней [7, с. 30].

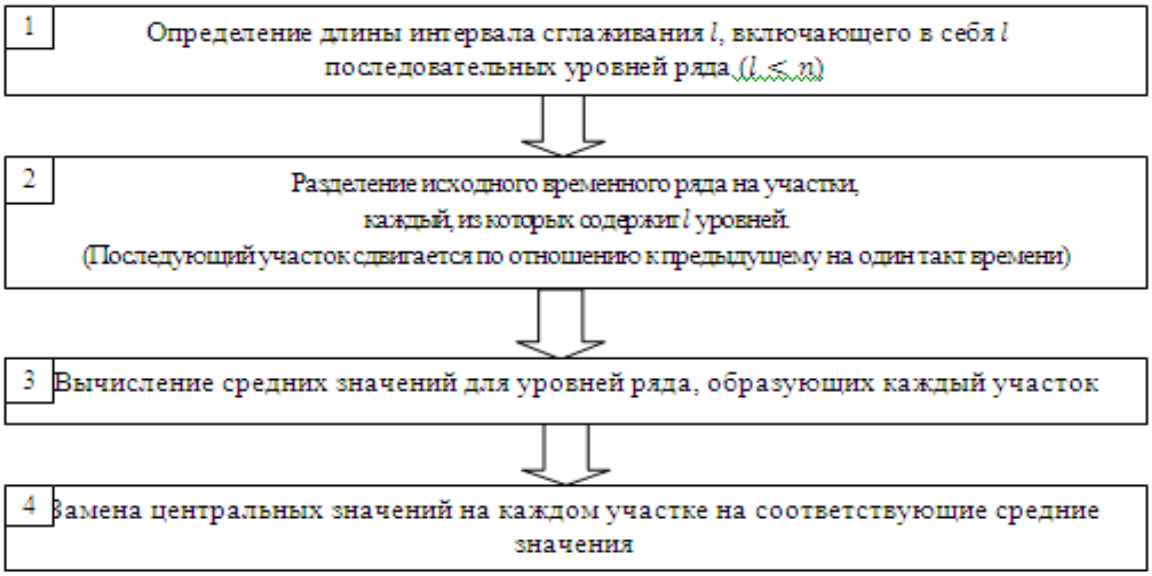

Рис 23. Схема сглаживания временного ряда по простой скользящей средней

Таким образом, на первом шаге необходимо определить длину интервала сглаживания l. При этом следует иметь в виду, что чем шире интервал сглаживания, тем в большей степени поглощаются колебания и тенденция развития носит более плавный, сглаженный характер. Однако чем шире интервал сглаживания, тем больше теряется информации о временном ряде.

Если длина интервала сглаживания $l = 2p + 1$ представляет собой нечётное число, тогда полученные значения скользящей средней приходятся на средний уровень интервала. При этом скользящая средняя $\hat{y}_t$ в момент времени t может быть определена по формуле

$$\hat{y}_t = \frac{\sum_{i=t-p}^{t+p} y_i}{2p+1} = \frac{y_{t-p} + y_{t-p+1} + \ldots + y_{t+p-1} + y_{t+p}}{2p+1},$$

где y_t – исходные уровни ряда. Если в качестве укрупнённого интервала используют период в три месяца, то первая подвижная трёхчлен-

ная средняя вычисляется как средняя арифметическая из данных за январь, февраль и март, вторая – как средняя арифметическая из данных за февраль, март, апрель и т. д.

Если подвижная средняя вычисляется по чётному числу членов ряда, то середина интервала будет находиться между двумя серединными уровнями. В этом случае рекомендуется выполнить последующее центрирование подвижных средних, т. е. вычислить из них двучленные средние. Центрированная скользящая средняя при $l = 2p$ может быть определена по формуле [27, с. 34]

$$\hat{y}_t = \frac{\frac{1}{2}y_{t-p} + y_{t-p+1} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+p-1} + \frac{1}{2}y_{t+p}}{2p} =$$

$$= \frac{\frac{1}{2}y_{t-p} + \sum_{i=t-p+1}^{t+p-1} y_i + \frac{1}{2}y_{t+p}}{2p}$$

Для сглаживания сезонных колебаний при работе с временными рядами квартальной динамики можно использовать четырёхчленную скользящую среднюю

$$\hat{y}_t = \frac{\frac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}y_{t+2}}{4}.$$

При исследовании временных рядов ежемесячной динамики для устранения сезонных колебаний, как правило, применяют 12-членную скользящую среднюю

$$\hat{y}_t = \frac{\frac{1}{2}y_{t-6} + y_{t-5} + \ldots + y_t + \ldots + y_{t+5} + \frac{1}{2}y_{t+6}}{12}.$$

Пример 4. Имеются статистические данные [14] об объемах туристских потоков. Используя метод скользящих средних, выясним, есть ли тенденция в данном временном ряду (табл. 20).

Так как полученные значения скользящей средней находятся в четном ряду, то методику следует дополнить центрированием ряда:

$y_1 = (1173{,}5 + 1203)/2 = 1188{,}25$; $y_2 = (1203 + 1273{,}75)/2 = 1226{,}25$ и т. д.

Расчёт центрированной скользящей средней представлен в последнем столбце табл. 19.

Таблица 19. Расчет центрированной скользящей средней

Год	Квар-тал	Объем туристских потоков, тыс. чел.	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя
2004	I	778	–	–	–
	II	1104 1739	4694	1173,5	–
	III	1739 896	4812	1203	1188,25
	IV	1073	4951	1273,75	1220,375
2005	I	896	4859	1214,75	1226,25
	II	1243 1760	4849	1212,25	1213,5
	III	1647 982	4894	1223,5	1217,875
	IV	1063 1769	4994	1248,5	1236
2006	I	941 995	5107	1276,75	1262,625
	II	1343 1775	5146	1286,5	1281,625
	III	1760 778	5187	1296,75	1291,75
	IV	1102 1739	5195	1298,75	1297,75
2007	I	982 896	5204	1301	1299,875
	II	1351 1647	5266	1316,5	1308,75
	III	1769 941	5279	1319,75	1318,125
	IV	1164 1760	5334	1333,5	1326,625
2008	I	995 982	5340	1335	1334,25
	II	1406 1769	5350	1337,5	1336,25
	III	1775 995	–	–	–
	IV	1174			–

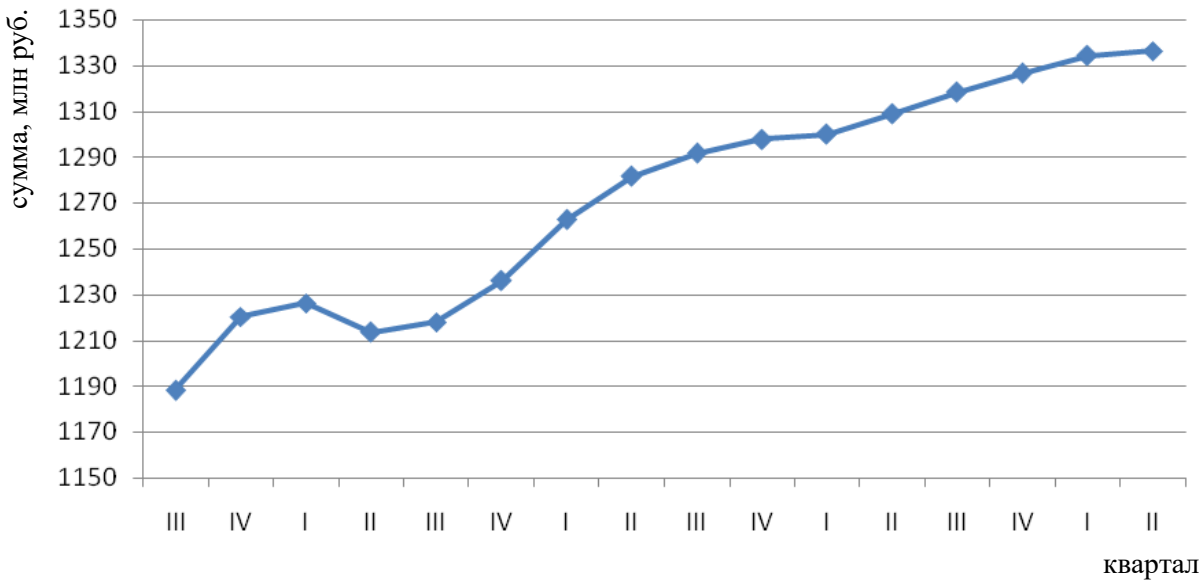

Рис. 24. Динамика центрированной скользящей средней объемов туристских потоков

Итак, по характеру изменения скользящих средних из четырех уровней можем сделать вывод о том, что показатель объемов туристских потоков зависел от квартала года, кроме того, с 2004 по 2008 гг. показатель имел тенденцию к непрерывному увеличению, что наглядно видно по значениям четырехчленных скользящих средних.

При использовании простой скользящей средней выравнивание на каждом участке ряда производится по прямой и таким образом выполняется аппроксимация неслучайной составляющей с помощью линейной функции от переменной времени *t*. Поэтому простую скользящую среднюю следует применять в тех случаях, когда графическое изображение динамического ряда напоминает прямую. Если для процесса характерно нелинейное развитие, то простая скользящая средняя может привести к существенным искажениям. Если тренд выравниваемого ряда имеет изгибы и для исследователя желательно сохранить волны, то целесообразно использовать взвешенную скользящую среднюю.

При построении взвешенной скользящей средней на каждом активном участке значение центрального уровня заменяется на расчетное, определяемое по формуле средней арифметической взвешенной. При этом взвешенная скользящая средняя приписывает каждому уровню вес, зависящий от удаления данного уровня до уровня, стоящего в середине активного участка. Поэтому весовые коэффициенты симметрич-

ны относительно центрального уровня на активном участке. В табл. 20 указаны веса для половины уровней активного участка и вес центрального уровня.

Расчет весовых коэффициентов осуществляют по следующей схеме [8, с. 205]. Для каждого активного участка подбирается полином вида $\hat{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots$, коэффициенты которого оцениваются с помощью метода наименьших квадратов. Начало отсчета (начало координат) переносят в середину активного участка. Тогда сглаженным значением для уровня, стоящего в середине активного участка, будет значение параметра a_0 подобранного полинома. Выражение для определения этого коэффициента (табл. 20) получают из системы нормальных уравнений, упрощенной за счет переноса начала координат в середину активного участка.

Таблица 20. Весовые коэффициенты для расчета взвешенной скользящей средней

Длина интервала сглаживания	**Весовой коэффициент**
5	$\frac{1}{35}[-3, +12, +17]$
7	$\frac{1}{21}[-2, +3, +6, +7]$
9	$\frac{1}{231}[-21, +14, +39, +54, +59]$
11	$\frac{1}{429}[-36, +9, +44, +69, +84, +89]$
13	$\frac{1}{143}[-11, 0, +9, +16, +21, +24, +25]$

Пример 5. Рассчитаем взвешенную скользящую среднюю для временного ряда y_t (табл. 21), используя сглаживание на каждом активном участке по полиному 2-го порядка. Согласно весовым коэффициентам при $l = 5$ вычислим скользящие средние:

$$\hat{y}_3 = \frac{1}{35}(-3 \cdot 510 + 12 \cdot 497 + 17 \cdot 504 + 12 \cdot 510 - 3 \cdot 509) = 502{,}7$$

$$\hat{y}_4 = \frac{1}{35}(-3 \cdot 497 + 12 \cdot 504 + 17 \cdot 510 + 12 \cdot 509 - 3 \cdot 503) = 509{,}3 \text{ и т. д.}$$

На рис. 25 представлены исходный временной ряд, динамика которого носит ярко выраженный нелинейный характер, и взвешенная скользящая средняя.

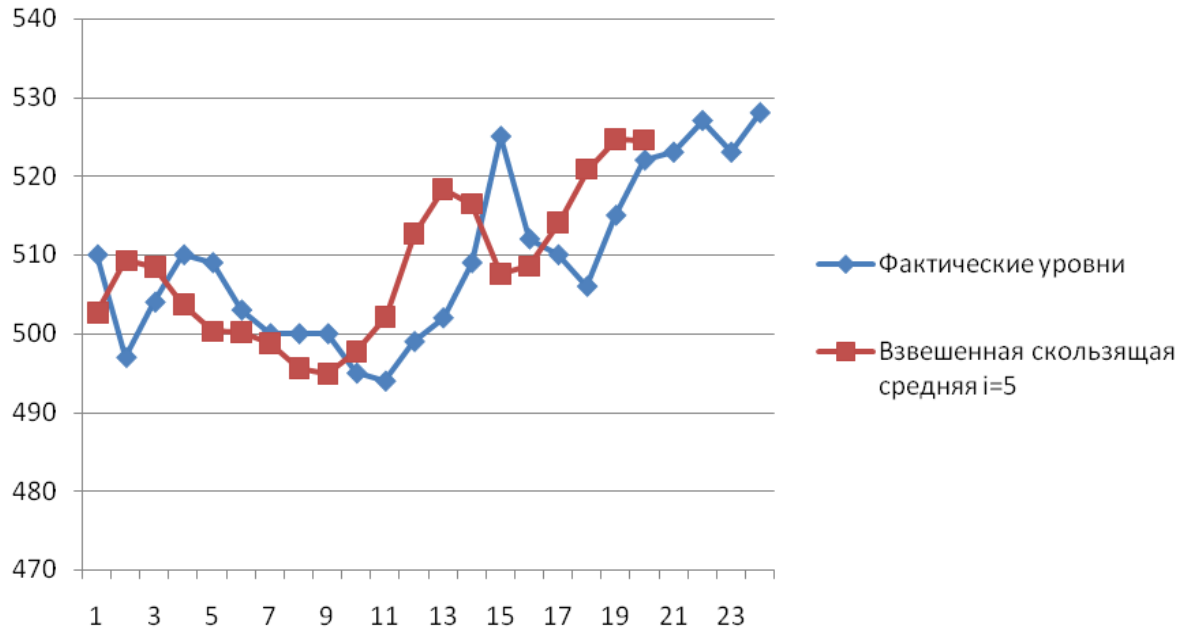

Рис. 25. Сглаживание временного ряда на основе взвешенной скользящей средней

В табл. 21 приведены результаты всех вычислений.

Таблица 21. Сглаживание ряда на основе взвешенной скользящей средней

t	y_t	**Взвешенная скользящая средняя**
1	510	–
2	497	–
3	504	502,7
4	510	509,3
5	509	508,5
6	503	503,7
7	500	500,3
8	500	500,2
9	500	498,8
10	495	495,6
11	494	494,9
12	499	497,8

Окончание табл. 21

t	y_t	**Взвешенная скользящая средняя**
13	502	502,1
14	509	512,7
15	525	518,3
16	512	516,5
17	510	507,6
18	506	508,6
19	515	514,1
20	522	520,9
21	523	524,7
22	527	524,6
23	523	–
24	528	–

Замена уровней временного ряда скользящими средними позволяет сгладить случайные и периодические колебания, определить существующую тенденцию в развитии процесса. Скользящие средние применяются при оценивании сезонной составляющей во временных рядах, в процедурах сезонной корректировки, часто на практике используются совместно с моделями кривых роста. Также они служат важным инструментом исследования в техническом анализе товарных и финансовых рынков.

5.4. О прогнозировании в экономике

Термин «прогнозирование» связан с теорией разработки прогнозов управляемых явлений (или наукой прогностикой), получившей свое развитие в нашей стране в середине XX века. *Прогноз* – это «научно обоснованное суждение о возможных состояниях объекта в будущем или об альтернативных путях и сроках достижения этих состояний» [18, с. 282]. Прогноз есть следствие действительности как единого целого; будущее, отраженное в прогнозе, – результат сложной комбинации причин и условий. Прогноз отражает реальные условия и противоречия, которые обусловливают изменение прогнозируемого явления. Прогноз представляет итог выводов эмпирических данных и обоснованных предположений. Под *экономическим прогнозированием* понимают «систему научных исследований качественного

и количественного характера, направленных на выяснение тенденции развития народного хозяйства или его частей (отраслей, регионов, предприятия и т. п.) и поиск оптимальных путей достижения целей этого развития» [18, с. 282].

Прогнозирование позволяет определить тенденции экономического развития (экономическую стратегию) на перспективу, имеет важное значение на этапе развития намеченного роста, при поиске возможностей и направлений дополнительных управляющих воздействий, нацеленных на устранение отклонений от планируемого пути развития экономической системы. Экономический прогноз может включать в сферу исследований следующие направления: прогнозы развития науки и техники; естественного движения населения, природных ресурсов и изменения окружающей среды, трудовых ресурсов и производительности труда; воспроизводства общественных благ и требуемых для этого капитальных вложений; уровня жизни населения; динамики народного хозяйства и реструктуризации производства; развития мировой экономики и внешнеэкономических связей и т. д.

Прогнозирование представляет собой аналитическую стадию процесса хозяйственного планирования, научно-исследовательское обоснование основных плановых показателей. Экономический прогноз не только интерпретирует закономерности и условия развития исследуемого явления, но и используется для поиска нужных решений. Прогнозирование в рамках планирования в условиях рыночной экономики – важное средство организации производства, распределения, объема и потребления материальных благ. Отличие прогноза (как элемента в процессе принятия и реализации экономических решений) от плана состоит в том, что плановые задания общеобязательны, характер составляющей план информации имеет директивный и адресный характер; прогноз же носит вариантный характер. Он может отражать и неуправляемые параметры, которые обусловлены стохастической природой исследуемых явлений. Вариантные результаты прогнозов могут сильно различаться из-за многообразия форм поведения экономических объектов и неоднозначности прогнозируемых параметров. Зачастую это обусловливает необходимость формирования прогноза в виде серии альтернатив, отражающих наиболее вероятные направления развития изучаемой системы. Прогнозирование составляет своеобразную преддирективную стадию плановой работы. Не ставя конкретных

заданий, прогноз может дать важнейшие показатели, необходимые для разработки планов. Главными функциями прогнозирования можно назвать квалифицированный анализ социально-экономических, научно-технических процессов и тенденций, исследование объективных причинно-следственных связей, определение ключевых проблем хозяйственного процесса, выявление альтернативы развития в перспективе, накопление информации и расчетов для выбора и принятия окончательного управленческого решения.

Прогностика выделяет такие понятия, как объект и прогнозный фон. *Объекты прогнозирования* – это те явления и процессы, на которые направлена познавательная и практическая деятельность человека. *Прогнозный фон* представляет собой «совокупность внешних по отношению к объекту условий, существенных для обоснованности прогноза» [4, с. 39].

В зависимости от сложности объектов прогнозы делятся на простые (парные зависимости) и сложные (прогнозы, учитывающие взаимосвязи и совместное влияние трех и большего числа признаков). Прогнозы по способу представления результатов подразделяются на точечные и интервальные. При точечном прогнозе получают единственный показатель. В случае интервального прогнозирования задаются нижняя и верхняя границы показателя. По методам разработки прогнозы делят на пассивные, в основе которых лежит изучение экономических процессов, обладающих большей инерционностью, и активные (целевые), учитывающие возможность некоторого воздействия на течение исследуемых явлений. В зависимости от временного горизонта прогнозы бывают оперативными, краткосрочными, перспективными (среднесрочными, долгосрочными и сверхдолгосрочными). Такое разделение связано с принятой классификацией планирования на оперативно-календарное (до месяца), текущее (годовое), перспективное (пятилетнее) и долгосрочное. По степени детерминированности различают детерминированные, стохастические (учитывающие случайную составляющую) и смешанные прогнозы. По критерию характера развития объекта во времени прогнозы могут быть дискретными (при скачкообразных изменениях процесса в фиксированные периоды времени), апериодическими (те, которые могут быть описаны непериодической функцией времени) и циклическими (те, которые могут быть отражены периодической функцией времени).

Методам прогнозирования свойственен вероятностный подход. Принято выделять три группы методов прогнозирования: экстраполяцию, моделирование, опрос экспертов. Это деление носит несколько условный характер, поскольку прогностические модели часто объединяют элементы нескольких методов в своеобразную прогнозируемую систему совместно с системами программирования, целеполагания и т. д.

Есть и иные варианты классификации видов и методов прогнозирования, например представленные в [2]. При классификации по целям прогнозирования авторы выделяют целевой, поисковый, нормативный и программный подходы к разработке прогнозов.

Целевой (генетический) прогноз состоит в целеполагании желаемых состояний, ответе на вопрос, что именно и почему желательно. При таком прогнозе по определенной шкале может строиться оценочная функция распределения предпочтительности по категориям: нежелательно, менее желательно, более желательно, рационально. Тем самым появляется возможность оценивать итоговый результат. Сама цель может быть формализована в виде критерия, сконструированного на основе параметров эффекта, зачастую являющихся ненаблюдаемыми переменными.

Поисковый прогноз заключается в определении объективно существующих тенденций развития путем анализа прошлых тенденций. Такой вид прогнозирования базируется на принципе развития из настоящего в будущее. Он позволяет охарактеризовать возможное состояние объекта в определенные моменты времени в будущем исходя из предположения о продолжении в будущем тенденции развития наблюдаемого объекта в прошлом. Мы получаем ответ на вопрос, что вероятнее всего произойдет при условии сохранения существующих тенденций и может стать основой для стратегического планирования.

Нормативный прогноз состоит в определении путей и сроков достижения возможных явлений и состояний, принятых в качестве цели. Решение прогнозной задачи должно обеспечивать реализацию конкретной цели (достижения определенного состояния) в заданные сроки. При этом ориентация прогноза во времени может происходить от будущего к настоящему. Непосредственно процесс прогнозирования начинается от конечного желаемого состояния и завершается настоящим состоянием. При данном подходе к рассмотрению прини-

мают только рациональные варианты прогноза, обеспечивающие попадание в требуемое конечное состояние из текущего исходного с учетом имеющихся ограничений на ресурсы, в том числе временные. Нормативное прогнозирование сходно с нормативными плановыми и проектными разработками. Их отличие состоит в том, что планы предполагают директивное выполнение мероприятий, в то время как прогнозирование – это только описание вероятностных альтернатив достижения заданных состояний.

Программный прогноз исследует возможные пути и меры по достижению поставленных целей и позволяет сформулировать гипотезу о возможных взаимовлияниях различных факторов, определить гипотетические сроки, очередность реализации промежуточных целей на пути к главной. Программный прогноз производят в рамках определенной проблемы в условиях поставленной цели.

При классификации по времени осуществления прогноза выделяют:

- прогноз в реальном масштабе времени, реализуемый настолько быстро, что существует возможность воздействовать на процесс во время его протекания. Цель такого прогноза – недопущение выхода объекта прогнозирования за пределы области управляемых или допустимых состояний;

- прогноз, не ограниченный по времени принятия решения. В данном случае время на прогнозирование и принятие решения практически не ограничено. Такой вид прогноза характерен для вопросов, касающихся появления концептуально новых рыночных, технологических и иных возможностей.

При классификации по степени определенности прогнозирование можно разделить на следующие типы:

- с детерминированными (определенными) условиями;

- со случайными, имеющими известное вероятностное распределение условиями;

- с неопределенными условиями.

Задачи с преобладанием значимости неопределенных условий составляют предмет теории игр. На практике менеджеры зачастую предпочитают принять решение и признать конкретные условия детерминированными (например, наихудшими), чем использовать сложный аппарат теории игр.

При классификации по степени формализации существующих условий можно выделить типы прогнозирования, использующие:

- интуитивные знания об объекте прогнозирования, которые существуют на уровне предчувствий, но недостаточно осознаны, чтобы быть вербально выраженными;

- предметное описание процесса или состояния языковыми средствами;

- описание математическими формулами.

В соответствии с этим выделяют высокую, среднюю и низкую степени формализации условий.

В случае классификации по степени достоверности знания можно разделить на достоверные (полученные из официальных источников) и знания с возможным умышленным искажением информации.

В реальности при прогнозировании и принятии решений присутствуют все три типа условий и информации одновременно. Качественный прогноз требует выделения той из этих частей, которая является определяющей в данном прогнозе, и выбора соответствующих методов разработки и верификации прогноза.

Процесс разработки прогнозов принято разделять на следующие этапы [2, с. 34]:

1. Анализ динамики моделируемого показателя и выявление тенденций изменения, циклической и сезонной составляющих и случайной компоненты.

2. Отбор основных факторов, его определяющих, и исследование тенденций их развития.

3. Обоснование метода прогнозирования и формы связи между явлениями.

4. Разработка прогноза и объективизация полученных результатов (расчет ошибок и доверительных интервалов прогнозов).

5. Содержательная интерпретация полученных результатов и их корректировка.

Отметим основные методические принципы, которые рекомендуют соблюдать при проведении прогнозирования [9, с. 27; 4, с. 42]:

1. Принцип системности требует рассмотрения объекта прогнозирования как системы взаимосвязанных элементов и характеристик в соответствии с целями и задачами исследования.

2. Принцип природной специфичности предполагает обязательный учет специфики природы объекта прогнозирования и закономерностей его развития.

3. Принцип оптимальности степени формализованности описания объекта требует использования формализованных моделей в тех соотношениях с неформальными интуитивными способами описания, которые при выполнении задачи прогноза обеспечивали бы ее решение с минимальными затратами.

4. Принцип минимизации размерности описания объекта предусматривает стремление к описанию объекта при минимальном числе переменных и параметров, обеспечивающих заданную точность и достоверность прогноза.

5. Принцип оптимального измерения показателей требует выбора для измерения каждого показателя такой шкалы, которая при минимальных затратах обеспечивала бы извлечение достаточной для прогноза информации из переменной.

6. Принцип дисконтирования данных предполагает при анализе объекта по ретроспективной информации большее значение придавать новой информации об объекте и меньшее – информации более ранней по времени. Данный принцип реализуется путем введения различных функций дисконтирования исходных данных и применения при построении модели объекта метода скользящей средней, метода экспоненциального сглаживания и т. п.

7. Принцип аналогичности предусматривает при анализе объекта постоянное сопоставление его свойств с известными в данной области сходными объектами и их моделями с целью отыскания объекта-аналога и использования при анализе и прогнозировании его модели или отдельных ее элементов.

8. Принцип согласованности означает необходимость согласования поисковых и нормативных прогнозов различной природы (признаков) и различного срока упреждения времени.

9. Принцип вариантности требует разработки вариантов прогноза, исходя из вариантов прогнозного фона.

10. Принцип непрерывности предполагает проведение корректировки прогноза по мере поступления новой информации об объекте прогнозирования.

11. Принцип верифицируемости означает потребность в достоверности, обоснованности и точности прогноза.

12. Принцип эффективности (или рентабельности) прогнозирования говорит о необходимости превышения экономического эффекта от использования прогноза над затратами по его разработке.

Нужно отметить, что в практическом анализе реальных объектов соблюсти все эти принципы обычно не удается, но каждое исследование должно быть направлено на максимальное приближение к соблюдению этих принципов. Степень такого приближения характеризует качество проведенного анализа.

Для прогнозирования экономических явлений нередко используют методы графической численной экстраполяции, основанные на построении интерполяционных многочленов Лагранжа или Ньютона (в случае равноотстоящих узлов). Экстраполяцию определяют как «метод, при котором прогнозируемые показатели рассчитываются как продолжение динамического ряда на будущее по выявленной закономерности развития» [4, с. 43]. Однако экстраполяция далеко за крайний узел известных значений малонадежна, в особенности если моделирующая функция быстро изменяется с ростом значений аргумента, и также нецелесообразна, если исходные данные получены со значительными погрешностями. В этой связи лучший результат дают статистические методы прогнозирования на основе корреляционно-регрессионных моделей, которые позволяют выделить детерминированную составляющую – тренд (регрессионную модель) экспериментальной зависимости, с помощью которого строится прогноз исследуемого экономического процесса с определенным доверительным интервалом. Такой подход предпринят при прогнозировании туристских потоков в [14]. Таким образом, для прогнозирования активно изменяющихся экономических явлений в настоящее время самыми распространенными считаются методы экстраполяции тенденций, использующие вероятностные подходы к изучению объектов прогнозирования (см., в частности, [5, 9]). В основе этих подходов лежит предположение о том, что рассматриваемый процесс сочетает в себе детерминированную и случайную составляющие $\tilde{y}(x) = f(\vec{a}, x) + \eta(x)$. Предполагается, что $f(\vec{a}, x)$ есть гладкая функция своих аргументов x (в большинстве случаев времени или факторной переменной) и вектора параметров $\vec{a}$, которые установлены и не изменяются на перио-

де упреждения прогноза. Детерминированная составляющая $f(\vec{a}, x)$ называется *трендом* и принимается за основу или тенденцию развития процесса в будущем. Считается также, что случайная составляющая $\eta(x)$ имеет нулевое математическое ожидание, и ее оценки, в частности дисперсия, характеризуют точность прогнозных значений. Чтобы снизить влияние случайной составляющей выборочных значений, т. е. приблизить исходный числовой ряд к тренду, полезной оказывается предварительная обработка исходных данных, получившая название процедуры сглаживания статистического ряда. Таким образом, процедура сглаживания направлена на минимизацию случайных отклонений экспериментальных точек от предполагаемого тренда исследуемого экономического процесса. Методы сглаживания получили распространение при исследовании динамических рядов, где в качестве независимой переменной выступает время.

При изменении зависимости результативного признака Y от одной или нескольких факторных переменных X детерминированную составляющую $f(\vec{a}, x)$ получают на основе корреляционно-регрессионного анализа. При этом далеко не каждое уравнение регрессии можно принять за модель развития изучаемого экономического явления. Согласно [8, с. 245], «*корреляционно-регрессионной моделью* системы взаимосвязанных признаков является такое уравнение регрессии, которое включает основные факторы, влияющие на вариацию результативного признака, и обладает высоким (не ниже 0,5) коэффициентом детерминации и коэффициентами регрессии, интерпретируемыми в соответствии с теоретическим знанием о природе связей в изучаемой системе». Таким образом, правильная интерпретация и применение в экономике моделей регрессии возможны лишь при понимании всех специфических черт исследуемого процесса.

Прогноз результативного признака Y на основе корреляционно-регрессионной модели осуществляется расчетом значений Y при подстановке в уравнение регрессии $y = y(x)$ ожидаемых значений факторных переменных X.

Использование регрессионной модели для прогнозирования, как правило, сопровождается построением доверительных интервалов прогностических значений результативного признака Y:

$$Y_{\text{рег}}(x) - t_p \sigma_{Y_{\text{рег}}(x)} < Y(x) < Y_{\text{рег}}(x) + t_p \sigma_{Y_{\text{рег}}(x)},$$

где $\sigma_{Y_{\text{рег}}(x)}$ – среднеквадратичная ошибка положения линии регрессии при значении $X = x$ [8, с. 214]; для линейной модели

$$\sigma_{Y_{\text{рег}}(x)} = s_{Y_{\text{ост}}} \left(1/n + \left(x - \bar{X}\right)^2 \Big/ \sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2\right)^{\frac{1}{2}};$$

t_p – значение t-критерия Стьюдента при уровне значимости p.

Средняя ошибка прогноза для индивидуального значения вычисляется по формуле [8, с. 215] $m_{Y(x)} = \left(\sigma^2_{Y_{\text{рег}}(x)} + s^2_{Y_{\text{ост}}}\right)^{\frac{1}{2}}$, где $s_{Y_{\text{ост}}} = \left(\left(\sum_{i=1}^{n}\left(y_i - y_{i\text{рег}}\right)^2\right)\Big/\left(n-2\right)\right)^{\frac{1}{2}}$ – оценка среднего квадратичного отклонения результативного признака Y от линии регрессии в генеральной совокупности с учетом степеней свободы вариации.

Отметим, что малый объем выборки может послужить причиной широкого доверительного интервала, что приводит к неопределенности прогноза по линии регрессии. Неопределенность прогноза по однофакторной модели также может быть обусловлена вариацией посторонних факторов. Следует учесть, что перенос закономерной связи $Y = Y(X)$, выявленной в варьирующей совокупности исходного конечномерного вектора информации (x_i, y_i) $(i = 1, 2, \ldots, n)$, взятого в статике, на динамику требует достаточных знаний об объекте исследования и возможностях его развития в будущем.

Прогнозирование можно осуществлять на основе уравнения регрессии $\tilde{y}_{x,y} = \tilde{y}(x, y, t)$ для взаимосвязанных динамических рядов. При этом так же, как и в случае однофакторной временной модели $\tilde{y} = \tilde{y}(t)$, следует предварительно проверить адекватность уравнений тренда экспериментальным данным.

Прогноз на основе аналитического выравнивания, конечно, не может дать точного совпадения прогнозных оценок с действительными результатами и носит вероятностный характер, как любое суждение о будущем. Различают перспективную экстраполяцию, дающую прогноз в будущее, и ретроспективную – в прошлое. Наряду с этим возможна интерполяция (расчет по уравнению тренда) неизвестных уровней внутри самого динамического ряда. Для прогнозных значений следует строить доверительный интервал. Точность и надежность прогнозов зависит от того, насколько инерционно изучаемое эконо-

мическое явление, насколько точно удалось выявить тенденцию его развития, что связано с качеством модели и устойчивостью ее параметров. Точность прогноза зависит также от периода экстраполяции: чем он уже, тем точнее прогноз. Прогноз по тренду основан на предположении, что его параметры сохраняются до прогнозируемого периода, что возможно, если рассматриваемая экономическая система развивается в достаточно стабильных условиях. Как правило, рекомендуется, чтобы срок прогноза не превышал 1/3 длины базового динамического ряда. Следует также иметь в виду, что в активно развивающихся экономических системах длинные динамические ряды становятся несопоставимыми, так как с течением времени может измениться методология расчета статистических показателей.

5.5. Прогнозирование рядов динамики на основе адаптивных методов

Если процесс имеет малую инерционность, то информативность уровней ряда по мере их удаления от периода прогнозирования склонна снижаться, так что наиболее ценную информацию будут содержать последние периоды. В этой связи представляют интерес современные адаптивные методы прогнозирования, где учитывается информационная ценность уровней ряда с определенными весомыми коэффициентами. Цель адаптивных методов заключается в построении самокорректирующихся в условиях изменений экономико-математических моделей. Такие модели предназначены прежде всего для краткосрочного прогнозирования. При этом прогнозная модель – это модель, аппроксимирующая тренд. Здесь прогнозы – это значения будущих членов временного ряда, а последовательность прогнозов $\{y_{t+\tau}\}$ ($\tau = 1, 2, \ldots, k$) составляет оценку тренда для периодов упреждения длиной τ.

У истоков адаптивного направления в прогнозировании лежит построение моделей экспоненциального сглаживания [22, с. 325 – 333], в основе которого лежит расчет экспоненциальных средних по рекуррентной формуле

$$S_t(y) = \alpha y_t + (1-\alpha) y_{t-1},$$

где S_t – значение экспоненциальной средней в момент времени t; $\{y_t\}$, $t = 1, \ldots, T$ – исходный динамический ряд; α – параметр сглаживания, $\alpha = \text{const}$, $0 < \alpha < 1$. В 1960-х гг. Р. Г. Брауном и Р. Ф. Майером были

разработаны полиномиальные модели прогнозирования на основе многократного экспоненциального сглаживания. Авторы принимают гипотезу, что исследуемый временной ряд $\{y_t\}$, $t = 1, \ldots, T$ представляет собой многочлен n-го порядка, а прогноз вперед на глубину τ выражается формулой

$$\tilde{y}(T+\tau) = \hat{a}_1 + \hat{a}_2\tau + \frac{1}{2!}\hat{a}_3\tau^2 + \ldots + \frac{1}{n!}\hat{a}_{n+1}\tau^n,$$

где $\hat{a}_1$, $\hat{a}_2$, $\hat{a}_{n+1}$ – параметры модели, которые необходимо рассчитать в результате многократного применения к исходному ряду оператора сглаживания $S_t^{[p]} = \alpha S_t^{[p-1]} + (1-\alpha) S_{t-1}^{[p]}$, где $p = 1, 2, \ldots, n$; $S_t^{[0]} = y_t$; $S_0^{[1]}, S_0^{[2]}, \ldots, S_0^{[n]}$ – начальные значения экспоненциальных средних соответствующего порядка. Предполагается, что прогнозирующий полином $\tilde{y}(T + \tau)$ представляет собой сумму $(n + 1)$ членов разложения в ряд Тейлора процесса $\{y_t\}$:

$$y(t+\tau) = \sum_{k=0}^{n} \frac{\tau^k y_t^{(k)}}{k!} = \sum_{k=0}^{n} \frac{\tau^k a_{k+1}}{k!}.$$

Доказывается, что коэффициенты прогнозирующего полинома связаны с экспоненциальными средними матричными уравнениями $S_t = Mb_t$, где M – матрица с компонентами $m_{ik} = \frac{\alpha^i}{(i-1)!}\sum_{j=0}^{\infty} j^k (1-\alpha)^j \frac{(i-1+j)!}{j!}$; b_t – вектор коэффициентов в разложении Тейлора. Из матричного уравнения находят коэффициенты прогнозирующего полинома. На практике обычно используют полиномы не выше второго порядка.

Приведем пример [15] построения адаптивной полиномиальной модели второго порядка ($n = 2$), где объект исследования – туристский поток по въезду иностранных граждан в Российскую Федерацию, условно обозначенный через $y(t)$. Его значения взяты из источников Всемирной туристской организации за 1992 – 2003 гг. Здесь используется метод экспонентного сглаживания, согласно которому временной ряд исходных данных представляется в виде $y(t) = f(t) + E(t)$, где $1 \le t \le T$ – переменная времени; $f(t)$ – детерминированная составляющая (тренд); $E(t)$ – случайная составляющая временного ряда.

Предполагается, что *f(t)* изменяется достаточно постепенно, эволюторно на определенном отрезке времени и математическое ожидание $M[E(t)] = 0$, а дисперсия $D[E(t)] = \sigma^2$. К данному ряду применяют оператор сглаживания $S_t(y) = cy_t + (1-c)S_{t-1}(y)$, где c – константа сглаживания такая, что $0 < c \leq 1$. При конечном значении $t = T$ вычисляется

$$S_T(y) = cy_T + c(1-c)y_{T-1} + c(1-c)^2 y_{T-2} + ... + c(1-c)^{T-2} y_2 + (1-c)^T y_1,$$

откуда следует, что в процессе сглаживания наблюдения учитываются с экспоненциально убывающими весами, т. е. последующие наблюдения воспринимаются с бóльшим доверием, чем предыдущие. Доказывается, что дисперсия $D\left[S_t(E)\right] = c\sigma^2 / (2-c) \angle \sigma^2$, т. е. при сглаживании роль случайной составляющей понижается. Применяют тройное экспоненциальное сглаживание. На основе значений сглаженных рядов вычисляют коэффициенты прогнозирующего квадратичного полинома

$$A_t = 3S_t^{(1)}(y) - 3S_t^{(2)}(y) + S_t^{(3)}(y);$$

$$C_t = \frac{c^2}{(1-c)^2}\left[S_t^{(1)}(y) - 2S_t^{(2)}(y) + S_t^{(3)}(y)\right];$$

$$B_t = \frac{c}{2(1-c)}\left[(6-5c)S_t^{(1)}(y) - 2(4-5c)S_t^{(2)}(y) + (4-3c)S_t^{(3)}(y)\right].$$

Здесь $S_t^{(1)}(y)$, $S_t^{(2)}(y)$, $S_t^{(3)}(y)$ – значения трех последовательных сглаженных рядов. Далее точечный прогноз на глубину τ осуществляется по формуле $y_{T+\tau} = A_T + B_T\tau + C_T\tau^2 / 2$.

Числовые расчеты проводились с помощью пакета прикладных программ по математической статистике. При $T = 12$ и $c = 0{,}8$ выполнялось тройное экспоненциальное сглаживание ряда $y(t)$: 0,67; 1,58; 0,92; 1,84; 1,90; 2,51; 2,98; 3,06; 2,60; 2,38; 3,11; 3,15. Сглаженные значения в конечной точке получены следующие: $S_T^{(1)}(y) = 3{,}115$; $S_T^{(2)}(y) = 3{,}068$; $S_T^{(3)}(y) = 3{,}017$. Дисперсия случайной составляющей $D[E(t)] = 0{,}125\sigma^2$. Построен прогнозирующий полином $y_{T+\tau} = 3{,}158 + 2{,}0248\tau - 0{,}032\tau^2$, дающий прогноз по въезду туристов в РФ. Более осторожный прогноз дает линейная модель в результате двукратного сглаживания. Аналогично построен прогноз по выезду граждан РФ за рубеж с туристской целью.

В том случае, когда в окрестности точки $t = T$ детерминированная составляющая временного ряда близка к постоянной, применяют аппа-

рат однократного экспоненциального сглаживания и прогноз определяется по формуле $\hat{y}_{T+\tau} = S_T^{(1)}(y)$. Такая модель отображает динамику в виде случайного процесса, не имеющего тенденции, её называют «наивной» («будет, как было»). Доверительный интервал прогноза получают по формуле $y_{t+\tau} \pm S_e t_p \sqrt{\frac{\alpha}{2-\alpha}}$, где S_e – средняя квадратичная ошибка аппроксимации исходного ряда, $S_e = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{y}_t - y_t\right)^2}$; t_p – значение t-критерия Стьюдента с уровнем значимости p и числом степеней свободы $(T-1)$; α – параметр сглаживания.

Если в окрестности точки T детерминированная составляющая линейная, то применяют двойное экспоненциальное сглаживание и точечный прогноз осуществляют по формуле $\hat{y}_{T+\tau} = \hat{a}_0^{(T)} + \hat{a}_1^{(T)}\tau$. Границы доверительного интервала прогноза следующие: $y_{t+\tau} \pm S_e t_\alpha \sqrt{\frac{\alpha\left[1+4(1-\alpha)+5(1-\alpha)^2+2\alpha(4-3\alpha)\tau+2\alpha^2\tau^2\right]}{(2-\alpha)^3}}$.

Если в окрестности точки T траектория динамики ряда близка к параболе, то строят трехпараметрическую адаптивную модель прогнозирования в виде квадратичного полинома $\hat{y}_{T+\tau} = \hat{a}_0^{(T)} + \hat{a}_1^{(\mathrm{T})}\tau + \frac{1}{2}\hat{a}_2^{(T)}\tau^2$. Границы доверительного интервала прогнозных значений вычисляют по формуле $\hat{y}(T+\tau) \pm t_{0,05}^{(n-1)} S_e \sqrt{2\alpha + 3\alpha^2 + 3\alpha^3\tau}$, где $t_{0,05}^{(n-1)}$ – значение t-критерия Стьюдента с числом степеней свободы $(n-1)$ и уровнем значимости 0,05.

Ниже, в соответствии с [22, с. 311 – 318], представлены основные этапы построения линейной адаптивной модели и полиномиальной модели второго порядка.

Алгоритм построения линейной адаптивной модели Брауна

1. По первым 5 – 10 точкам временного ряда методом наименьших квадратов оценивают начальные значения параметров $\hat{a}_1$ и $\hat{a}_0$ линейной модели $\hat{y}_t = \hat{a}_0 + \hat{a}_1 t$.

2. На основе $\hat{a}_1$ и $\hat{a}_0$ вычисляют начальные значения экспоненциальных средних, которые соответствуют моменту времени $t = 0$:

$$S_0^{(1)} = a_{0(0)} - \frac{\beta}{\alpha} a_{1(0)},\ S_0^{(2)} = a_{0(0)} - \frac{2\beta}{\alpha} a_{1(0)}.$$

3. С учетом выбранного значения параметра сглаживания α или коэффициента дисконтирования данных β = 1 – α вычисляют значения экспоненциальных средних в последующий момент времени:

$$S_t^{(1)} = \alpha y_t + \beta S_{t-1}^{(1)},\ S_t^{(2)} = \alpha S_t^{(1)} + \beta S_{t-1}^{(2)}.$$

4. Корректируют параметры линейной модели $a_{0(t)}$ и $a_{1(t)}$ по формулам

$$a_{0(t)} = 2S_t^{(1)} - S_t^{(2)},$$

$$a_{1(t)} = \frac{\alpha}{\beta}(S_t^{(1)} - S_t^{(2)}).$$

5. По модели со скорректированными параметрами $a_{0(t)}$ и $a_{1(t)}$ находят прогноз на следующий момент времени:

$$\hat{y}_t(\tau) = a_{0(t)} + a_{1(t)}\tau = \hat{y}_{(t+1)} = a_{0(t)} + a_{1(t)}.$$

6. Возврат к пункту 3, если $t < T$. Если $t = T$, то модель с параметрами $a_{0(t)}$ и $a_{1(t)}$ принимают за основу для прогнозирования на τ = 1, 2, … шагов вперёд.

7. Расчет показателей точности модели: средней квадратичной ошибки аппроксимации $S_e = \left(\frac{1}{T-1} \sum_{t=1}^{T} e_t^2 \right)^{\frac{1}{2}}$, где $e_t = y_t - \hat{y}_t$ – остатки; средней относительной ошибки аппроксимации: $\varepsilon_e = \frac{1}{T} \sum_{t=1}^{T} \frac{|e_t|}{y_t} 100\%$.

8. Точечный и интервальный прогнозы.

Алгоритм расчета параметров квадратичного прогнозирующего полинома

Адаптивная полиномиальная модель второго порядка имеет вид

$$\hat{y}_t(\tau) = a_{0(t)} + a_{1(t)}\tau + \frac{a_{2(t)}}{2}\tau^2.$$

1. По первым 5 – 10 точкам исходных данных находят оценки $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$ коэффициентов квадратичной модели по методу наименьших квадратов. Таким образом получают начальные значения параметров $\hat{a}_{0,0}$, $\hat{a}_{1,0}$, $\hat{a}_{2,0}$ квадратичной модели, которые соответствуют моменту времени $t = 0$.

2. Вычисляют начальные значения экспоненциальных средних (с учетом выбранного значения параметра сглаживания α или коэффициента β):

$$S_0 = \hat{a}_{0,0} - \frac{\beta}{\alpha}\hat{a}_{1,0} + \frac{\beta(2-\alpha)}{2\alpha^2}\hat{a}_{2,0};$$

$$S_0^1 = \hat{a}_{0,0} - \frac{2\beta}{\alpha}\hat{a}_{1,0} + \frac{\beta(3-2\alpha)}{\alpha^2}\hat{a}_{2,0};$$

$$S_0^2 = \hat{a}_{0,0} - \frac{3\beta}{\alpha}\hat{a}_{1,0} + \frac{3\beta(4-3\alpha)}{2\alpha^2}\hat{a}_{2,0}.$$

3. Рассчитывают значения экспоненциальных средних в последующий момент времени *t*:

$$S_t = \alpha y_t + \beta S_{t-1};\ S_t^1 = \alpha S_t + \beta S_{t-1}^1;\ S_t^2 = \alpha S_t^1 + \beta S_{t-1}^2.$$

4. Корректируют параметры квадратичной модели по формулам

$$\hat{a}_{0,T} = 3S_T - 3S_T^1 + S_T^2;$$

$$\hat{a}_{1,T} = \frac{\alpha}{2\beta^2}\left[(6-5\alpha)S_T - 2(5-4\alpha)S_T^1 + (4-3\alpha)S_T^2\right];$$

$$\hat{a}_{2,T} = \frac{\alpha^2}{\beta^2}\left[S_T - 2S_T^1 + S_T^2\right].$$

Отметим, что в формулах участвует коэффициент β дисконтирования данных, отражающий бóльшую степень доверия более поздним наблюдениям.

5. По модели Брауна $\hat{y}_t(\tau) = a_{0(t)} + a_{1(t)}\tau + \frac{a_{2(t)}}{2}\tau^2$ со скорректированными коэффициентами $a_{0(t)}$, $a_{1(t)}$, $a_{2(t)}$ находят прогноз на следующий момент времени τ = 1: $\hat{y}_t(1) = a_{0(t)} + a_{1(t)} + \frac{a_{2(t)}}{2}$.

6. Переход к пункту 3, если $t < T$. Если $t = T$, то построенную модель $\hat{y}(T+\tau) = a_{0(T)} + a_{1(T)}\tau + \frac{a_{2(T)}}{2}\tau^2$ принимают за основу прогнозирования будущих значений ряда $y(t)$.

7. Расчет показателей точности модели: вычисление среднего квадратичного отклонения S_e и средней относительной ошибки аппроксимации ε_e.

8. Точечный и интервальный прогнозы.

При построении квадратичного полинома так же, как и в случае линейной модели, оптимальное значение параметра сглаживания α находят многократным построением модели при разных значениях α и выбором наилучшего значения в целях повышения точности аппроксимации исходного ряда и снижения ошибки прогнозирования. В примере ниже за наилучшую принимается модель с наименьшей средней относительной ошибкой аппроксимации ε_e.

Пример 6. Расчет линейной адаптивной модели прогнозирования процентной ставки кредитных организаций. Ниже рассматривается динамический ряд значений средних процентных ставок кредитных организаций России по краткосрочным кредитам в долларах США (% годовых). Данные показатели играют важную роль в финансовой сфере деятельности предприятий, а также реальных физических лиц. С точки зрения экономической теории увеличение процентной ставки на кредиты может привести к росту инфляции, ибо при этом деньги обесцениваются. Исходя из вышесказанного научно обоснованный прогноз таких процентных ставок имеет важное значение в управлении финансами. Исходные данные представлены в табл. 22 (источник – официальный сайт Центрального Банка РФ: http://www. cbr.ru).

На первом этапе работы по первым десяти точкам находят начальные коэффициенты $\hat{a}_0$, $\hat{a}_1$ для линейной модели согласно методу наименьших квадратов. Таким образом строится модель линейной регрессии $\tilde{y}(t) = 12{,}66 - 0{,}21035t$ со следующими статистическими характеристиками: коэффициентом парной корреляции $r_{yt} = -0{,}75$; стандартной ошибкой регрессии $S = 0{,}59$; критерием $F_{\text{набл}} = 10{,}42$. Согласно таблице определяется значение $F_{\text{крит}} = (\alpha, \nu_1, \nu_2)$, где $\nu_1 = 1$ и $\nu_2 = n - 2$ – количество степеней свободы; α – уровень значимости. В данном случае $\nu_1 = 1$ – количество факторов, включенных в модель, $\nu_2 = 8$, $\alpha = 0{,}05$, $F_{\text{крит}} = 5{,}32$. Так как $F_{\text{набл}} > F_{\text{крит}}$, на этом основании при заданном уровне значимости $\alpha = 0{,}05$ построенная линейная модель является адекватной экспериментальным данным, т. е. статистически значима. Коэффициент линейной корреляции $r_{yt} = -0{,}75$ указывает на значимую обратную связь, это означает снижение процентной ставки с течением времени. Вычисленные коэффициенты линейной модели выбираются в качестве начальных оценок при построении адаптивной модели и корректируются в ходе реализации алгоритма, изложенного выше. При этом используются программные средства Microsoft Excel.

Табл. 23 иллюстрирует алгоритм расчета с параметром сглаживания α = 0,3, соответствующим наименьшей средней относительной ошибке аппроксимации исходного ряда.

Таблица 22. Ставки по краткосрочным кредитам нефинансовым организациям

Год	**Квартал**	**Условное обозначение времени *t***	**Процентная ставка нефинансовым организациям y_t, %**
2000	I	1	12,200
	II	2	12,133
	III	3	11,666
	IV	4	11,766
2001	I	5	13,133
	II	6	11,000
	III	7	11,333
	IV	8	10,766
2002	I	9	10,666
	II	10	10,366
	III	11	10,566
	IV	12	10,266
2003	I	13	11,366
	II	14	9,000
	III	15	8,800
	IV	16	8,666
2004	I	17	8,800
	II	18	7,666
	III	19	8,166
	IV	20	8,400
2005	I	21	7,866
	II	22	8,600
	III	23	9,033
	IV	24	8,833
2006	I	25	8,633
	II	26	8,366
	III	27	8,566
	IV	28	8,500
2007	I	29	8,733
	II	30	8,866
	III	31	8,400

Таблица 23. Расчет параметров линейной адаптивной модели

t	y_t	$\hat{a}_0$	$\hat{a}_1$	$S_t^{(1)}$	$S_t^{(2)}$	$\hat{y}_t$	e_t
0	–	12,660	–0,210	13,151	13,642	–	–
1	12,200	12,322	–0,233	12,866	13,409	12,450	–0,317
2	12,133	12,112	–0,229	12,646	13,180	12,090	–0,424
3	11,666	11,772	–0,248	12,352	12,931	11,883	–0,117
4	11,766	11,647	–0,227	12,176	12,705	11,524	1,609
5	13,133	12,294	–0,073	12,463	12,632	11,421	–0,421
6	11,000	11,599	–0,182	12,024	12,450	12,221	–0,888
7	11,333	11,374	–0,190	11,817	12,260	11,416	–0,650
8	10,766	10,971	–0,228	11,502	12,032	11,184	–0,518
9	10,666	10,704	–0,234	11,251	11,798	10,743	–0,377
10	10,366	10,417	–0,244	10,985	11,554	10,469	0,097
11	10,566	10,373	–0,208	10,860	11,346	10,173	0,093
12	10,266	10,216	–0,199	10,682	11,147	10,165	1,201
13	11,366	10,705	–0,078	10,887	11,069	10,017	–1,017
14	9,000	9,797	–0,224	10,321	10,844	10,627	–1,827
15	8,800	9,179	–0,294	9,865	10,550	9,573	–0,907
16	8,666	8,773	–0,314	9,505	10,237	8,885	–0,085
17	8,800	8,633	–0,283	9,293	9,954	8,460	–0,794
18	7,666	8,001	–0,345	8,805	9,609	8,350	–0,184
19	8,166	7,916	–0,299	8,613	9,310	7,657	0,743
20	8,400	8,017	–0,228	8,549	9,082	7,618	0,248
21	7,866	7,828	–0,221	8,344	8,861	7,788	0,812
22	8,600	8,113	–0,132	8,421	8,729	7,607	1,426
23	9,033	8,518	–0,037	8,605	8,692	7,981	0,852
24	8,833	8,660	–0,006	8,673	8,686	8,480	0,153
25	8,633	8,644	–0,007	8,661	8,679	8,655	–0,289
26	8,366	8,498	–0,032	8,573	8,647	8,636	–0,070
27	8,566	8,517	–0,023	8,571	8,624	8,467	0,033
28	8,500	8,497	–0,022	8,549	8,602	8,494	0,239
29	8,733	8,607	0,001	8,604	8,602	8,475	0,391
30	8,866	8,739	0,024	8,683	8,627	8,607	–0,207
31	8,400	8,578	–0,009	8,598	8,618	8,763	0,751

Здесь t – условное обозначение времени; y_t – фактическое значение процентной ставки; $\hat{a}_0$, $\hat{a}_1$ – коэффициенты модели; $S_t^{(1)}$, $S_t^{(2)}$ – экспоненциальные средние; $\hat{y}_t$ – расчетное значение по модели; e_t – отклонение от фактического значения.

Таким образом, получены следующие значения параметров линейной адаптивной модели: $\hat{a}_0 = 8{,}578$; $\hat{a}_1 = -0{,}009$. Здесь среднеквад-

ратическая ошибка аппроксимации S_e = 0,697; средняя относительная ошибка аппроксимации ε_e = 0,099, что составляет 9,9 %. Так построена линейная адаптивная модель: $\tilde{y}_t(\tau) = 8{,}578 - 0{,}009\tau$ (с параметром сглаживания $\alpha = 0{,}3$).

Для выбора лучшей модели была вычислена средняя относительная ошибка аппроксимации для различных параметров сглаживания, результаты расчета приведены в табл. 24.

Таблица 24. Значения (α, ε_e) для линейной модели

α	**ε_e**
0,1	13,3
0,2	10,7
0,3	9,95
0,4	10,0
0,5	10,9
0,6	11,7
0,7	12,9
0,8	14,2
0,9	15,7

В ряде случаев экономико-математические модели прогнозирования могут быть полезным инструментом исследования. При этом, конечно, для увеличения точности прогнозов экономического развития в изменяющихся условиях, условиях неопределенности или неполной информации необходима работа по совершенствованию моделей. Важную роль в этом должны сыграть адаптивные методы прогнозирования. К примеру, в работе [14] адаптивные модели Брауна используют в целях прогнозирования туристских ресурсов Владимирской области. Отличие адаптивных моделей от других прогностических моделей состоит в том, что они отражают текущие свойства ряда и способны непрерывно учитывать эволюцию динамических характеристик изучаемых процессов.

Контрольные вопросы

1. Сформулируйте определение временного ряда.
2. Приведите примеры моментных и интервальных временных рядов.

3. Каковы основные задачи математического моделирования экономических процессов, представленных одномерными временными рядами?

4. Что понимается под аналитическим выравниванием временных рядов? Какие основные функции используют для моделирования трендовой составляющей временного ряда? Для описания каких процессов делается выбор в пользу каждой из них?

5. Укажите формулы сглаживания сезонных колебаний при работе с временными рядами поквартальной и месячной динамики.

6. Какие модели применяются для описания периодических колебаний случайных процессов?

7. Что представляют собой модели авторегрессии и скользящего среднего? Для прогнозирования каких процессов их используют?

8. По каким критериям, на ваш взгляд, можно осуществить выбор наилучшей модели тренда?

9. Что понимают под термином «прогнозирование» в сфере экономики?

10. Каковы основные типы прогнозов, используемых в экономических исследованиях?

11. Укажите основные методические принципы, которые рекомендуют соблюдать при проведении прогнозирования.

12. Каковы способы определения взаимосвязи между двумя временными рядами? Какие при этом могут возникнуть проблемы и каковы способы их устранения?

13. Какие методы используют для изучения динамики взаимосвязанных экономических процессов?

14. Как проверить наличие автокорреляции остатков в модели временного ряда? Укажите способы устранения автокорреляции остатков.

15. В каких целях используют адаптивные модели прогнозирования? Какое характерное отличие адаптивных методов от построения моделей кривых роста? Какие типы адаптивных моделей вам известны?

16. Опишите основные этапы расчета адаптивных полиномиальных моделей Брауна первого и второго порядка.

17. Как выбрать лучшую модель экспоненциального сглаживания в целях прогнозирования?

18. По каким направлениям выполняют оценку качества моделей временных рядов? Каковы оценки точности моделей?

19. Какую роль играет кросскорреляционная функция для анализа взаимосвязанных динамических рядов?

Глава 6. ПРОГНОЗИРОВАНИЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ, ПОДВЕРЖЕННЫХ СЕЗОННЫМ КОЛЕБАНИЯМ

6.1. Понятие сезонных колебаний и сезонной составляющей

Сезонные колебания – это повторяющиеся в каждом временном периоде колебания, связанные с изменением времени года. Такие изменения непосредственно могут быть связаны с колебаниями других факторов: политических, экономических, социальных, природных. Если в течение года имеется только одно повышение (снижение) уровня за несколько лет, то говорят об одном сезонном цикле; если в течение периода наблюдают несколько минимумов и максимумов, то статистическую модель с учётом сезонности выбирают согласно полученному циклическому процессу.

Отметим, что временной ряд не всегда содержит сезонную (циклическую) составляющую. Проверку на наличие или отсутствие сезонных колебаний проводят с помощью какого-либо критерия или визуально при построении графика. При подтверждении наличия сезонного процесса выделяется сезонная составляющая. Процедуры расчета значений сезонной компоненты зависят от принятой модели временного ряда, содержащей сезонность в аддитивной или мультипликативной форме.

Аддитивную модель $y_t = u_t + v_t + \varepsilon_t$ применяют в том случае, когда амплитуда сезонных колебаний со временем не меняется. Если происходят существенные сезонные изменения, то используют мультипликативную модель: $y_t = u_t\, v_t\, \varepsilon_t$. При этом для аддитивной модели характеристики сезонности будут измеряться в абсолютных величинах, а для мультипликативной – в относительных величинах.

Временной ряд, в котором наблюдаются и тренд, и сезонные колебания, будем называть *тренд-сезонным временным рядом*. Статистические процедуры оценивания сезонности и корректировки временных рядов представляют большой интерес в сфере экономики, причем оценки сезонности представляют собой лишь промежуточный этап исследования. В дальнейшем они используются для моделирования тренд-сезонных процессов. Методика построения прогнозирования с помощью тренд-сезонных моделей включает в себя следующие основные этапы [22, с. 80]:

1. Анализ динамики сезонной волны. Оценивание сезонной составляющей (аддитивной или мультипликативной в зависимости от характера сезонности).

2. Десезонализация исходных данных.

3. Расчет параметров тренда на основе преобразованного временного ряда.

4. Моделирование динамики исходного ряда с учетом трендовой и сезонной составляющих.

5. Использование построенной модели в целях прогнозирования в случае, если получены удовлетворительные характеристики точности и адекватности модели.

Порядок выполнения данных этапов и их состав может изменяться в зависимости от постановки задач, методов решения, используемых программных средств. Остановимся подробнее на расчёте индексов сезонности и построении сезонной волны.

Индексом сезонности I_s называют относительный показатель, который используют для расчета сезонной составляющей. При исчислении индексов применяют разные методы, выбор которых зависит от характера общей тенденции временного ряда. Расчет индексов сезонности применим лишь тогда, когда тренд исключен из динамического ряда или имеет постоянный уровень. Методика расчета индексов сезонности в деталях изложена в [27, с. 140 – 158].

Если ряд динамики не имеет выраженной тенденции развития, то индексы сезонности исчисляют непосредственно по эмпирическим данным без их предварительного выравнивания. Моделирование циклических колебаний осуществляется по методике, аналогичной методике моделирования сезонной составляющей.

Для расчета индекса сезонности необходимо иметь данные по периодам не менее чем за три года. Сущность метода заключается в расчете по одноименным периодам (месяцам или кварталам) среднего значения $\overline{y_s}$ и для всего анализируемого ряда общего среднего уровня ряда – $\overline{y}$. По этим данным определяют индекс сезонности $I_s = \frac{\overline{y}_s}{\overline{y}} 100\ \%$.

В качестве среднего уровня ряда может быть использована также средняя арифметическая взвешенная, мода или другая структурная средняя. Их применяют для временных рядов за достаточно длительный период времени или для элиминирования случайной составляющей.

Если ряд динамики *имеет тренд* (нестационарный ряд динамики), то порядок расчета включает в себя этапы:

- определения по внутригодовым уровням ряда (месячным, квартальным) за несколько лет расчетных (выровненных) уровней по методикам скользящей средней или аналитического выравнивания;

- определения относительной величины фактических значений уровней ряда y_t и выравненных (расчетных) значений $\tilde{y}_t$;

- усреднения полученных показателей сезонности за весь исследуемый период по формуле $I_s = \frac{\sum_t \frac{y_t}{\tilde{y}_t}}{n}$.

Для элиминирования сезонной составляющей в аддитивных моделях выравненные и скорректированные уровни ряда динамики вычитают из исходных значений ряда. Если ряд динамики имеет длительный период (15 – 20 лет), то сезонные колебания выявляют либо с учетом единой качественной особенности периода (ряд сначала делят на качественно однородные периоды), либо согласно методике многократного скользящего выравнивания.

Для получения устойчивой тенденции сезонных колебаний, на которых бы не отражались особенности развития процессов в конкретные периоды, индекс сезонности рассчитывают по формуле $\bar{I}_s = \frac{\sum_s I_s}{t}$, где I_s – индексы сезонности; t – число лет. На основе полученного индекса сезонности рассчитывают коэффициент сезонности $K_s = \sqrt{\frac{\sum_s (I_s - 1)^2}{n}}$, где n – число периодов. Коэффициент изменяется от 0 до 1. Если функция для измерения сезонных колебаний имеет синусоидальную форму, то используют метод Фурье. В гармониках Фурье исходный ряд – это не первичный ряд за несколько лет, а усредненные значения месячных уровней, из которых исключены тренд и случайная компонента.

Пример 7. Имеются данные Федеральной службы государственной статистики об объеме платных услуг населению (млн руб.) по Владимирской области за 2005 – 2011 гг., представленные в табл. 18 и на рис. 21 (см. главу 5).

Для расчета индекса сезонности вычисляем среднемесячные значения, затем рассчитываем среднегодовое значение. По полученным значениям определяем индекс сезонности, результаты представим в табл. 25.

Таблица 25. Расчет индекса сезонности объема платных услуг населению

Месяц	**Среднемесячное значение, млн руб.**	**Индекс сезонности, %**
Январь	2140,96	98,72
Февраль	2164,01	99,79
Март	2228,86	102,78
Апрель	2164,67	99,82
Май	2015,46	92,94
Июнь	2115,99	97,57
Июль	2123,90	97,94
Август	1990,36	91,78
Сентябрь	1985,11	91,54
Октябрь	2191,41	101,05
Ноябрь	2350,01	108,36
Декабрь	2553,33	117,74
Общее среднее	**2168,67**	**100**

Индекс сезонности на диаграмме, которая представляет собой графическое изображение сезонной волны (рис. 26).

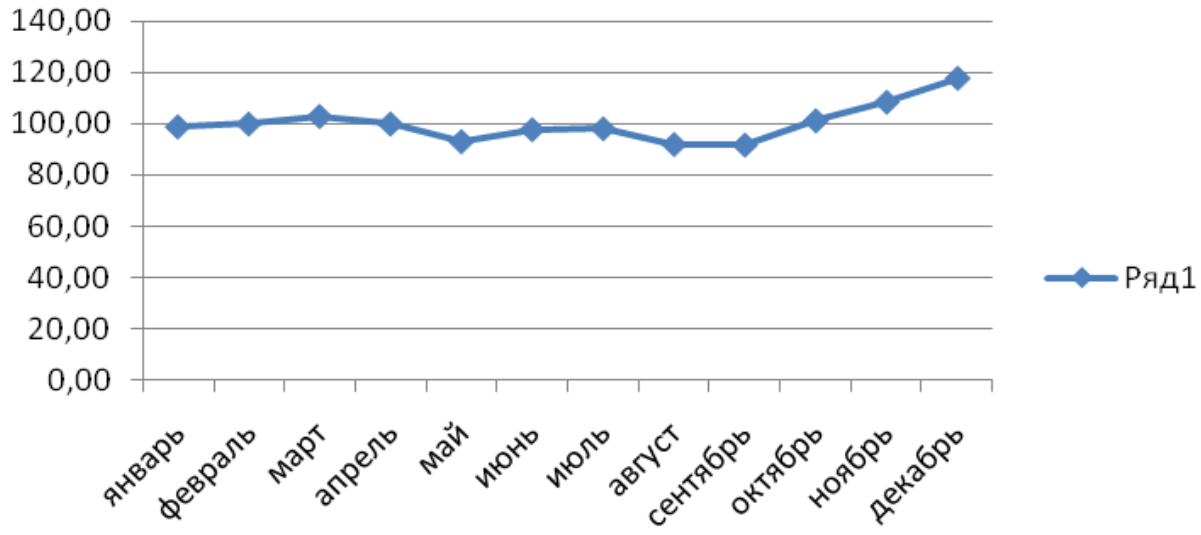

Рис. 26. Сезонная волна объема платных услуг населению (млн руб.)

Сезонную компоненту также можно изобразить на круговой диаграмме (рис. 27).

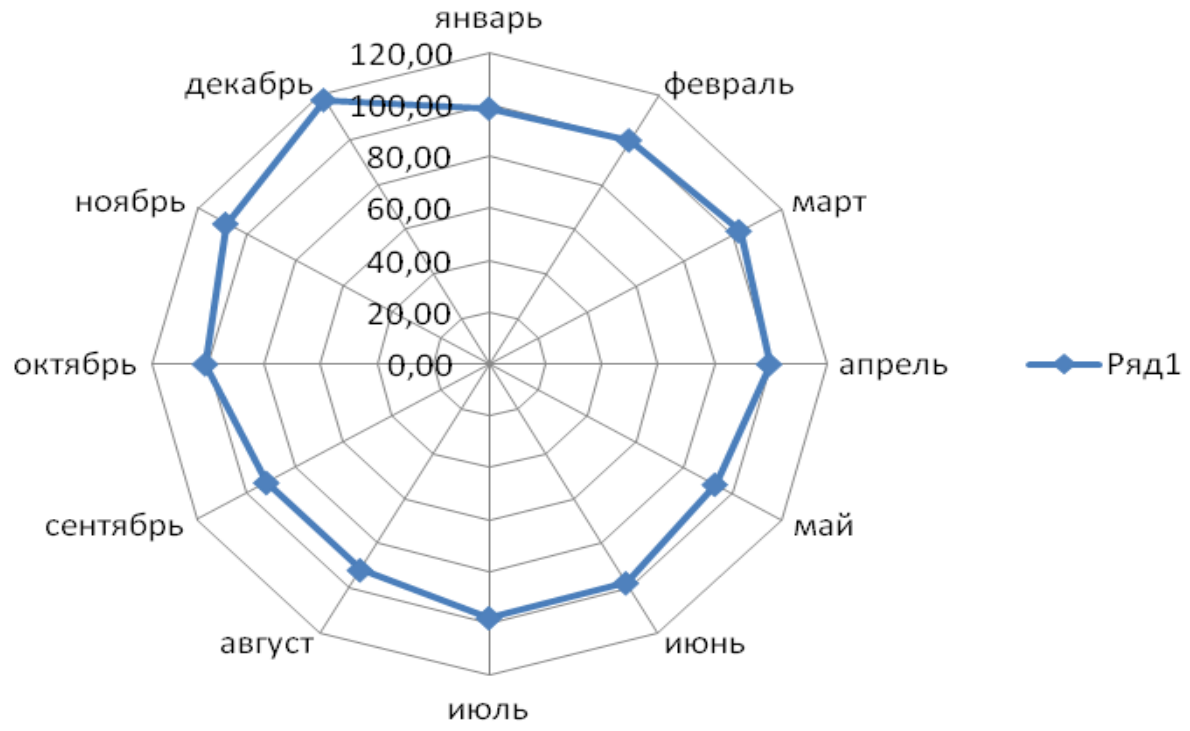

Рис. 27. Индекс сезонности объема платных услуг населению

Среднее значение индекса сезонности составляет 100 %.

Таким образом, наиболее высокие объемы платных услуг наблюдались в ноябре и декабре, а низкие объемы – в мае, августе и сентябре.

6.2. Методы выявления сезонных колебаний

Остановимся более подробно на методологических вопросах выделения сезонной составляющей. Проверку на наличие или отсутствие сезонных колебаний осуществляют на основе определенного критерия (например, дисперсионного, гармонического), а также на основе построения сезонной волны и расчетов индексов сезонности, на базе построения автокорреляционной функции процесса, в некоторых случаях визуально при построении графика.

Если ряд динамики имеет длительный период (15 – 25 лет), то сезонные колебания выявляют либо с учетом единой качественной

особенности периода (ряд сначала делят на качественно однородные периоды), либо согласно методике многократного скользящего выравнивания.

Рассмотрим методику применения дисперсионного критерия, согласно которому проверяют нулевую гипотезу о том, что во временном ряде, из которого отфильтрован тренд, отсутствуют сезонные колебания. В случае справедливости данной гипотезы *F*-статистика будет иметь *F*-распределение [22, с. 339]:

$$F = m\frac{n-T_0}{T_0}\frac{\sum_{j=1}^{T_0}(\overline{e}_j-\overline{e})^2}{\sum_{j=1}^{T_0}\sum_{i=1}^{m}(e_{ij}-\overline{e}_j)^2};\quad \overline{e}_j=\frac{1}{m}\sum_{i=1}^{m}e_{ij};\quad \overline{e}=\frac{1}{T_0}\sum_{j=1}^{T_0}\overline{e}_j=\frac{1}{n}\sum_{j=1}^{T_0}\sum_{i=1}^{m}e_{ij}.$$

Здесь использованы следующие обозначения: e_{ij} – значение остаточного ряда после выделения из него тренда, $i = 1, \ldots, m$; $j = 1, \ldots, T_0$; m – число лет наблюдений; T_0 – период колебаний ($T_0 = 4$ для ряда квартальных данных, $T_0 = 12$ для ряда месячных данных); n – количество наблюдений во временном ряде, $n = mT_0$.

Для выявления сезонных колебаний на основе автокорреляционной функции [22, с. 303]: рассчитывают коэффициенты автокорреляции и строят коррелограмму – график зависимости значений коэффициента автокорреляции от величины лага τ. Значения *коэффициентов автокорреляции* для рядов большой протяжённости можно вычислить по приближённой формуле

$$r(\tau)=\frac{\sum_{t=1}^{n-\tau}(y_t-\overline{y})(y_{t+\tau}-\overline{y})}{\sum_{t=1}^{n}(y_t-\overline{y})^2}.$$

Расчетные значения коэффициентов автокорреляции сравниваются с табличными значениями с заданным уровнем значимости, и в случае, когда *расчетное* значение больше *табличного*, соответствующий коэффициент признают значимым. Если самым высоким оказался статистически значимый коэффициент автокорреляции первого порядка, то ряд содержит только тенденцию. Если наиболее высок коэффициент автокорреляции II, III и следующих порядков, то ряд содержит сезонные или циклические колебания (в два, три и более периода времени). Автокорреляция первого порядка в ряду остатков может быть также проверена по критерию Дарбина – Уотсона [22, с. 319].

Алгоритм оценивания сезонной компоненты можно представить в виде последовательности следующих этапов расчёта.

I этап. Сглаживание исходного временного ряда с помощью процедуры скользящей средней при четной длине интервала сглаживания для предварительного оценивания тенденции развития.

II этап. Вычисление уровней временного ряда e_t, представляющих собой отношения/разности фактических уровней y_t и сглаженных значений y'_t, полученных на предыдущем шаге (соответственно для случая мультипликативной/аддитивной сезонности).

III этап. Усреднение значений уровней e_t для одноименных месяцев (кварталов) с целью элиминирования влияния случайной составляющей и определения предварительных значений сезонной компоненты.

IV этап. Проведение корректировки первоначальных значений сезонной составляющей с учетом мультипликативного или аддитивного характера сезонности для того, чтобы суммарное воздействие сезонности на динамику было нейтральным.

6.3. Построение тренд-сезонной аддитивной модели

На практике часто требуется выделить во временных рядах сезонную компоненту, случайную составляющую, трендовую компоненту, т. е. провести *декомпозицию* ряда, разложение его на составные части, или структурные составляющие. Кроме того, сезонные и случайные колебания зачастую затрудняют анализ тенденции развития показателя, и их влияние должно быть элиминировано.

Алгоритм расчета тренд-сезонной аддитивной модели состоит из следующих этапов [22, с. 340 – 342]:

1. Сглаживаем исходный временной ряд методом центрированной скользящей средней, используя весовые коэффициенты:

$\frac{1}{4}(\frac{1}{2}, 1, 1, 1, \frac{1}{2})$ – для квартальных данных; $\frac{1}{12}(\frac{1}{2}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \frac{1}{2})$ – для месячных данных, в общем случае по формуле

$$y_t^{'} = \frac{1}{T_0}\left[\frac{y_{t-T_0/2}}{2} + y_{t-\frac{T_0}{2}+1} + ... + y_{t-\frac{T_0}{2}-1} + \frac{y_{t+T_0/2}}{2}\right]. \qquad (6)$$

В результате получаем новый временной ряд, являющийся предварительным значением тренда.

2. Из исходного временного ряда вычитаем сглаженные значения:

$e_{ij} = y_{ij} - y'_{ij}$.

3. Усредняем полученные значения e_{ij} за все годы по каждому месяцу (кварталу): $\bar{e}_j = \frac{1}{m}\sum_{i=1}^{m} e_{ij}$, где $\bar{e}_j$ – сезонная компонента, «невыправленная» сезонная волна.

4. Корректируем средние значения $\bar{e}_j$, увеличивая или уменьшая их на одно и то же число так, чтобы их сумма была равна нулю – получим «выправленную» сезонную волну. Корректирующее число рассчитывают следующим образом: сумма оценок сезонных компонент делится на 12 (при месячных данных) или на 4 (при квартальных данных). На эту величину корректируется среднее значение $\bar{e}_j$ за каждый месяц (квартал) так, чтобы их сумма равнялась нулю. Получим S_j – сезонную волну, т. е. значения сезонной компоненты для каждого месяца (квартала).

5. Проведем десезонализацию исходных данных: вычитаем соответствующие значения сезонной компоненты из фактических значений ряда за каждый месяц (квартал), т. е. $u_{ij} = y_{ij} - S_j$.

Вычисленные таким образом значения ряда состоят из тренда и случайной компоненты.

6. Подбираем для полученного ряда кривую роста, аппроксимирующую тренд. Находим параметры уравнения кривой и подставляем в него последовательно значения условного времени t. Полученные оценки и будут значениями тренда.

7. Определяем значения случайной компоненты $E = y_{ij} - U - S_i$, которые можно использовать для определения точности и адекватности систематических компонент – тренда и сезонной волны.

8. Осуществляем прогнозирование тренд-сезонного экономического процесса путем сложения (для аддитивной модели) значений тренда, рассчитанных по уравнению тренда для каждого момента времени прогнозного периода, с соответствующим месячным (квартальным) значением сезонной компоненты $\hat{y}_{ij} = U + S_j$.

Рассмотрим реализацию алгоритма построения тренд-сезонной модели на примере временного ряда данных объема платных услуг населению за семь лет, представленных выше в табл. 18 и на рис. 21.

Графики ежемесячной динамики тренд-сезонного временного ряда показаны на рис. 28.

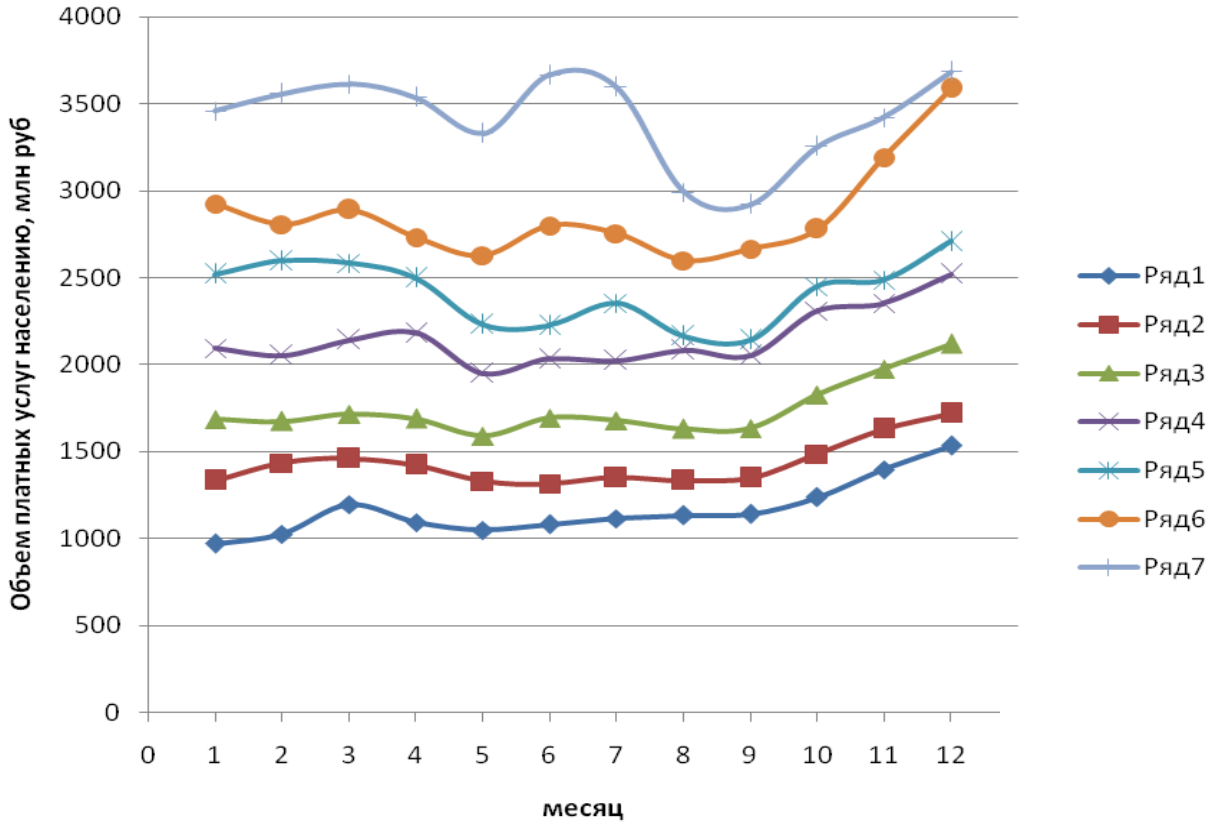

Рис. 28. Графики ежемесячной динамики объемов платных услуг

Визуальный анализ ежемесячной динамики объема платных услуг населению свидетельствует о наличии возрастающего тренда и сезонных колебаний. Поскольку амплитуда сезонных колебаний с течением временем изменяется незначительно, для описания динамики временного ряда выбрана модель с аддитивной сезонностью. Для проверки наличия во временном ряде сезонных колебаний применим F-критерий. Последовательность вычислений будет следующей:

1. Проведем сглаживание исходного временного ряда методом центрированной скользящей средней по формуле (6) при $T_0 = 12$, $m = 7$, $n = 84$. В результате получаем новый временной ряд, являющийся предварительным значением тренда (табл. 26).

Таблица 26. Сглаженный ряд

Месяц	Год						
	2005	2006	2007	2008	2009	2010	2011
Январь	–	1330,642	1589,904	1957,446	2346,329	2606,575	3263,038
Февраль	–	1348,854	1615,904	1990,654	2363,433	2641,379	3314,567
Март	–	1365,879	1640,121	2026,913	2370,479	2681,342	3341,617
Апрель	–	1384,971	1666,225	2064,346	2380,146	2716,908	3371,921
Май	–	1405,013	1694,917	2100,079	2391,646	2760,071	3401,108
Июнь	–	1422,288	1726,175	2132,442	2404,988	2826,154	3414,704
Июль	1179,471	1444,617	1760,013	2166,813	2429,642	2885,121	–
Август	1211,667	1469,275	1792,8	2207,2	2455,238	2938,804	–
Сентябрь	1239,692	1489,942	1826,271	2248,221	2476,817	3000,271	–
Октябрь	1264,379	1511,717	1864,454	2279,833	2499,338	3063,867	–
Ноябрь	1289,754	1533,679	1899,925	2304,775	2525,588	3126,613	–
Декабрь	1311,096	1560,408	1928,992	2324,517	2566,042	3191,863	–

2. Из исходного временного ряда вычтем сглаженные значения $e_{ij} = y_{ij} - y'_{ij}$.

Получим ряд, из которого удален тренд. Поскольку первые и последние шесть наблюдений (1-е полугодие 2005 года и 2-е полугодие 2011 года) исходного временного ряда не сглажены, то в дальнейшем будем рассматривать только те значения остатков, которые представлены целым числом лет (табл. 27).

Таблица 27. Значения остатков

Месяц	Остаток				
	2006	2007	2008	2009	2010
Январь	2,758333	95,69583	136,6542	173,4708	317,825
Февраль	85,74583	58,29583	61,94583	232,7667	164,5208
Март	94,52083	76,67917	114,7875	212,1208	209,4583
Апрель	38,22917	23,175	116,5542	118,5542	14,09167
Май	–75,7125	–104,717	–150,079	–160,846	–131,571
Июнь	–107,988	–31,275	–99,7417	–179,288	–27,2542
Июль	–94,2167	–82,3125	–144,013	–76,3417	–132,221
Август	–135,375	–162,2	–124,7	–292,738	–340,604
Сентябрь	–141,642	–193,471	–197,121	–336,617	–336,671
Октябрь	–26,1167	–36,8542	27,86667	–48,7375	–283,067
Ноябрь	94,62083	74,975	47,625	–40,0875	64,5875
Декабрь	157,1917	192,2083	195,8833	141,4583	395,9375

3. По данным табл. 27 рассчитаем средние значения для каждого месяца и для ряда в целом по формулам:

$$\bar{e}_j = \frac{1}{m}\sum_{i=1}^{m} e_{ij}\ ;\ \ \bar{e} = \frac{1}{T_0}\sum_{j=1}^{T_0}\bar{e}_j = \frac{1}{n}\sum_{j=1}^{T_0}\sum_{i=1}^{m} e_{ij}\ .$$

Получим ряд $\bar{e}_j$ и общую среднюю $\bar{e} = -9{,}22285$.

Для выявления сезонных колебаний применим F-критерий

$$F = m\frac{n-T_0}{T_0}\frac{\sum_{j=1}^{T_0}(\bar{e}_j - \bar{e})^2}{\sum_{j=1}^{T_0}\sum_{i=1}^{m}(e_{ij} - \bar{e}_j)^2} = 16{,}066.$$

Табличное значение F-критерия на уровне значимости 0,05 равно 1,95. Так как $F_{\text{расч}} > F_{\text{табл}}$, то с вероятностью 0,95 можно утверждать, что в данном временном ряде присутствуют сезонные колебания.

4. Проведем корректировку значений $\bar{e}_j$ так, чтобы их сумма стала равной нулю. Их среднее значение $\bar{e} = -9{,}22285$ вычтем из каждого уровня $\bar{e}_j$ и получим их скорректированные значения S_i – составляющие сезонной волны, т. е. значения сезонной компоненты для каждого месяца. Результаты расчета приведены в табл. 28.

Таблица 28. Расчет сезонной компоненты S_i

Месяц	e_t					$\bar{e}_i$	S_i
	2006	**2007**	**2008**	**2009**	**2010**		
Январь	2,758333	95,69583	136,6542	173,4708	317,825	145,2808	154,5037
Февраль	85,74583	58,29583	61,94583	232,7667	164,5208	120,655	129,8778
Март	94,52083	76,67917	114,7875	212,1208	209,4583	141,5133	150,7362
Апрель	38,22917	23,175	116,5542	118,5542	14,09167	62,12083	71,34368
Май	–75,7125	–104,717	–150,079	–160,846	–131,571	–124,585	–115,362
Июнь	–107,988	–31,275	–99,7417	–179,288	–27,2542	–89,1092	–79,8863
Июль	–94,2167	–82,3125	–144,013	–76,3417	–132,221	–105,821	–96,598
Август	–135,375	–162,2	–124,7	–292,738	–340,604	–211,123	–201,9
Сентябрь	–141,642	–193,471	–197,121	–336,617	–336,671	–241,104	–231,881
Октябрь	–26,1167	–36,8542	27,86667	–48,7375	–283,067	–73,3817	–64,1588
Ноябрь	94,62083	74,975	47,625	–40,0875	64,5875	48,34417	57,56701
Декабрь	157,1917	192,2083	195,8833	141,4583	395,9375	216,5358	225,7587
						–110,674	0
Среднее значение						**–9,22285**	

5. Проведем десезонализацию исходных данных путем вычитания из фактических наблюдений соответствующих значений сезонной компоненты (табл. 29):

$u_{ij} = y_{ij} - S_i$.

Таблица 29. Десезонализированные исходные данные.

Месяц	**Год**						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Январь	815,5963	1178,896	1531,096	1939,596	2365,296	2769,896	3304,796
Февраль	895,3222	1304,722	1544,322	1922,722	2466,322	2676,022	3429,522
Март	1046,464	1309,664	1566,064	1990,964	2431,864	2740,064	3461,764
Апрель	1022,556	1351,856	1618,056	2109,556	2427,356	2659,656	3464,256
Май	1164,962	1444,662	1705,562	2065,362	2346,162	2743,862	3445,162
Июнь	1161,686	1394,186	1774,786	2112,586	2305,586	2878,786	3743,486
Июль	1210,398	1446,998	1774,298	2119,398	2449,898	2849,498	3692,998
Август	1335,3	1535,8	1832,5	2284,4	2364,4	2800,1	3193,3
Сентябрь	1372,081	1580,181	1864,681	2282,981	2372,081	2895,481	3151,481
Октябрь	1299,659	1549,759	1891,759	2371,859	2514,759	2844,959	3316,259
Ноябрь	1339,833	1570,733	1917,333	2294,833	2427,933	3133,633	3362,833
Декабрь	1308,141	1491,841	1895,441	2294,641	2481,741	3362,041	3459,141

6. На основе преобразованного временного ряда рассчитаем уравнение линейной трендовой модели (см. рис. 29, 30).

По уравнению **линейного тренда**: $U = 30{,}60t + 867{,}8$ определяем значения случайной компоненты по формуле: $E = Y_{ij} - U - S_{ij}$.

7. Строим уравнение тренд-сезонной аддитивной модели прогнозирования

$\tilde{y}(84 + \tau) = (30{,}60(84 + \tau) + 867{,}8) + S(\tau)$, $\tau = 1, 2, \ldots, 12$.

Вычисляем прогнозные значения исследуемого показателя на очередные 12 месяцев, а также среднюю относительную ошибку прогноза $E_{отн} = 6{,}977$. Результаты прогноза покажем на рис. 30, где представлены исходные данные y_t (что соответствует условному времени $t = 1, 2, \ldots, 84$) и прогнозные значения показателя $\breve{y}_t$ ($t = 1, 2, \ldots, 84$) на 12 месяцев будущего года.

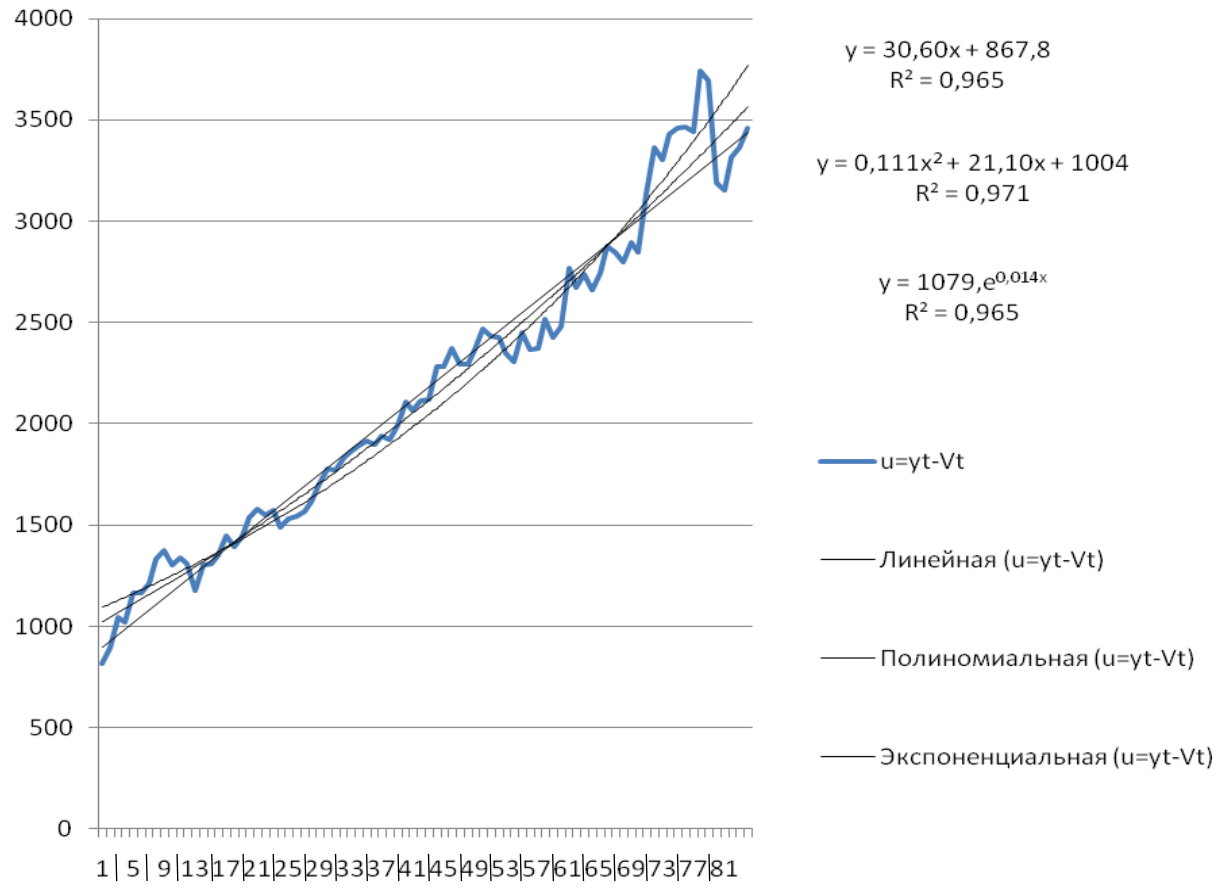

Рис. 29. Графики трендовых моделей и десезонализированных исходных данных

Рис. 30. Результаты прогнозирования с участием линейного тренда

8. На основе табл. 28, 29 строим модель-параболу (рис. 29). По уравнению **параболического тренда** $U = 0{,}111t^2 + 21{,}10t + 1004$ определяем значения случайной компоненты.

9. Строим уравнение тренд-сезонной аддитивной модели прогнозирования

$\tilde{y}(84 + \tau) = (0{,}111(\tau + 84)^2 + 21{,}10(\tau + 84) + 1004) + S(\tau)$,

$\tau = 1, 2, \ldots, 12$.

Вычисляем прогнозные значения исследуемого показателя на очередные 12 месяцев будущего года, а также ошибку прогноза $E_{\text{отн}} = 3{,}359$.

10. На основе табл. 29, 30 строим модель-экспоненту (см. рис. 29). По уравнению **экспоненциального тренда** $U = 1079e^{0{,}014t}$ определяем значения случайной компоненты.

11. Строим уравнение тренд-сезонной аддитивной модели прогнозирования

$\tilde{y}(84 + \tau) = (1079e^{0{,}014(\tau+84)}) + S(\tau)$, $\tau = 1, 2, \ldots, 12$.

Точечный прогноз значений исследуемого показателя на очередные 12 месяцев, а также расчёт ошибки прогноза $E_{\text{отн}} = 2{,}559$ представлены в табл. 30.

Таблица 30. Расчёт прогноза с участием экспоненциального тренда

t, условное время	**S_t сезонная составляющая**	**$\check{y}_t$ прогноз**	**y_t факт**	**Ошибка прогноза, %**
85	154,5037	3701,264	3597,2	0,028929251
86	129,8778	3726,642	3748,5	0,005831045
87	150,7362	3798,209	3789,1	0,002404133
88	71,34368	3770,241	3736,6	0,009003038
89	–115,362	3635,684	3779,4	0,03802623
90	–79,8863	3724,043	3845,7	0,031634432
91	–96,598	3760,961	3766,5	0,001470504
92	–201,9	3710,044	3854,7	0,037527048
93	–231,881	3735,216	3905,1	0,043503098
94	–64,1588	3958,869	4105,9	0,035809808
95	57,56701	4137,313	4226,3	0,021055574
96	225,7587	4363,023	4601,8	0,051887821
				0,307081981
			$E_{\text{отн}}$ =	**2,55901651**

По модели с экспоненциальным трендом получена наименьшая ошибка прогнозирования, равная 2,559 %. Также из графиков (рис. 31) видно, что модель с экспоненциальным трендом ближе остальных при-

мыкает к линии фактических значений. Поэтому для построения прогноза на 2013 год за основу взята модель с экспоненциальным трендом (табл. 31).

Таблица 31. Расчет прогноза

τ	S_τ	Прогноз по годам	
		2012	2013
1	154,5037	3701,264	4350,097
2	129,8778	3726,642	4384,622
3	150,7362	3798,209	4465,466
4	71,34368	3770,241	4446,904
5	–115,362	3635,684	4321,887
6	–79,8863	3724,043	4419,921
7	–96,598	3760,961	4466,65
8	–201,9	3710,044	4425,682
9	–231,881	3735,216	4460,943
10	–64,1588	3958,869	4694,827
11	57,56701	4137,313	4883,648
12	225,7587	4363,023	5119,88

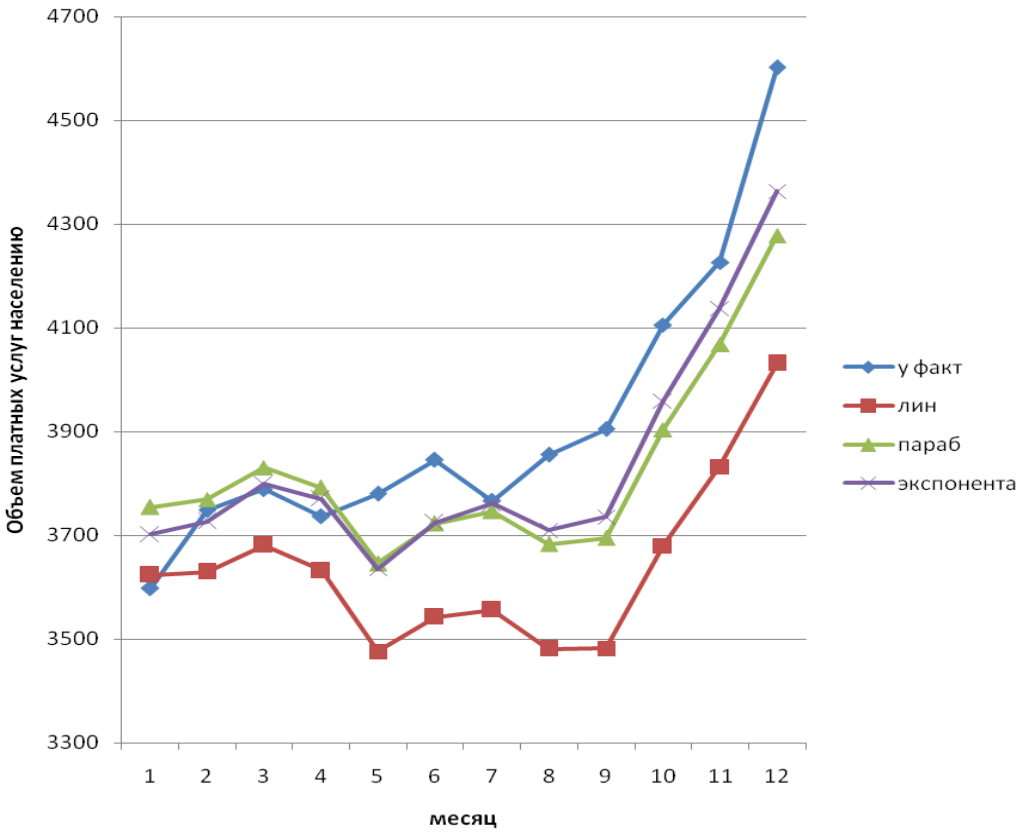

Рис. 31. Динамика объема платных услуг населению на 2012 год, млн руб. (фактические данные; прогноз с линейным трендом, с параболическим трендом, с экспоненциальным трендом)

Проанализируем ряд остатков по десезонализированным данным и трендовой модели экспоненты. Анализ остатков на нормальный закон распределения по трем известным критериям показал, что остатки распределены нормально на уровне значимости 0,05. Приведем распечатку процедуры программы STADIA.

Интервальный вариационный ряд:

X-лев.	X-станд	Частота	%	Накопл.	%
–278	–2,557	3	3,571	3	3,571
–177,3	–1,871	6	7,143	9	10,71
–76,55	–1,186	21	25	30	35,71
24,17	–0,5005	19	22,62	49	58,33
124,9	0,185	21	25	70	83,33
225,6	0,8704	8	9,524	78	92,86
326,3	1,556	4	4,762	82	97,62
427,1	2,241	2	2,381	84	100
527,8	2,927				

Колмогоров = 0,08487, Значимость = 0,168, степ. своб. = 84
Гипотеза 0: <Распределение не отличается от нормального>
Омега-квадрат = 0,09171, Значимость = 0,1438, степ. своб = 84
Гипотеза 0: <Распределение не отличается от нормального>
Хи-квадрат = 4,353, Значимость = 0,4998, степ. своб = 5
Гипотеза 0: <Распределение не отличается от нормального>

Проверим случайность возникновения отдельных отклонений от тренда при уровне вероятности 0,95 по формуле [22, с. 321]:

$$p > \left[\frac{2}{3}(n-2) - 1{,}96\sqrt{\frac{16n-29}{90}}\right] \qquad (7)$$

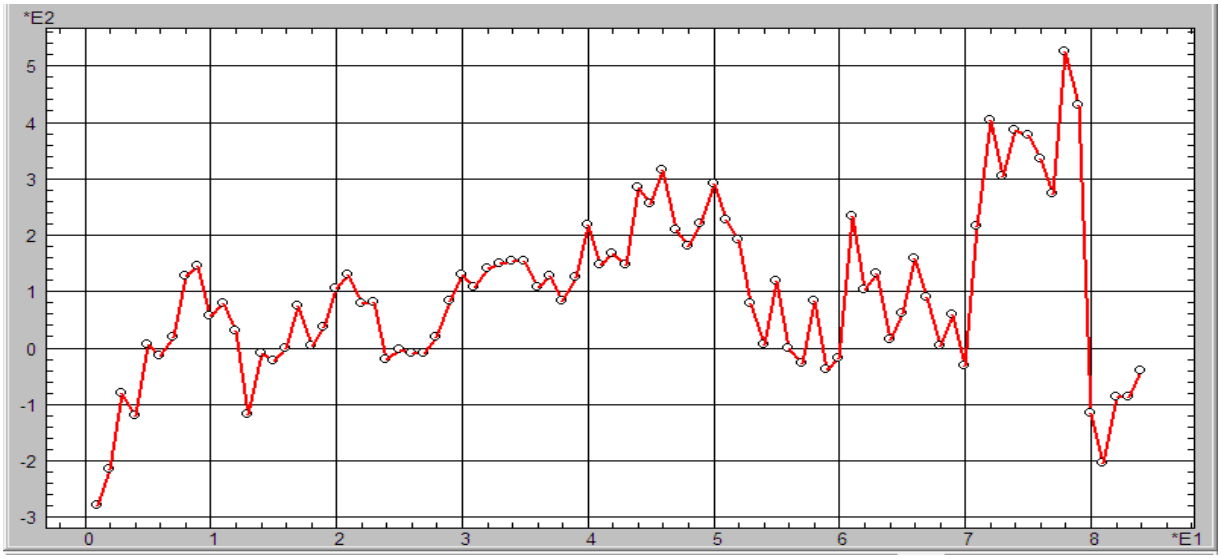

Рис. 32. График остатков по трендовой модели-экспоненте

Анализируя график остатков (см. рис. 32), вычисляем число поворотных точек, в данном случае p = 54; количество наблюдений n = 84. Таким образом, правая часть неравенства (7) равна 47, поэтому по критерию поворотных точек ряд обладает свойством случайности. Следовательно, анализ ряда остатков позволяет строить доверительный интервал прогноза по выбранной трендовой модели. Учитывая сезонный характер исходного ряда в аддитивной форме, мы строим точечный и параллельно интервальный прогнозы согласно формуле [22, с. 324], где учитывается стандартная ошибка регрессии (несмещенная оценка стандартного отклонения остатков), которая составила 177,92. Результаты моделирования представлены на рис. 33.

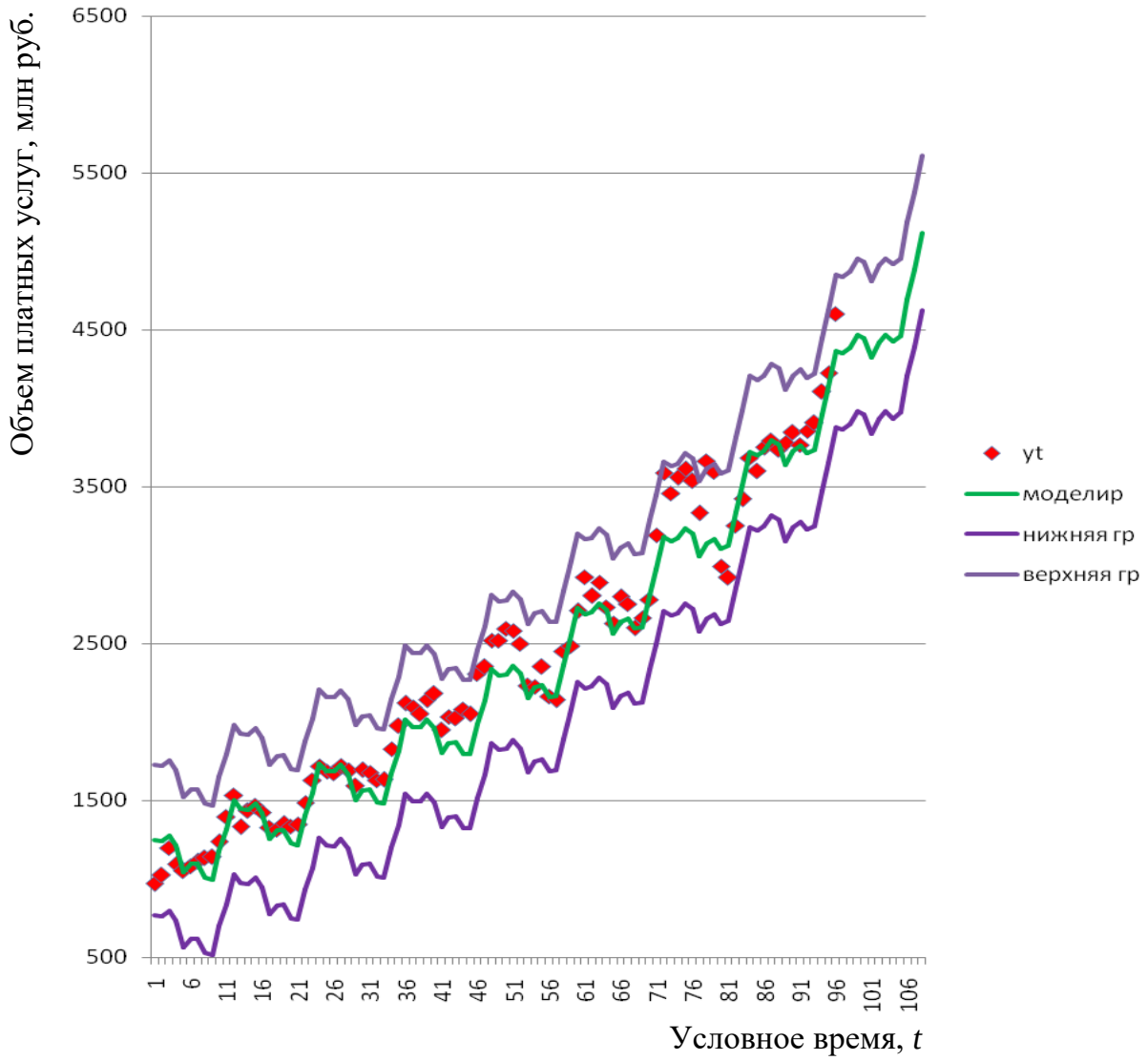

Рис. 33. График исходных данных; зона доверительного интервала; график моделируемых данных с прогнозом

Анализ остатков на автокорреляцию показал, что коэффициенты автокорреляции, имеющие порядок, кратный периоду $T = 12$, незначительны, что означает, что сезонная составляющая достаточно хорошо выявлена и присутствует в модели (т. е. ее в остатках не оказалось).

Из уравнения модели следует, что трендовая составляющая показывает тенденцию возрастания показателя за год примерно в 1,2 раза. Сезонная компонента показывает, что самые высокие объемы платных услуг наблюдались в декабре. В свою очередь, сезонность отрицательно влияет на показатель в августе, сентябре (см. табл. 32).

6.4. Построение тренд-сезонной мультипликативной модели

Следуя [22, с. 348], укажем основные этапы построения тренд-сезонной мультипликативной модели.

1. Сглаживаем исходный временной ряд методом центрированной скользящей средней, используя весовые коэффициенты, либо по формуле (6). В результате получаем временной ряд, являющийся предварительным значением тренда.

2. Делим значения исходного временного ряда на соответствующие сглаженные значения: $e_{ij} = \dfrac{y_{ij}}{\tilde{y}_{ij}}$.

3. Усредняем полученные значения e_{ij} за все годы по каждому месяцу (кварталу): $\bar{e}_j = \dfrac{1}{m}\sum_{i=1}^{m} e_{ij}$, где $\bar{e}_j$ – сезонная компонента, «невыправленная» сезонная волна.

Корректируем средние значения $\bar{e}_j$ сезонных компонент за каждый месяц так, чтобы их сумма, деленная на 12, равнялась 1.

4. Проведем десезонализацию исходных данных. Исходные уровни временного ряда делим на скорректированные значения сезонной волны: $u_{ij} = \dfrac{y_{ij}}{S_j}$. Вычисленные таким образом значения ряда состоят из тренда и случайной компоненты.

5. Подбираем для полученного ряда кривую роста, аппроксимирующую тренд. Находим параметры уравнения кривой и подставляем в него последовательно значения условного времени t. Полученные оценки и будут значениями тренда.

6. Определяем значения случайной компоненты $\varepsilon_{ij} = y_{ij} - u_{ij}S_j$.

7. Осуществляем прогнозирование тренд-сезонного экономического процесса путем умножения (для мультипликативной модели) значений тренда, рассчитанных по уравнению тренда для каждого месяца прогнозного периода, на соответствующие месячные значения сезонной компоненты: $\hat{y}_{ij} = y'_{ij} S_j$. Следуя данному алгоритму, приведём пример расчета тренд-сезонной мультипликативной модели.

Пример 8. Имеются следующие данные Федеральной службы государственной статистики о денежных доходах на душу населения в месяц (руб.) по Владимирской области за 2005 – 2011 гг. (источник: http//vladimirstat.gks.ru) (табл. 32).

Таблица 32. Среднедушевые денежные доходы

Месяц	**Год**						
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Январь	2717,8	4258	4287,7	5833,4	7473,4	7910,7	9842,7
Февраль	3743,7	4848,6	5980,2	7980,7	10315,9	11630,6	12648,9
Март	4239,4	5365,3	6569,3	8208,4	10684,1	11646,3	12696,8
Апрель	4120,4	5395,8	6411,5	9072,9	11281,2	12572,1	13697,7
Май	3992,6	5354	6540,8	9150,4	10341,1	11390,5	12704,4
Июнь	4043,9	6063	7639,4	9944,5	10874	12662	13186,5
Июль	3948,2	5389,3	6665,7	10011,2	10906	12972,1	14172,2
Август	4074,9	5426,7	7208,3	10710,2	10218,9	12109,2	12839,7
Сентябрь	4249,9	5299,9	6962	10016,6	10497,8	12285,6	13296,5
Октябрь	4050,3	5468,9	7138,2	10576,6	10536,4	12211,5	13506,9
Ноябрь	4110,7	5953,8	7855,5	10006,8	10929	13011	13794,9
Декабрь	6166,8	9092,3	11594,6	13256,1	16833,2	19199,1	21451,4

Визуальный анализ рис. 34 свидетельствует о наличии возрастающего тренда и присутствии сезонных колебаний. Так как амплитуда сезонных колебаний изменяется с течением времени, увеличивается с ростом тренда, то для описания динамики временного ряда выбираем модель с мультипликативной сезонностью. Для проверки наличия во временном ряде сезонных колебаний применяем *F*-критерий.

Последовательность вычислений будет следующая:

1. Проведем сглаживание исходного временного ряда методом центрированной скользящей средней по формуле (6) при $T_0 = 12$, $m = 7$, $n = 84$. В результате получаем временной ряд, являющийся предварительным значением тренда (табл. 33).

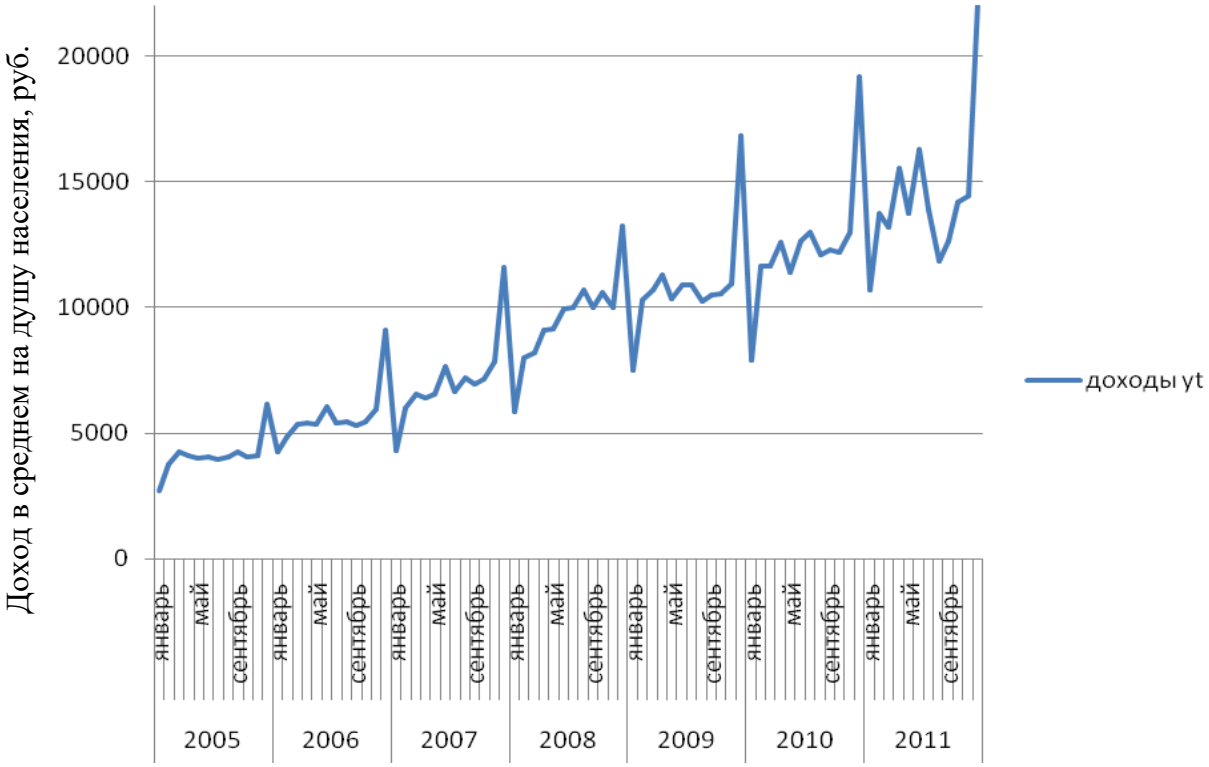

Рис. 34. Ежемесячная динамика денежных доходов

Таблица 33. Сглаженные значения

Месяц	Год						
	2005	2006	2007	2008	2009	2010	2011
Январь	–	4883,838	6224,833	8273,946	10499,55	11563,88	13784,69
Февраль	–	5000,208	6352,25	8559,254	10516,36	11728,73	13811,9
Март	–	5100,283	6495,738	8832,442	10515,94	11881,98	13814,9
Апрель	–	5203,142	6634,546	9102,983	10534,32	12026,27	13911,75
Май	–	5339,046	6783,338	9335,888	10571,07	12182,82	14054,3
Июнь	–	5537,738	6966,838	9494,754	10758,54	12368,15	14236,62
Июль	4185,725	5660,871	7135,504	9632,317	10925,8	12581,75	–
Август	4295,938	5709,258	7283,263	9797,95	10998,8	12785,35	–
Сентябрь	4388,888	5806,575	7434,913	9998,404	11093,68	12938,54	–
Октябрь	4488,942	5899,063	7614,1	10193,57	11187,55	13126,77	–
Ноябрь	4598,808	5990,833	7833,725	10335,2	11285,07	13348,66	–
Декабрь	4739,663	6105,967	8038,504	10423,54	11403,29	13596,84	–

2. Делим значения исходного временного ряда на соответствующие сглаженные значения, получаем значения остатков (табл. 34) $e_{ij} = \frac{y_{ij}}{\tilde{y}_{ij}}$.

Таблица 34. Значения остатков

Месяц	e_t по годам				
	2006	2007	2008	2009	2010
Январь	0,871855	0,688806	0,705032	0,711783	0,684087
Февраль	0,96968	0,94143	0,932406	0,980938	0,991633
Март	1,051961	1,011325	0,929347	1,015991	0,980165
Апрель	1,037027	0,966381	0,996695	1,0709	1,045386
Май	1,002801	0,964245	0,980132	0,978246	0,934964
Июнь	1,094851	1,096538	1,047368	1,010732	1,023759
Июль	0,952027	0,93416	1,039335	0,998187	1,031025
Август	0,950509	0,989708	1,093106	0,929092	0,947116
Сентябрь	0,912741	0,936393	1,00182	0,946287	0,949536
Октябрь	0,92708	0,937498	1,037576	0,941797	0,930275
Ноябрь	0,993818	1,00278	0,968225	0,968448	0,974705
Декабрь	1,489084	1,442383	1,271747	1,47617	1,412027

3. Для ряда остатков рассчитаем средние значения для каждого месяца и для ряда в целом по формулам

$$\bar{e}_i = \frac{1}{5}\sum_{j=1}^{5} e_{ij};\ \ \bar{e} = \frac{1}{T_0}\sum_{i=1}^{T_0} \bar{e}_i \ (i = 1, 2, ..., 12)$$

Получим ряд $\bar{e}_i$ (табл. 36).

Общая средняя $\bar{e} = 1{,}001685$.

Для выявления сезонных колебаний вычислим F-критерий

$$F = m\frac{n - T_0}{T_0}\frac{\sum_{j=1}^{T_0}(\bar{e}_j - \bar{e})^2}{\sum_{j=1}^{T_0}\sum_{i=1}^{m}(e_{ij} - \bar{e}_j)^2} = 5\frac{60-12}{12}\frac{11{,}90018}{1{,}459392} = 163{,}08.$$

Табличное значение F-критерия на уровне значимости 0,05 равно 2,705. Так как $F_{\text{расч}} > F_{\text{критич}}$, то с вероятностью 0,95 можно утверждать, что в данном временном ряде присутствуют сезонные колебания.

4. Проведем корректировку $\bar{e}_j$ так, чтобы их сумма стала равной единице по формуле

$$S_i = \bar{e}_i k (i = 1, 2, \ldots, m),\ k = \frac{m}{\sum_{i=1}^{m} \bar{e}_i},$$

где m – число фаз в полном сезонном цикле ($m = 12$ для месячной динамики); $k = 0{,}998318$.

Результаты расчета приведены в табл. 35.

Таблица 35. Расчет сезонной компоненты S_i

Месяц	**e_t по годам**					**$\bar{e}_i$**	**S_i**
	2006	**2007**	**2008**	**2009**	**2010**		
Январь	0,871855	0,688806	0,705032	0,711783	0,684087	0,732313	0,731080606
Февраль	0,96968	0,94143	0,932406	0,980938	0,991633	0,963217	0,961596918
Март	1,051961	1,011325	0,929347	1,015991	0,980165	0,997758	0,996078949
Апрель	1,037027	0,966381	0,996695	1,0709	1,045386	1,023278	1,021556436
Май	1,002801	0,964245	0,980132	0,978246	0,934964	0,972078	0,970442127
Июнь	1,094851	1,096538	1,047368	1,010732	1,023759	1,05465	1,052875258
Июль	0,952027	0,93416	1,039335	0,998187	1,031025	0,990947	0,989279507
Август	0,950509	0,989708	1,093106	0,929092	0,947116	0,981906	0,980254027
Сентябрь	0,912741	0,936393	1,00182	0,946287	0,949536	0,949355	0,947758086
Октябрь	0,92708	0,937498	1,037576	0,941797	0,930275	0,954845	0,953238285
Ноябрь	0,993818	1,00278	0,968225	0,968448	0,974705	0,981595	0,979943753
Декабрь	1,489084	1,442383	1,271747	1,47617	1,412027	1,418282	1,415896049
Сумма						**12,02022**	**12**
k						**0,998318**	

5. Проведем десезонализацию исходных данных (табл. 36): исходные уровни временного ряда делим на соответствующие скорректированные значения сезонной волны, т. е. $u_{ij} = \frac{y_{ij}}{V_i}$.

Таблица 36. Десезонализированные исходные данные

	u_{ij}						
Месяц	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Январь	3717,511	5824,255	5864,88	7979,148	10222,4	10820,56	14596,58
Февраль	3893,211	5042,237	6219,03	8299,423	10727,88	12095,09	14305,72
Март	4256,088	5386,42	6595,16	8240,712	10726,16	11692,15	13249,1
Апрель	4033,453	5281,94	6276,207	8881,448	11043,15	12306,81	15210,97
Май	4114,207	5517,073	6740,021	9429,104	10656,07	11737,43	14167,85
Июнь	3840,816	5758,517	7255,75	9445,089	10327,91	12026,12	15443,16
Июль	3990,985	5447,702	6737,934	10119,69	11024,18	13112,67	14033,26
Август	4156,984	5536,014	7353,502	10925,94	10424,75	12353,12	12090,26
Сентябрь	4484,161	5592,039	7345,756	10568,73	11076,46	12962,8	13310,71
Октябрь	4248,99	5737,18	7488,369	11095,44	11053,27	12810,54	14903,05
Ноябрь	4194,833	6075,655	8016,276	10211,61	11152,68	13277,29	14733,02
Декабрь	4355,404	6421,587	8188,878	9362,34	11888,73	13559,68	15642,53

6. На основе десезонализированных исходных данных рассчитаем уравнения следующих трендовых моделей: параболической и линейной. Параметры моделей получены в MS Excel:

$U = 0{,}039t^2 + 138{,}9t + 3121$; $U = 142{,}3t + 3073$.

7. Определяем значения случайной компоненты (с участием каждого тренда U) по формуле $E = Y_{ij} - US_i$.

8. Строим уравнения тренд-сезонных мультипликативных моделей прогнозирования с участием линейного и параболического трендов:

$\tilde{y}(84 + \tau) = (0{,}039(84 + \tau)^2 + 138{,}9(84 + \tau) + 3121)\ S(\tau)$,

$\tilde{y}(84 + \tau) = (142{,}3(84 + \tau) + 3073)\ S(\tau)$, где $\tau = 1, 2, \ldots, 12$.

9. По каждой модели рассчитываем точечный прогноз исследуемого показателя на следующие 12 месяцев 2012 г. и сравниваем с фактическими данными. Точность моделей оцениваем по средней относительной ошибке прогнозирования, которая по модели-параболе составила $E_{отн} = 5{,}440$ %, по линейной модели несколько больше: $E_{отн} = 5{,}744$ %. Таким образом, более точно исходные данные представлены моделью с участием параболического тренда, которая была принята за основу в целях прогнозирования на 2013 год (табл. 37). Модель с параболической траекторией тренда показывает рост показателя с постоянным ускорением, которое корректируется с каждым месяцем сезонными коэффициентами, а также демонстрирует значительное повышение доходов в декабре.

Таблица 37. Расчет прогноза

τ	**Сезонная компонента $S(\tau)$**	**Прогноз на 2012 год**		**Прогноз на 2013 год с участием параболического тренда**
		с участием линейного тренда	**с участием параболического тренда**	
1	0,731081	11089,4	11119,206	12400,04166
2	0,961597	14722,82	14765,171	16450,76545
3	0,996079	15392,51	15439,713	17186,68386
4	1,021556	15931,58	15983,493	17776,10355
5	0,970442	15272,53	15325,241	17029,06515
6	1,052875	16719,66	16780,621	18630,16021
7	0,98928	15850,53	15911,432	17650,18184
8	0,980254	15845,41	15909,421	17633,22507
9	0,947758	15455	15520,497	17188,04307
10	0,953238	15680,01	15749,597	17427,67808
11	0,979944	16258,74	16334,168	18060,17815
12	1,415896	23693,32	23808,043	26303,23682

6.5. Прогнозирование динамики туристских потоков на основе адаптивных моделей Хольта – Уинтерса

В настоящее время прогнозы на основе адаптивных моделей разрабатываются также для исследования динамики сезонных экономических процессов. В случае линейного тренда для прогнозирования тренд-сезонных процессов используют аддитивную и мультипликативную модели Хольта – Уинтерса.

Туризм – один из наиболее динамично развивающихся секторов экономики, переживающий интенсивные изменения именно в последние годы. Это означает, что информативность показателей туризма, в том числе и фактор сезонности, по мере их удаления от периода прогнозирования склонны снижаться. Поэтому авторы пособия сделали акцент на выбор адаптивных сезонных моделей (использующих корректирование с течением времени сезонной компоненты) в целях прогнозирования объемов туристских потоков, представленных одномерными временными рядами. На основании статистической информации Федерального агентства по туризму сформированы следующие временные ряды: $X(t)$ – ряд поквартальных данных по выезду российских граждан за рубеж с целью туризма и $Y(t)$ – ряд поквартальных данных по въезду иностранных туристов в Россию за период с 2005 по 2013 г. Здесь t – условный показатель времени, принимающий значения 1, 2, …, 36.

Графическая иллюстрация динамики временных рядов $X(t)$ и $Y(t)$ представлена на рис. 35 и 36. Здесь отчетливо видны внутригодовые сезонные колебания, характерные для поквартальной динамики объемов туристских потоков. Пики активности приходятся в обоих случаях на III квартал, а самые низкие уровни – на I квартал ежегодно. При этом для динамики $X(t)$ в рассматриваемом периоде заметна возрастающая тенденция. Причем амплитуда сезонных колебаний изменяется с течением времени, что приводит к выводу о мультипликативном характере сезонности. Кроме того, анализ значений автокорреляционной функции (АКФ) для обоих рядов $X(t)$ и $Y(t)$ (рис. 37, 38) показывает наличие сезонных колебаний периодичностью в четыре квартала. Значения АКФ характеризуют тесноту статистической связи между уровнями рядов, разделенными l временных тактов.

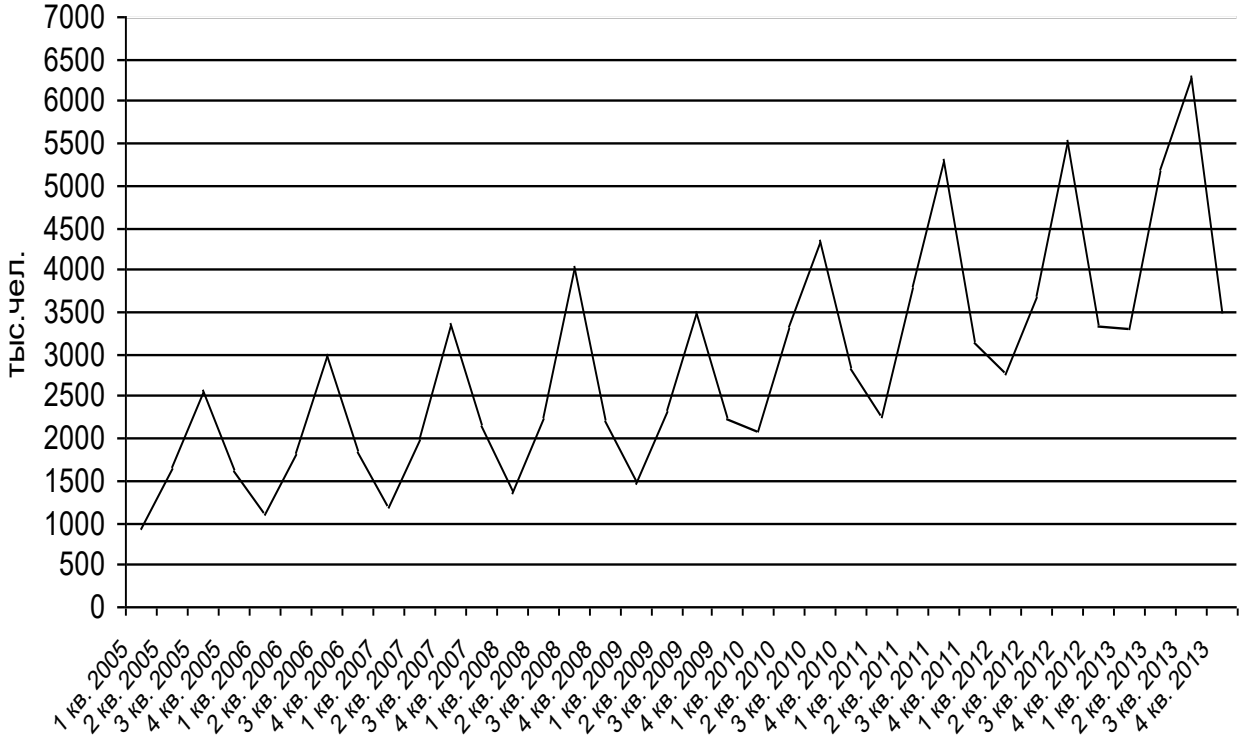

Рис. 35. Выезд российских граждан за рубеж с целью туризма

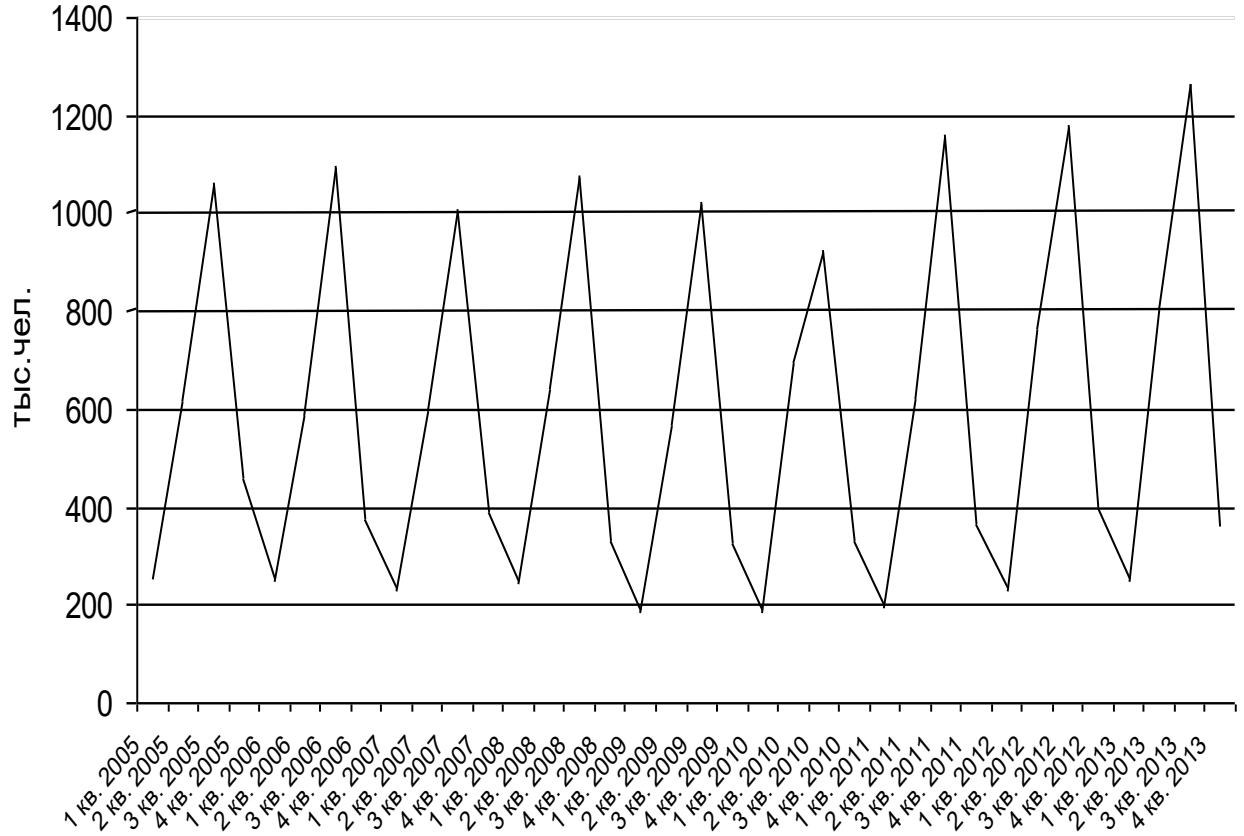

Рис. 36. Въезд иностранных туристов в Россию

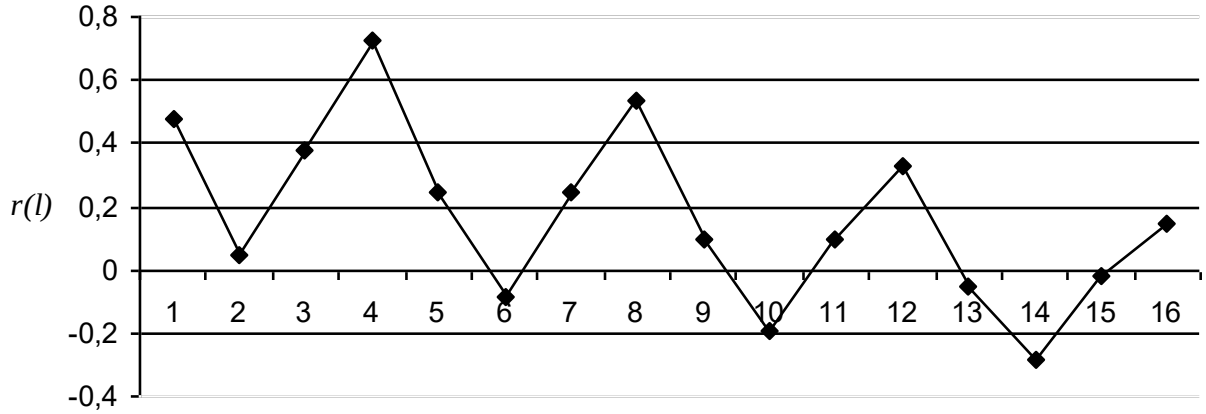

Рис. 37. Автокорреляционная функция ряда *X(t)*

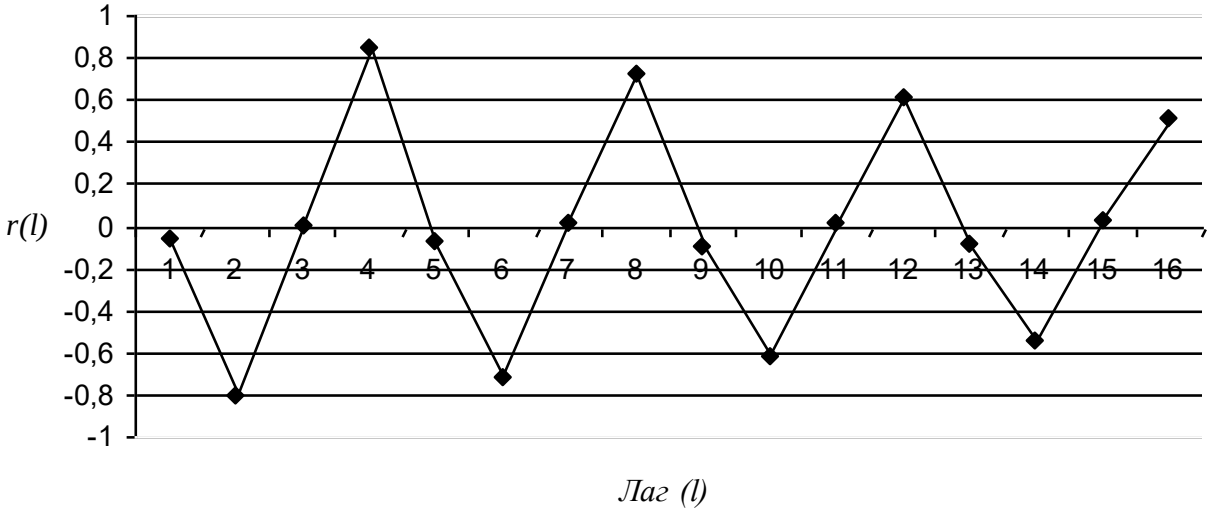

Рис. 38. Автокорреляционная функция ряда *Y(t)*

Таким образом, визуальный анализ исходных данных позволяет представить структуру временного ряда в виде $X(t) = U(t) \cdot S(t) + e(t)$, где $U(t)$ – трендовая составляющая; $S(t)$ – сезонная компонента; $e(t)$ – случайная компонента. Аналогичную структуру предполагаем и для ряда $Y(t)$.

Дальнейшая статистическая обработка данных показала, что динамику объемов туристских потоков можно описать с помощью моделей, сочетающих в себе линейный рост с мультипликативным сезонным эффектом. Ориентируясь на решение задачи прогнозирования, чтобы в большей степени учитывать свежую информацию, в качестве расчетной модели нами была выбрана адаптивная модель Хольта –

Уинтерса. Данная модель представляет собой объединение двухпараметрической модели линейного роста Хольта и сезонной модели Уинтерса. Прогноз по модели Хольта – Уинтерса на τ шагов вперед определяется выражением [1, с. 99 – 100]

$$Y_t(\tau) = (a_t + b_t\tau)F_{t+\tau-L}. \quad (8)$$

Здесь коэффициенты тренда a_t и b_t и оценка сезонности $F_{t+\tau-L}$ рассчитываются последовательно по рекуррентным формулам

$$\begin{gathered} a_t = \alpha_1 Y_t/F_{t-L} + (1 - \alpha_1)(a_{t-1} + b_{t-1}), \\ b_t = \alpha_3(a_t - a_{t-1}) + (1 - \alpha_3)b_{t-1}, \\ F_t = \alpha_2 Y_t/a_t + (1 - \alpha_2)F_{t-L}, \end{gathered} \quad (9)$$

где F_{t-L} – массив постоянно обновляемых сезонных коэффициентов размерностью L, в нашем случае $L = 4$ для квартальных данных. Значение коэффициента сезонности, которое приписывают, например, моменту времени $(t + 1)$, вычисляют сезон назад – в момент времени $(t + 1 - 4)$. В данном методе используют три параметра экспоненциального сглаживания α_1, α_2, α_3 такие, что $0 < \alpha_1, \alpha_2, \alpha_3 < 1$. Прогноз по модели (8) – это функция прошлых и текущих данных, параметров адаптации α_1, α_2, α_3, а также начальных значений коэффициентов a_0 и b_0 и сезонного фактора. Влияние начальных условий на прогнозную оценку зависит от величины весов и длины ряда. Алгоритм построения адаптивных моделей для исследуемых временных рядов можно представить в виде следующей последовательности шагов.

1. По исходным данным оцениваем с помощью метода наименьших квадратов значения a_0 и b_0 начальной линейной модели для ряда $X(t)$: $\tilde{X}(t) = 1410{,}6 + 71{,}333t$; для ряда $Y(t)$: $\tilde{Y}(t) = 499{,}75 + 19{,}345t$.

2. Рассчитываем начальный массив сезонных коэффициентов на основе отношений фактических уровней ряда к соответствующим значениям по линейной модели:

$F_{-3} = (X(1)/\tilde{X}(1) + X(5)/\tilde{X}(5))/2 = 0{,}62521;$

$F_{-2} = (X(2)/\tilde{X}(2) + X(6)/\tilde{X}(6))/2 = 1{,}03477;$

$F_{-1} = (X(3)/\tilde{X}(3) + X(7)/\tilde{X}(7))/2 = 1{,}56754;$

$F_0 = (X(4)/\tilde{X}(4) + X(8)/\tilde{X}(8))/2 = 0{,}93521.$

Аналогично рассчитываем начальные сезонные коэффициенты для ряда $Y(t)$: $F_{-3} = 0{,}45835$; $F_{-2} = 1{,}04636$; $F_{-1} = 1{,}81021$; $F_0 = 0{,}68278$.

3. Устанавливаем параметры сглаживания α_1, α_2, α_3.

4. Находим прогноз на первый шаг по модели (1), где $t = 0$, $\tau = 1$, $L = 4$.

5. Находим величину отклонения $e(t + 1) = Y(t + 1) – \tilde{Y}(t + 1)$ и относительную ошибку $|e(t + 1)| / Y(t + 1)$.

6. Корректируем коэффициенты модели a_t, b_t и сезонный коэффициент F_t по формулам (9).

7. Находим прогноз на следующий момент времени $\tilde{Y}(t + 1) = (a_t + b_t)F_{t+1–L}$.

8. Возвращаемся к п. 5, расчеты повторяем вплоть до конечной точки $Y(T)$ исходного ряда.

9. На выходе получаем коэффициенты a_T и b_T искомой модели, а также массив сезонных коэффициентов $F = (F_{T–3}, F_{T–2}, F_{T–1}, F_T)$.

10. Рассчитываем среднюю по модулю относительную ошибку аппроксимации исходного ряда $\bar{e} = (|e(1)| + … + |e(T)|)/T$.

11. Коэффициенты a_T и b_T модели и массив F используем для расчета прогноза по формуле (8) при $\tau = 1, 2, …$.

Реализация представленного алгоритма была выполнена на основе пакета прикладных программ STADIA и MS Excel. Согласно данному алгоритму адаптация к информации (за счет параметров сглаживания) происходит итеративно с получением каждой новой фактической точки ряда. Модель при этом постоянно впитывает динамику новых уровней ряда, приспосабливается к ним и на выходе отражает развитие процесса к конечному моменту наблюдений. Для выбора лучшей модели (в целях прогнозирования) поиск компромиссных значений констант сглаживания (α_1, α_2, α_3) был выполнен следующим образом. На первом этапе расчетов последний участок исходных временных рядов, поквартальные данные за 2013 г., мы оставили необработанным. По укороченным рядам $\{X(t), t = 1, …, 32\}$ и $\{Y(t), t = 1, …, 32\}$ был построен ряд моделей Хольта – Уинтерса и рассчитан прогноз на 2013 г. для различных наборов (α_1, α_2, α_3). При этом всякий раз при переборе возможных значений параметров (α_1, α_2, α_3) на сетке значений от 0 до 1 с шагом 0,1 находились отклонения e_i прогнозируемых уровней от фактических данных за 2013 год и вычислялась средняя по модулю относительная ошибка прогнозирования $\bar{\varepsilon} = \frac{1}{k}\sum_{i=1}^{k}\frac{|e_i|}{y_{(\text{факт})i}}$. Таким образом, критерием сравнения моделей служила ошибка $\bar{\varepsilon}$. Вектор значений, свой по въезду и выезду, при котором была получена ошибка $\bar{\varepsilon}$, не превышающая 0,1, взят нами за основу на втором этапе расчетов моделей по исходным данным, включая

2013 г. В конечном итоге были построены следующие адаптивные модели прогнозирования объемов выездного и въездного туристских потоков соответственно:

$\tilde{X}(T + \tau) = (4273{,}613 + 92{,}157\tau)F_{T+\tau-4}$ (при $\alpha_1 = 0{,}1$; $\alpha_2 = 0{,}6$, $\alpha_3 = 0{,}1$; $\bar{e} = 0{,}07$);

$\tilde{Y}(T + \tau) = (721{,}221 + 7{,}056\tau)F_{T+\tau-4}$ (при $\alpha_1 = 0{,}1$; $\alpha_2 = 0{,}9$, $\alpha_3 = 0{,}4$; $\bar{e} = 0{,}10$).

Невысокое значение ошибки $\bar{e}$ свидетельствует о хорошей точности моделей. Статистический анализ ряда остатков показал случайный характер их колебаний около нуля и отсутствие значимой автокорреляции, что подтверждает адекватность построенных моделей.

Для вычисления точечного прогноза на τ шагов вперед в уравнение модели подставляют значения $\tau = 1, 2, \ldots$. Массивы сезонных коэффициентов (сформировавшихся к концу наблюдений при $T = 36$) для прогнозирования по кварталам за 2014 – 2015 гг. представлены в табл. 38. Расчет прогноза на 2014 – 2015 годы по построенным моделям при $\tau = 1, 2, \ldots, 8$ представлен в табл. 39.

Таблица 38. Коэффициенты сезонности

Квартал	**Коэффициент сезонности ряда *X*(*t*)**	**Коэффициент сезонности ряда *Y*(*t*)**
I	0,8012	0,3548
II	1,1795	1,1302
III	1,4938	1,7496
IV	0,8418	0,5000

Таблица 39. Прогноз объемов туристских потоков на 2014 – 2015 гг.

Год	**Квартал**	**Прогноз по выезду *X*(*t*), тыс. чел.**	**Прогноз по въезду *Y*(*t*), тыс. чел.**
2014	I	3497,8	258,4
	II	5258,3	831,1
	III	6797,1	1298,8
	IV	3908,1	374,7
2015	I	3793,1	268,4
	II	5693,1	863,0
	III	7347,8	1348,2
	IV	4218,4	388,8

Расчёты показали, что для моделирования динамики туристских потоков в целях краткосрочного прогнозирования целесообразно использовать адаптивную методику, сочетающую в себе (с поступлением новой информации) корректировку тренда и сезонных коэффициентов. Прогноз сезонной волны дает наибольшее значение по въезду и выезду в третьем квартале каждого года. Результаты прогнозирования показывают рост объемов въездных туристских потоков, хотя и невысокий по сравнению с динамикой выезда.

6.6. Учёт фактора сезонности на основе модели регрессии с фиктивными переменными

Термин «фиктивные переменные» используется как противоположность понятию «значащие переменные», показывающему уровень количественного показателя, принимающего значения из непрерывного интервала. Как правило, фиктивная переменная – это индикаторная переменная, отражающая качественную характеристику. Чаще всего применяют бинарные фиктивные переменные, принимающие два значения, 0 и 1, в зависимости от определенного условия.

Ниже приводится пример построения модели регрессии с фиктивными переменными (отвечающими за сезонные факторы).

Пример 9. Имеются следующие поквартальные данные (табл. 40) грузооборота автомобильного транспорта по Владимирской области за 2004 – 2011 гг. (млн т · км).

Таблица 40. Грузооборот автомобильного транспорта

Квартал	**Год**							
	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
I	81,8	193,4	152,2	196,4	135	171,3	217,3	229
II	93,5	247,2	189,3	277,7	217,2	188,8	257,7	288,6
III	103,7	285,3	227,2	267,7	349,5	221,7	311,7	333
IV	86,9	187,2	204,1	180	318,3	225,7	277	356,8

Визуальный анализ графика (рис. 39) свидетельствует о наличии тренда, сезонности и случайной составляющей. Сезонные колебания демонстрируют периодические подъемы в III квартале и спады в IV квартале. Построим модель вида [7, с. 95]

$$y_t = a_0 + a_1 t + \sum_{i=2}^{4} c_i D_i + \varepsilon_t$$

где $D_i = \begin{cases} 1, \text{ если наблюдение принадлежит } i\text{-ому кварталу (кв. II, III, IV)} \\ 0, \text{ если наблюдение принадлежит I кварталу} \end{cases}$

Эталонным будем брать I квартал, тогда фиктивные переменные (D_2, D_3, D_4) позволяют оценить разницу в уровнях сезонности между эталонным кварталом и остальными (табл. 41).

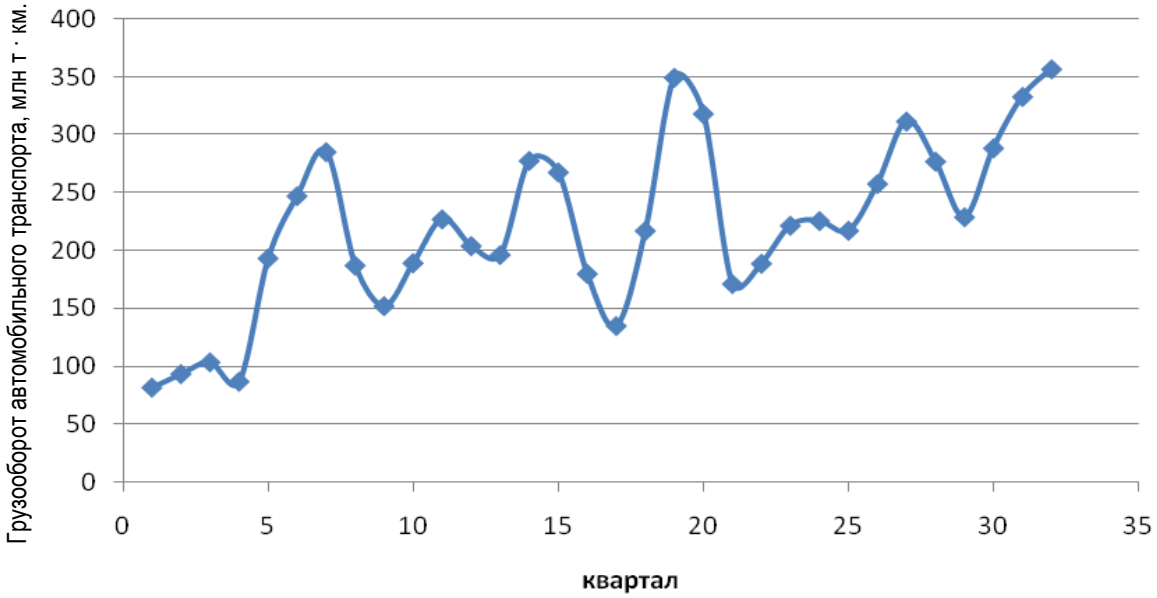

Рис. 39. График динамики грузооборота автомобильного транспорта

Таблица 41. Матрица данных

Yt	***t***	***D2***	***D3***	***D4***
81,8	1	0	0	0
93,5	2	1	0	0
103,7	3	0	1	0
86,9	4	0	0	1
193,4	5	0	0	0
247,2	6	1	0	0
285,3	7	0	1	0
187,2	8	0	0	1
152,2	9	0	0	0
189,3	10	1	0	0
227,2	11	0	1	0

Окончание табл. 41

Yt	***t***	***D*2**	***D*3**	***D*4**
204,1	12	0	0	1
196,4	13	0	0	0
277,7	14	1	0	0
267,7	15	0	1	0
180	16	0	0	1
135	17	0	0	0
217,2	18	1	0	0
349,5	19	0	1	0
318,3	20	0	0	1
171,3	21	0	0	0
188,8	22	1	0	0
221,7	23	0	1	0
225,7	24	0	0	1
217,3	25	0	0	0
257,7	26	1	0	0
311,7	27	0	1	0
277	28	0	0	1
229	29	0	0	0
288,6	30	1	0	0
333	31	0	1	0
356,8	32	0	0	1

Используя программу STADIA, строим модель множественной линейной регрессии. Приведём протокол расчёта:

Коэфф.	a_0	a_1	a_2	a_3	a_4
Значение	95,65	5,094	42,86	80,24	42,17
Ст.ошиб.	23,24	0,981	25,45	25,51	25,6
Значим.	0,0005	0	0,1001	0,0042	0,1075

Множеств R	R^2	R^2 прив	Ст. ошиб.	F	Значим
0,77233	0,59649	0,53671	50,863	9,978	0,0001

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>.

В данном случае уравнение модели принимает вид $y_t = 95{,}65 + 5{,}094t + 42{,}86D_2 + 80{,}24D_3 + 42{,}17D_4$.

Статистические характеристики модели следующие: $R^2 = 0{,}60$; F(набл) = 9,98. Трендовая составляющая модели показывает, что примерный рост грузооборота с течением времени составляет около 5 млн т · км. Запишем уравнения, характеризующие динамику уровней

ряда по четырём кварталам: $y1(t) = 95{,}65 + 5{,}094t$; $y2(t) = 138{,}51 + 5{,}094t$; $y3(t) = 175{,}89 + 5{,}094t$; $y4(t) = 137{,}82 + 5{,}094t$.

Различия в регрессионных уравнениях вызваны влиянием сезонных эффектов, свободные члены уравнения показывают возрастание объемов грузооборота во втором и третьем кварталах (весенне-летний период) и значительное понижение в зимний период.

При $t = 33, D_2 = 0, D_3 = 0, D_4 = 0$: $y = 263{,}7$; соответственно: $y_{\text{факт}} = 226{,}4$.

При $t = 34, D_2 = 1, D_3 = 0, D_4 = 0$: $y = 311{,}7$; соответственно: $y_{\text{факт}} = 246{,}6$.

При $t = 35, D_2 = 0, D_3 = 1, D_4 = 0$: $y = 354{,}2$; соответственно: $y_{\text{факт}} = 277{,}0$.

При $t = 36, D_2 = 0, D_3 = 0, D_4 = 1$: $y = 321{,}2$; соответственно: $y_{\text{факт}} = 266{,}9$.

В результате мы получили, что прогнозные значения по нашей модели несколько отличаются от фактических значений за тот же промежуток времени.

6.7. Моделирование периодических колебаний временного ряда инвестиций в основной капитал

Сезонная и циклическая компоненты являются периодическими с разными периодами. Обычно в экономических временных рядах на длиннопериодическую циклическую составляющую накладывается сезонная волна или волны более высоких частот. При этом выдвигают предположение о том, что эти волны не влияют друг на друга, и используют аддитивную модель их совместного существования.

С другой стороны, получила распространение методика расчёта периодических составляющих временного ряда (в том числе сезонных и длиннопериодических) и определения их достоверности и затем включения в модель. Процесс моделирования сезонной и циклической компонент выполняется в такой последовательности [13, с. 127 – 129]:

Шаг 1. Выбор модели для выявления периодических составляющих. Для выявления периодических составляющих временного ряда используют следующую модель:

$$y_t = a_0 + a_1 t + a_2 \sin(\frac{2\pi t}{T}) + a_3 \cos(\frac{2\pi t}{T}) + e_t,$$

где a_0, a_1, a_2, a_3 – коэффициенты, определяемые методом наименьших квадратов; $a_0 + a_1 t$ – тренд, приводящий временной ряд к стационарному виду с нулевым средним значением без искажения периодических составляющих остатка; $\sin(\frac{2\pi t}{T})$, $\cos(\frac{2\pi t}{T})$ – тригонометрические функ-

ции периодической составляющей временного ряда, сумма которых позволяет учесть начальную фазу колебания; t – порядковый номер времени; T – период или длина волны циклической составляющей временного ряда; $\frac{2\pi t}{T}$ – аргумент тригонометрической функции, выраженный в радианах; $\frac{1}{T}$ – частота, численно равная количеству колебаний в единицу времени;

$$a_2 \sin(\frac{2\pi t}{T}) + a_3 \cos(\frac{2\pi t}{T}) = c \cdot \cos(\frac{2\pi t}{T} + \varphi),$$

где φ – начальная фаза колебаний;

$$\varphi = \text{arctg}(\frac{a_1}{a_2});$$

$$c = \sqrt{a_2^2 + a_3^2}$$

Шаг 2. Создается база данных.

$X_1 = t$ – время;

$$X_2 = \sin(\frac{2\pi t}{T});$$

$$X_3 = \cos(\frac{2\pi t}{T});$$

где T – период циклической составляющей.

Шаг 3. Вычисление ошибки S_e данной модели. Переход к шагу 2 с изменением периода T.

Шаг 4. Построение графика зависимости ошибки модели S_e временного ряда от периода T циклического фактора, включенного в модель.

Шаг 5. Определение периодических составляющих временного ряда. Определяются такие значения периодов T, при которых ошибка модели S_e имеет глобальный и локальный минимумы. Обнаруженные периоды характеризуют периодические составляющие временного ряда, среди которых надо исключить ложные периоды. Ложная периодичность относится к периодам (эхам), которые кратны величине основного периода. Основной период имеет наименьшую ошибку модели, а кратные или ложные периоды имеют увеличенную ошибку модели. Например, квартальная периодическая составляющая три месяца порождает эхо через шесть месяцев или годовой период.

Шаг 6. Определение достоверных периодических составляющих временного ряда. Для определения достоверных периодических составляющих необходимо построить модель с участием всех обнаруженных периодических составляющих и по критерию Стьюдента определить достоверность коэффициентов, стоящих перед тригонометрическими функциями.

Шаг 7. Построение модели временного ряда.

Окончательная модель временного ряда с включением всех компонент может иметь следующий вид:

$$y_t = a_0 + a_1 t + a_2 \sin(\frac{2\pi t}{T_1}) + a_3 \cos(\frac{2\pi t}{T_1}) + a_4 \sin(\frac{2\pi t}{T_2}) + a_5 \cos(\frac{2\pi t}{T_2}) + e_t,$$

где $(a_0 + a_1 t)$ – тренд; $a_2 \sin(\frac{2\pi t}{T_1}) + a_3 \cos(\frac{2\pi t}{T_1})$ – сезонная компонента; T_1 – период сезонной компоненты; $a_4 \sin(\frac{2\pi t}{T_2}) + a_5 \cos(\frac{2\pi t}{T_2})$ – циклическая компонента; T_2 – период циклической компоненты; e_t – остатки компоненты.

Коэффициенты модели рассчитываются методом наименьших квадратов.

Существуют следующие проблемы предложенной модели временного ряда:

– обнаруженные периодические колебания, составляющие временной ряд, должны иметь экономическое обоснование и механизм их возникновения;

– если возникают трудности в экономической интерпретации тенденции временного ряда (периодических составляющих), то можно использовать алгоритмические методы. Если временной ряд имеет некоторую тенденцию, то эти методы способны ее воспроизвести;

– для построения доверительных интервалов регрессии и прогноза необходимо считать более точным не среднее значение временного интервала, а его последнее, более свежее значение.

Пример 10. Имеются поквартальные данные инвестиций в основной капитал (млн руб.) по Владимирской области за 2003 – 2011 гг., представленные в табл. 18 и на рис. 22 (см. главу 5).

Визуальный анализ графика исходных данных позволяет судить о наличии периодических колебаний в данном ряду (рис. 22).

Корреляционный анализ исходных данных выделяет значимый коэффициент автокорреляции с лагом l = 4, что указывает на присутствие периодических колебаний в данном ряду (см. рис. 40 и протокол расчёта по программе STADIA).

Коэффициенты автокорреляции:

Сдвиг	Корреляция	Крит. значение
0	1	0,3297
1	0,3032	0,3347
2	0,238	0,3398
3	0,2007	0,3451
4	**0,7259**	**0,3507**
5	0,1696	0,3565
6	0,1201	0,3628
7	0,08218	0,3692
8	0,5342	0,3762

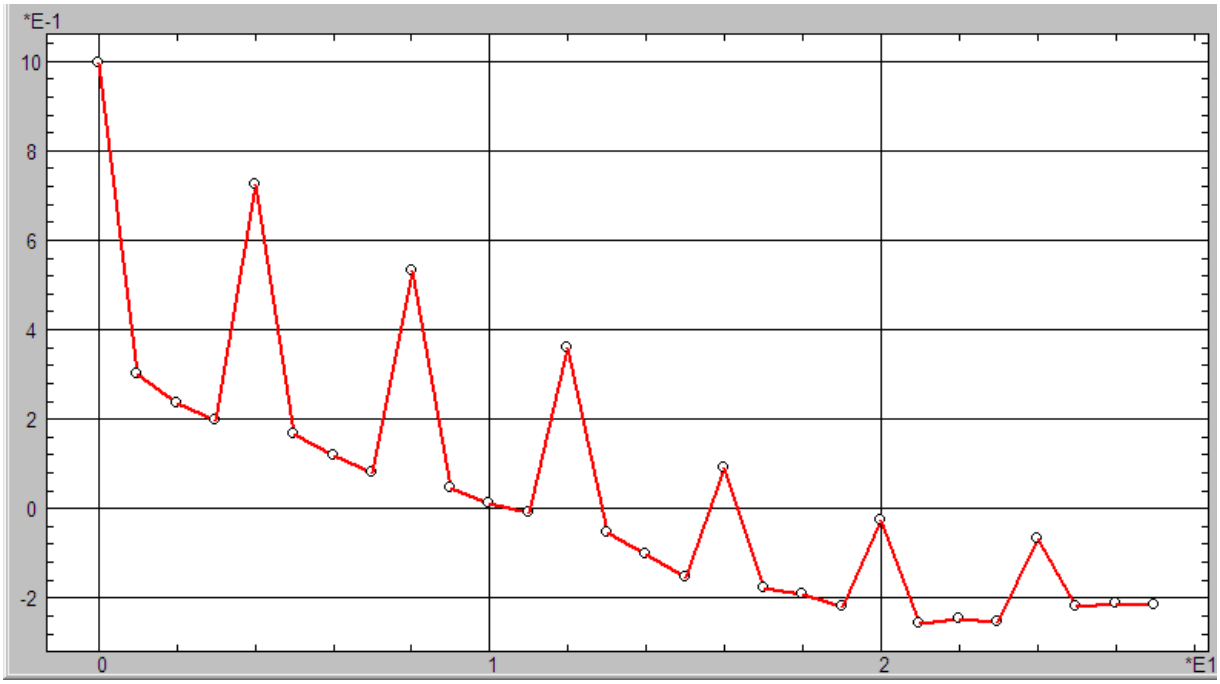

Рис. 40. График автокорреляционной функции

С целью определения достоверных периодических составляющих временного ряда рассчитаем модели вида

$$y_t = a_0 + a_1 t + a_2 \sin(\frac{2\pi t}{T}) + a_3 \cos(\frac{2\pi t}{T}) + e_t,$$

где a_0, a_1, a_2, a_3 – коэффициенты для периодов $T = 1, 2, \ldots, 16$, определяемые методом наименьших квадратов. Для каждой модели вычислим

ошибку регрессии S_e и F-критерий и построим графики зависимости ошибки регрессии от периода колебаний T и F-критерия (рис. 41, 42).

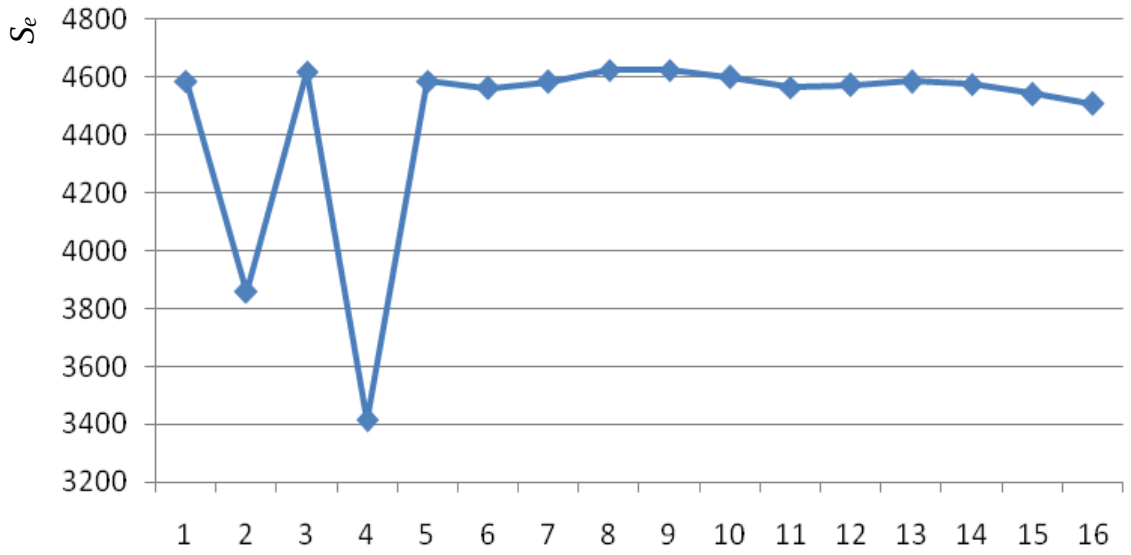

Рис. 41. График зависимости стандартной ошибки регрессии от периода T

Визуальный анализ рис. 41 показывает наличие двух выраженных локальных минимумов ошибок моделей при периодах $T = 2$ и $T = 4$. Также на рис. 42 отчетливо видны два максимума при $T = 2$ и $T = 4$. В этой связи в модель временного ряда наряду с линейной составляющей включаем соответствующие периодические составляющие.

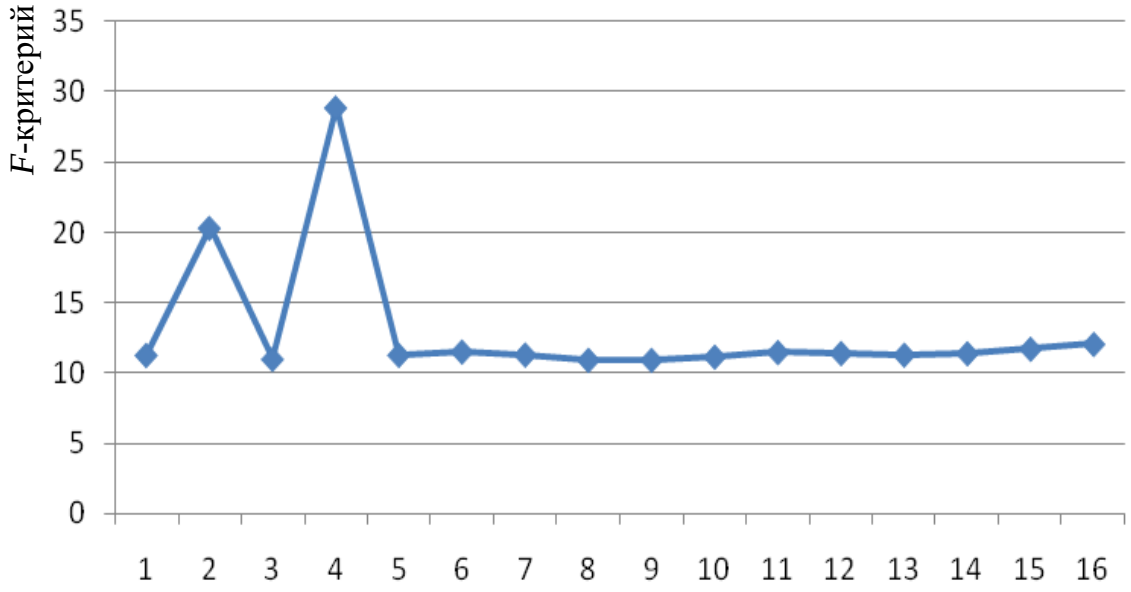

Рис. 42 График зависимости F-критерия от периода T

Строим следующую многофакторную модель общей нелинейной регрессии:

$$y(t) = a_0 + a_1 \cdot t + a_2 \cdot \sin(\frac{2\pi t}{2}) + a_3 \cdot \cos(\frac{2\pi t}{2}) + a_4 \cdot \sin(\frac{2\pi t}{4}) +$$

$$+ a_5 \cdot \cos(\frac{2\pi t}{4}).$$

Протокол расчета на основе программы STADIA приведен ниже.

Модель множественной регрессии:

Коэффиц.	a_0	a_1	a_2	a_3	a_4	a_5
Значение	**855,7**	**387,7**	**–6,09E4**	**372**	**–2024**	**3626**
Ст.ошиб.	811,2	38,28	2,402E4	811,1	561,7	561,7
Значим.	0,3003	0	0,0158	0,6541	0,0014	0

Множеств R	R^2	R^2прив	Ст. ошиб.	F	Значим
0,93702	0,87801	0,85768	2371,8	43,19	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

$X_{прогн}$	$Y_{прогн}$	Ст. ошиб	Довер. инт
37	9325	2354	4722
38	1,596E4	2364	4742
39	1,373E4	2375	4764
40	2,43E4	2386	4786

Таким образом, по критерию Фишера модель статистически значима. Прогнозные значения по данной модели отличаются от фактических данных за тот же период времени. Прогноз на I квартал 2012 года составил 9325 млн руб. Фактическое значение равно 7007,9 млн руб. Результаты моделирования в зоне доверительного интервала представлены на рис. 43.

В целях прогнозирования инвестиций было продолжено исследование на основе расчета мультипликативной модели Хольта – Уинтерса. Для выбора наилучшей модели была вычислена средняя относительная ошибка аппроксимации исходных данных для различных параметров сглаживания на сетке значений с шагом 0,01 в диапазоне от 0 до 1, и в итоге были выбраны параметры сглаживания (α_1 = 0,50; α_2 = 0,61; α_3 = 0,16), дающие минимальную ошибку. Небольшое значение ошибки (14,79 %) свидетельствует о неплохой аппроксимации исходных данных. Получили следующую модель:

$y(40 + \tau) = (16616{,}534 + 474{,}269\tau)F_\tau$.

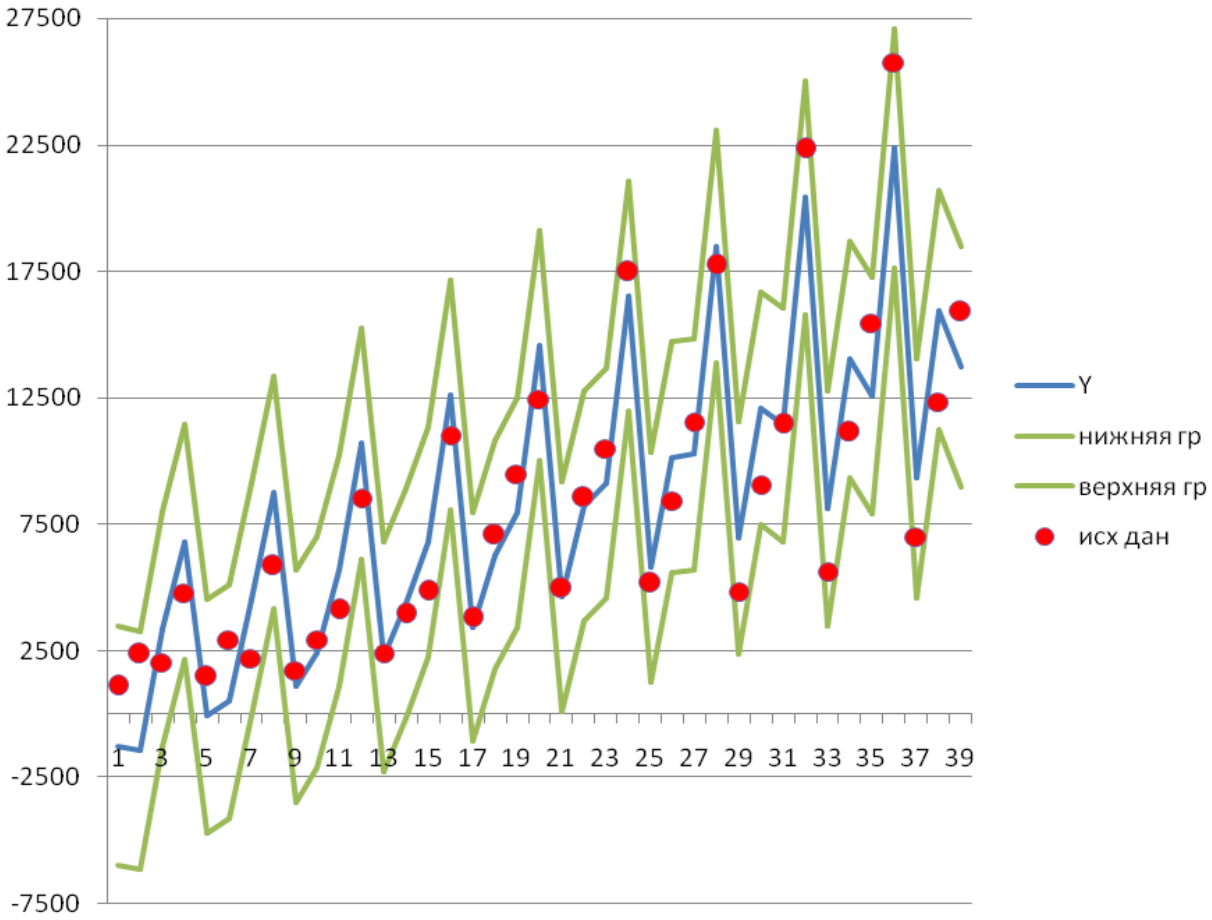

Рис. 43. Динамика инвестиций в зоне доверительного интервала (млн руб.)

Расчёты по модели показали, что к началу 2013 года вектор сезонных коэффициентов сформировался так, что снижение уровней наблюдается в I квартале и значительное повышение в IV квартале (табл. 43, рис. 44). Расчет прогноза предполагает общую сумму инвестиций к концу 2013 г., равную 73028,97 млн руб., что на 12 % больше по сравнению с 2012 г.

Таблица 43. Расчет прогноза на 2013 год

τ	***F***	**Прогноз на 2013 год**
1,000	0,452	7723,378
2,000	0,816	14336,568
3,000	1,031	18607,234
4,000	1,748	32361,793
Итого		**73028,973**

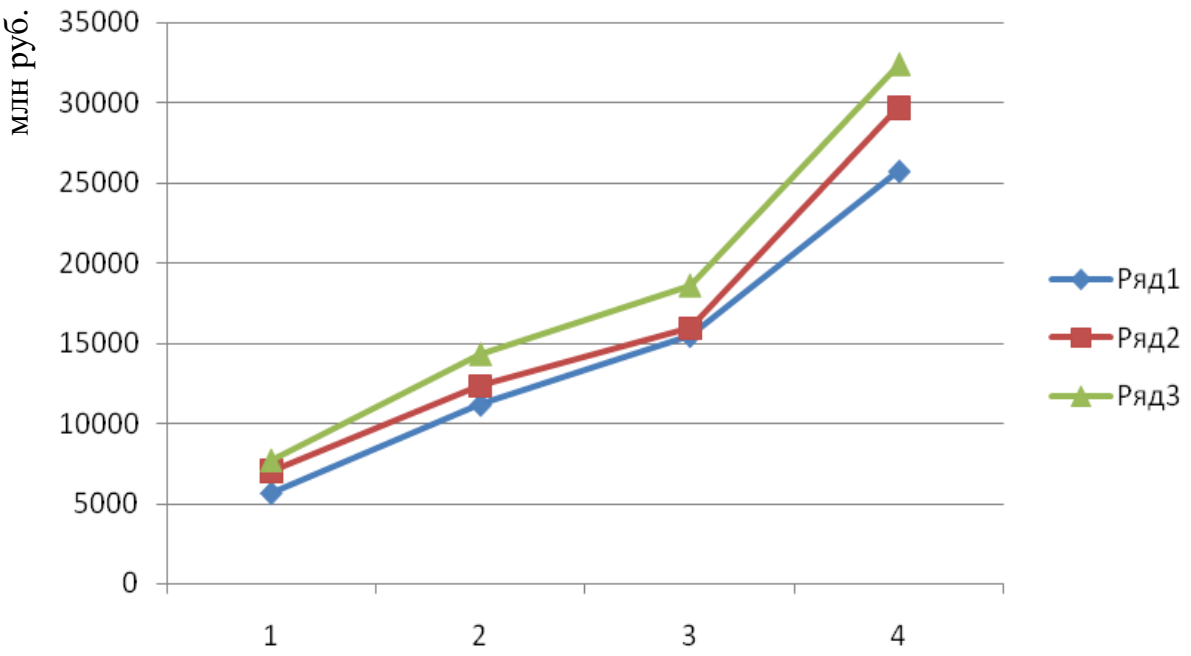

Рис. 44. Объёмы инвестиций (ряд 1 – инвестиции за 2011 г.; ряд 2 – инвестиции за 2012 г.; ряд 3 – прогноз на 2013 г. – поквартальные данные)

6.8. Выводы по результатам прогнозирования временных рядов

Временные ряды – объект изучения специального раздела математической статистики, посвященного анализу случайных процессов. Во временном ряде содержится информация об особенностях и закономерностях протекания изучаемого процесса, а статистический анализ позволяет выявить закономерности и использовать их для оценки характеристик процесса в будущем, т. е. для прогнозирования.

В общем случае каждый уровень временного ряда можно представить как функцию четырех компонент $f(t)$, $s(t)$, $u(t)$, $\varepsilon(t)$, отражающих закономерность и случайность развития. Здесь $f(t)$ – тренд (долговременная тенденция) развития; $s(t)$ – сезонная компонента; $u(t)$ – циклическая компонента; $\varepsilon(t)$ – остаточная компонента. Одна из важнейших задач исследования экономического временного ряда – выявление основной тенденции изучаемого процесса, выраженной неслучайной составляющей $f(t)$ (тренда либо тренда с циклической или сезонной компонентой). Случайная компонента $\varepsilon(t)$ – обязательная составная часть любого временного ряда в экономике, так как случайные отклонения неизбежно сопутствуют любому экономическому яв-

лению. При прогнозировании таких процессов следует учитывать фактор сезонности $s(t)$ и, насколько это возможно, снижать ее отрицательные воздействия.

Статистические процедуры оценивания сезонности и корректировки временных рядов имеют большое значение в сфере экономики; причем оценки сезонности представляют собой лишь промежуточный этап исследования. В дальнейшем их используют для моделирования тренд-сезонных процессов.

Для исследования и прогнозирования тренд-сезонных экономических процессов независимо от причин, порождающих сезонные колебания, необходимо уметь решать следующие задачи:

- определять наличие тренда во временном ряде;
- выявлять присутствие во временном ряде сезонных колебаний;
- осуществлять фильтрацию временного ряда, т. е. разделять ряд на составляющие его компоненты – тренд, сезонную и случайную компоненты;
- анализировать динамику сезонной волны, т. е. выявлять, изменяется ли со временем ее амплитуда и происходит ли перемещение точек экстремума сезонной волны;
- составлять прогноз тренд-сезонного экономического процесса.

Порядок решения перечисленных задач и их состав может изменяться в зависимости от цели исследования, методов решения, используемых пакетов прикладных программ и других факторов.

В настоящее время прогнозы на основе адаптивных моделей разрабатываются для исследования динамики сезонных экономических процессов. Для прогнозирования тренд-сезонных процессов используются аддитивная и мультипликативная модели Хольта – Уинтерса. Если амплитуда колебаний с течением времени претерпевает существенные изменения, то отдают предпочтение мультипликативной модели.

Сформулируем выводы по нашим результатам моделирования:

1. Сезонность отрицательно влияет на экономические процессы, поэтому ее необходимо уметь измерять и анализировать, с тем чтобы снижать ее отрицательные воздействия. Для определения наличия сезонности эффективно работают следующие методы: расчет сезонной волны, дисперсионный, гармонический, автокорреляционный анализа.

2. В целях прогнозирования объема платных услуг населению по Владимирской области целесообразно использовать тренд-сезонную аддитивную модель с участием экспоненциального тренда. Наша модель показала тенденцию возрастания показателя с каждым годом примерно в 1,2 раза. Сезонный вектор указывает, что в течение года предполагаются высокие объемы услуг в декабре, сезонность отрицательно влияет на август, сентябрь.

3. Для прогнозирования среднедушевых денежных доходов по нашему региону выбор сделан в пользу мультипликативной тренд-сезонной модели с параболической траекторией тренда. Предполагается рост показателя с постоянным ускорением, которое корректируется с каждым месяцем сезонными коэффициентами, в течение года модель демонстрирует понижение доходов.

4. Динамика инвестиций в основной капитал весьма подвержена сезонным колебаниям, что доказано расчетом АКФ. Прогноз инвестиций в основной капитал точнее отражает будущую динамику ряда по модели Хольта – Уинтерса с линейной тенденцией и сезонностью в мультипликативной форме. К началу 2013 г. вектор сезонных коэффициентов сформировался так, что снижение уровней наблюдается в I квартале и значительное повышение – в IV квартале; расчет прогноза предполагает общую сумму инвестиций к концу 2013 г. в размере 73028,97 млн руб., что на 12 % больше по сравнению с 2012 г.

5. Модели регрессии по одноименным кварталам и с участием фиктивных переменных целесообразны для прогнозирования грузооборота автомобильного транспорта. Построенная модель демонстрирует рост грузооборота с каждым кварталом примерно на 5 млн т · км; сезонность выражена в возрастании объемов грузооборота в III и IV кварталах (летне-осенний период) и значительном их понижении в зимний период.

В ряде случаев математические модели прогнозирования могут быть полезным инструментом исследования. При этом, конечно, для увеличения точности прогнозов развития случайных процессов в изменяющихся условиях, условиях неопределенности или неполной информации необходима работа по совершенствованию моделей.

Контрольные вопросы

1. Когда говорят о наличии сезонных колебаний в экономических процессах? Какие вы знаете методы выявления сезонных колебаний?

2. Поясните расчет индексов сезонности и построение сезонной волны.

3. Что понимают под тренд-сезонным временным рядом?

4. Какие этапы исследования включает в себя методика прогнозирования с помощью тренд-сезонных моделей?

5. Укажите уравнение аддитивной модели прогнозирования. В каких случаях ее следует строить? Поясните основные шаги алгоритма получения аддитивной модели.

6. Что представляет собой мультипликативная модель прогнозирования? Когда ее используют? Как рассчитывают скорректированную сезонную компоненту в мультипликативной модели?

7. В каком случае отдают предпочтение мультипликативной модели прогнозирования?

8. Чем отличается алгоритм построения мультипликативной модели от алгоритма аддитивной модели прогнозирования?

9. На чем базируется аналитический метод построения адекватной модели периодических составляющих? Как выглядит уравнение такой модели? Поясните все входящие в него параметры.

10. Какие вы знаете адаптивные модели прогнозирования тренд-сезонных процессов?

11. В чём сущность методики построения моделей Хольта – Уинтерса?

12. Чем отличаются модели Хольта – Уинтерса от адаптивных моделей Брауна?

ЗАКЛЮЧЕНИЕ

Современные компьютерные технологии и развитые методы прикладной математической статистики позволяют анализировать сложные многофакторные системы экспериментальных данных. При помощи этих методов изучают связи между различными факторными переменными, выявляются новые, ранее неизвестные связи, уточняют или отвергают гипотезы о существовании закономерностей в развитии явлений и процессов.

Однако при всем многообразии существующих подходов многие вопросы теории и практики статистических исследований разработаны недостаточно. Так, сложность изучения региональной экономики обусловлена многосторонним характером социально-экономического статуса регионов, массой причин, определяющих нестационарность и случайность факторов, влияющих на развитие регионов. Следует обратить внимание на комплексный анализ данных так, чтобы изучить числовую информацию со всех сторон на основе комбинации многомерных методов математической статистики. В этой связи настоящее пособие представляет систематизированное изложение механизмов статистико-математического моделирования. Решение прикладных задач проиллюстрировано на практических примерах, уделено внимание оценке качества моделей прогнозирования временных рядов.

Издание нацелено на формирование профессиональных компетенций учащихся в области математического моделирования, развитие их общей математической культуры и навыков применения компьютерных технологий. Очень важно, исходя из предмета и конечной цели исследования, сформулировать задачу на языке построения абстрактных математических моделей с учетом применения компьютерных технологий. Здесь совершенно необходимо знание и умение использовать основополагающие методы и принципы теории вероятностей и математической статистики. Высокий уровень образования, позволяющий умело владеть математическим аппаратом как инструментом для исследований, приводит к глубокому осмыслению процессов, происходящих в природе и обществе.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Арженовский, С. В. Методы социально-экономического прогнозирования : учеб. пособие / С. В. Арженовский. – М. ; Ростов н/Д. : Наука-Спектр, 2009. – 236 с. – ISBN 978-591131-941-0.

2. Белянский, В. П. Прогнозирование в индустрии гостеприимства и туризма : учебник / В. П. Белянский. – М. : РЭА им. Плеханова, 2005. – 278 с. – ISBN 5-7307-0540-9.

3. Вероятность и математическая статистика : энциклопедия / под ред. Ю. В. Прохорова. – М. : Большая рос. энцикл., 2003. – 912 с. – ISBN 5-7107-7433-2.

4. Владимирова, Л. П. Прогнозирование и планирование в условиях рынка : учеб. пособие / Л. П. Владимирова. – М. : Дашков и К°, 2005. – 400 с. – ISBN 5-94798-613-2.

5. Гришин, А. Ф. Статистические модели: построение, оценка, анализ : учеб. пособие / А. Ф. Гришин, Е. В. Кочерова. – М. : Финансы и статистика, 2005. – 416 с. – ISBN 5-279-02941-6.

6. Домбровский, В. В. Эконометрика : учебник / В. В. Домбровский. – М. : Новый учебник, 2004. – 342 с. – ISBN 5-8393-0400-X.

7. Дуброва, Т. А. Прогнозирование социально-экономических процессов. Статистические методы и модели : учеб. пособие / Т. А. Дуброва. – М. : Маркет ДС, 2007. – 192 с. – ISBN 978-5-7958-0170-4.

8. Елисеева, И. И. Общая теория статистики : учебник / И. И. Елисеева, М. М. Юзбашев. – М. : Финансы и статистика, 1995. – 368 с. – ISBN 5-279-01181-9.

9. Ерохина, Л. И. Предприятия в среде сервиса. Управление прогнозируемыми процессами (теория и практика) : учеб. пособие / Л. И. Ерохина. – М. : Флинта, 2005. – 245 с. – ISBN 5-89502-760-1.

10. Наука РАН, РАСХН, РАМН в цифрах: 2013 : стат. сб. / И. В. Зиновьева [и др.] ; под ред. Л. Э. Миндели. – М. : Институт проблем развития наук РАН, 2014. – 240 с. – ISBN 978-5-91294-077-4.

11. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. – М. : ФИЗМАТЛИТ, 2006. – 816 с. – ISBN 5-9221-0707-0.

12. Кулаичев, А. П. Методы и средства комплексного анализа данных / А. П. Кулаичев. – М. : ФОРУМ : ИНФРА-М, 2006. – 512 с. – ISBN 5-8199-0234-3.

13. Курбыко, И. Ф. Дополнительные главы математической статистики : учеб. пособие / И. Ф. Курбыко, А. С. Левизов, С. В. Левизов. – Владимир : Изд-во ВлГУ, 2011. – 135 с. – ISBN 978-5-9984-0188-6.

14. Левизов, А. С. Механизм прогнозирования туристской инфраструктуры как определяющего фактора экономического развития туризма в регионе : монография / А. С. Левизов, В. Ф. Архипова. – Владимир : Собор, 2010. – 200 с. – ISBN 978-50904418-77-9.

15. Левизов, А. С. О моделировании туристских потоков методами математической статистики / А. С. Левизов, И. Ф. Курбыко // Математика. Экономика. Образование : тр. XIII Междунар. конф. – Ростов н/Д. : ЦВВР, 2005.

16. Левизов, А. С. Прогнозирование динамики туристских потоков на основе адаптивных моделей / А. С. Левизов, И. Ф. Курбыко // Динамика сложных систем. – 2014. – № 2.

17. Левизов, А. С. Статистическая оценка инновационного развития российских регионов / А. С. Левизов, И. Ф. Курбыко // Вестник университета (ГУУ). Серия «Развитие отраслевого и регионального управления». – 2015. – № 8.

18. Лопатников, Л. И. Экономико-математический словарь : Словарь современной экономической науки / Л. И. Лопатников. – М. : Дело, 2003. – 520 с. – ISBN 5-7749-0275-7.

19. Математический энциклопедический словарь / под ред. Ю. В. Прохорова. – М. : Совет. энцикл., 1988. – 845 с. – ISBN 5-85270-278-1.

20. Медведева, Н. Б. Статистическое изучение влияния финансирования науки на уровень инновационного развития экономики Российской Федерации / Н. Б. Медведева // Вестник университета (ГУУ). Серия «Развитие отраслевого и регионального управления». – 2015. – № 8.

21. Наука, инновации и информационное общество / Федеральная служба государственной статистики [Электронный ресурс]. – Режим доступа: http: //www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations/science/# (дата обращения: 27.04.2016).

22. Орлова, И. В. Экономико-математические методы и модели: компьютерное моделирование : учеб. пособие / И. В. Орлова, В. А. Половников. – М. : Вуз. учеб., 2014. – 388 с. – ISBN 978-5-9558-0208-4.

23. Радченко, С. Г. Устойчивые методы оценивания статистических моделей : монография / С. Г. Радченко. – Киев : Санспарель, 2005. – 504 с. – ISBN 966-96574-0-7.

24. Регионы России. Социально-экономические показатели : стат. сб. / Росстат. – М., 2014. – 900 с.

25. Регионы России. Социально-экономические показатели : стат. сб. / Росстат. – М., 2016. – 1326 с.

26. Салин, В. Н. Курс теории статистики для подготовки специалистов финансово-экономического профиля : учебник / В. Н. Салин, Э. Ю. Чурилова. – М. : Финансы и статистика, 2006. – 480 с. – ISBN 5-279-03063-5.

27. Эконометрика : учеб. пособие в схемах и табл. / под ред. С. А. Орехова. – М. : Эксмо, 2008. – 224 с. – ISBN 978-5-699-26764-4.

Учебное издание

КУРБЫКО Инна Федоровна
ЛЕВИЗОВ Алексей Сергеевич
ЛЕВИЗОВ Сергей Всеволодович

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ

Учебное пособие

Редактор Е. С. Глазкова
Технический редактор А. В. Родина
Корректор Н. В. Пустовойтова
Компьютерная верстка Е. А. Герасиной
Выпускающий редактор А. А. Амирсейидова

Подписано в печать 16.11.18.
Формат 60×84/16. Усл. печ. л. 10,70. Тираж 50 экз.
Заказ
Издательство
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых.
600000, Владимир, ул. Горького, 87.