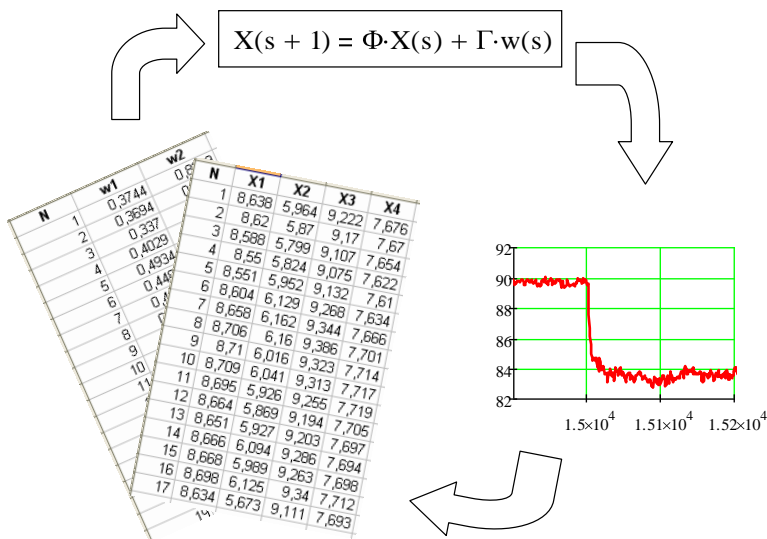


Е. Г. Крушель, А. Э. Панфилов, И. В. Степанченко

Обработка экспериментальной информации. Лабораторный практикум



МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
КАМЫШИНСКИЙ ТЕХНОЛОГИЧЕСКИЙ ИНСТИТУТ (ФИЛИАЛ)
ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО БЮДЖЕТНОГО ОБРАЗОВАТЕЛЬНОГО
УЧРЕЖДЕНИЯ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ВОЛГОГРАДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Е. Г. Крушель, А. Э. Панфилов, И. В. Степанченко

Обработка экспериментальной информации. Лабораторный практикум

Учебное пособие



Волгоград
2014

УДК 519.2(075.8)

К 84

Рецензенты: завкафедрой «Информационная безопасность автоматизированных систем» д. ф.-м. н., профессор Саратовского государственного технического университета имени Ю. А. Гагарина В. В. Байбурин; коллектив кафедры «Системотехника» Саратовского государственного технического университета имени Гагарина Ю. А. (завкафедрой д. т. н., профессор Ю. Б. Томашевский)

Крушель, Е. Г. ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНОЙ ИНФОРМАЦИИ. ЛАБОРАТОРНЫЙ ПРАКТИКУМ: учеб. пособие / Е. Г. Крушель, А. Э. Панфилов, И. В. Степанченко. – Волгоград: ИУНЛ ВолгГТУ, 2014. – 88 с.

ISBN 978-5-9948-1390-4

Изложены практические вопросы обработки экспериментальной информации с примерами такой обработки в программных средах Microsoft Excel и Mathcad. Призвано содействовать успешной подготовке выпускников к решению задач анализа данных.

Предназначено для студентов ВПО, обучающихся по направлению 230100 «Информатика и вычислительная техника».

Ил. 50. Табл. 6. Библиогр.: 6 назв.

Печатается по решению редакционно-издательского совета

Волгоградского государственного технического университета

Учебное издание

Елена Георгиевна Крушель, Александр Эдуардович Панфилов,

Илья Викторович Степанченко

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНОЙ ИНФОРМАЦИИ. ЛАБОРАТОРНЫЙ ПРАКТИКУМ

Учебное пособие

Редактор Попова Л. В. Компьютерная верстка: Попова Л. В.

Темплан 2014 г., поз. № 11К. Подписано в печать 18.07.2014 г. Формат 60×84 ¹/₁₆.

Бумага листовая. Печать офсетная. Усл. печ. л. 5,11. Уч.-изд. л. 5,38.

Тираж 100 экз. Заказ №

Волгоградский государственный технический университет

400131, г. Волгоград, пр. Ленина, 28, корп. 1.

Отпечатано в КТИ; 403874, г. Камышин, ул. Ленина, 5.

ISBN 978-5-9948-1390-4

© Волгоградский
государственный
технический
университет, 2014

ВВЕДЕНИЕ

Учебная дисциплина «Обработка экспериментальной информации» (ОЭИ) не входит в перечень дисциплин базовой части образовательного стандарта нового поколения (ФГОС-3) для направления 230100 – «Информатика и вычислительная техника». Но многие вузы, ведущие подготовку бакалавров по этому направлению, считают необходимым включить дисциплину ОЭИ в учебный план в составе дисциплин, перечень которых вуз устанавливает самостоятельно (вариативная часть), поскольку компьютерная обработка данных относится к научно-исследовательской деятельности выпускников. В частности, изучение основ ОЭИ предусмотрено учебным планом, разработанным в Волгоградском государственном техническом университете (ВолгГТУ).

Согласно этому учебному плану на изучение дисциплины ОЭИ отводится 144 учебных часа со следующим распределением: 18 лекционных часов, 18 часов практических занятий, 18 часов лабораторных работ и 90 часов на выполнение курсовой работы. Повидимому, и в других вузах число часов аудиторных занятий, отводимых для изучения дисциплины, будет ограниченным.

Для того чтобы «вместить» в эти часы основные понятия сложной и наукоемкой дисциплины «Обработка экспериментальной информации», целесообразно использовать время, отведенное на практические и лабораторные занятия, для предоставления материалов, не только закрепляющих, но и *дополняющих* курс лекций как разделами теории обработки данных, так и аспектами использования теории в прикладных задачах. Закрепление навыков обработки данных, полученных студентами, происходит при выполнении курсовой работы. Такая концепция распределения материала по ОЭИ между видами занятий прошла трехлетнюю проверку и успешно используется в Камышинском технологическом институте, входящем в ВолгГТУ в качестве филиала.

В курсе лекций и на практических занятиях по дисциплине ОЭИ рассматриваются следующие вопросы: обработка одномерных рядов наблюдений; области практического применения основных законов распределения экспериментальных данных; задачи статистической проверки гипотез; методы и алгоритмы анализа многомерных рядов наблюдений (дисперсионный, корреляционный, кластерный и факторный анализы).

Из-за ограниченности учебного времени, отведенного на лекционные и практические занятия, вне рассмотрения остался важный раздел ОЭИ, относящийся к обработке временных рядов. Изучение теоретических основ и практических аспектов этого раздела предусмотрено в лабораторных работах, методические материалы по которым представлены в данном учебном пособии.

Для того чтобы показать практическую направленность теории ОЭИ, в лабораторных работах используются *реальные* экспериментальные данные, полученные в ходе НИР по изучению пассажиропотоков городского общественного автотранспорта.

Для решения задач профессиональной обработки данных успешно применяются пакеты прикладных программ (например, Statgraphics, Statistica, SPSS). Но выпускники бакалавриата в большинстве случаев будут работать на рядовых предприятиях, не располагающих этими пакетами. Поэтому мы сочли целесообразным научить студентов решать распространенные задачи ОЭИ с помощью программных средств массового применения – в частности, в среде табличного процессора Microsoft Excel (оставляя изучение пакетов с более широкими функциональными возможностями обучающимся в магистратуре).

В пособии представлены материалы по четырем лабораторным работам с различным уровнем сложности.

В лабораторной работе № 1 «**Первичная обработка экспериментальных данных**» рассматриваются возможности группы статистических функций табличного процессора Microsoft Excel и пакета «Анализ данных», включенного в Microsoft Excel в качестве надстройки. Изучаются практические методы расчета описательных статистик, построения гистограмм, сглаживания и группировки экспериментальных данных.

В разделе «Необходимые теоретические сведения» лабораторной работы № 2 «**Временные ряды. Предварительный анализ данных**» рассматриваются основные понятия теории временных рядов (трендовая модель; структурный анализ временного ряда; методы предварительной обработки временного ряда; способы диагностики наличия тренда). Теоретические знания используются для обработки реального временного ряда по следующим направлениям: выявление и устранение аномальных наблюдений по критерию Ирвина; диагностика наличия тренда методом проверки разности средних значений и методом Фостера-Стьюарта; сглаживание временного ряда различными методами.

В лабораторной работе № 3 «**Временные ряды. Выбор трендовой модели и оценка ее адекватности**» рассматриваются вопросы формирования набора трендовых моделей для последующего выбора адекватной модели; способы оценки параметров полиномиальных и экспоненциальных моделей; алгоритмы проверки адекватности моделей по критериям серий, пиков и по критерию проверки остаточной компоненты трендовой модели на соответствие нормальному закону распределения.

В лабораторной работе № 4 «**Вероятностный анализ динамической системы и его применение**» рассматривается возможность использования экспериментальных данных для экологической экспертизы проектов создания новых предприятий, в частности – для прогнозирования изменений экологической ситуации и показателей качества жизни в городе после ввода нового предприятия, эксплуатация которого приведет к дополнительному загрязнению атмосферы выбросами вредных веществ.

В приложениях приведены ряды наблюдений, которые должны быть обработаны в лабораторных работах, и описана структура листов рабочих книг Microsoft Excel для проведения расчетов.

Базой при подготовке теоретической части к лабораторным работам № 2, 3 послужили материалы из [2, 3], переработанные и дополненные нами в части примеров с использованием реальных экспериментальных данных, описания способов реализации расчетов базовыми средствами Microsoft Excel, а также надстройки «Анализ данных». Сценарии и методики проведения лабораторных работ № 1, 4 являются оригинальными. Все лабораторные работы опробованы в ходе проведения лабораторного практикума у студентов направления «Информатика и вычислительная техника» Камышинского технологического института (филиала) ВолгГТУ.

Авторы, разрабатывая электронные средства поддержки дисциплины «Обработка экспериментальной информации», учитывали трудности ее дистанционного изучения (в частности, студентами-заочниками). Поэтому презентации к лекциям, практическим занятиям и лабораторным работам выполнены с таким уровнем подробности, чтобы гарантировать возможность понимания материала студентами вне аудитории. Разработанные средства могут быть переданы безвозмездно всем заинтересованным читателям по запросу, направляемому по адресу elena_krushel@yandex.ru.

ЛАБОРАТОРНАЯ РАБОТА № 1. ПЕРВИЧНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Время выполнения работы: 2 аудиторных часа.

Цели работы:

- Изучить возможности среды Microsoft Excel по статистической обработке реальных данных.
- Провести расчет статистических характеристик экспериментальных данных (экспериментальные данные получены при обследовании пассажиропотока общественного муниципального автотранспорта).
- Построить различные гистограммы распределения экспериментальных данных.

Общее задание на лабораторную работу:

Для выполнения данной работы необходимо выполнить 4 этапа:

1. Рассчитать статистические характеристики экспериментальных данных.
2. Построить гистограммы распределения частот посадки пассажиров.
3. Выполнить сглаживание экспериментальных данных.
4. Выполнить группировку экспериментальных данных.

1.1. Введение. Средства статистической обработки в Microsoft Excel

Среда Microsoft Excel имеет встроенные средства по статистической обработке различных данных. В частности, в ней имеются (упорядочено по частоте практического использования):

- встроенные статистические функции;
- надстройка «Пакет анализа»;
- средства для «ручного» программирования процедур обработки данных.

Кратко рассмотрим эти средства.

Встроенные статистические функции

Список статистических функций и их описание можно посмотреть через «Мастер функций», выбрав в нем категорию «Ста-

тистические». «Мастер функций» открывается при нажатии на кнопку «fx» в строке формул (рис. 1.1).

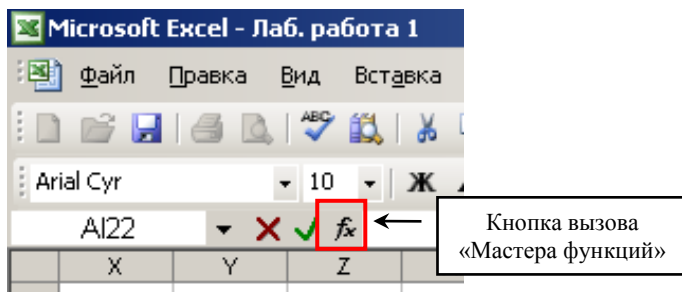


Рис. 1.1. Кнопка вызова «Мастера функций»

Функции, необходимые для выполнения лабораторной работы, изучаются студентами самостоятельно (описание функций можно найти, например, в [1], или в справочной системе Microsoft Excel).

Надстройка «Пакет анализа»

В состав Microsoft Excel входит набор средств анализа данных (в виде надстройки «Пакет анализа»), предназначенный для решения задач обработки данных, математической статистики и ее приложений.

Чтобы данный пакет был доступен, необходимо его подключить (по умолчанию «Пакет анализа» отключен). Подключение пакета в различных версиях Microsoft Excel производится по-разному из-за различий графического интерфейса программы Microsoft Excel.

Для подключения пакета в *Microsoft Excel 2003* следует выбрать позицию меню «Сервис» => «Надстройки...». Появится окно «Надстройки», в котором надо установить флажок «Пакет анализа» и нажать кнопку «ОК» (рис. 1.2).

После этого в главном меню «Сервис» появится позиция меню «Анализ данных...», при выборе которой открывается окно «Анализ данных» (рис. 1.3).

Замечание. Если надстройка «Пакет анализа» отсутствует в списке «Доступные надстройки:» (рис. 1.2), то для проведения ее поиска следует нажать кнопку «Обзор...». В случае появления сообщения о том, что «Пакет анализа» не установлен на компьютере и предложения установить его, следует согласиться и нажать кнопку «Да».

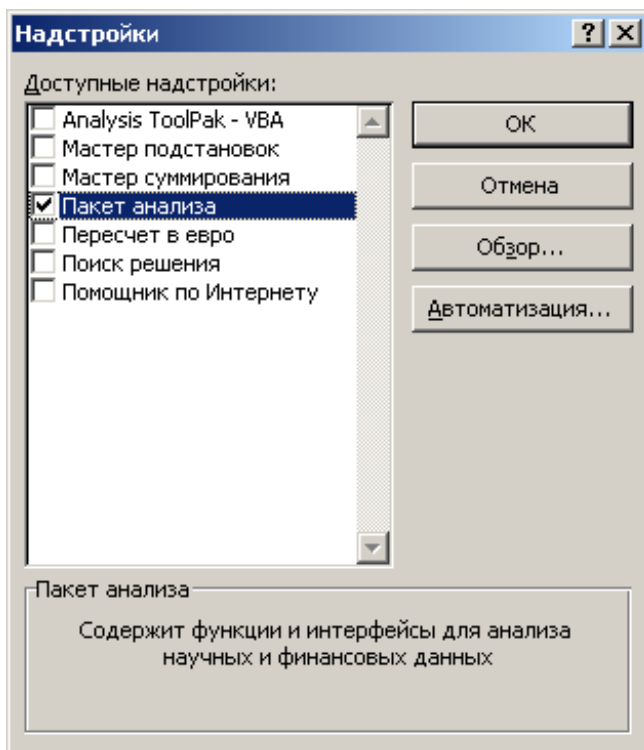


Рис. 1.2. Окно выбора надстройки «Пакет анализа»

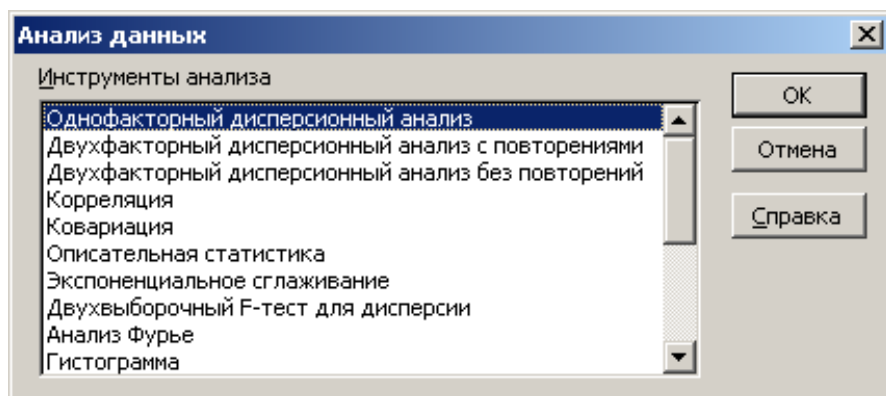


Рис. 1.3. Окно «Анализ данных»

Для подключения пакета в *Microsoft Excel 2007* следует нажать круглую кнопку «Microsoft Office» (в левом верхнем углу главного окна), а затем кнопку «Параметры Excel» (рис. 1.4).

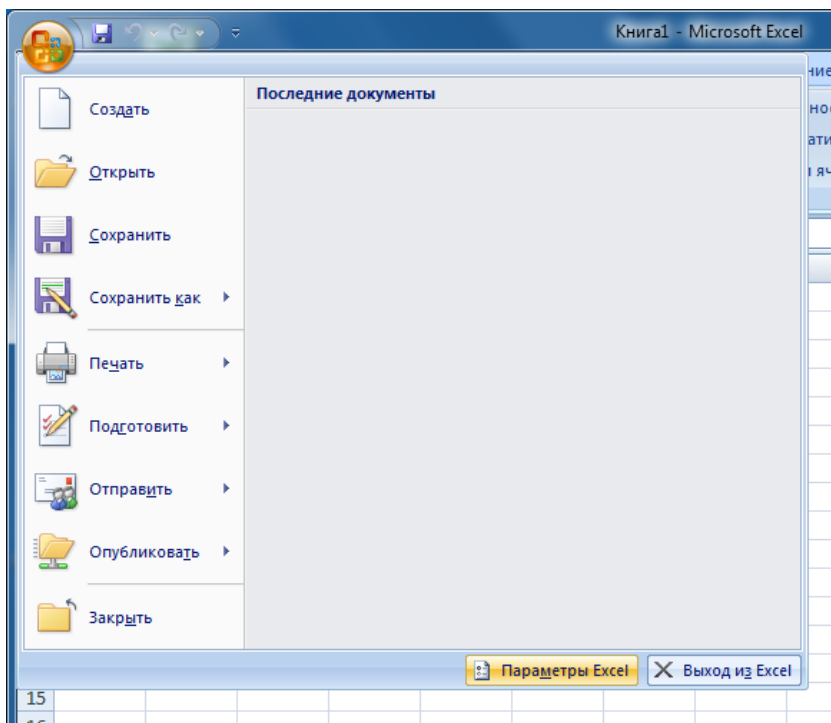


Рис. 1.4. Открытие параметров программы в Microsoft Excel 2007

В появившемся окне «Параметры Excel», в списке команд левой части окна выбрать команду «Надстройки», в списке «Управление» выбрать позицию «Надстройки Excel» и нажать кнопку «Перейти...» (рис 1.5). После этого появится окно «Надстройки», в котором надо установить флажок «Пакет анализа» и нажать кнопку «ОК» (рис. 1.2).

В результате подключения надстройки на вкладке «Данные» в группе «Анализ» станет доступна команда «Анализ данных». При выборе этой команды будет открываться окно «Анализ данных» (рис. 1.3).

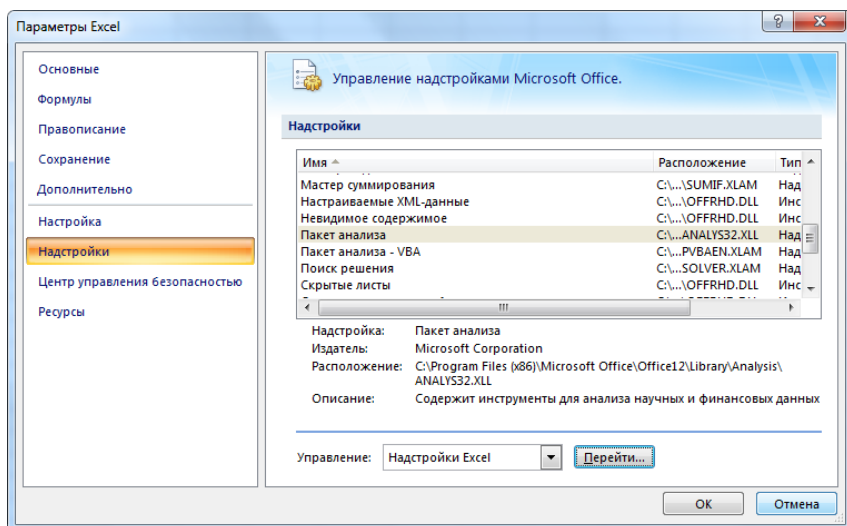


Рис. 1.5. Окно выбора надстроек в Microsoft Excel 2007 и старше

Для подключения пакета в *Microsoft Excel 2010* следует на вкладке «Файл» выбрать команду «Параметры» для открытия окна «Параметры Excel» (см. рис. 1.5). Дальнейшие действия по подключению надстройки аналогичны действиям, описанным для Microsoft Excel 2007.

Описания инструментов анализа, перечисленных в окне «Анализ данных», можно посмотреть в справочной системе Microsoft Excel. В ходе выполнения этой и последующих лабораторных работ некоторые из них будут изучены и использованы.

Средства для «ручного» программирования процедур обработки данных

В случаях, когда для обработки информации необходимо выполнить действия, которым нет аналога во встроенных функциях Microsoft Excel и надстройке «Пакет анализа», остается возможность реализовать их «вручную». При этом необходимые расчеты можно выполнить прямо на листе Microsoft Excel, используя встроенные функции для получения промежуточных результатов. Дру-

гую возможность предоставляет встроенный язык программирования Visual Basic for Application.

Задания

1. Просмотреть список встроенных статистических функций Microsoft Excel. Разобраться с назначением некоторых из них (не менее 10 функций).
2. Подключить надстройку «Пакет анализа».

1.2. Расчет статистических характеристик экспериментальных данных

Рассчитаем статистические характеристики экспериментальных данных о посадке пассажиров в автобусы на остановке общественного муниципального автотранспорта в течение каждой минуты периода наблюдений (с 6:00 до 21:00), находящихся в файле, структура которого описана в приложении 2. Для этого воспользуемся соответствующим инструментом из «Пакета анализа».

Откройте окно «Анализ данных» (позиция меню «Сервис» => «Анализ данных...»), выберите строку «Описательная статистика» и нажмите кнопку «ОК». Откроется окно «Описательная статистика». Назначение полей этого окна можно узнать в справочной системе Microsoft Excel.

В окне «Описательная статистика» следует задать следующие параметры (рис. 1.6):

- «Входной интервал» – диапазон ячеек с суммарным пассажиропотоком входящих пассажиров (столбец В);
- «Параметры вывода» – выбрать позицию «Выходной интервал» и указать ячейку (какую пожелаете), с которой будет вставлен блок рассчитанных статистических характеристик;
- установить флажок «Итоговая статистика»;
- установить флажок «Уровень надежности» со значением 95 %;
- установить флажок «К-ый наименьший» со значением 2;
- установить флажок «К-ый наибольший» со значением 2.

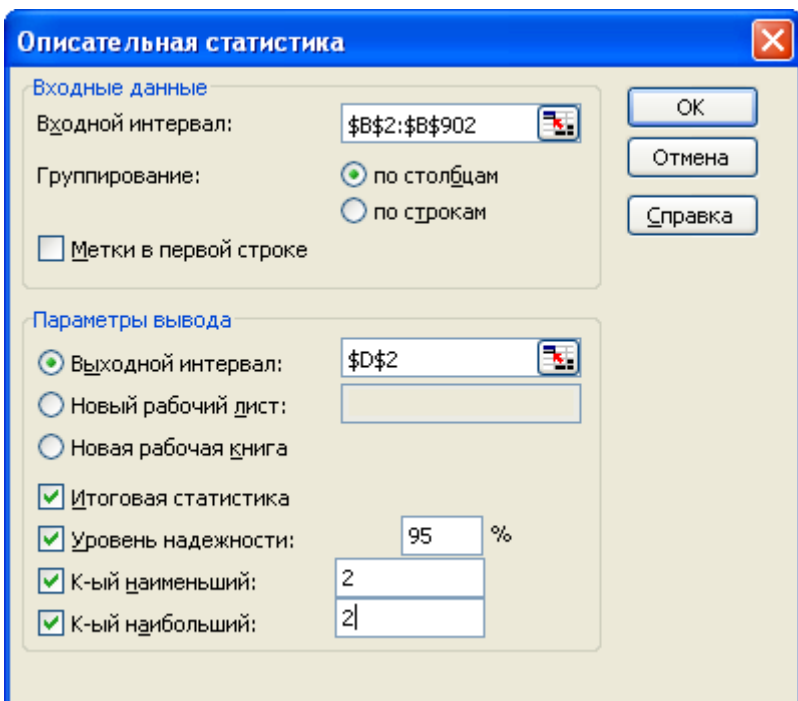


Рис. 1.6. Окно «Описательная статистика»

Нажмите кнопку «ОК». На текущем листе, начиная с ячейки из поля «Выходной интервал» появятся рассчитанные статистические параметры.

Задания

3. Разобраться со смыслом показателей, полученных с помощью инструмента «Описательная статистика».
4. К таблице показателей, полученных с помощью инструмента «Описательная статистика», добавить справа еще один столбец, в котором рассчитать те же самые показатели с использованием встроенных статистических функций Microsoft Excel. Проверить их совпадение (рис. 1.7).

Рекомендация

Соответствие статистических показателей, рассчитываемых инструментом «Описательная статистика», встроенным статистическим функциям Microsoft Excel приведено в табл. 1.1.

	A	B	C	D	E	F
1	Время	Суммарный пассажиропоток входящих пассажиров		Описательная статистика		С использованием формул
2	6:00	0		Столбец1		
3	6:01	0				
4	6:02	0		Среднее	1,873474	1,8734739
5	6:03	2		Стандартная ошибка	0,096195	0,0961946
6	6:04	0		Медиана	1	1
7	6:05	0		Мода	0	0
8	6:06	1		Стандартное отклонение	2,887439	2,8874395
9	6:07	0		Дисперсия выборки	8,337307	8,3373067
10	6:08	0		Экссесс	16,30832	16,308324
11	6:09	1		Асимметричность	3,025809	3,0258088
12	6:10	0		Интервал	31	31
13	6:11	0		Минимум	0	0
14	6:12	0		Максимум	31	31
15	6:13	0		Сумма	1688	1688
16	6:14	1		Счет	901	901
17	6:15	0		Наибольший(2)	19	19
18	6:16	0		Наименьший(2)	0	0
19	6:17	0		Уровень надежности(95,0%)	0,188792	0,1885379
20	6:18	0				

Рис. 1.7. Результаты расчета описательных статистик с использованием встроенных статистических функций Microsoft Excel

Замечание. Как видно из рис. 1.7, значения для уровня надежности (среднего), рассчитанные с помощью инструмента «Описательная статистика» и по формулам Microsoft Excel не совпадают (хотя отличие только в четвертом знаке после запятой). Дело в том, что в инструменте «Описательная статистика» при вычислении уровня значимости по формуле, приведенной в приложении 1, используется квантиль распределения Стьюдента, а во встроенной функции (ДОВЕРИТ(...)) используется квантиль стандартного нормального распределения. В данной формуле квантиль нормального распределения следует использовать, если известно точное значение дисперсии генеральной совокупности, а

квантиль Стьюдента – если точное значение дисперсии не известно, а вместо него используется значение оценки дисперсии, рассчитанное по значениям выборки. В данном примере точное значение дисперсии не известно, а значит правильнее использовать значение, полученное через инструмент «Описательная статистика».

Таблица 1.1

Соответствие статистических показателей встроенным функциям

Статистический показатель	Встроенная функция Microsoft Excel
Среднее	СРЗНАЧ(диапазон)
Стандартная ошибка (среднего)	Расчет по формуле из приложения 1
Медиана	МЕДИАНА(диапазон)
Мода	МОДА(диапазон)
Стандартное отклонение	СТАНДОТКЛОН(диапазон)
Дисперсия выборки	ДИСП(диапазон)
Эксцесс	ЭКСЦЕСС(диапазон)
Асимметричность	СКОС(диапазон)
Интервал	Расчет по формуле из приложения 1
Минимум	МИН(диапазон)
Максимум	МАКС(диапазон)
Сумма	СУММ(диапазон)
Счет	СЧЁТ(диапазон)
Наибольшее (k)	НАИБОЛЬШИЙ(диапазон; k)
Наименьшее (k)	НАИМЕНЬШИЙ(диапазон; k)
Уровень надежности	ДОВЕРИТ(альфа; станд_откл; размер) или по формуле из приложения 1

Задания

- По экспериментальным данным построить диаграмму зависимости количества входящих пассажиров от времени суток.
- (Задание повышенной сложности) Разработать программу для расчета тех же статистик, что и в п. 3 задания, но без использования встроенных статистических функций Microsoft Excel. Эти расчеты можно составить как в среде Visual Basic for Microsoft Excel, так и в любой другой среде программирования.

1.3. Построение гистограммы распределения частот посадки пассажиров

«Пакет анализа» позволяет быстро вычислить выборочные и интегральные частоты попадания данных в указанные интервалы (в терминологии Microsoft Excel – «карманы») значений, а также строить гистограммы распределения значений в выборке экспериментальных данных для каждого «кармана». Освоим эту возможность.

Построим гистограмму частот количества пассажиров, входящих в автобус (количество случаев, когда в автобус входило заданное количество пассажиров).

Для этого предварительно необходимо сформировать список интервалов значений количества пассажиров, входящих в автобус (границ интервалов, интервалы карманов). Выберем размер интервала, равный 1, то есть, определим частоту вхождения 0, 1, 2, ... пассажиров в автобус за 1 минуту. Из полученных ранее данных известно, что минимальное количество пассажиров, входящих в автобус (за 1 минуту), составляет 0, максимальное количество – 31. В свободных ячейках листа сформируйте список со значениями границ интервалов (от 0 до 31 с шагом 1).

Следует отметить, что для определения количества «карманов» гистограммы k рекомендуется использовать формулу Стерджесса:

$$k = \lceil \sqrt{1 + 3.32 \cdot \lg_{10}(n)} \rceil,$$

где n – число экспериментальных данных;

$\lceil \dots \rceil$ – операция округления вверх.

Откройте окно «Анализ данных» (позиция меню «Сервис» => «Анализ данных...»), выберите строку «Гистограмма» и нажмите кнопку «ОК». Откроется окно «Гистограмма». Назначение полей этого окна можно узнать из помощи по Microsoft Excel.

В окне «Гистограмма» следует задать следующие параметры (рис. 1.8):

– «Входной интервал» – диапазон ячеек с суммарным пассажиропотоком входящих пассажиров (столбец В);

- «Интервал карманов» – сформированный ранее диапазон ячеек со значениями границ интервалов;
- «Параметры вывода» – выбрать позицию «Выходной интервал» и указать ячейку (какую пожелаете), с которой будет вставлен блок рассчитанных частот;
- установить флажок «Интегральный процент»;
- установить флажок «Вывод графика».

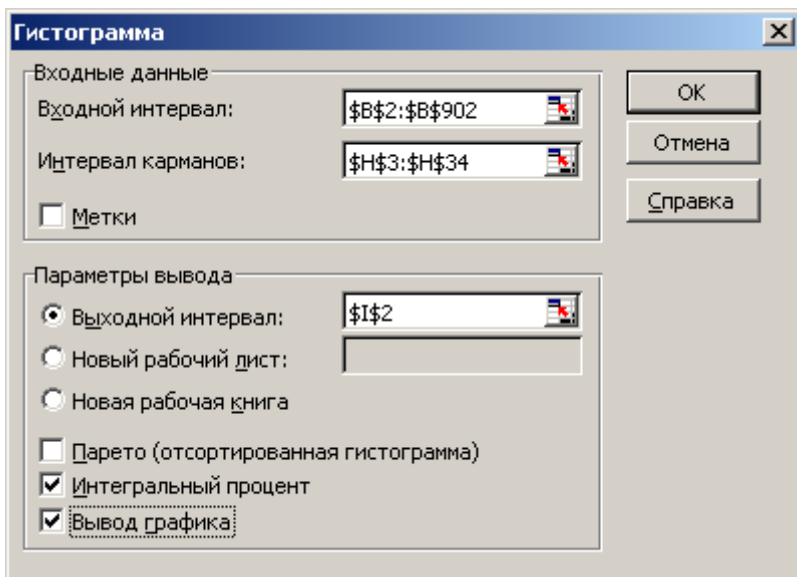


Рис. 1.8. Окно «Гистограмма»

Нажмите кнопку «ОК». На текущем листе, начиная с ячейки из поля «Выходной интервал» появятся рассчитанные выборочные и интегральные частоты посадки пассажиров в автобус, а также будет построена гистограмма частот.

Задания

7. Разобраться с процессом построения гистограммы частот с помощью инструмента «Гистограмма».
8. Построить таблицы частот и гистограммы для интервалов посадки пассажиров с шагом 3 и с шагом 5.

1.4. Сглаживание экспериментальных данных

Расчет скользящего среднего

Скользящее среднее используется для расчета значений в прогнозируемом периоде на основе среднего значения, рассчитанного по указанному числу предшествующих периодов. Скользящее среднее, в отличие от простого среднего для всей выборки, содержит сведения о тенденциях изменения данных. Определим зависимость среднего значения пассажиров, входящих в автобусы в минуту, используя предысторию продолжительностью 20 минут (в данном случае – по 20-ти одноминутным отсчетам).

Откройте окно «Анализ данных» (позиция меню «Сервис» => «Анализ данных...»), выберите строку «Скользящее среднее» и нажмите кнопку «ОК». Откроется окно «Скользящее среднее». Назначение полей этого окна можно узнать в справочной системе Microsoft Excel.

В окне «Скользящее среднее» следует задать следующие параметры (рис. 1.9):

- «Входной интервал» – диапазон ячеек с суммарным пассажиропотоком входящих пассажиров (столбец В);

- «Интервал» – значение 20 (минут);

- «Выходной интервал» – указать ячейку (какую пожелаете), с которой будет вставлен блок рассчитанных значений скользящего среднего;

- установить флажок «Вывод графика»;

- установить флажок «Стандартные погрешности».

Нажмите кнопку «ОК». На текущем листе, начиная с ячейки из поля «Выходной интервал» появятся рассчитанные значения скользящего среднего и стандартных погрешностей среднего, а также будет построена диаграмма скользящего среднего.

Сглаживание кривой на диаграмме

На диаграмме скользящего среднего для кривой экспериментальных данных построим сглаженную кривую (тренд), отражающую тенденцию изменений экспериментальных данных.

Для этого выделите на диаграмме кривую экспериментальных данных, вызовите контекстное меню (правой кнопкой мыши), выберите позицию «Добавить линию тренда...». В появившемся окне выберите вид тренда – «Линейная фильтрация» с количеством точек (поле «Точки») – 20 (рис. 1.10).

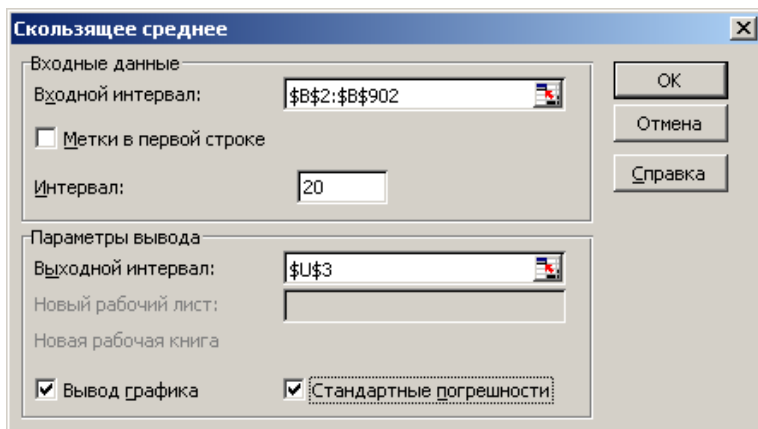


Рис. 1.9. Окно «Скользящее среднее»

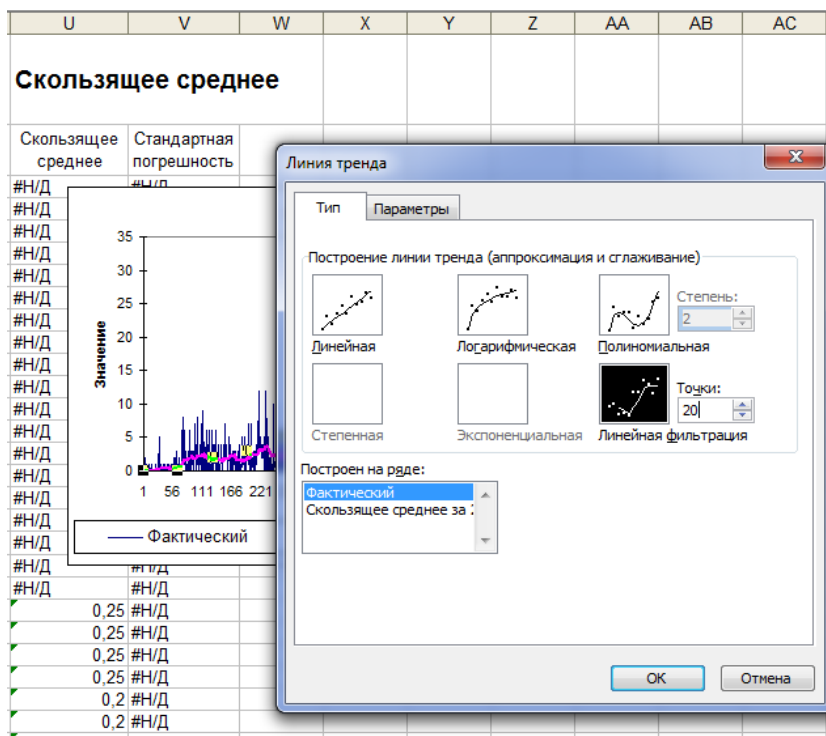


Рис. 1.10. Окно задания линии тренда для графика

Задания

9. Разобраться с процессом расчета скользящего среднего с помощью инструмента «Скользящее среднее». Почему для первых 20-ти значений скользящее среднее не рассчитано (имеет значения «#Н/Д»)?
10. Как соотносятся между собой графики скользящего среднего и тренда (для удобства сравнения растяните диаграмму и уберите линии с кривой экспериментальных данных)?

1.5. Группировка исходных данных

Предположим, что для анализа пассажиропотока необходимо получить распределение пассажиров, входящих в автобусы не по минутно, а за бóльшие интервалы времени (например, за 20 минут). Для этого необходимо сгруппировать (просуммировать) исходные экспериментальные данные. В качестве интервала группировки возьмем, например, 20 минут.

Средств автоматизированного группирования данных в среде Microsoft Excel нет. Поэтому обработку данных будем проводить «вручную», с использованием встроенных функций Microsoft Excel.

После группировки следует получить данные в виде таблицы и графика в виде, как представлено на рис. 1.11.

Для расчета числа вошедших пассажиров за 20 минут можно использовать встроенную функцию «СУММЕСЛИ» и «СУММ», в которой в качестве критерия указать выражение «<= & соответствующая_ячейка_интервала_времени». Конкретный вид формулы определите самостоятельно.

Для определения числа вошедших пассажиров за 1 минуту (третий столбец таблицы на рис. 1.11) следует второй столбец разделить на 20.



Рис. 1.11. Группировка исходных данных для 20-минутного интервала

Задание

11. Проведите группировку данных, как на рис. 1.11, для интервала времени 20 мин; 10 мин; 1 час.

1.6. Контрольные вопросы

1. Какие средства предоставляет Microsoft Excel для статистической обработки данных?
2. Как просмотреть список и использовать встроенные статистические функции Microsoft Excel?
3. Какую информацию предоставляет инструмент «Описательная статистика» надстройки «Анализ данных»?
4. Как можно построить гистограмму распределения частот некоторых данных?
5. Для чего используется инструмент «Скользящее среднее» надстройки «Анализа данных»?
6. Почему при использовании инструмента «Скользящее среднее» несколько первых значений равны «#Н/Д»?

ЛАБОРАТОРНАЯ РАБОТА № 2. ВРЕМЕННЫЕ РЯДЫ. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ

Время выполнения работы: 4 аудиторных часа.

Цели работы:

- Получить навыки по выявлению аномальных данных.
- Ознакомиться с простыми приемами диагностики тренда временного ряда.
- Изучить способы сглаживания временных рядов.

Общее задание на лабораторную работу:

Провести предварительный анализ экспериментальных данных для последующего построения модели тренда в лабораторной работе № 3.

Для выполнения данной работы необходимо выполнить 4 этапа:

1. Сгруппировать экспериментальные данные по 20-минутным интервалам.
2. Выявить и заменить аномальные данные в экспериментальных данных.
3. Определить наличие тренда у экспериментальных данных несколькими методами.
4. Сгладить полученный ряд различными методами сглаживания.

В качестве исходных экспериментальных данных используются данные из файла лабораторной работы № 1 (структура файла описана в приложении 2), содержащие значения (в виде временного ряда) о количестве пассажиров, входящих в автобус, на одной остановке пассажирского транспорта.

2.1. Понятие временных рядов

Обычно для представления динамических процессов, протекающих в технических, экономических и др. системах, используют последовательность (ряд) значений некоторого показателя, упорядоченного по времени. Подобный ряд значений показателя отражает ход развития изучаемого процесса. Последовательность наблюдений одного показателя (признака), упорядоченная в зависимости от последовательно возрастающих или убывающих значе-

ний другого показателя, называется *динамическим рядом*, или *рядом динамики*. Если в качестве признака, в зависимости от которого происходит упорядочивание, берется время, то такой динамический ряд называется *временным рядом* [2, 3].

Отдельные элементы рядов динамики, являющиеся значениями наблюдаемого показателя, называются *уровнями ряда*. Каждому уровню ряда соответствуют момент или интервал времени, к которым он относится. Временные ряды, в которых значения показателя относятся к определенным моментам времени, называются *моментными* (например, температура в печи при закалке стальных изделий, фиксируемая посекундно). Если уровни временного ряда образуются суммированием, усреднением или каким-либо другим методом агрегирования за некоторый промежуток времени, то такие ряды называют *интервальными* временными рядами (например, ряд наблюдений за количеством обслуживаемых покупателей в магазине по часам, ряд значений усредненной температуры окружающего воздуха в течение суток). Под длиной временного ряда понимают время, прошедшее от начального момента наблюдений до конечного, или число уровней ряда.

Если во временном ряду проявляется длительная закономерность изменения уровней, то говорят, что имеет место *тренд* (тенденция) ряда. Тренд характеризует общее направление развития рассматриваемого процесса. Если математическая модель развития изучаемой системы описывается с учетом тренда ее основных показателей, то такая модель называется *трендовой моделью*.

Для выявления наличия тренда временных рядов, определения параметров тренда, а также для анализа трендовых моделей используется аппарат теории вероятностей и математической статистики. Однако следует помнить, что этот аппарат предназначен для обработки простых статистических совокупностей, и поэтому применение методов теории вероятности и математической статистики требует определенных поправок. Временные ряды отличаются от простых статистических совокупностей тем, что уровни временного ряда упорядочены во времени и их перестановка недопустима, тогда как элементы статистической совокупности не являются упорядоченными и перестановка этих элементов не изменяет значений оценок статистических показателей – среднего зна-

чения, дисперсии и т. д. Кроме того, уровни временного ряда зависят друг от друга, а элементы статистической совокупности являются независимыми.

2.2. Структурный состав временного ряда

При рассмотрении временного ряда показателя Y_t , состоящего из n уровней ($Y_1, Y_2, Y_3, \dots, Y_n$), в его составе, в общем случае, можно выделить три структурных элемента:

- тренд;
- колебания (циклические, сезонные и др.);
- остаточная компонента.

Основным структурным элементом временного ряда является **тренд** U_t , характеризующий наличие общего направления изменения наблюдаемого показателя в течение продолжительного времени. Кроме тренда во временных рядах могут наблюдаться близкие к повторяющимся **колебания** V_t относительно основной тенденции (тренда). Колебания с относительно небольшим периодом (например, день, месяц, год), обусловленные влиянием природно-климатических условий на рассматриваемый показатель, обычно называются **сезонными колебаниями**. Колебания с достаточно большим периодом (несколько лет) называются **циклическими**. Сезонные колебания заметно проявляются в сельском хозяйстве, в добывающих отраслях, а также в потреблении энергоносителей. Под сезонностью понимают значимое различие в поведении рассматриваемой системы в течение регулярно повторяющихся временных периодов, обусловленных влиянием природных факторов или особенностей технологического процесса. Примером, иллюстрирующим сезонные колебания, являются графики расхода воды на полив в оросительной системе, в которых прослеживается влияние дневного и ночного времени. Большие циклические колебания обусловлены общими спадами и подъемами в рассматриваемом процессе. Примером, иллюстрирующим циклические колебания, является временной ряд интенсивности солнечной активности, имеющий повторяющиеся циклы с периодом около 11 лет.

Тренд и колебания (различного периода) называются **регулярными**, или **систематическими компонентами**. Они характеризуют основное изменение показателя временного ряда «условно

очищенное» от возможных случайных факторов. Если из временного ряда выделить и убрать систематические компоненты, то оставшаяся часть будет составлять *остаточную компоненту* e_t . Она является неотъемлемой (обязательной) частью любого временного ряда, так как такие процессы всегда сопровождаются небольшими изменениями, вызванными слабыми влияниями большого числа случайных (или неучтенных неслучайных) факторов. Если регулярная компонента временного ряда выделена правильно, что является целью при построении трендовой модели, то остаточная компонента будет обладать следующими свойствами [3]:

- случайностью изменения своих значений;
- соответствием нормальному закону распределения;
- равенством нулю математического ожидания;
- независимостью значений уровней друг от друга, т. е. отсутствием существенной автокорреляции.

Проверка наличия этих четырех свойств у остаточной последовательности является основой проверки на адекватность трендовой модели.

В зависимости от вида связи между компонентами временного ряда может быть построена *аддитивная* модель, имеющая вид:

$$Y_t = U_t + V_t + e_t,$$

или, если компоненты ряда умножаются, то получаем *мультипликативную* модель:

$$Y_t = U_t \times V_t \times e_t.$$

Существует также и *смешанная* модель вида:

$$Y_t = U_t \times V_t + e_t.$$

В большинстве случаев при анализе временных рядов наличием колебательной компоненты V_t пренебрегают, а если требуется ее учитывать, то для ее выделения применяют специальные методы, основанные на спектральном анализе.

Изучение соотношения между закономерностью и случайностью в формировании значений уровней ряда, оценка количественной меры их влияния есть *основная цель статистического анализа временных рядов*. Трендовая модель, построенная по значениям показателя в прошлом, используется для прогнозирования значений этого показателя в будущем, а учет случайности позво-

ляет определить степень отклонения будущих значений от выявленного закономерного развития.

2.3. Этапы построения прогноза по временным рядам

Экстраполяционное прогнозирование процессов (прогнозирование «на будущее» при условии сохранения поведения «как в прошлом»), представленных одномерными временными рядами, сводится к выполнению следующих основных этапов [2]:

- 1) предварительный анализ данных;
- 2) построение моделей: формирование набора аппроксимирующих функций (кривых роста) и численное оценивание параметров этих моделей;
- 3) проверка адекватности моделей и оценка их точности;
- 4) выбор лучшей модели.

В данной лабораторной работе рассматривается первый из этапов. Остальные этапы рассмотрены в лабораторной работе № 3.

Задание

1. Подготовка экспериментальных данных. Проведите группировку временного ряда для интервала 20 минут (см. лабораторную работу № 1).

2.4. Предварительный анализ данных временного ряда

К процедурам предварительного анализа относятся:

- выявление аномальных наблюдений;
- проверка наличия тренда;
- сглаживание временного ряда.

2.4.1. Выявление аномальных наблюдений

Аномальными уровнями считаются такие значения временного ряда, которые «сильно выбиваются» из общей тенденции. Их наличие существенно искажает основные статистические характеристики ряда, в том числе и соответствующую трендовую модель. Причинами аномальных значений часто являются ошибки технического характера (сбой при фиксации и передаче данных, ошибки при вычислениях).

Так как наличие аномальных наблюдений приводит к искажению результатов анализа данных, то необходимо убедиться в отсут-

ствии аномалий. Для выявления аномальных значений ряда разработаны различные критерии, например, критерий Ирвина [4, 5].

Согласно критерию Ирвина, аномальной считается точка Y_t , отстоящая от предыдущей точки Y_{t-1} на величину, большую среднеквадратичного отклонения. Для всех точек временного ряда рассчитывается **критерий Ирвина**:

$$\lambda_t = \frac{|Y_t - Y_{t-1}|}{\sigma},$$

где σ – оценка среднеквадратичного отклонения временного ряда,

$$\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{t=1}^n (Y_t - \bar{Y})^2};$$

\bar{Y} – среднее значение временного ряда, $\bar{Y} = \frac{1}{n} \cdot \sum_{t=1}^n Y_t$.

Точка ряда Y_t считается аномальной, если выполняется условие $\lambda_t > \lambda_{таб}$. Табличные значения $\lambda_{таб}$ уменьшаются с ростом длины ряда, их значения приведены в табл. 2.1 для доверительных вероятностей p равных 0.95 и 0.99.

Таблица 2.1

Критические значения параметра $\lambda_{таб}$ [3]

n		10	20	30	40	50	100	200	500	1000
$\lambda_{таб}$	$p = 0.95$	1.46	1.27	1.20	1.15	1.11	1.02	0.95	0.87	0.83
	$p = 0.99$	2.03	1.80	1.70	1.63	1.60	1.47	1.38	1.28	1.22

Задание

2. Выявление аномальных данных. После группировки проверьте полученный временной ряд на аномальные значения по критерию Ирвина. Обнаруженные аномальные значения замените путем интерполирования (усреднения) по соседним точкам (рис. 2.1, 2.2).

Рекомендации:

1. После однократной замены аномальных точек временной ряд все равно может содержать другие аномальные точки. Поэтому процедуру обнаружения и замены аномалий необходимо повторять циклически, до тех пор, пока аномальные точки не перестанут обнаруживаться.

2. Так как аномальные точки могут располагаться подряд (блоком), а их замена осуществляется путем усреднения их соседей, то одновременно заменять аномальные точки во всем блоке нецелесообразно. За одну итерацию «проверки-замены» следует заменять только по одной точке в каждом блоке. Заменяемая точка выбирается по максимальному значению критерия Ирвина в пределах данного блока.

	H	I	J	K	L	M	N	O	P	Q	R	S
1	Группировка по 20 мин		Критерий Ирвина 1	Аномалии 1			Без аномалий 1	Критерий Ирвина 2	Аномалии 2			Без аномалий 2
2	6:00	5			кол точек	46	5					5
3	6:20	9	0,167295	0	крит. знач.	1,12	9	0,165885	0			9
4	6:40	9	0	0	СКО	23,91	9	0	0	СКО	24,11	9
5	7:00	34	1,045592	0	кол. аномалий	6	34	1,036784	0	кол. аномалий	4	34
6	7:20	43	0,376413	0			43	0,373242	0			43
7	7:40	37	0,250942	0			37	0,248828	0			37
8	8:00	38	0,041824	0			38	0,041471	0			38
9	8:20	40	0,083647	0			40	0,082943	0			40
10	8:40	32	0,33459	0			32	0,331771	0			32
11	9:00	44	0,501884	0			44	0,497656	0			44
12	9:20	61	0,711003	0			61	0,705013	0			61
13	9:40	47	0,585532	0			47	0,580599	0			47
14	10:00	49	0,083647	0			49	0,082943	0			49
15	10:20	54	0,209118	0			54	0,207357	0			54
16	10:40	63	0,376413	0			63	0,373242	0			63
17	11:00	51	0,501884	0			51	0,497656	0			51
18	11:20	69	0,752826	0			69	0,746484	0			69
19	11:40	114	1,882066	1			114	1,866211	1			75,5
20	12:00	61	2,216655	1			82	1,327083	1			82
21	12:20	50	0,460061	0			50	1,327083	1			50
22	12:40	43	0,292766	0			43	0,290299	0			43

Рис. 2.1. Выявление и замена аномальных данных по критерию Ирвина

2.4.2. Проверка наличия тренда

Для определения наличия тренда временного ряда используются различные методы. В данной лабораторной работе рассмотрен метод проверки разностей средних уровней [3, 4] и метод Фостера-Стьюарта [4, 5].

Метод проверки разностей средних уровней

Согласно этому методу временной ряд разбивается на две примерно равные по числу уровней части n_1 и n_2 ($n = n_1 + n_2$), каждая из которых рассматривается как некоторая самостоятельная выборочная совокупность, имеющая нормальное распределение.

	Н	И	Критерий Ирвина 1	Аномалии 1	Л	М	Н
1		Группировка по 20 мин					Без аномалий 1
2	6.00	5			кол точек	=СЧЁТ(И2:И47)	=И2
3	6.20	9=ABS(И3-И2)/\$M\$4	=ЕСЛИ(И3>\$M\$3;1;0)	крит. знач.	1,12		=И3
4	6.40	9=ABS(И4-И3)/\$M\$4	=ЕСЛИ(И4>\$M\$3;1;0)	СКО	=СТАНДОТКЛОН(И2:И47)		=И4
5	7.00	34=ABS(И5-И4)/\$M\$4	=ЕСЛИ(И5>\$M\$3;1;0)	кол. аномалий	=СУММ(К3:К47)		=И5
6	7.20	43=ABS(И6-И5)/\$M\$4	=ЕСЛИ(И6>\$M\$3;1;0)				=И6
7	7.40	37=ABS(И7-И6)/\$M\$4	=ЕСЛИ(И7>\$M\$3;1;0)				=И7
8	8.00	38=ABS(И8-И7)/\$M\$4	=ЕСЛИ(И8>\$M\$3;1;0)				=И8
9	8.20	40=ABS(И9-И8)/\$M\$4	=ЕСЛИ(И9>\$M\$3;1;0)				=И9
10	8.40	32=ABS(И10-И9)/\$M\$4	=ЕСЛИ(И10>\$M\$3;1;0)				=И10
11	9.00	44=ABS(И11-И10)/\$M\$4	=ЕСЛИ(И11>\$M\$3;1;0)				=И11
12	9.20	61=ABS(И12-И11)/\$M\$4	=ЕСЛИ(И12>\$M\$3;1;0)				=И12
13	9.40	47=ABS(И13-И12)/\$M\$4	=ЕСЛИ(И13>\$M\$3;1;0)				=И13
14	10.00	49=ABS(И14-И13)/\$M\$4	=ЕСЛИ(И14>\$M\$3;1;0)				=И14
15	10.20	54=ABS(И15-И14)/\$M\$4	=ЕСЛИ(И15>\$M\$3;1;0)				=И15
16	10.40	63=ABS(И16-И15)/\$M\$4	=ЕСЛИ(И16>\$M\$3;1;0)				=И16
17	11.00	51=ABS(И17-И16)/\$M\$4	=ЕСЛИ(И17>\$M\$3;1;0)				=И17
18	11.20	69=ABS(И18-И17)/\$M\$4	=ЕСЛИ(И18>\$M\$3;1;0)				=И18
19	11.40	114=ABS(И19-И18)/\$M\$4	=ЕСЛИ(И19>\$M\$3;1;0)				=И19
20	12.00	61=ABS(И20-И19)/\$M\$4	=ЕСЛИ(И20>\$M\$3;1;0)				=И19+И21)/2
21	12.20	50=ABS(И21-И20)/\$M\$4	=ЕСЛИ(И21>\$M\$3;1;0)				=И21
22	12.40	43=ABS(И22-И21)/\$M\$4	=ЕСЛИ(И22>\$M\$3;1;0)				=И22

Рис. 2.2. Выявление и замена аномальных данных по критерию Ирвина (режим отображения формул)

Если временной ряд имеет тенденцию к тренду, то средние, вычисленные для каждой совокупности, должны существенно (значимо) различаться между собой. Если же расхождение несущественно (случайно), то временной ряд не имеет тенденции к тренду. Таким образом, проверка наличия тренда в исследуемом ряду сводится к проверке гипотезы о равенстве средних двух нормально распределенных совокупностей, проверяемой с помощью t-критерия Стьюдента. Заметим, что метод применим только для случая, когда обе части выборок имеют одинаковую дисперсию, поэтому рассматриваемый метод диагностики наличия тренда нужно предварить проверкой гипотезы о равенстве дисперсий, для чего используется критерий Фишера.

Для каждой из частей ряда вычисляются среднее значение и оценка дисперсии:

$$\bar{Y}_1 = \frac{\sum_{t=1}^{n_1} Y_t}{n_1}; \quad S_1 = \frac{\sum_{t=1}^{n_1} (Y_t - \bar{Y}_1)^2}{n_1 - 1}; \quad \bar{Y}_2 = \frac{\sum_{t=1}^{n_2} Y_t}{n_2}; \quad S_2 = \frac{\sum_{t=1}^{n_2} (Y_t - \bar{Y}_2)^2}{n_2 - 1}.$$

Далее, рассчитывается F-критерий Фишера:

$$F = \begin{cases} S_1/S_2, & \text{если } S_1 > S_2 \\ S_2/S_1, & \text{если } S_2 > S_1 \end{cases}.$$

Если полученное значение F меньше табличного значения $F_{таб}$ (со степенями свободы $n_i - 1$ и $n_k - 1$, где i – индекс части ряда с большей дисперсией, k – индекс части с меньшей дисперсией), то гипотеза об однородности дисперсий не отвергается (принимается) и переходят к следующему этапу расчета. Если $F \geq F_{таб}$, то гипотеза об однородности дисперсий отклоняется и метод не дает ответа на вопрос о наличии или отсутствии тренда.

Окончательная проверка гипотезы об отсутствии тренда (точнее, о равенстве средних значений двух выборок) производится с использованием t -критерия Стьюдента, вычисляемого по формуле:

$$t = \frac{|\bar{Y}_1 - \bar{Y}_2|}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

где σ – среднеквадратичное отклонение разности средних:

$$\sigma = \sqrt{\frac{(n_1 - 1) \cdot S_1 + (n_2 - 1) \cdot S_2}{n_1 + n_2 - 2}}.$$

Если расчетное значение t меньше табличного значения $t_{таб}$, то гипотеза о равенстве средних значений не отвергается (принимается), т. е. тренда нет, в противном случае – тренд есть. Для определения табличного значения $t_{таб}$ число степеней свободы принимается равным $n_1 + n_2 - 2$.

Задания

3. Определение наличия тренда.

3.1. После удаления аномальных данных проверьте полученный временной ряд на наличие тренда методом проверки разностей средних. Сделайте вывод о наличии или отсутствии тренда временного ряда (рис. 2.3).

Рекомендации:

1. Для расчета $F_{таб}$ критерия Фишера можно использовать встроенную Excel-функцию `ФРАСПОВР(вероятность; степени_свободы1; степени_свободы2)`.

2. Для расчета $t_{таб}$ критерия Стьюдента можно использовать встроенную Excel-функцию СТЬЮДРАСПОБР (вероятность ; степени_свободы).

	A	B	C	D	E	F	G	H	I	J	
	Время	Исходные данные без аномалий		Метод проверки разностей средних							
2	6:00	5		Объем 1	23		Объем 2	23			
3	6:20	9		от	6:00		от	13:40			
4	6:40	9		до	13:20		до	21:00			
5	7:00	26		Среднее1	45,2173913		Среднее2	26,021739			
6	7:20	43		Дисперсия1	403,0187747		Дисперсия2	415,26087			
7	7:40	37		Гипотеза об однородности дисперсии - критерий Фишера							
8	8:00	38		Критерий	1,030375991		=Н6/Е6				
9	8:20	40		F табл	2,047770309		=ФРАСПОБР(0,05;Н2-1;Е2-1)				
10	8:40	32		Вывод	дисперсии одинаковы, проверять дальше						
11	9:00	44		Гипотеза об отсутствии тренда - t-критерий Стьюдента							
12	9:20	61									
13	9:40	47		СКО разности средних	20,227205		=КОРЕНЬ((Е2-1)*Е6+(Н2-1)*Н6)/(Е2+Н2-2)				
14	10:00	49		t-критерий		3,218221383		=ABS(Е5-Н5)/(F13*КОРЕНЬ(1/Е2+1/Н2))			
15	10:20	54		t табличный		2,015367547		=СТЬЮДРАСПОБР(0,05;Е2+Н2-2)			
16	10:40	63		Вывод	тренд есть						
17	11:00	51									

Рис. 2.3. Проверка наличия тренда методом разностей средних уровней

Задание

3.2. Проверьте еще раз временной ряд на наличие тренда методом проверки разностей средних с использованием инструментов из пакета «Анализ данных» – «Двухвыборочный F-тест для дисперсии» и «Двухвыборочный t-тест с одинаковыми дисперсиями» (рис. 2.4, 2.5).

Замечание. Режим работы инструмента «Двухвыборочный F-тест для дисперсий» служит для проверки гипотезы H_0 о равенстве дисперсий двух нормальных распределений ($D_1 = D_2$). При этом, в качестве альтернативной рассматривается гипотеза $H_1: D_1 < D_2$, если $S_1 < S_2$; или гипотеза $H_1: D_1 > D_2$, если $S_1 > S_2$. Из-за этого условие принятия гипотезы H_0 отличается:

- если дисперсия первой части больше дисперсии второй части ($S_1 > S_2$), то гипотеза H_0 не отвергается при условии $F < F_{кр}$;
- если дисперсия первой части меньше дисперсии второй части ($S_1 < S_2$), то гипотеза H_0 не отвергается при условии $F > F_{кр}$.

На рис. 2.4 дисперсия первой части (403,01) меньше дисперсии второй части (420,33), поэтому условием не отвержения гипотезы H_0 является $F > F_{кр}$ ($0,9588 > 0,4883$).

	A	B	C	D	E	F	G	H	I	J
5	7:00	26		Среднее1	45,2173913		Среднее2	26,021739		
6	7:20	43		Дисперсия1	403,0187747		Дисперсия2	415,26087		
7	7:40	37		Гипотеза об однородности дисперсии - критерий Фишера						
8	8:00	38		Критерий	1,030375991		=H5/E6			
9	8:20	40		F табл	2,047770309		=FПАСЛОБР(0,05,H2-1,E2-1)			
10	8:40	32		Вывод	дисперсии одинаковы, проверять дальше					
11	9:00	44								
12	9:20	61		Гипотеза об отсутствии тренда - t-критерий Стьюдента						
13	9:40	47		СКО разности средних	20,227205		=КОРЕНЬ(((E2-1)*E6+(H2-1)*H6)/(E2+H2-2))			
14	10:00	49		t-критерий	3,218221383		=ABS(E6-H6)/(F13*КОРЕНЬ(1/E2+1/H2))			
15	10:20	54		t табличный	2,015367547		=СТЪЮДПАСЛОБР(0,05,E2+H2-2)			
16	10:40	63		Вывод	тренд есть					
17	11:00	51								
18	11:20	69		Проверка посредством "Пакета анализа"						
19	11:40	75,5		Двухвыборочный F-тест для дисперсии						
20	12:00	82								
21	12:20	62,5		<i>Переменная 1</i>		<i>Переменная 2</i>				
22	12:40	43		Среднее	45,2173913		27,91304348			
23	13:00	55		Дисперсия	403,0187747		420,333004			
24	13:20	45		Наблюдения	23		23			
25	13:40	44		df	22		22			
26	14:00	58,5		F	0,958808304					
27	14:20	73		P(F<=f) однос	0,461152257					
28	14:40	60,5		F критическое	0,488336019		Вывод	проверять дальше		
29	15:00	48								
30	15:20	37								
31	15:40	27								
32	16:00	30								
33	16:20	25								
34	16:40	37								
35	17:00	24,5								
36	17:20	12								
37	17:40	17								
38	18:00	27								
39	18:20	25								
40	18:40	9								
41	19:00	16								
42	19:20	17								
43	19:40	4								

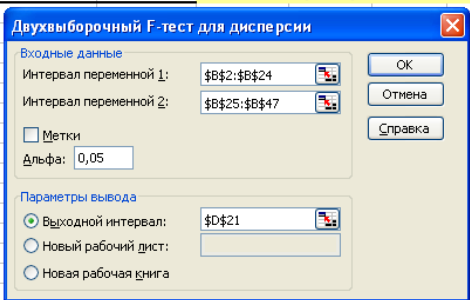


Рис. 2.4. Задание параметров для инструмента «Двухвыборочный F-тест для дисперсии»

Метод Фостера-Стьюарта

Этот метод дает более надежные результаты по сравнению с предыдущим [3, 4]. Метод позволяет установить наличие как самого тренда, так и наличие тренда дисперсии: при отсутствии тренда дисперсии разброс уровней ряда постоянен, при наличии тренда дисперсии дисперсия увеличивается или уменьшается.

Согласно методу выполняется сравнение каждого уровня ряда с предыдущим и формируются две последовательности:

C	D	E	F	G	H	I	J	K	L
12		Гипотеза об отсутствии тренда - t-критерий Стьюдента							
13	СКО разности средних		20,28979767	=КОРЕНЬ((E2-1)*E6+(H2-1)*H6)/(E2+H2-2)					
14	t-критерий		2,892187469	=ABS(E5-H5)/(F13*КОРЕНЬ(1/E2+1/H2))					
15	t табличный		2,015367547	=СТЮДРАСПОБР(0,05,E2+H2-2)					
16	Вывод	тренд есть							
17									
18	Проверка посредством "Пакета анализа"								
19	Двухвыборочный F-тест для дисперсии								
20									
21		Переменная 1	Переменная 2						
22	Среднее	45,2173913	27,91304348						
23	Дисперсия	403,0187747	420,333004						
24	Наблюдения	23	23						
25	df	22	22						
26	F	0,958808304							
27	P(F<=f) однос	0,461152257							
28	F критическое	0,488336019							
29									
30									
31	Двухвыборочный t-тест с одинаковыми дисперсиями								
32		Переменная 1	Переменная 2						
33	Среднее	45,2173913	27,91304348						
34	Дисперсия	403,0187747	420,333004						
35	Наблюдения	23	23						
36	Объединенная	411,6758893							
37	Гипотетическое	0							
38	df	44							
39	t-статистика	2,892187469							
40	P(T<=t) однос	0,002962989							
41	t критическое	1,680229977							
42	P(T<=t) двухс	0,005925978							
43	t критическое	2,015367547							
44	Вывод	тренд есть							
45									

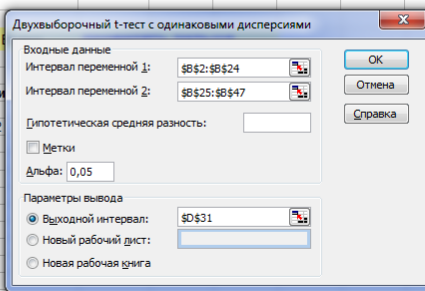


Рис. 2.5. Задание параметров для инструмента «Двухвыборочный t-тест с одинаковыми дисперсиями»

$$k_t = \begin{cases} 1, & \text{если } Y_t \text{ больше всех предыдущих уровней} \\ 0, & \text{в противном случае;} \end{cases}$$

$$l_t = \begin{cases} 1, & \text{если } Y_t \text{ меньше всех предыдущих уровней} \\ 0, & \text{в противном случае;} \end{cases}$$

$$t = 2, 3, \dots, n.$$

Вычисляются величины s и d , характеризующие изменения среднего значения и дисперсии временного ряда:

$$s = \sum_{t=2}^n (k_t + l_t), \quad d = \sum_{t=2}^n (k_t - l_t).$$

Величина s характеризует изменение дисперсии временного ряда, она может принимать значение от 0 (когда все уровни ряда равны) до $n - 1$ (ряд монотонно изменяется). Величина d характеризует из-

менение среднего значения временного ряда и изменяется от $-(n-1)$ (когда ряд монотонно убывает) до $+(n-1)$ (когда ряд монотонно возрастает). Величины s и d являются случайными с математическим ожиданием равным μ для значения s и равным 0 для значения d .

Далее проверяются гипотезы о случайности отклонения величин s и d от их математических ожиданий с помощью критерия Стьюдента:

$$t_s = \frac{|s - \mu|}{\sigma_1}, \quad \sigma_1 = \sqrt{2 \cdot \ln(n) - 3.4253},$$

$$t_d = \frac{|d - 0|}{\sigma_2}, \quad \sigma_2 = \sqrt{2 \cdot \ln(n) - 0.8456}, \quad \mu = \sigma_2^2,$$

где μ – оценка математического ожидания величины s для случайного временного ряда;

σ_1 – оценка среднеквадратического отклонения s для случайного временного ряда;

σ_2 – оценка среднеквадратического отклонения d для случайного временного ряда.

Формулы для σ_1 и σ_2 применимы при $n \geq 50$. Значения σ_1 и σ_2 при $n < 50$ приведены в табл. 2.2.

Таблица 2.2

Значения σ_1 и σ_2 для критерия Фостера-Стьюарта [5]

n	10	15	20	25	30	35	40	45	50
σ_1	1,288	1,512	1,677	1,791	1,882	1,956	2,019	2,072	2,121
σ_2	1,964	2,153	2,279	2,373	2,447	2,509	2,561	2,606	2,645

Полученные значения t_s и t_d необходимо сравнить с табличными значениями критерия Стьюдента $t_{\text{таб}}$ с n степенями свободы. Если $t_{\text{таб}}$ больше расчетного значения, то соответствующий тренд отсутствует: например, если $t_s > t_{\text{таб}}$, а $t_d < t_{\text{таб}}$, то тренд дисперсии есть, а тренда ряда нет.

Задание

3.3. Проверьте временной ряд без аномалий на наличие тренда методом Фостера-Стьюарта. Сделайте вывод о наличии или отсутствии тренда временного ряда (рис. 2.6).

	L	M	N	O	P	Q	R	S	T	U	
1	Время	Исходные данные без аномалий	Метод Фостера-Стьюарта								
2	6:00	5	k	l	k-l			s (дисперсия)	d (среднее)		
3	6:20	9	1	0	1			12	6		
4	6:40	9	0	0	0	СКО		2,057178357	2,609920074		
5	7:00	26	1	0	1	мат ожидание		6,811682793			
6	7:20	43	1	0	1						
7	7:40	37	0	0	0	<i>Гипотеза о случайности отклонений s и d от их мат. ожиданий</i>					
8	8:00	38	0	0	0						
9	8:20	40	0	0	0	t-критерий		2,522055119	2,298920974		
10	8:40	32	0	0	0	t табл		2,014103359	2,014103359		
11	9:00	44	1	0	1	Вывод		есть тренд	есть тренд		
12	9:20	61	1	0	1						
13	9:40	47	0	0	0						
14	10:00	49	0	0	0						
15	10:20	54	0	0	0						
16	10:40	63	1	0	1						

Рис. 2.6. Проверка наличия тренда методом Фостера-Стьюарта

2.4.3. Сглаживание временного ряда

Довольно часто уровни временного ряда колеблются, так что тенденция развития процесса (регулярная составляющая) скрыта случайными отклонениями. Сглаживание временного ряда позволяет отфильтровать мелкие случайные колебания и выявить основную тенденцию изменения исследуемой величины. При механическом сглаживании выравнивание отдельных уровней производится с использованием значений соседних уровней. Для сглаживания используются следующие методы [3, 5]:

- **Простая (среднеарифметическая) скользящая средняя:**

$$\tilde{Y}_t = \frac{\sum_{i=t-p}^{t+p} Y_i}{2 \cdot p + 1}, \quad p < t < n - p.$$

Сглаженное значение \tilde{Y}_t является среднеарифметическим из $2 \cdot p + 1$ соседних точек. Наиболее часто используется сглаживание по 5 точкам:

$$\tilde{Y}_t = \frac{Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2}}{5}.$$

- **Взвешенная (средневзвешенная) скользящая средняя:**

$$\tilde{Y}_t = \frac{\sum_{i=t-p}^{t+p} \alpha_i \cdot Y_i}{\sum_{i=t-p}^{t+p} \alpha_i}, \quad p < t < n - p.$$

В этом методе каждая из точек входит в общую сумму с весовым коэффициентом α_i . Для сглаживания по 5 точкам используют весовые коэффициенты $(-3, 12, 17, 12, -3)$. Для сглаживания по 7 точкам используются коэффициенты $(-2, 3, 6, 7, 6, 3, -2)$ или $(5, -30, 75, 131, 75, -30, 5)$.

- **Экспоненциальное сглаживание**

В этом методе для сглаживания текущей точки используются все предшествующие точки, причем значения весовых коэффициентов убывают по экспоненте по мере удаления от текущей точки. Формулу экспоненциального сглаживания можно записать в виде выражения, в котором текущая точка зависит от всех предыдущих точек:

$$\tilde{Y}_t = \frac{\sum_{i=1}^t \alpha_i \cdot Y_i}{\sum_{i=1}^t \alpha_i}.$$

Но в таком виде формула неудобна для использования, поскольку для каждой точки необходим свой набор весовых коэффициентов. Используя рекуррентные соотношения, можно получить выражение для текущей сглаженной точки как функцию от текущей несглаженной точки и предыдущей сглаженной:

$$\tilde{Y}_t = \rho \cdot Y_t + (1 - \rho) \cdot \tilde{Y}_{t-1}, \quad 0 < \rho < 1,$$

где ρ – параметр сглаживания.

Фиктивное начальное значение сглаженного ряда (\tilde{Y}_1) принимают равным первой точке или среднеарифметическому первых трех точек:

$$\tilde{Y}_1 = Y_1, \text{ или } \tilde{Y}_1 = (Y_1 + Y_2 + Y_3) / 3.$$

▪ **Среднехронологическая средняя:**

$$\tilde{Y}_t = \frac{\frac{Y_{t-T/2}}{2} + \sum_{i=t-T/2+1}^{t+T/2-1} Y_i + \frac{Y_{t+T/2}}{2}}{T}, \quad \frac{T}{2} < t < n - \frac{T}{2}.$$

Эта формула используется для моментных временных рядов. Обычно период сглаживания принимают равным одному году, т. е. $T = 4$ квартала или $T = 12$ месяцев.

При сглаживании временного ряда по $2 \cdot p + 1$ соседним точкам (методом простой скользящей средней, взвешенной скользящей средней, среднехронологическим) p точек в начале и в конце ряда остаются несглаженными. Эти точки следует либо исключить из рассмотрения (часто нежелательно), либо использовать специальные формулы сглаживания для крайних точек [2, 5]. В частности, можно использовать формулы:

$$\tilde{Y}_1 = (5 \cdot Y_1 + 2 \cdot Y_2 - Y_3) / 6, \quad \tilde{Y}_n = (5 \cdot Y_n + 2 \cdot Y_{n-1} - Y_{n-2}) / 6.$$

Заметим, что при экспоненциальном сглаживании все точки ряда подвергаются сглаживанию.

Задания

4. Сглаживание временного ряда.

4.1. Выполните сглаживание временного ряда, полученного после удаления аномальных данных, следующими методами:

- среднеарифметическая по 5 точкам;
- средневзвешенная по 5 точкам;
- средневзвешенная по 7 точкам;
- среднехронологическая по 12 точкам;
- экспоненциальное сглаживание.

На одной диаграмме постройте графики исходного ряда и сглаженные ряды.

4.2. Выполните экспоненциальное сглаживание временного ряда с использованием инструмента «Экспоненциальное сглаживание» из пакета «Анализ данных» (рис. 2.7).

***Замечание.** В случае если в окне «Экспоненциальное сглаживание» не задан «фактор затухания» (рис. 2.7), то он, по умолчанию, принимается равным 0.3.*

Время	Исходные данные без аномалий	Сглаживание			Средне-взвешенное по 12 точкам	Экспоненциальное	Альфа для эксп. сглаживания	Экспоненциальное из "Пакета анализа"
		Средне-арифметическое по 5 точкам	Средне-взвешенное по 5 точкам	Средне-взвешенное по 7 точкам				
6:00	5					5	0,3	
6:20	9					7,8		
6:40	9	18,4	12,257143			8,64		
7:00	26	24,8	26,514286	26		20,792		
7:20	43	30,6	38,457143	34,38095		36,3376		
7:40	37	36,8	40,085714	41		36,80128		
8:00	38	38	38,428571	38,71429	34,41667	37,640384		
8:20	40	38,2	36,485714	35	38,125	39,292115		
8:40	32	43	35,857143	40,80952	42,25	34,187635		
9:00	44	44,8	45,8	45,39095	45,54167	44,95632		
9:20	61							
9:40	47							
10:00	49							
10:20	54							
10:40	63							
11:00	51							
11:20	69							
11:40	75,5							
12:00	82							
12:20	62,5							
12:40	43							
13:00	55							
13:20	45							
13:40	44							

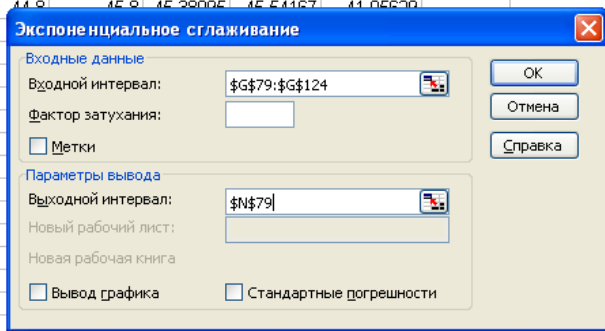


Рис. 2.7. Задание параметров для инструмента «Экспоненциальное сглаживание»

2.5. Контрольные вопросы

1. Поясните, в чем состоят характерные отличия временных рядов от статистических выборок?
2. Что такое тренд, трендовая модель?
3. В чем особенность моментных временных рядов, интервальных временных рядов?
4. Какова общая структура временного ряда?
5. На чем основан критерий Ирвина по определению аномальных значений ряда?
6. Для каких целей может быть использован метод Фостера-Стюарта?
7. Когда метод сравнения разностей средних уровней не дает ответа на вопрос о наличии тренда?
8. Какой метод позволяет определить тренд дисперсии?
9. Влияние каких компонент временного ряда устраняется с помощью методов сглаживания?

ЛАБОРАТОРНАЯ РАБОТА № 3. ВРЕМЕННЫЕ РЯДЫ. ВЫБОР ТРЕНДОВОЙ МОДЕЛИ И ОЦЕНКА ЕЕ АДЕКВАТНОСТИ

Время выполнения работы: 6 аудиторных часов.

Цели работы:

- Получить навыки по выбору трендовых моделей и определению их параметров методом наименьших квадратов.
- Получить навыки по использованию различных показателей для определения адекватности трендовых моделей.

Общее задание на лабораторную работу:

Провести подбор трендовых моделей для экспериментальных данных и определить их адекватность. Для выполнения данной работы необходимо выполнить 3 этапа:

1. Определить подходящие виды трендовых моделей по характеру изменения экспериментальных данных.
2. Определить параметры отобранных трендовых моделей методом наименьших квадратов.
3. Определить адекватность выбранных трендовых моделей.

В качестве исходных экспериментальных данных используются данные из файла, структура которого описана в приложении 3. Файл содержит данные (в виде временного ряда) о стоимости нефти на мировом рынке [6].

3.1. Этапы построения прогноза по временным рядам

Экстраполяционное прогнозирование процессов (прогнозирование «на будущее» при условии сохранения поведения «как в прошлом»), представленных одномерными временными рядами, сводится к выполнению следующих основных этапов [2]:

- 1) предварительный анализ данных;
- 2) построение моделей: формирование набора аппроксимирующих функций (кривых роста) и численное оценивание параметров моделей;
- 3) проверка адекватности моделей и оценка их точности;
- 4) выбор лучшей модели.

В данной лабораторной работе рассматриваются второй и третий этапы (первый этап рассмотрен в лабораторной работе № 2).

Задания

1. Проведите предварительный анализ временного ряда.
 - 1.1. Проверьте исходный временной ряд с ценами нефти на аномальные значения по критерию Ирвина (см. лабораторную работу № 2). Обнаруженные аномальные значения замените путем интерполирования (усреднения) по соседним точкам.
 - 1.2. Выполните сглаживание временного ряда методом простой скользящей средней по трем точкам (см. лабораторную работу № 2).

3.2. Формирование набора аппроксимирующих функций (кривых роста)

На практике для описания тенденции развития исследуемого явления широко используются *модели кривых роста*, представляющие собой различные гладкие функции времени. При таком подходе изменение исследуемого показателя связывают лишь с течением времени; считается, что влияние других факторов несущественно или косвенно сказывается через фактор времени [2, 3].

Правильно выбранная модель кривой роста должна соответствовать характеру изменения тенденции исследуемого явления. Кривая роста позволяет получить выровненные значения уровней динамического ряда. Это те уровни, которые наблюдались бы в случае полного совпадения динамики явления с кривой.

Прогнозирование на основе модели кривой роста базируется на *экстраполяции*, т. е. на продлении в будущее тенденции, наблюдавшейся в прошлом. При этом предполагается, что:

- во временном ряду присутствует тренд;
- характер развития показателя обладает свойством инерционности;
- сложившаяся тенденция не должна претерпевать существенных изменений в течение периода упреждения.

Процедура разработки прогноза с использованием кривых роста включает в себя выбор одной или нескольких кривых, форма

которых соответствует характеру изменения временного ряда, и оценку параметров выбранных кривых.

В литературе описано несколько десятков кривых роста, многие из которых широко применяются для аппроксимации временных рядов (особенно в экономике). Кривые роста условно могут быть разделены на три класса в зависимости от того, какой тип динамики развития они хорошо описывают.

К типу I относятся функции, используемые для описания процессов с монотонным характером тенденции развития и отсутствием пределов роста. Эти условия справедливы для многих экономических показателей, например, для большинства натуральных показателей промышленного производства.

К типу II относятся кривые, описывающие процесс, который имеет предел роста в исследуемом периоде. С такими процессами часто сталкиваются в демографии, при изучении потребностей в товарах и услугах (в расчете на душу населения), при исследовании эффективности использования ресурсов и т. д. Примерами показателей, для которых могут быть указаны пределы роста, являются среднедушевое потребление определенных продуктов питания, расход удобрений на единицу площади и т. п.

Функции, относящиеся ко II типу, называются **кривыми насыщения**. Если кривые насыщения имеют точки перегиба, то они относятся к III типу кривых роста.

Кривые III типа – **S-образные кривые**, описывают как бы два последовательных процесса: один с ускорением развития, другой – с замедлением. S-образные кривые находят применение в демографических исследованиях, в страховых расчетах, при решении задач прогнозирования научно-технического прогресса, при определении спроса на новый вид продукции.

Среди кривых роста I типа, прежде всего, следует выделить **класс полиномов**:

$$U_t = a_0 + a_1 \cdot t + a_2 \cdot t^2 + a_3 \cdot t^3 + \dots + a_p \cdot t^p,$$

где a_i – параметры многочлена, $i = 0, 1, \dots, p$;

t – независимая переменная (время);

p – степень полинома.

Коэффициенты полиномов малых степеней (обычно до 3-й степени) могут иметь конкретную интерпретацию в зависимости от содержания динамического ряда. Например, их можно трактовать как начальный уровень ряда при $t = 0$ (a_0), скорость роста (a_1), ускорение роста (a_2), изменение ускорения (a_3).

Обычно в анализе временных рядов применяются полиномы не выше третьего порядка. Использовать для определения тренда полиномы высоких степеней нецелесообразно, поскольку полученные таким образом аппроксимирующие функции будут отражать случайные отклонения (что противоречит смыслу тенденции).

Полином первой степени на графике изображается прямой (рис. 3.1) и используется для описания процессов, развивающихся во времени равномерно. Полином второй степени применим в тех случаях, когда процесс развивается равноускоренно (т. е. имеется равноускоренный прирост или равноускоренное снижение уровней).

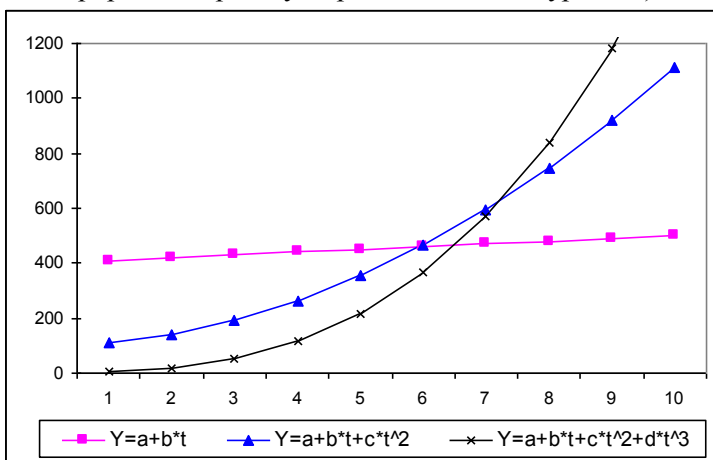


Рис. 3.1. Пример кривых из класса полиномов (I тип)

К кривым II типа (рис 3.2) можно отнести *простую экспоненту*:

$$U_t = a \cdot b^t$$

и *модифицированную экспоненту*:

$$U_t - k = a \cdot b^t \text{ или } U_t = k + a \cdot b^t,$$

где $a < 0$; $0 < b < 1$ – параметры экспоненты;

k – асимптота, значение которой считается известным.

К кривым роста III типа относятся кривая Гомперца и логистическая кривая (Перла-Рида).

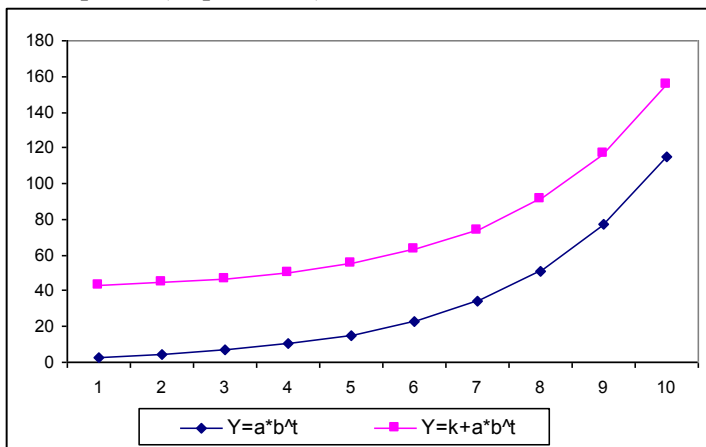


Рис. 3.2. Пример кривых II типа

Кривая Гомперца имеет вид:

$$U_t = k \cdot a^{b^t}.$$

Если $a > 1$, асимптота k лежит ниже кривой, а сама кривая изменяется монотонно: при $b < 1$ – монотонно убывает; при $b > 1$ – монотонно возрастает (рис. 3.3).

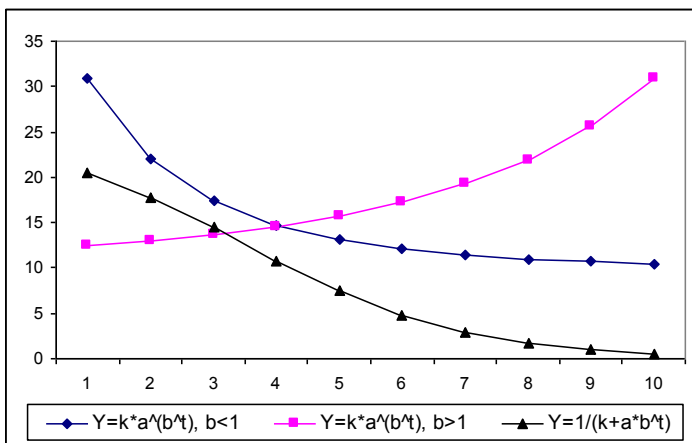


Рис. 3.3. Пример кривых III типа

Логистическая кривая имеет вид:

$$U_t = \frac{1}{k + a \cdot b^t}.$$

При $t \rightarrow -\infty$ логистическая кривая стремится к нулю, а при $t \rightarrow +\infty$ – к асимптоте, равной значению параметра k . Кривая симметрична относительно точки перегиба с координатами: $t = \ln(b/a)$; $U_t = k/2$ (рис. 3.3).

Кривые роста III типа менее распространены на практике, чем кривые I типа, и потому, далее, рассматриваться не будут.

Для выбора кривой роста можно использовать **метод характеристик прироста**, основанный на использовании характерных свойств рассмотренных выше кривых. Процедура выбора кривых с использованием этого метода включает выравнивание ряда Y_t (обычно с помощью простой скользящей средней по трем точкам) и определение средних приростов и производных величин:

$$\Delta Y_t = \frac{Y_{t+1} - Y_{t-1}}{2} \text{ – первый средний прирост;}$$

$$\Delta^2 Y_t = \frac{\Delta Y_{t+1} - \Delta Y_{t-1}}{2} \text{ – второй средний прирост;}$$

$$\frac{\Delta Y_t}{Y_t}, \ln \Delta Y_t \text{ – производные величины.}$$

В соответствии с характером изменений средних приростов и производных показателей выбирается вид кривой роста с помощью табл. 3.1.

Таблица 3.1

Выбор кривых роста по характеру приростов [3]

Показатель	Характер изменения	Кривая роста
ΔY_t	Примерно постоянный	Полином первого порядка ($U_t = a_0 + a_1 \cdot t$)
ΔY_t	Примерно линейный	Полином второго порядка ($U_t = a_0 + a_1 \cdot t + a_2 \cdot t^2$)
$\Delta^2 Y_t$	Примерно линейный	Полином третьего порядка ($U_t = a_0 + a_1 \cdot t + a_2 \cdot t^2 + a_3 \cdot t^3$)
$\Delta Y_t / Y_t$	Примерно постоянный	Экспонента ($U_t = a \cdot b^t$)
$\ln \Delta Y_t$	Примерно линейный	Модифицированная экспонента ($U_t = \frac{1}{k + a \cdot b^t}$)

На практике отбирают две-три кривые роста и окончательный вывод делают исходя из значений критерия, в качестве которого принимают сумму квадратов отклонений фактических значений уровней от расчетных. Из рассматриваемых кривых предпочтение будет отдано той, которой соответствует минимальное значение критерия. Выбор такого критерия удобен еще и потому, что параметры выбранной кривой роста можно найти с помощью метода наименьших квадратов.

Задание

2. Определите подходящие виды трендовых моделей. По сглаженному временному ряду вычислите величины, указанные в табл. 3.1. На основании характера изменения полученных значений выберите наиболее подходящие виды трендовых моделей.

3.3. Численное оценивание параметров моделей

Оценки параметров полиномов выполняют методом наименьших квадратов [2, 5], путем минимизации выражения:

$$\sum_{t=1}^n (Y_t - U_t)^2 \rightarrow \min .$$

В случае кривых роста, не относящихся к полиномам (экспонента, логистическая кривая и т. п.), для определения параметров кривых можно выделить два подхода:

- попытаться свести уравнение кривой к виду полинома и применить метод наименьших квадратов;
- использовать специальные процедуры, специфические для данного вида кривой.

Если в качестве U_t использовать полином вида $U_t = a_0 + a_1 \cdot t$, то для определения параметров a_0 и a_1 получим систему линейных (относительно параметров a_0 и a_1) уравнений:

$$\begin{cases} a_0 \cdot n + a_1 \cdot \sum_{t=1}^n t = \sum_{t=1}^n Y_t \\ a_0 \cdot \sum_{t=1}^n t + a_1 \cdot \sum_{t=1}^n t^2 = \sum_{t=1}^n t \cdot Y_t \end{cases} .$$

Для квадратичного тренда $U_t = a_0 + a_1 \cdot t + a_2 \cdot t^2$ получим систему:

$$\left\{ \begin{array}{l} a_0 \cdot n + a_1 \cdot \sum_{t=1}^n t + a_2 \cdot \sum_{t=1}^n t^2 = \sum_{t=1}^n Y_t \\ a_0 \cdot \sum_{t=1}^n t + a_1 \cdot \sum_{t=1}^n t^2 + a_2 \cdot \sum_{t=1}^n t^3 = \sum_{t=1}^n t \cdot Y_t \\ a_0 \cdot \sum_{t=1}^n t^2 + a_1 \cdot \sum_{t=1}^n t^3 + a_2 \cdot \sum_{t=1}^n t^4 = \sum_{t=1}^n t^2 \cdot Y_t \end{array} \right. .$$

Для полинома 3-го порядка (кубического тренда) $U_t = a_0 + a_1 \cdot t + a_2 \cdot t^2 + a_3 \cdot t^3$ система уравнений будет выглядеть в виде:

$$\left\{ \begin{array}{l} a_0 \cdot n + a_1 \cdot \sum_{t=1}^n t + a_2 \cdot \sum_{t=1}^n t^2 + a_3 \cdot \sum_{t=1}^n t^3 = \sum_{t=1}^n Y_t \\ a_0 \cdot \sum_{t=1}^n t + a_1 \cdot \sum_{t=1}^n t^2 + a_2 \cdot \sum_{t=1}^n t^3 + a_3 \cdot \sum_{t=1}^n t^4 = \sum_{t=1}^n t \cdot Y_t \\ a_0 \cdot \sum_{t=1}^n t^2 + a_1 \cdot \sum_{t=1}^n t^3 + a_2 \cdot \sum_{t=1}^n t^4 + a_3 \cdot \sum_{t=1}^n t^5 = \sum_{t=1}^n t^2 \cdot Y_t \\ a_0 \cdot \sum_{t=1}^n t^3 + a_1 \cdot \sum_{t=1}^n t^4 + a_2 \cdot \sum_{t=1}^n t^5 + a_3 \cdot \sum_{t=1}^n t^6 = \sum_{t=1}^n t^3 \cdot Y_t \end{array} \right. .$$

При использовании кривых роста, не являющихся полиномами, можно попытаться с помощью преобразований и замен переменных привести их к полиномиальному виду и затем определить параметры кривых. Для экспоненты вида $U_t = a \cdot b^t$ это достигается путем логарифмирования:

$$\ln U_t = \ln a + t \cdot \ln b .$$

Используя замену переменных $z_t = \ln U_t$, $p = \ln a$, $s = \ln b$, получим линейное уравнение:

$$z_t = p + t \cdot s .$$

Параметры p и s можно определить из системы линейных уравнений (как и для случая линейного тренда):

$$\left\{ \begin{array}{l} p \cdot n + s \cdot \sum_{t=1}^n t = \sum_{t=1}^n \ln Y_t \\ p \cdot \sum_{t=1}^n t + s \cdot \sum_{t=1}^n t^2 = \sum_{t=1}^n t \cdot \ln Y_t \end{array} \right\},$$

а затем определить параметры исходной экспоненты (значение основания при возведении в степень, конечно же, должно соответствовать основанию логарифма, используемого при линейаризации):

$$a = e^p, \quad b = e^s.$$

Задания

3. Определите параметры отобранных трендовых моделей:
 - 3.1. Рассчитайте параметры выбранных трендовых моделей, а также полинома первого порядка методом наименьших квадратов.
 - 3.2. Для каждой выбранной трендовой модели постройте график исходного ряда и выбранной модели (рис. 3.4).
 - 3.3. Для каждой выбранной трендовой модели рассчитайте численный критерий отклонения: $\sum_{t=1}^n (Y_t - U_t)^2$.

Рекомендация

Для расчета параметров трендовой модели следует найти решение соответствующей системы линейных (относительно искомым параметров a_i) уравнений. В программе Microsoft Excel это можно выполнить путем решения соответствующего матричного уравнения.

Например, для линейной трендовой модели система уравнений, записанная в матричном виде, будет иметь вид:

$$M \cdot A = G,$$

$$\text{где } M = \begin{vmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{vmatrix}, \quad A = \begin{vmatrix} a_0 \\ a_1 \end{vmatrix}, \quad G = \begin{vmatrix} \sum_{t=1}^n Y_t \\ \sum_{t=1}^n t \cdot Y_t \end{vmatrix}.$$

Тогда вектор искомым параметров A можно рассчитать по матричной формуле:

$$A = M^{-1} \cdot G.$$

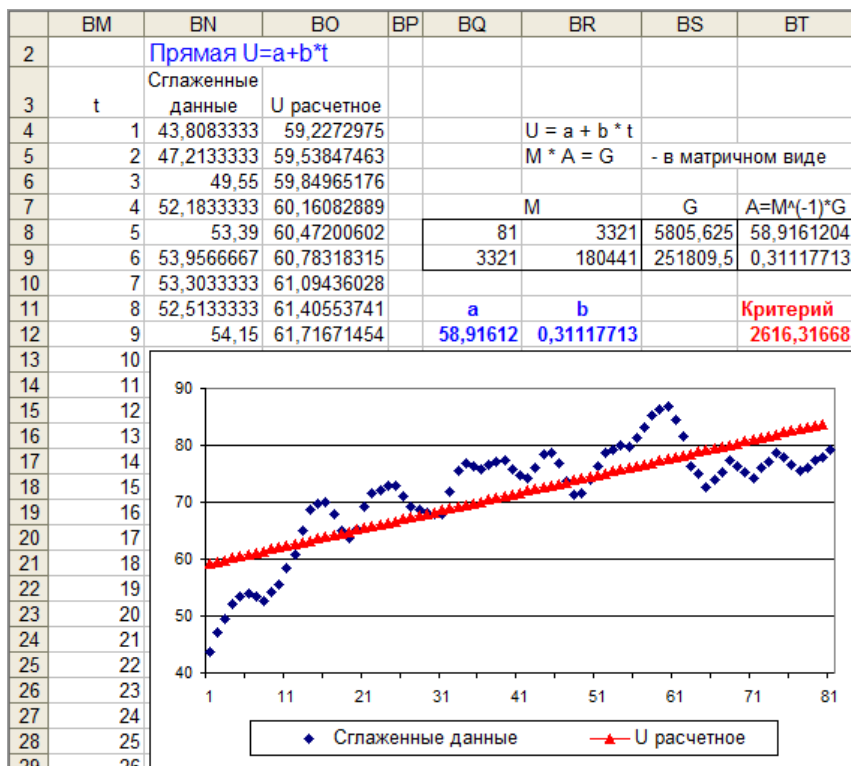


Рис. 3.4. Расчет параметров трендовой модели для экспериментальных данных

Пример вычисления вектора искомых параметров A для линейного тренда представлен на рис. 3.4. Здесь в ячейках записаны следующие значения и формулы:

– в ячейки BM4:BM84 (столбец «t») записаны последовательные значения переменной t от 1 до 81 (81 – количество уровней имеющегося временного ряда);

– в ячейки BN4:BN84 (столбец «Сглаженные данные») записаны значения сглаженного временного ряда, для которого определяется трендовая модель;

– в ячейки BQ8:BR9 записаны формулы, соответствующие элементам матрицы M , а именно:

- в ячейку BQ8 формула «=BM84»;
- в ячейку BQ9 и BR8 формула «=СУММ(BM4:BM84)»;
- в ячейку BR9 формула «=СУММКВ(BM4:BM84)»;

– в ячейки BS8:BS9 записаны формулы, соответствующие элементам вектора G , а именно:

- в ячейку BS8 формула «=СУММ(BN4:BN84)»;
- в ячейку BS9 формула «=СУММПРОИЗВ(BM4:BM84; BN4:BN84)»;

– в ячейки BT8:BT9 записаны формулы нахождения искомых параметров A «=МУМНОЖ(МОБР(BQ8:BR9); BS8:BS9)». Эти формулы отличаются от остальных формул тем, что здесь используется операция обращения матрицы и значит формулы должны быть записаны как матричные формулы (в терминах Microsoft Excel – формулы массива). Чтобы указать, что это формулы массива, следует выполнить следующее:

- 1) в ячейку BT8 записать формулу «=МУМНОЖ(МОБР(Q8:BR9); BS8:BS9)»;
 - 2) выделить блок ячеек BT8:BT9, начиная с ячейки BT8;
 - 3) нажать клавишу F2, а затем комбинацию клавиш CTRL+SHIFT+ENTER (после этого в строке формул текущая формула будет взята в фигурные скобки – признак того, что формула является формулой массива);
- в ячейки BQ12:BR12 записаны формулы, ссылающиеся на значения рассчитанных значений параметров вектора A , а именно:
- в ячейку BQ12 записана формула «=BT8»;
 - в ячейку BR12 записана формула «=BT9»;
- в ячейки BO4:BO84 (столбец «U расчетное») записаны формулы расчета прогнозных значений временного ряда по линейной трендовой модели U_t «=\$BQ\$12 + \$BR\$12 * BM4»;
- в ячейку BT12 записана формула расчета критерия отклонения исходного ряда и линейной трендовой модели: «=СУММКВРАЗН(BN4:BN96; BO4:BO96)».

3.4. Проверка адекватности моделей и оценка их точности

После определения вида и параметров трендовой модели следует обязательно проверить ее на адекватность, т. е. соответствие модели исследуемому процессу. Трендовая модель считается **адекватной**, если она правильно отражает систематические компоненты временного ряда. Это требование эквивалентно требованию, предъявляемым к остаточной компоненте:

- 1) случайность;

- 2) соответствие нормальному закону распределения;
- 3) равенство нулю математического ожидания;
- 4) независимость значений (отсутствие автокорреляции).

3.4.1. Проверка на случайность

Для проверки случайности остаточной компоненты временно-го ряда можно использовать критерий серий и критерий пиков [3].

Критерий серий

Для ряда остаточной компоненты $e_t = Y_t - U_t$ находят медиану e_{med} . Значения e_t сравнивают с e_{med} и формируется последовательность: если $e_t > e_{med}$, то ставится «+», если $e_t < e_{med}$, то ставится «-», при $e_t = e_{med}$ – значение опускается. Последовательность подряд идущих «+» или «-» называется *серией*. Для того чтобы последовательность e_t была случайной, протяженность самой длинной серии не должна быть слишком длинной, а число серий – слишком малым.

Обозначим K_{max} – протяженность самой длинной серии, а v – общее число серий. Остаточная последовательность e_t признается случайной, если:

$$K_{max} < \left[3.3 \cdot (\lg n + 1) \right]; \quad v > \left[\frac{1}{2} \cdot \left(+1 - 1.96 \cdot \sqrt{n-1} \right) \right],$$

где скобки $\lfloor \dots \rfloor$ означают целую часть числа.

Если хотя бы одно из этих неравенств нарушается, то остаточная компонента считается не случайной и, значит, трендовая модель неадекватна.

Критерий пиков

Точка e_t считается *пиковой точкой*, если она больше или меньше своих соседей, т. е. выполняется одно из условий:

$$e_{t-1} < e_t > e_{t+1} \quad \text{или} \quad e_{t-1} > e_t < e_{t+1}.$$

Общее число пиковых точек p для случайной последовательности характеризуется математическим ожиданием числа пиковых точек и их дисперсией:

$$m_p = \frac{2}{3} \cdot (n-2), \quad D_p = \frac{16 \cdot n - 29}{90}$$

и их не должно быть слишком мало.

Если выполняется следующее неравенство, то остаточная компонента считается случайной, если не выполняется, то – не случайной:

$$p > \lfloor m_p - 1.96 \cdot \sqrt{D_p} \rfloor,$$

где скобки $\lfloor \dots \rfloor$ означают целую часть числа.

Задания

4. Для каждой выбранной трендовой модели проверьте ее адекватность исходным данным:

4.1. Проверьте остаточную компоненту на случайность по критерию серий (рис. 3.5) и критерию пиков (рис. 3.6).

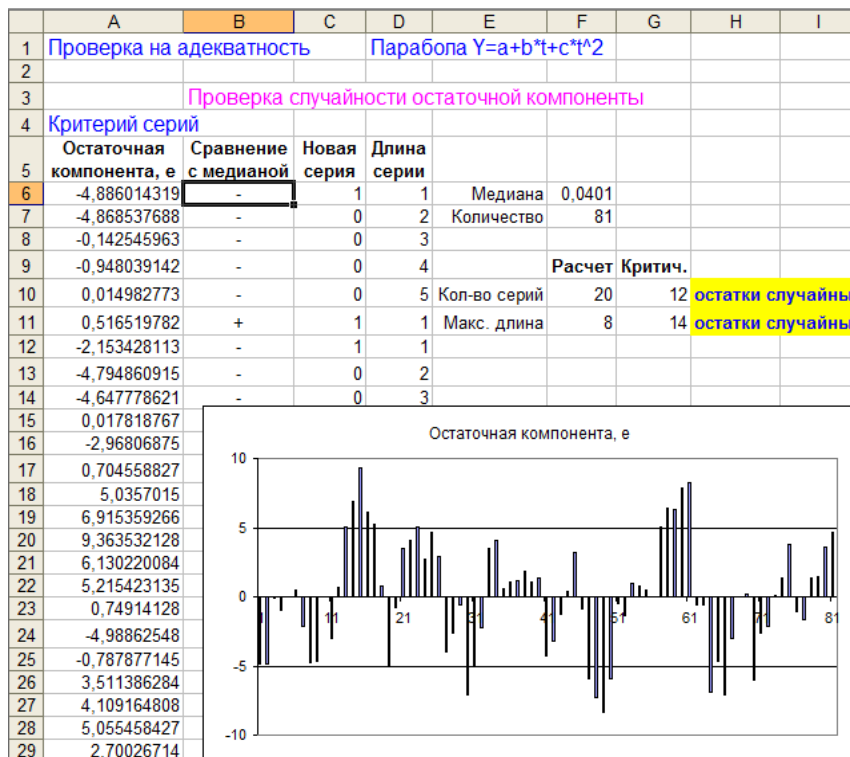


Рис. 3.5. Проверка случайности остаточной компоненты по критерию серий

Рекомендации:

1. На рис. 3.5 представлен пример вычисления критерия серий при проверке остаточной компоненты на случайность. Здесь в ячейках записаны следующие значения и формулы:

– в ячейки А6:А86 (столбец «Остаточная компонента, е») записаны последовательные значения остаточной компоненты e ;

– в ячейку F6 записана формула расчета значения медианы для ряда остаточной компоненты e_{med} «=МЕДИАНА(А6:А86)»;

– в ячейку F7 записана формула расчета количества значений ряда остаточной компоненты n «=СЧЁТ(А6:А86)»;

– в ячейки В6:В86 (столбец «Сравнение с медианой») записаны формулы определения знака при сравнении соответствующего значения остаточной компоненты с ее медианой «=ЕСЛИ(А6<F6; "-" ; ЕСЛИ(А6=F6; 0; "+"))»;

– в ячейки С6:С86 (столбец «Новая серия») записаны формулы для установки признака начала новой серии в ячейках В6:В86 (признаком начала серии является отличие в знаках в ячейках В6:В86 для текущей и предыдущей строк): в ячейку С6 записано значение «1» (начало первой серии), в ячейку С7 записана формула «=ЕСЛИ(ИЛИ(В7=В6; В7=0); 0; 1)», в остальные ячейки (С8:С86) скопирована формула из ячейки С7;

– в ячейки D6:D86 (столбец «Длина серии») записаны формулы для определения текущей длины текущей серии (длина серии увеличивается на 1, если знаки в ячейках В6:В86 для текущей и предыдущей строк не отличаются, иначе длина принимается равной 1): в ячейку D6 записано значение «1» (текущая длина первой серии), в ячейку D7 записана формула «=ЕСЛИ(ИЛИ(В7=В6; В6=0); D6+1; ЕСЛИ(В7=0; D6; 1))», в остальные ячейки (D8:D86) скопирована формула из ячейки D7;

– в ячейку F10 записана формула расчета общего числа серий ν , которое соответствует количеству единиц в ячейках столбца «Новая серия» - «=СУММ(С6:С86)»;

– в ячейку G10 записана формула расчета критического значения, с которым следует сравнивать значение ν - «=ОКРУГЛВНИЗ(0,2* (F7+1-1,96*КОРЕНЬ(F7-1)); 0)»;

- в ячейку F11 записана формула расчета протяженности самой длинной серии K_{max} , которая соответствует максимальному значению из ячеек столбца «Длина серии» – «=МАКС(D6:D86)»;
- в ячейку G11 записана формула расчета критического значения с которым следует сравнивать значение K_{max} – «=ОКРУГЛВНИЗ(3,3*LN(F7+1); 0)»;
- в ячейку H10 записана формула, выдающая результат сравнения расчетного и критического значений для параметра ν – «=ЕСЛИ(F10>G10; "остатки случайны"; "остатки не случайны")»;
- в ячейку H11 записана формула, выдающая результат сравнения расчетного и критического значений для параметра K_{max} – «=ЕСЛИ(F11<G11; "остатки случайны"; "остатки не случайны")».

	К	L	M	N	O
1					
2					
3					
4	Критерий пиков				
5	Остаточная компонента, e	Пики			
6	-4,886014319				
7	-4,868537688	0	Количество	81	
8	-0,142545963	1	Мат. ожидание	52,6666667	
9	-0,948039142	1	Дисперсия	14,0777778	
10	0,014982773	0			
11	0,516519782	1	Кол-во пиков	38	
12	-2,153428113	0	Крит. значение	45	
13	-4,794860915	1			остатки не случайны
14	-4,647778621	0			
15	0,017818767	1			

Рис. 3.6. Проверка случайности остаточной компоненты по критерию пиков

2. На рис. 3.6 представлен пример вычисления критерия пиков при проверке остаточной компоненты на случайность. Здесь в ячейках записаны следующие значения и формулы:

- в ячейки K6:K86 (столбец «Остаточная компонента, e») записаны последовательные значения остаточной компоненты e ;
- в ячейку N7 записана формула расчета количества значений ряда остаточной компоненты n «=СЧЁТ(K6:K86)»;

- в ячейку N8 записана формула расчета математического ожидания числа пиковых точек m_p « $=2/3*(N7-2)$ »;
- в ячейку N9 записана формула расчета дисперсии числа пиковых точек D_p « $=(16*N7-29)/90$ »;
- в ячейки L7:L86 (столбец «Пики») записаны формулы для определения, является ли текущая точка остаточной компоненты пиковой: в ячейку L7 записана формула « $=ЕСЛИ(И(K7<K6; K7<K8); 1; ЕСЛИ(И(K7>K6; K7>K8); 1; 0))$ », в остальные ячейки (L8:L86) скопирована формула из ячейки L7;
- в ячейку N11 записана формула расчета количества пиковых точек в ряду остаточной компоненты p , которое соответствует количеству единиц в ячейках столбца «Пики» - « $=СУММ(L6:L86)$ »;
- в ячейку N12 записана формула расчета критического значения, с которым следует сравнивать значение p – « $=ОКРУГЛВНИЗ(N8-1,96*КОРЕНЬ(N9);0)$ »;
- в ячейку N13 записана формула, выдающая результат сравнения расчетного и критического значений для параметра p – « $=ЕСЛИ(N11>N12; "остатки случайны"; "остатки не случайны")$ ».

3.4.2. Проверка на нормальное распределение

Проверка соответствия распределения остаточной компоненты нормальному закону выполняется с помощью показателей асимметрии и эксцесса или с помощью RS-критерия [4].

Выборочные характеристики *асимметрии* и *эксцесса*:

$$A = \frac{\frac{1}{n} \cdot \sum_{i=1}^n e_i^3}{\sqrt{\left(\frac{1}{n} \cdot \sum_{i=1}^n e_i^2\right)^3}}, \quad E = \frac{\frac{1}{n} \cdot \sum_{i=1}^n e_i^4}{\left(\frac{1}{n} \cdot \sum_{i=1}^n e_i^2\right)^2} - 3.$$

Соответствующие среднеквадратические ошибки:

$$\sigma_A = \sqrt{\frac{6 \cdot (n-2)}{(n+1) \cdot (n+3)}}, \quad \sigma_E = \sqrt{\frac{24 \cdot (n-2) \cdot (n-3)}{(n+1)^2 \cdot (n+3) \cdot (n+5)}}.$$

Если одновременно выполняются следующие неравенства:

$$|A| < 1.5 \cdot \sigma_A, \quad \left| E + \frac{6}{n+1} \right| < 1.5 \cdot \sigma_E,$$

то гипотеза о нормальном характере распределения остаточной компоненты принимается.

Если выполняется хотя бы одно из неравенств:

$$|A| \geq 2 \cdot \sigma_A, \quad \left| E + \frac{6}{n+1} \right| \geq 2 \cdot \sigma_E,$$

то гипотеза о нормальном характере распределения отвергается и трендовая модель признается неадекватной. В других случаях – о нормальности распределения нельзя сделать никакого вывода.

RS-критерий (*критерий Девиды-Хартли-Пирсона*) – один из самых простых критериев проверки нормальности закона распределения случайной величины, он характеризует отношение размаха вариаций к стандартному отклонению R/S :

$$R = e_{\max} - e_{\min}, \quad S = \sqrt{\frac{1}{n-1} \cdot \sum_{t=1}^n e_t^2}.$$

Значение R/S сравнивается с табличными нижней и верхней границами данного отношения, и если это значение не попадает в интервал между критическими границами, то гипотеза о нормальности распределения отвергается; в противном случае эта гипотеза принимается. Табличные интервалы RS -критерия приведены в табл. 3.2.

Таблица 3.2

Нижняя и верхняя границы интервала RS -критерия [4]

n	10		20		30		50		80	
R/S	2.67	3.69	3.18	4.49	3.47	4.89	3.83	5.35	4.15	5.73

Задание

4.2. Проверьте остаточную компоненту на нормальное распределение с помощью показателей асимметрии и эксцесса и с помощью RS -критерия (рис. 3.7).

Рекомендации:

1. На рис. 3.7 представлен пример проверки на нормальное распределение остаточной компоненты. Здесь в ячейках записаны следующие значения и формулы:

– в ячейки Р6:Р86 (столбец «e²») записаны последовательные значения квадратов остаточной компоненты *e*: в ячейку Р6 записана формула «=А6*А6», в остальные ячейки (Р7:Р86) скопирована формула из ячейки Р6;

– в ячейки Q6:Q86 (столбец «e³») записаны последовательные значения кубов остаточной компоненты *e*: в ячейку Q6 записана формула «=А6*Р6», в остальные ячейки (Q7:Q86) скопирована формула из ячейки Q6;

	P	Q	R	S	T	U
1						
2						
3	Проверка остаточной компоненты на нормальное распределение					
4						
5	e ²	e ³		Ассиметрия	Эксцесс	
6	23,87314	-116,644	Значение	0,000888865	-2,9694055	
7	23,70266	-115,397	Ошибки	0,262326764	0,0551777	
8	0,020319	-0,0029	Крит. значение	0,393490146	2,8962347	
9	0,898778	-0,85208		распределение ненормальное		
10	0,000224	3,36E-06				
11	0,266793	0,137804		RS-критерий		
12	4,637253	-9,98599	Размах	17,6956414		
13	22,99069	-110,237	Станд. ошибка	4,03041669		
14	21,60185	-100,401	RS-критерий	4,390524047		
15	0,000318	5,66E-06	Нижн. граница	4,27		
16	8,809432	-26,147	Верхн. граница	5,86		
17	0,496403	0,349745		распределение нормальное		
18	25,35829	127,6968				
19	47,82219	330,7077				
20	87,67573	820,9546				

Рис. 3.7. Проверка остаточной компоненты на нормальное распределение

– в ячейку S6 записана формула расчета выборочного значения асимметрии *A* для ряда остаточной компоненты *e* «=(1/F7*СУММ(Q6:Q86))/КОРЕНЬ(1/F7*СУММ(Р6:Р86)^3)»;

– в ячейку T6 записана формула расчета выборочного значения эксцесса *E* для ряда остаточной компоненты *e* «=(1/F7*СУММКВ(Р6:Р86))/(1/F7*СУММ(Р6:Р86)^2)-3»;

– в ячейку S7 записана формула расчета среднеквадратической ошибки асимметрии σ_A для ряда остаточной компоненты e «=КОРЕНЬ(6*(N7-2)/((N7+1)*(N7+3)))»;

– в ячейку T7 записана формула расчета среднеквадратической ошибки эксцесса σ_E для ряда остаточной компоненты e «=КОРЕНЬ((24*(N7-2)*(N7-3))/((N7+1)^2*(N7+3)*(N7+5)))»;

– в ячейку S8 записана формула расчета критического значения, с которым следует сравнивать значение асимметрии A – «=1,5*S7»;

– в ячейку T8 записана формула расчета критического значения, с которым следует сравнивать значение эксцесса E – «=ABS(T6+6/(N7+1))»;

– в ячейку S9 записана формула, выдающая результат сравнения расчетного и критического значений для параметров A и E – «=ЕСЛИ(И(ABS(S6)<S8; T8<S8); "распределение нормальное"; ЕСЛИ(ИЛИ(ABS(S6)>=2*S7; T8>=2*S7); "распределение ненормальное"; "не определено"))».

2. На рис. 3.7 в ячейках S12:S17 представлен пример использования RS-критерия для проверки соответствия остаточной компоненты нормальному распределению. Здесь в ячейках записаны следующие формулы:

– в ячейку S12 записана формула расчета размаха R для ряда остаточной компоненты e «=МАКС(А6:А86)-МИН(А6:А86)»;

– в ячейку S13 записана формула расчета значения стандартного отклонения S для ряда остаточной компоненты e «=СТАНДОТКЛОН(А6:К86)»;

– в ячейку S14 записана формула расчета RS-критерия «=S12/S13»;

– в ячейки S15, S16 записаны значения верхней и нижней критических границ для RS-критерия, согласно табл. 3.2;

– в ячейку S17 записана формула, выдающая результат сравнения расчетного значения RS-критерия с его критическими границами - «=ЕСЛИ(И(S14<=S16; S14>=S15); "распределение нормальное"; "распределение ненормальное"))».

3.4.3. Проверка равенства математического ожидания нулю

Проверка равенства математического ожидания остаточной компоненты нулю при условии, что она распределена по нормальному закону, осуществляется на основе t -критерия Стьюдента [4]. Его расчетное значение определяется формулой:

$$t = \frac{|m_e|}{S} \cdot \sqrt{n},$$

где m_e – оценка математического ожидания остаточной компоненты; S – оценка среднеквадратического отклонения остаточной компоненты.

Если расчетное значение t меньше табличного значения с уровнем значимости α и числом степеней свободы $n - 1$, то гипотеза о равенстве нулю математического ожидания принимается, в противном случае эта гипотеза отвергается и модель считается неадекватной.

Задание

4.3. Проверьте равенство нулю математического ожидания остаточной компоненты по t -критерию Стьюдента (рис. 3.8).

	W	X	Y	Z	AA
1					
2					
3	Проверка равенства мат. ожидания остаточной компоненты 0				
4					
5	Остаточная компонента, e				
6	-4,886014319				
7	-4,868537688	Количество	81		
8	-0,142545963	Среднее	0,0090741		
9	-0,948039142	СКО	4,0304167		
10	0,014982773	t -критерий	0,0202626		
11	0,516519782	Крит. значение	1,9900634		
12	-2,153428113		мат ожидание = 0		
13	-4,794860915				
14	-4,647778621				
15	0,017818767				
16	-2,96806875				

Рис. 3.8. Проверка равенства математического ожидания остаточной компоненты нулю

Рекомендация

На рис. 3.8 представлен пример проверки гипотезы о равенстве нулю математического ожидания остаточной компоненты. Здесь в ячейках записаны следующие значения и формулы:

– в ячейки W6:W86 (столбец «Остаточная компонента, е») записаны последовательные значения остаточной компоненты e ;

– в ячейку Y7 записана формула расчета количества точек ряда остаточной компоненты n ;

– в ячейку Y8 записана формула расчета среднего значения ряда остаточной компоненты m_e «=СРЗНАЧ(W6:W86)»;

– в ячейку Y9 записана формула расчета стандартного отклонения (СКО) ряда остаточной компоненты S «=СТАНДОТКЛОН(W6:W86)»;

– в ячейку Y10 записана формула расчета t -критерия Стьюдента «=ABS(Y8)/Y9*КОРЕНЬ(Y7)»;

– в ячейку Y11 записана формула расчета критического значения t -критерия Стьюдента для уровня значимости 0.05 «=СТЮДРАСПОБР(0,05; Y7-1)»;

– в ячейку Y12 записана формула, выдающая результат сравнения расчетного и критического значений t -критерия Стьюдента – «=ЕСЛИ(Y10<Y11; "мат ожидание = 0"; "мат ожидание НЕ = 0")».

3.4.4. Проверка на независимость

Проверка на независимость значений остаточной компоненты выполняется с помощью d -критерия Дарбина-Уотсона [4], расчетное значение которого определяется по формуле:

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}.$$

Если полученное значение находится в интервале от 2 до 4, то это свидетельствует об отрицательной связи и необходимо сделать преобразование $d = 4 - d$.

Расчетное значение критерия d необходимо сравнить с критическими значениями, приведенными в табл. 3.3, зависящими от количества уровней ряда n . Если $d > d_2$, то гипотеза о независимости

сти уровней остаточной последовательности принимается. Если $d < d_1$, то эта гипотеза отвергается. При значениях $d_1 < d < d_2$ нельзя сделать тот или иной вывод.

Таблица 3.3. Критические значения d-критерия Дарбина-Уотсона [4]

n	15	20	30	80
d_1	1,08	1,2	1,35	1,61
d_2	1,36	1,41	1,49	1,66

Модель будет считаться адекватной, если все четыре вышеприведенные проверки свойств остаточной компоненты дали положительные результаты.

Задания

- 4.4. Проверьте остаточную компоненту на независимость по критерию Дарбина-Уотсона (рис. 3.9).
- 4.5. На основе выполненных заданий 4.1-4.4 сделайте мотивированный вывод об адекватности трендовой модели.

	AC	AD	AE	AF
1				
2				
3	Проверка остаточной компоненты на независимость			
4	d-критерий Дарбина-Уотсона			
5	Разница остаточной компоненты, $e(t) - e(t-1)$			
6				
7	0,017476631	d-критерий Дарбина-Уотсона	0,6145731	
8	4,725991726	Преобразованный критерий	0,6145731	
9	-0,80549318	Критическое d1	1,65	
10	0,963021915	Критическое d2	1,69	
11	0,50153701		зависимы	
12	-2,669947896			
13	-2,641432801			
14	0,147082294			
15	4,665597388			
16	-2,985887517			
17	3,672627577			
18	4,331142672			

Рис. 3.9. Проверка остаточной компоненты на независимость

Рекомендация

На рис. 3.9 представлен пример проверки остаточной компоненты на независимость с помощью d-критерия Дарбина-Уотсона. Здесь в ячейках записаны следующие значения и формулы:

– в ячейки АС7:АС86 (столбец «Разница остаточной компоненты, $e(t) - e(t-1)$ ») записаны последовательные значения отклонений ряда остаточной компоненты e ;

– в ячейку АЕ7 записана формула расчета d-критерия Дарбина-Уотсона «=СУММКВ(АС7:АС86)/СУММ(Р6:Р86)» (в ячейках Р6:Р86 расположены квадраты значений ряда остаточной компоненты);

– в ячейку АЕ8 записана формула расчета преобразованного критерия Дарбина-Уотсона, при условии, что сам d-критерий Дарбина-Уотсона превышает значение 2 «=ЕСЛИ(И(АЕ7>2; АЕ7<4); 4-АЕ7; АЕ7)»;

– в ячейки АЕ9, АЕ10 записаны значения верхней и нижней критических границ для d-критерия Дарбина-Уотсона, согласно табл. 3.3;

– в ячейку АЕ11 записана формула, выдающая результат сравнения преобразованного критерия Дарбина-Уотсона с его критическими границами «=ЕСЛИ(АЕ8>АЕ10; "независимы"; ЕСЛИ(АЕ8<АЕ9; "зависимы"; "не определено"))».

3.5. Контрольные вопросы

1. Охарактеризуйте типы кривых роста, наиболее часто используемые на практике при построении трендовых моделей.
2. При удовлетворении каких требований модель считается адекватной?
3. Каков порядок вычисления критерия серий?
4. Какая точка считается пиковой?
5. Что характеризуют показатели асимметрии и эксцесса?
6. С помощью какого критерия можно проверить независимость уровней остаточной компоненты временного ряда?
7. Как выбирается адекватная трендовая модель из нескольких вариантов?

ЛАБОРАТОРНАЯ РАБОТА № 4. ВЕРОЯТНОСТНЫЙ АНАЛИЗ ДИНАМИЧЕСКОЙ СИСТЕМЫ И ЕГО ПРИМЕНЕНИЕ

Время выполнения работы: 6 аудиторных часов.

Цели работы:

- Получить навыки по проведению расчета статистических характеристик экспериментальных данных в пакете Mathcad.
- Получить навыки по построению динамической модели описания экспериментальных данных и проведению ее проверки на адекватность.
- Получить навыки по выполнению прогнозирования выхода динамической модели.

Общее задание на лабораторную работу:

Для выполнения данной работы необходимо выполнить 3 этапа:

1. Провести обработку экспериментальных данных с целью построения модели.
2. Построить модель описания экспериментальных данных и проверить ее на адекватность.
3. Выполнить прогнозирование выхода модели при заданных входах.

4.1. Содержательная постановка задачи

Представьте себе, что Вы являетесь экспертом экологической части проекта создания нового крупного промышленного предприятия. Экспертиза проводится по заданию мэра небольшого промышленного города N-ска, который интуитивно считает, что ввод в действие нового предприятия приведет к неблагоприятным последствиям. Мэр утверждает, что загрязнение атмосферы и грунтовых вод вредными выбросами предприятий, уже действующих на территории города, привело к ухудшению условий жизни и показателей благополучия горожан по сравнению с другими городами, расположенными в сходных природных условиях и сравнимыми с N-ском по численности.

В числе показателей, характеризующих благополучие города, мэр назвал такие, значения которых хуже, чем в городах с аналогичными условиями жизни и работы:

- 1) отклонение средней продолжительности жизни (в годах) от восточноевропейской (составляющей 72 года);
- 2) отклонение среднего количества рождений детей на 1000 жителей от среднего показателя по стране (96 рождений);
- 3) отклонение среднего числа дней нетрудоспособности в году от норматива, исходя из которого планируются медицинские услуги и лекарственное обеспечение (20 дней в году);
- 4) отклонение уровня кислотности грунтовых вод от нормы, (норма принимается за 100 %).

Предприятия, работающие на территории города, загрязняют атмосферу и грунт различными веществами, многие из которых специфичны только для данного предприятия и поэтому не представляют новой экологической опасности для города. Но имеются вещества, выбрасываемые всеми предприятиями (в том числе и проектируемыми). Эти вещества относятся к группе продуктов сгорания жидкого и газообразного топлива, причем известно, что для N-ска наибольшую опасность представляют выбросы сернистого газа SO₂ (последствия – кислотные дожди, аллергические заболевания, астма) и окисла углерода CO (последствия – отравления, удушье, аллергия).

Согласно экологической части проекта, в которой рассчитаны выбросы вредных веществ в атмосферу и в грунт, ввод нового предприятия приведет к увеличению приземных концентраций сернистого газа и окисла углерода в 2 раза и к увеличению в 2 раза разброса этих показателей вокруг среднего значения.

Перечень вопросов, которые должен выяснить эксперт по заказу мэра:

1. *Каковы будут изменения показателей благополучия города после ввода предприятия?*

2. *Насколько быстро закончится период изменения этих показателей от существующих до установившихся (в результате ввода в действие нового предприятия) значений?*

4.2. Схема проведения экспертизы, составляющей тему лабораторной работы

1. Получение исходных (указанных мэром) данных, позволяющих построить модель связи показателей, характеризующих экологическое благополучие города, с приземными концентрациями вредных веществ. Эти данные должны содержать результаты многолетних наблюдений за показателями и концентрациями, имеющиеся в экологических службах N-ска и других городов с близкими условиями.

2. Обработка полученных данных с целью построения модели. Построение модели.

3. Проверка модели путем сравнения реконструированных (расчетным путем) значений показателей экологического благополучия с фактическими (предоставленными в качестве исходных данных).

4. Если точность реконструкции окажется удовлетворительной – прогноз динамики изменения показателей экологического благополучия, ожидаемой после ввода нового предприятия в эксплуатацию.

4.2.1. Исходные данные для построения модели связи показателей экологического благополучия с приземными концентрациями вредных веществ (выполнение п. 1 схемы из п. 4.2)

Расчеты будем проводить в среде Mathcad.

Требуемые исходные данные получены по результатам 15-летних наблюдений за показателями и концентрациями, регулярно проводимых ежемесячно в 100 городах с аналогичными природными условиями и сходной численностью населения. Наблюдения упорядочены во времени. За период наблюдения ввод новых предприятий не проводился, поэтому можно считать, что исходные данные относятся к стационарным.

Исходные данные хранятся в файле, структура которого описана в приложении 4. Порядок распаковки (загрузки) исходных данных:

1. Прочитать файл с исходными данными в массив Z .
2. Распаковать массив, который представляет собой структуру вида $Z = (X \ w)$, где X – результаты наблюдений за показателями бла-

гополучия, w – результаты наблюдения за приземными концентрациями вредных веществ. Распаковка производится присваиванием:

$$X := Z_{0,0} \quad w := Z_{0,1}.$$

3. Рассчитать число элементов в массивах (проверить, одинаковое ли это число, и, если нет – выбрать меньшее из них в качестве числа исходных данных). Функция Mathcad для определения числа строк в некотором массиве y : $\text{rows}(y)$.

4. Определить диапазон изменения индексов в массивах X и w (с учетом особенностей нумерации Mathcad максимальное значение индекса s_{\max} на 1 меньше числа строк в массивах):

$$s := 0..s_{\max}.$$

5. Принять во внимание следующие обозначения, использованные в исходных данных:

а) показатели экологического благополучия города в каждом s -м наблюдении представлены 4-мерным вектором X_s с компонентами:

– $(X_s)_0$ – отклонение средней продолжительности жизни (в годах) от средневропейской (72 года);

– $(X_s)_1$ – отклонение среднего количества рождений детей на 1000 жителей от среднего показателя по стране (96 рождений);

– $(X_s)_2$ – отклонение среднего числа дней нетрудоспособности в году от норматива, исходя из которого планируются медицинские услуги и лекарственное обеспечение (20 дней в году);

– $(X_s)_3$ – отклонение уровня кислотности грунтовых вод от нормы (норма принимается за 100 %);

б) значения приземных концентраций, измеренных одновременно с измерениями показателей экологического благополучия, представлены в каждом s -м наблюдении 2-мерным вектором w_s с компонентами:

– $(w_s)_0$ – приземная концентрация сернистого газа (мг в 10 м^3 воздуха);

– $(w_s)_1$ – приземная концентрация окисла углерода (мг в 10 м^3 воздуха).

Фрагменты графиков показателей экологического благополучия и концентраций вредных веществ приведены на рис. 4.1, 4.2. Постройте их по своим данным.

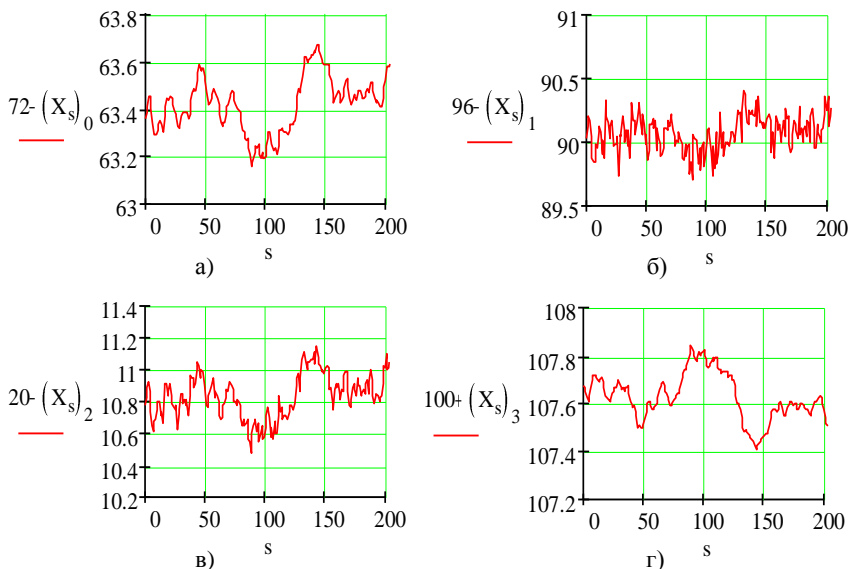


Рис. 4.1. Исходные данные. Колебания показателей экологического благополучия: а) продолжительность жизни; б) число рождений; в) число дней нетрудоспособности; г) кислотность грунтовых вод

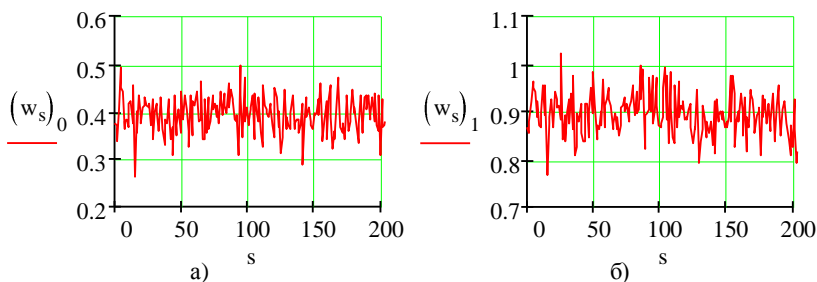


Рис. 4.2. Исходные данные. Колебания приземных концентраций вредных веществ: а) сернистого газа; б) окисла углерода

Выполнение п. 1 схемы из п. 4.2 завершено.

4.2.2. Обработка полученных данных с целью построения модели. Построение модели связи показателей экологического благополучия со значениями приземных концентраций вредных веществ (выполнение п. 2 схемы из п. 4.2)

Обработка исходных данных

Схема обработки исходных данных строится исходя из конечной цели – получения модели связи показателей благополучия с концентрациями. Опишем схему «с конца»:

1. Модель связи ищем в виде:

$$X_{s+1} = \Phi \cdot X_s + \Gamma \cdot w_s, \quad (4.1)$$

где Φ – 4×4 матрица, характеризующая инерционность изменения показателей экологического благополучия;

Γ – 4×2 матрица, характеризующая степень влияния вредных веществ на показатели благополучия.

Гипотезы, заложенные в модель (4.1), которые *нужно проверить*:

- Можно ли считать, что Φ и Γ – константные матрицы?
- Можно ли считать, что по сравнению с инерционностью процессов, характеризующих экологическое благополучие, значения концентраций без больших потерь точности допустимо считать независимыми во времени?
- Можно ли считать, что статистики приземных концентраций стационарны (математическое ожидание и матрица ковариаций не зависят от времени, корреляционная матрица зависит только от сдвига во времени между значениями концентраций)?

2. Если гипотезы окажутся приемлемыми, то по заданным наблюдениям X_s и w_s , $s = 0, \dots, s_{\max}$, нужно рассчитать Φ и Γ .

3. Если этот расчет будет успешным, то появится возможность дать ответ на вопросы мэра путем подстановки новых значений приземных концентраций.

4. Для расчета матрицы Φ используем соотношение:

$$\Phi = \text{CORR}_X(1) \cdot \text{CORR}_X(0)^{-1}, \quad (4.2)$$

где $\text{CORR}_X(t)$ – корреляционная матрица, рассчитанная по данным об X_s .

5. Для расчета матрицы Γ используем соотношение:

$$\Gamma = \text{CORR}_{XW}(1) \cdot \chi^{-1}, \quad (4.3)$$

где $CORR_{\chi w}(t)$ – взаимная корреляционная функция между показателями благосостояния и приземными концентрациями вредных веществ;

χ – матрица ковариаций приземных концентраций w_s .

Мы видим, что для получения модели нужно выполнить следующее:

1. Рассчитать оценки статистик для приземных концентраций вредных веществ w_s (математического ожидания, матрицы ковариаций и корреляционной матрицы). Проверить, можно ли считать w_s чисто случайной последовательностью. На основании осмотра графиков можно сделать предположение о том, что колебания значений приземных концентраций имеют гораздо более высокочастотный характер, чем показатели экологического благополучия (это естественно – ведь показатели благополучия инерционны, в их значениях содержится «память» о прошлом). Поэтому при построении модели можно предположить, что последовательность w_s – чисто случайная (отсутствует корреляция значений во времени). Но это предположение следует подтвердить (или опровергнуть) расчетами.

2. Рассчитать оценки статистик для показателей экологического благополучия X_s (математического ожидания, матрицы ковариаций и корреляционной матрицы). Проверить, можно ли считать их соответствующими стационарному режиму.

3. Рассчитать оценку взаимной корреляционной матрицы между X_s и w_s .

После этого все данные для формул (4.2) и (4.3), а значит и модели (4.1), окажутся полученными.

Рабочие формулы и проверки гипотез (в формулах будем использовать удвоение обозначений статистик, чтобы подчеркнуть, что это – не истинные статистики, а их *оценки* по доступным данным):

1. Оценка математического ожидания входного воздействия:

$$mm_w := \frac{1}{s_{\max} + 1} \cdot \sum_{k=0}^{s_{\max}} w_k. \quad (4.4)$$

Проверка гипотезы о постоянстве математического ожидания. С практической (инженерной, а не научной) точки зрения лучше всего это сделать по графику: нанести значения w_s и mm_w и визуально оценить, можно ли считать математическое ожидание постоянным (рис. 4.3).

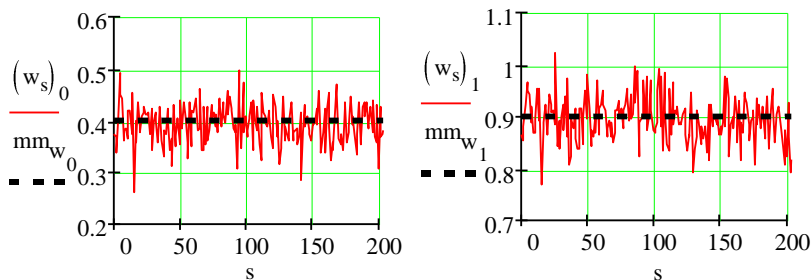


Рис. 4.3. К проверке гипотезы о постоянстве математического ожидания приземных концентраций

2. Центрированная (математическим ожиданием) последовательность значений приземных концентраций:

$$w0_s := w_s - mm_w. \quad (4.5)$$

3. Матрица ковариаций входного воздействия (должна получиться симметричной):

$$\chi\chi := \frac{1}{s_{\max}} \cdot \sum_{k=0}^{s_{\max}} \left(w0_k \cdot w0_k^T \right). \quad (4.6)$$

Проверка гипотезы о постоянстве матрицы ковариаций: нужно рассчитать 2 варианта оценки матрицы ковариаций (по первой и по второй половине экспериментальной выборки) и оценить в % расхождение между этими оценками. Приемлемое расхождение $\pm 10\%$. Соответствующие формулы приведены на рис. 4.4 (групповые параллельные вычисления в Mathcad – значок стрелочки в последней формуле, проводятся при нажатии клавиш <Ctrl>+<минус>; если последняя формула выдает ошибку «Это значение должно быть матрицей скалярных элементов», то следует щелкнуть правой кнопкой мыши на ошибке и в контекстном меню выбрать позицию «Абсолютное значение»).

$$\text{middle} := \text{ceil}\left(\frac{s_{\max}}{2}\right) \quad - \text{число элементов в половине выборки}$$

$$\chi\chi^1 := \frac{1}{\text{middle} - 1} \cdot \sum_{k=0}^{\text{middle}-1} \left(w_{0k}^0 \cdot w_{0k}^{0T} \right) \quad - \text{оценки матрицы ковариации по первой и второй половинам выборки}$$

$$\chi\chi^2 := \frac{1}{s_{\max} - \text{middle}} \cdot \sum_{k=\text{middle}}^{s_{\max}} \left(w_{0k}^0 \cdot w_{0k}^{0T} \right)$$

Результаты вычисления оценок матриц ковариации

$$\chi\chi^1 = \begin{pmatrix} 1.596 \times 10^{-3} & 1.109 \times 10^{-3} \\ 1.109 \times 10^{-3} & 2.063 \times 10^{-3} \end{pmatrix} \quad \chi\chi^2 = \begin{pmatrix} 1.589 \times 10^{-3} & 1.065 \times 10^{-3} \\ 1.065 \times 10^{-3} & 2.032 \times 10^{-3} \end{pmatrix}$$

$$dx := \frac{\overrightarrow{|\chi\chi^2 - \chi\chi^1|}}{\chi\chi} = \begin{pmatrix} 0.482 & 4.044 \\ 4.044 & 1.498 \end{pmatrix} \% \quad - \text{процент расхождения оценок (приемлемый, т.к. меньше 10\%)}$$

Рис. 4.4. Проверка гипотезы о постоянстве матрицы ковариаций

4. Корреляционная матрица приземных концентраций приведена на рис. 4.5 (для общности формулы укажем формулы для ее оценки при «опережающих» и «отстающих» сдвигах времени, хотя для расчета по (4.2) нужен только «опережающий» сдвиг на 1 такт):

$$\text{crr}_{w(t)} := \frac{1}{s_{\max} - |t|} \cdot \sum_{k=0}^{s_{\max}-|t|} \left(w_{0k+|t|}^0 \cdot w_{0k}^{0T} \right) \quad - \text{опережающий сдвиг на } t \text{ тактов}$$

$$\text{crrr}_{w(t)} := \frac{1}{s_{\max} - |t|} \cdot \sum_{k=0}^{s_{\max}-|t|} \left(w_{0k}^0 \cdot w_{0k+|t|}^{0T} \right) \quad - \text{отстающий сдвиг на } t \text{ тактов}$$

$$\text{CORR}_{w(t)} := \text{if}(t \geq 0, \text{crr}_{w(t)}, \text{crrr}_{w(t)}) \quad - \text{обобщающая формула для корреляционной матрицы}$$

Рис. 4.5. Формулы расчета корреляционной матрицы приземных концентраций

Сравните значения $CORR_w(0)$ и матрицы ковариаций $\chi\chi$ (должны совпасть).

Проверка гипотезы о независимости значений приземных концентраций во времени. Постройте графики зависимости элементов корреляционной матрицы от величины сдвига t . Примеры таких графиков, показанных на рис. 4.6, иллюстрируют спад корреляции практически до 0 за 1 такт, что позволяет принять гипотезу о независимости значения концентрации вредных веществ в данный момент времени от значений концентраций в другие моменты времени.

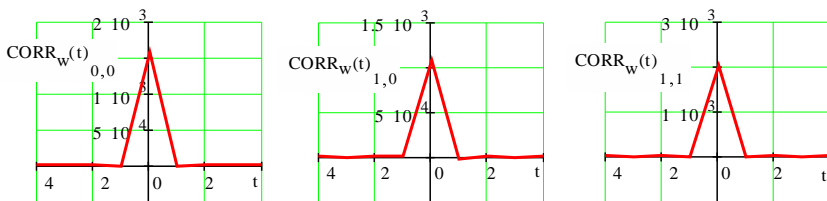


Рис. 4.6. Элементы корреляционной матрицы входного воздействия (приземных концентраций вредных веществ)

5. Оценка математического ожидания показателей экологического благополучия:

$$mm_X := \frac{1}{s_{\max} + 1} \cdot \sum_{k=0}^{s_{\max}} X_k. \quad (4.7)$$

Проверка гипотезы о постоянстве математического ожидания – аналогично п. 1 данного этапа.

6. Центрированная (математическим ожиданием) последовательность значений показателей экологического благополучия:

$$X0_s := X_s - mm_X. \quad (4.8)$$

7. Оценки корреляционной матрицы и (одновременно) матрицы ковариаций показателей экологического благополучия приведены на рис. 4.7:

$$\begin{aligned}
\text{crr}_X(t) &:= \frac{1}{s_{\max} - |t|} \cdot \sum_{k=0}^{s_{\max} - |t|} \left(\text{XO}_{k+|t|} \cdot \text{XO}_k^T \right) && \text{- опережающий сдвиг на } t \text{ тактов} \\
\text{crr}_X(t) &:= \frac{1}{s_{\max} - |t|} \cdot \sum_{k=0}^{s_{\max} - |t|} \left(\text{XO}_k \cdot \text{XO}_{k+|t|}^T \right) && \text{- отстающий сдвиг на } t \text{ тактов} \\
\text{CORR}_X(t) &:= \text{if}(t \geq 0, \text{crr}_X(t), \text{crr}_X(t)) && \text{- обобщающая формула для} \\
&&& \text{корреляционной матрицы}
\end{aligned}$$

Рис. 4.7. Формулы расчета корреляционной матрицы показателей экологического благополучия

Проверку гипотезы о постоянстве матрицы ковариаций и о зависимости значений корреляционной матрицы только от сдвига t сделаем дальше (по конечному результату разработки модели).

По корреляционной матрице можно оценить период времени, в течение которого прошлые значения вектора состояния сказываются на текущем значении (а текущее – на будущие значения состояний). Для этого нужно построить графики, как на рис. 4.8 и 4.9 (на графиках корреляция нормирована среднеквадратическими отклонениями соответствующих показателей).

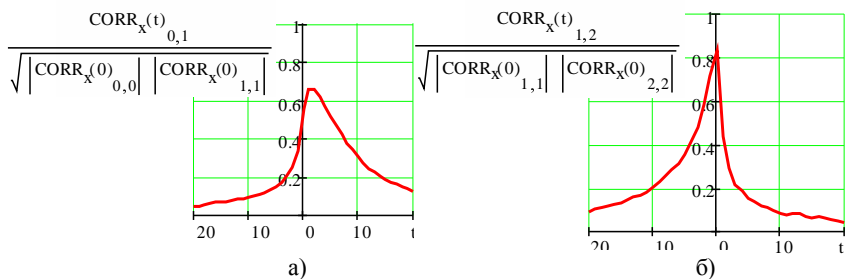


Рис. 4.8. Взаимные корреляции между показателями экологического благополучия: а) корреляция показателей «продолжительность жизни» и «число рождений»; б) корреляция показателей «число рождений» и «число дней нетрудоспособности».

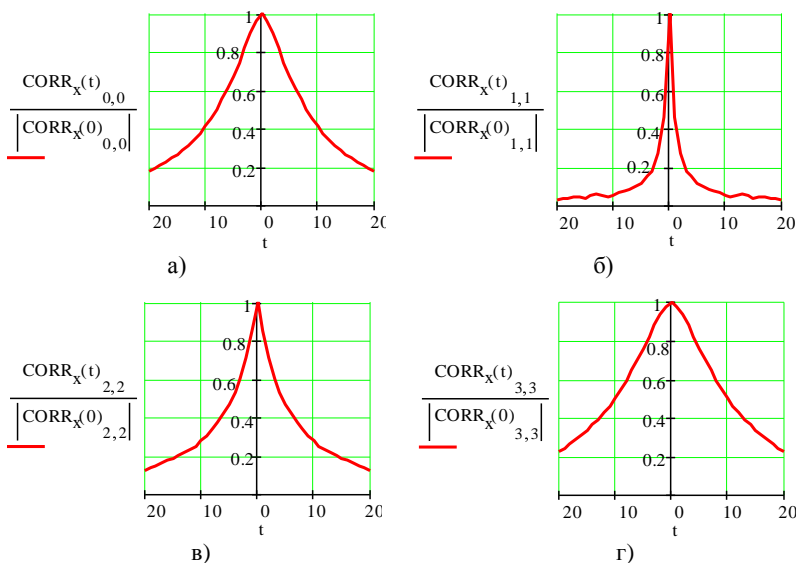


Рис. 4.9. Автокорреляционные функции показателей экологического благополучия: а) показатель «продолжительность жизни»; б) показатель «число рождений»; в) показатель «число дней нетрудоспособности»; г) показатель «кислотность грунтовых вод»

Корреляция считается *значимой*, если ее *нормированное* значение по абсолютной величине превышает 0.3. Используя это правило, заключаем, что период заметной временной связи между показателями различен; наиболее инерционна кислотность грунтовых вод (18 месяцев), наименее инерционен показатель «число рождений на 1000 жителей» (2 месяца).

Графики нормированных взаимных корреляционных функций (рис. 4.8) иллюстрируют сильную взаимозависимость показателей (0.7 ... 0.8 от максимально возможного значения 1).

Предварительная обработка исходных данных завершена.

Построение модели связи показателей экологического благополучия со значениями приземных концентраций вредных веществ

В результате выполнения п. 4.2.1 получены все необходимые данные для расчета параметров Φ и Γ модели (4.1) по формулам (4.2) и (4.3).

Соответствующие расчеты завершают выполнение п. 2 расчетной схемы из п. 4.2.

4.2.3. Проверка модели путем сравнения реконструированных (расчетным путем) значений показателей благополучия с фактическими (выполнение п. 3 схемы из п. 4.2)

Полагая, что модель (4.1) верна, рассчитаем статистики показателей благополучия по формулам вероятностного анализа и сравним с оценками, полученными при выполнении п. 4.2.1. Если согласование результатов будет удовлетворительным (расхождение до $\pm 10\%$), то модель можно будет считать приемлемой.

Рабочие формулы:

1. Математическое ожидание показателей благополучия:

$$m_{x_0} := X_0$$

$$m_{x_{s+1}} := \Phi \cdot m_{x_s} + \Gamma \cdot mm_w \quad (4.9)$$

Результаты расчетов показывают незначительное смещение (завышение) математических ожиданий (рис. 4.10), но это смещение не превышает допустимую величину. Есть возможность устранить смещение, см. п. 3 ниже.

$$\frac{m_{x_{s \max}} - mm_x}{mm_x} = \begin{pmatrix} 7.221 \\ 5.667 \\ 6.711 \\ 7.668 \end{pmatrix} \%$$

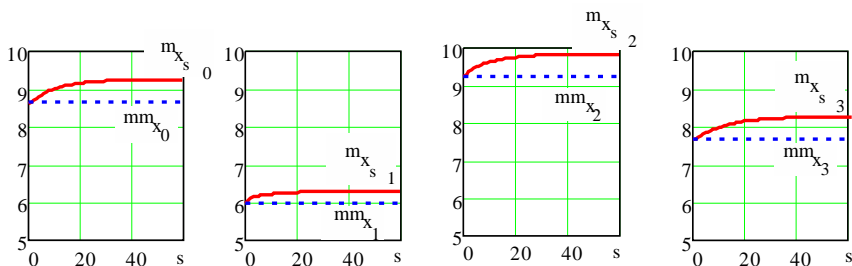


Рис. 4.10. Удовлетворительная точность расчета математического ожидания

2. Матрица ковариаций показателей благополучия:

$$\text{cov}_{X_0} := \text{CORR}_X(0)$$

$$\text{cov}_{X_{s+1}} := \Phi \cdot \text{cov}_{X_s} \cdot \Phi^T + \Gamma \cdot \chi \chi \cdot \Gamma^T \quad (4.10)$$

Расчеты по модели показывают хорошее совпадение с результатами обработки исходных данных (рис. 4.11):

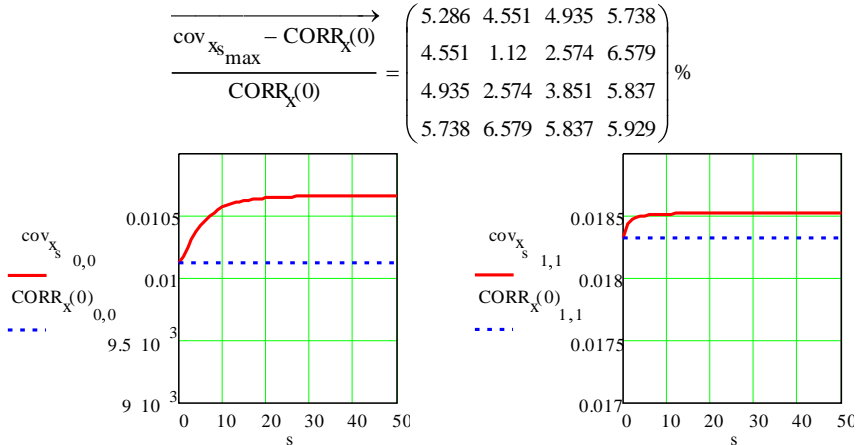


Рис. 4.11. Хорошее совпадение модели расчета матрицы ковариаций и результата расчета ее оценки по исходным данным

3. Реконструкция исходных данных путем расчета по модели из одних и тех же начальных условий с устранением смещения математического ожидания (рис. 4.12):

$$XX_0 := X_0 \quad \text{- совпадающие начальные условия}$$

$$XX_{s+1} := \Phi \cdot XX_s + \Gamma \cdot w_s \quad \text{- расчет показателей по модели}$$

$$XX_s := XX_s + mm_x - m_{x_s} \quad \text{- поправка на смещение математического ожидания}$$

Рис. 4.12. Устранение смещения математического ожидания

На рис. 4.13 для сравнения приведены графики исходных данных и их реконструкции по модели, иллюстрирующие хорошее совпадение. Если получится так, то можно считать, что все гипотезы, принятые при выборе модели, оправданы.

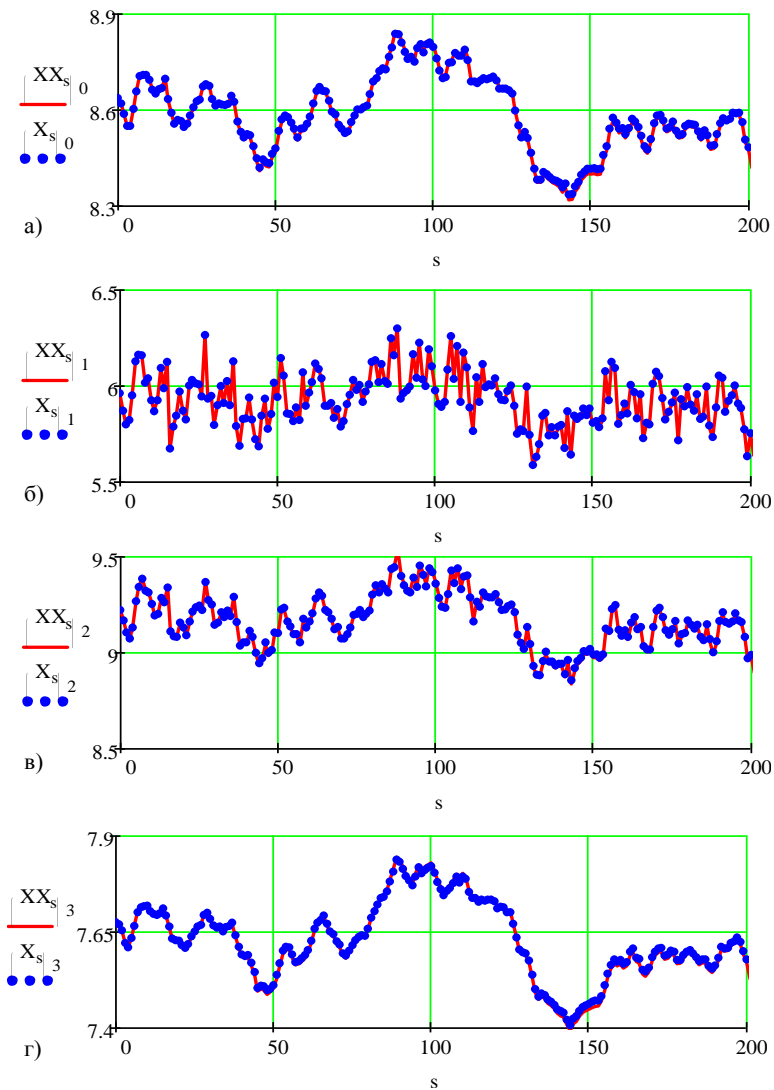


Рис. 4.13. Сравнение исходных данных о показателях экологического благополучия с данными, реконструированными по разработанной модели: а) «продолжительность жизни»; б) «количество рождений»; в) «число дней нетрудоспособности»; г) «кислотность грунтовых вод»

Выполнение п. 3 схемы из п. 4.2 завершено.

4.2.4. Прогноз динамики изменения показателей благополучия, ожидаемой после ввода нового предприятия в эксплуатацию (выполнение п. 4 схемы из п. 4.2)

Схема расчетов:

1. Расширяем диапазон тактов времени до нового значения $s_{\text{new_max}}$. Например:

$$s_{\text{new_max}} := s_{\text{max}} + 10000. \quad (4.11)$$

2. Для интервала «будущего» времени $s := s_{\text{max}} + 1 .. s_{\text{new_max}}$ генерируем новые значения приземных концентраций, которые ожидаются после ввода проектируемого предприятия. Для этого:

- Рассчитываем ожидаемые значения математического ожидания значений приземных концентраций (согласно исходным данным в 2 раза больше, чем до внедрения предприятия, формула (4.4)) и среднего квадратического отклонения этих значений (также в 2 раза больше, чем до внедрения). Соответственно дисперсия (диагональные элементы матрицы ковариаций, формула (4.6)) увеличится в 4 раза:

$$\begin{aligned} mm_{w_new} &:= 2mm_w \\ \chi\chi_{w_new_{0,0}} &:= 4xx_{0,0} \quad \chi\chi_{w_new_{1,1}} := 4xx_{1,1} \end{aligned} \quad (4.12)$$

- Рассчитываем недиагональные элементы матрицы ковариаций из условия равенства нормированных (соответствующими дисперсиями) коэффициентов корреляции между выбросами окисла серы и окисла углерода:

$$\begin{aligned} \chi\chi_{w_new_{0,1}} &:= \frac{xx_{0,1}}{\sqrt{xx_{0,0} \cdot xx_{1,1}}} \cdot \sqrt{\chi\chi_{w_new_{0,0}} \cdot \chi\chi_{w_new_{1,1}}} \\ \chi\chi_{w_new_{1,0}} &:= \chi\chi_{w_new_{0,1}} \end{aligned} \quad (4.13)$$

- Рассчитываем матрицу G линейного преобразования машинных случайных чисел в вектор w_{new_s} , имеющий математическое ожидание и матрицу ковариаций согласно (4.12) и (4.13). Матрица преобразователя рассчитывается итерационно (рис. 4.14), начиная

с произвольной матрицы (желательно, чтобы она была меньше произведения $\chi\chi_{w_new} \cdot \chi\chi_{w_new}^T$).

$\nu_{max} := 1000$ - максимальное число итераций

$\nu := 0.. \nu_{max}$ - счетчик итераций

$\lambda := 0.1$ - параметр, обеспечивающий сходимость итераций

$G_0 := 0.5 \cdot \chi\chi_{w_new} \cdot \chi\chi_{w_new}^T$ - начальное приближение

$G_{\nu+1} := G_{\nu} + \lambda \cdot (\chi\chi_{w_new} - G_{\nu} \cdot G_{\nu}^T)$ - итерационная процедура

Рис. 4.14. Расчет матрицы линейного преобразования G для вектора w_{new}

- Сходимость можно установить проверкой (разность $\chi\chi_{w_new} - G_{\nu_{max}} \cdot G_{\nu_{max}}^T$ должна быть близка к нулевой матрице). Удобно проверить сходимость по графикам, как на рис. 4.15.

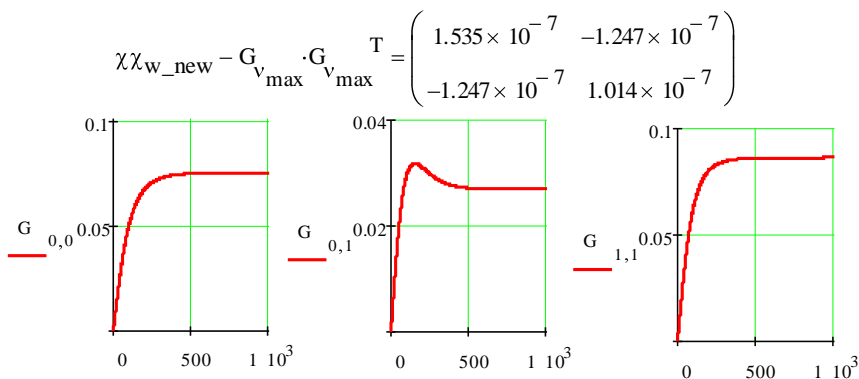


Рис. 4.15. Иллюстрация сходимости итерационного расчета преобразователя G

- Формируем последовательность значений приземных концентраций, которая ожидается после ввода нового предприятия в действие (рис. 4.16).

$$s := s_{\max} + 1 \dots s_{\text{new_max}}$$

- счетчик времени

$$w0_{\text{new}_s} := G_{V_{\max}} \cdot \begin{pmatrix} \text{norm}(1,0,1)_0 \\ \text{norm}(1,0,1)_0 \end{pmatrix}$$

- случайный вектор с рассчитанной матрицей ковариаций, формируемый из машинных случайных чисел с нулевым мат. ожиданием, единичной дисперсией и отсутствием корреляции во времени

$$w_{\text{new}_s} := w0_{\text{new}_s} + \text{num}_{w_{\text{new}}}$$

- результат генерации значений приземных концентраций, имитирующей новую экологическую ситуацию

Рис. 4.16. Формирование значений приземных концентраций после ввода нового предприятия

Проверка качества расчета преобразователя G – по факту близости оценки матрицы ковариаций:

$$\text{cov}_{w_{\text{new}}} := \frac{1}{s_{\text{new_max}} - s_{\max} - 1} \cdot \sum_{k=s_{\max}+1}^{s_{\text{new_max}}} \left(w0_{\text{new}_k} \cdot w0_{\text{new}_k}^T \right)$$

к значению $\chi\chi_{w_{\text{new}}}$.

3. Используем результаты генерации будущих значений приземных концентраций, чтобы получить имитатор будущих значений вектора показателей экологического благополучия XX_s (рис. 4.17).

$$s := 0 \dots s_{\text{new_max}}$$

- счетчик времени

$$XX_0 := X_0$$

- начальные условия

Продление траектории значений вектора экологического благополучия на период эксплуатации нового предприятия

$$XX_{s+1} := \text{if} \left(s \leq s_{\max}, \Phi \cdot XX_s + \Gamma \cdot w_s, \Phi \cdot XX_s + \Gamma \cdot w_{\text{new}_s} \right)$$

Рис. 4.17. Формирование будущих значений показателей экологического благополучия после ввода нового предприятия

4. Строим графики, иллюстрирующие динамику изменения показателей экологического благополучия, рассчитываем статистики этих показателей для новых условий по формулам, аналогичным (4.7), (4.8) и рис. 4.7, оформляем экспертное заключение и направляем его Заказчику (в данном случае – преподавателю). Примеры графиков показаны на рис. 4.18 и 4.19.

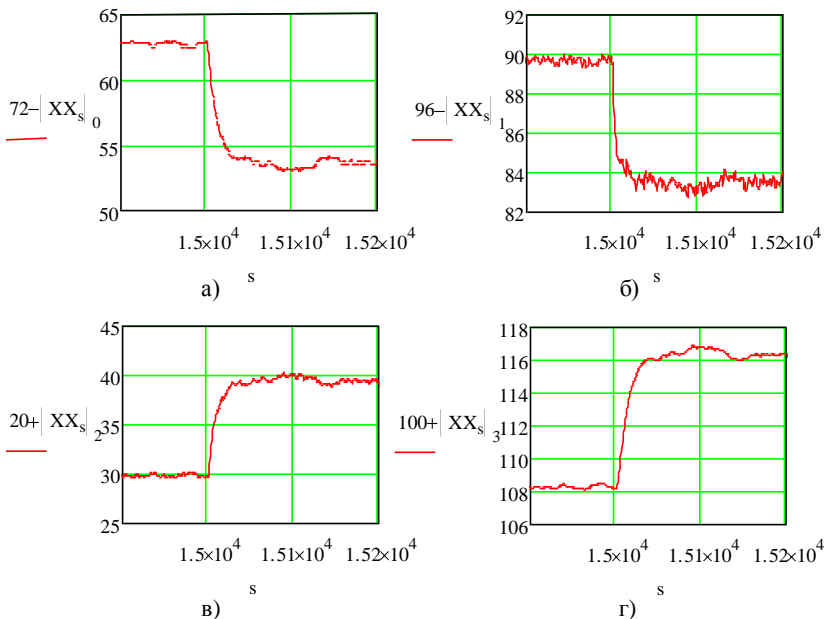


Рис. 4.18. Изменения показателей экологического благополучия, ожидаемых после ввода нового предприятия в эксплуатацию: а) показатель «продолжительность жизни»; б) показатель «число рождений»; в) показатель «число дней нетрудоспособности»; г) показатель «кислотность грунтовых вод»

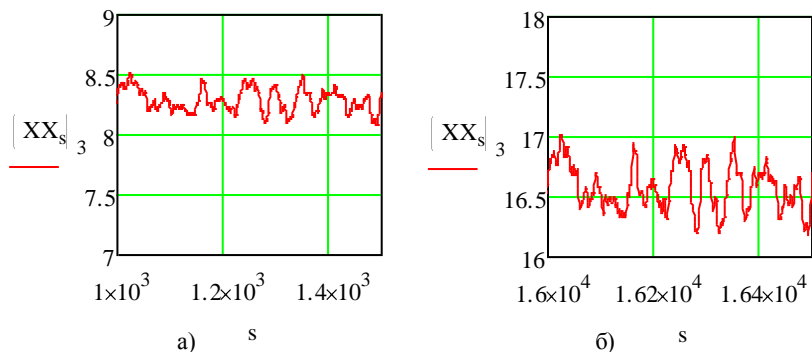


Рис. 4.19. Иллюстрация увеличения степени разброса значений показателей экологического благополучия (показатель «кислотность грунтовых вод»): а) до ввода нового предприятия в эксплуатацию; б) после ввода нового предприятия в эксплуатацию

Задание

Составить отчет по данной работе, который должен содержать следующее:

1. Иметь все расчеты, необходимые для ответа на вопросы, поставленные в п. 4.1.
2. Содержать ответы на вопросы Заказчика.
3. Содержать комментарии к результатам расчета.
4. Иметь дополнительные (к требуемым Заказчиком) выводы, которые Вам удалось сделать.

4.3. Контрольные вопросы

1. Какие этапы входят в схему проведения экспертизы в данной лабораторной работе?
2. Как выполняется проверка того, что вектор приземных концентраций вредных веществ является чисто случайной последовательностью?
3. Как проверить гипотезу о стационарности статистик приземных концентраций вредных веществ?
4. Как проверить гипотезу о постоянстве матрицы ковариации?
5. По графикам рис. 4.9 определите период заметной временной связи между показателями экологического благополучия.
6. Почему после внедрения нового предприятия дисперсия приземных концентраций увеличится в 4 раза?

Список использованной и рекомендуемой литературы

1. Блаттнер, П. Использование Microsoft Office Excel 2003. Специальное издание / П. Блаттнер. – Пер. с англ. – М.: Издательский дом «Вильямс», 2005. – 864 с.
2. Орлова, И. В. Экономико-математические методы и модели: компьютерное моделирование: учеб. пособие / И. В. Орлова, В. А. Половников. – 3-е изд., перераб. и доп. – М.: Вузовский учебник: ИНФРА-М, 2012. – 478 с.
3. Федосеев, В. В. Экономико-математические методы и прикладные модели: учеб. для бакалавров / В. В. Федосеев, А. Н. Гармаш, И. В. Орлова; под ред. В. В. Федосеева. – 3-е изд., перераб. и доп. – М.: Издательство Юрайт, 2012. – 328 с.
4. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.
5. MachineLearning.ru - Профессиональный информационно-аналитический ресурс. Статьи категории «Прикладная статистика» [Электронный ресурс]. – Режим доступа: [http://www.machinelearning.ru/wiki/index.php?title=Категория: Прикладная_статистика,свободный](http://www.machinelearning.ru/wiki/index.php?title=Категория:Прикладная_статистика,свободный). – Загл. с экрана.
6. Сайт «OilCapital.ru». Цены на нефть [Электронный ресурс]. – Режим доступа: http://www.oilcapital.ru/stat/stat_1/stat_1.shtml,свободный. – Загл. с экрана.

ПРИЛОЖЕНИЕ 1. ФОРМУЛЫ РАСЧЕТА НЕКОТОРЫХ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК

Среднее значение (оценка математического ожидания) – среднее арифметическое значение для выборки

$$m = \frac{1}{N} \cdot \sum_{i=1}^N x_i,$$

где N – объем выборки.

Медиана – значение элемента отсортированной выборки, который делит ее на две равные части

$$med = \begin{cases} x_{k+1}, & \text{при нечетном } N = 2 \cdot k + 1 \\ \frac{x_k + x_{k+1}}{2}, & \text{при четном } N = 2 \cdot k \end{cases}$$

Мода – значение элемента выборки, которой имеет наибольшую частоту

$$mod = x_j, \quad \text{при } j = \arg(\max(n_j)),$$

где n_j – частота j -го значения в выборке.

Дисперсия выборки – показывает разброс случайной величины от ее среднего значения (имеет размерность равную квадрату размерности случайной величины). Формула для оценки дисперсии по экспериментальным данным:

$$D = \frac{1}{N-1} \cdot \sum_{i=1}^N (x_i - m)^2.$$

Стандартное отклонение – показывает разброс случайной величины от ее среднего значения (имеет размерность равную размерности случайной величины)

$$\sigma = \sqrt{D}.$$

Стандартная ошибка (среднего) – величина отклонения среднего значения от истинного математического ожидания:

$$Em = \sqrt{\frac{D}{N}} = \frac{\sigma}{\sqrt{N}}.$$

Эксцесс – оценка «крутости» («остроконечности», подъема кривой) функции плотности распределения данных по сравнению с функцией нормального распределения:

$$E = \frac{\mu_4}{\sigma^4} - 3,$$

где μ_4 – центральный момент 4-го порядка:

$$\mu_4 = \frac{(N^2 - 2 \cdot N + 3) \cdot \sum_{i=1}^N (x_i - m)^4 - 3 \cdot (2 \cdot N - 3) \cdot \sum_{i=1}^N (x_i - m)^2}{(N-1) \cdot (N-2) \cdot (N-3)}.$$

Асимметричность – оценка «кособокости» (несимметричности) функции плотности распределения данных по сравнению с функцией нормального распределения:

$$A = \frac{\mu_3}{\sigma^3},$$

где μ_3 – центральный момент 3-го порядка

$$\mu_3 = \frac{N}{(N-1) \cdot (N-2)} \cdot \sum_{i=1}^N (x_i - m)^3.$$

Интервал – разброс значений от минимума до максимума
 $int = max - min$.

Уровень надежности (среднего) – величина половины доверительного интервала для значения среднего при заданной надежности:

$$E(\gamma) = \beta_{1-\gamma/2} \cdot \sqrt{\frac{D}{N}},$$

где β – квантиль уровня значимости $1 - \gamma / 2$ стандартного нормального распределения (с параметрами $m = 0, D = 1$), если известна дисперсия; если дисперсия неизвестна (используется ее оценка), то в качестве β следует использовать квантиль распределения Стьюдента того же уровня значимости с $N - 1$ степенями свободы.

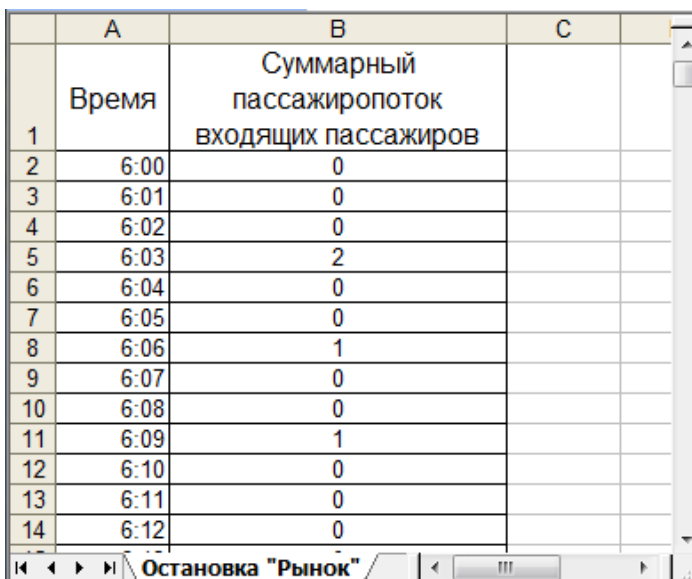
ПРИЛОЖЕНИЕ 2.

СТРУКТУРА ФАЙЛА С ЭКСПЕРИМЕНТАЛЬНЫМИ ДАНЫМИ ДЛЯ ЛАБОРАТОРНЫХ РАБОТ № 1, 2

Экспериментальные данные для лабораторных работ № 1, 2 содержат сведения о количестве пассажиров, входящих в автобус на некоторой автобусной остановке.

Данные находятся на одном листе файла (книги) формата Microsoft Excel и имеют следующую структуру (рис. П2.1):

- в столбце «А» содержатся отсчеты времени. В примере время изменяется с 6:00 до 21:00 с шагом 1 минута;
- в столбце «В» содержатся значения суммарного пассажиропотока входящих в автобусы пассажиров, относящихся к соответствующему отсчету времени (количество пассажиров, вошедших в автобусы в течение текущего интервала времени).



	А	В	С
	Время	Суммарный пассажиропоток входящих пассажиров	
1			
2	6:00	0	
3	6:01	0	
4	6:02	0	
5	6:03	2	
6	6:04	0	
7	6:05	0	
8	6:06	1	
9	6:07	0	
10	6:08	0	
11	6:09	1	
12	6:10	0	
13	6:11	0	
14	6:12	0	

Рис. П2.1. Фрагмент файла с данными для лабораторных работ № 1, 2

ПРИЛОЖЕНИЕ 3. СТРУКТУРА ФАЙЛА С ЭКСПЕРИМЕНТАЛЬНЫМИ ДАННЫМИ ДЛЯ ЛАБОРАТОРНОЙ РАБОТЫ № 3

Экспериментальные данные для лабораторной работы № 3 содержат сведения по динамике стоимости нефти на мировом рынке (данные взяты с сайта OilCapital.ru [6]).

Данные находятся на одном листе файла (книги) формата Microsoft Excel и имеют следующую структуру (рис. ПЗ.1):

- в столбце «А» содержатся даты анализируемого периода времени. В примере период дат охватывает интервал с марта 2009 по сентябрь 2010 с периодом 1 неделя;
- в столбце «В» содержатся значения цены нефти (доллар США за баррель), относящиеся к соответствующей дате из столбца «А».

	A	B	C	D	E
1	Цена на нефть, доллар за баррель				
2	02.03.2009	44,59			
3	09.03.2009	45,65			
4	16.03.2009	51,4			
5	23.03.2009	51,6			
6	30.03.2009	53,55			
7	06.04.2009	55,02			
8	13.04.2009	53,3			
9	20.04.2009	51,59			
10	27.04.2009	52,65			
11	04.05.2009	58,21			
12	11.05.2009	56,1			
13	18.05.2009	60,63			
14	25.05.2009	65,8			
15	01.06.2009	68,5			
16	08.06.2009	71,75			
17	15.06.2009	69,3			

Рис. ПЗ.1. Фрагмент файла с данными для лабораторной работы № 3

ПРИЛОЖЕНИЕ 4. СТРУКТУРА ФАЙЛА С ЭКСПЕРИМЕНТАЛЬНЫМИ ДАННЫМИ ДЛЯ ЛАБОРАТОРНОЙ РАБОТЫ № 4

Экспериментальные данные для лабораторной работы № 4 содержат сведения о показателях экологического благополучия города и приземных концентраций вредных веществ.

Данные находятся на текстовом файле (рис. П4.1), в виде, пригодном для загрузки в среду Mathcad (для этого данные предварительно были сформированы в среде Mathcad в форме матриц и записаны в текстовый файл).

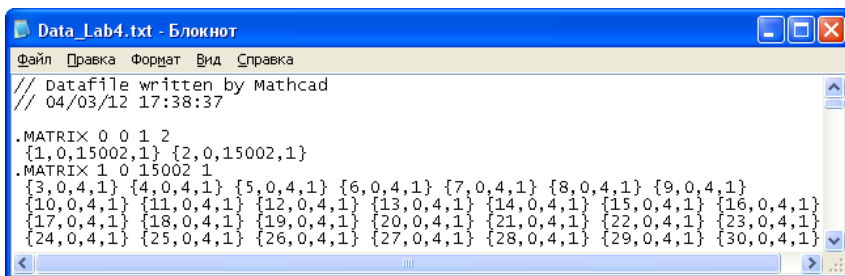


Рис. П4.1. Фрагмент файла с данными для лабораторной работы № 4

Чтение (загрузка) файла с данными в среду Mathcad показано на рис. П4.2.

$$Z := \text{READPRN}(\text{"Data_Lab4.txt"}) \quad Z = (\{15002,1\} \quad \{15002,1\})$$

"Распаковка" исходных данных

$$X := Z_{0,0}$$

$$X =$$

	0
0	[4, 1]
1	[4, 1]
2	[4, 1]
3	...

$$w := Z_{0,1}$$

$$w =$$

	0
0	[2, 1]
1	[2, 1]
2	[2, 1]
3	...

$$X_0 = \begin{pmatrix} 8.638 \\ 5.964 \\ 9.222 \\ 7.676 \end{pmatrix} \quad \begin{array}{l} \text{Показатели} \\ \text{экологического} \\ \text{благополучия для} \\ \text{0-го месяца} \end{array} \quad w_0 = \begin{pmatrix} 0.374 \\ 0.865 \end{pmatrix} \quad \begin{array}{l} \text{Концентрация} \\ \text{вредных} \\ \text{веществ для} \\ \text{0-го месяца} \end{array}$$

Рис. П4.2. Чтение файла с данными в среде Mathcad

После чтения файла с данными в среде Mathcad будет получена матрица размером 1×2 , где первый компонент (X) – матрица показателей экологического благополучия, второй компонент (w) – матрица концентраций вредных веществ. Компонентами матриц X и w являются вектора с соответствующими значениями, относящимися к определенному номеру месяца (s).

Вектор показателей экологического благополучия X_s содержит 4 компоненты:

- $(X_s)_0$ – отклонение средней продолжительности жизни (в годах) от среднеевропейской (72 года);
- $(X_s)_1$ – отклонение среднего количества рождений детей на 1000 жителей от среднего показателя по стране (96 рождений);
- $(X_s)_2$ – отклонение среднего числа дней нетрудоспособности в году от норматива, исходя из которого планируются медицинские услуги и лекарственное обеспечение (20 дней в году);
- $(X_s)_3$ – отклонение уровня кислотности грунтовых вод от нормы (норма принимается за 100 %).

Вектор значений приземных концентраций вредных веществ w_s содержит 2 компоненты:

- $(w_s)_0$ – приземная концентрация сернистого газа (мг в 10 м^3 воздуха);
- $(w_s)_1$ – приземная концентрация окисла углерода (мг в 10 м^3 воздуха).

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.....	3
ЛАБОРАТОРНАЯ РАБОТА № 1. ПЕРВИЧНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ.....	6
1.1. Введение. Средства статистической обработки в Microsoft Excel	6
1.2. Расчет статистических характеристик экспериментальных данных	11
1.3. Построение гистограммы распределения частот посадки пассажиров	15
1.4. Сглаживание экспериментальных данных.....	17
1.5. Группировка исходных данных	19
1.6. Контрольные вопросы.....	20
ЛАБОРАТОРНАЯ РАБОТА № 2. ВРЕМЕННЫЕ РЯДЫ. ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ	21
2.1. Понятие временных рядов.....	21
2.2. Структурный состав временного ряда.....	23
2.3. Этапы построения прогноза по временным рядам.....	25
2.4. Предварительный анализ данных временного ряда.....	25
2.5. Контрольные вопросы.....	37
ЛАБОРАТОРНАЯ РАБОТА № 3. ВРЕМЕННЫЕ РЯДЫ. ВЫБОР ТРЕНДОВОЙ МОДЕЛИ И ОЦЕНКА ЕЕ АДЕКВАТНОСТИ.....	38
3.1. Этапы построения прогноза по временным рядам.....	38
3.2. Формирование набора аппроксимирующих функций (кривых роста).....	39
3.3 Численное оценивание параметров моделей.....	44
3.4. Проверка адекватности моделей и оценка их точности.....	48
3.5. Контрольные вопросы.....	60
ЛАБОРАТОРНАЯ РАБОТА № 4. ВЕРОЯТНОСТНЫЙ АНАЛИЗ ДИНАМИЧЕСКОЙ СИСТЕМЫ И ЕГО ПРИМЕНЕНИЕ	61
4.1. Содержательная постановка задачи.....	61
4.2. Схема проведения экспертизы, составляющей тему лабораторной работы	63

4.3. Контрольные вопросы.....	80
Список использованной и рекомендуемой литературы	80
ПРИЛОЖЕНИЕ 1. Формулы расчета некоторых статистических характеристик.....	81
ПРИЛОЖЕНИЕ 2. Структура файла с экспериментальными данными для лабораторных работ № 1, 2	83
ПРИЛОЖЕНИЕ 3. Структура файла с экспериментальными данными для лабораторной работы № 3	84
ПРИЛОЖЕНИЕ 4. Структура файла с экспериментальными данными для лабораторной работы № 4	85