Министерство науки и высшего образования Российской Федерации
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

О.К. АЛЬСОВА

ИССЛЕДОВАНИЕ ВРЕМЕННЫХ РЯДОВ В СРЕДЕ R

Утверждено редакционно-издательским советом университета
в качестве учебного пособия

НОВОСИБИРСК
2021

УДК 519.246.8:004(075.8)
А 579

Рецензенты:

Ю.А. Котов, канд. физ.-мат. наук, доцент

А.В. Гаврилов, канд. техн. наук, доцент

Работа подготовлена на кафедре вычислительной техники
для студентов и магистрантов АВТФ по дисциплинам
«Методы анализа данных», «Компьютерные технологии анализа
и обработки данных», «Интеллектуальный анализ данных
и методы машинного обучения»

Альсова О.К.

А 579 Исследование временных рядов в среде R: учебное пособие / О.К. Альсова. – Новосибирск: Изд-во НГТУ, 2021. – 88 с.

ISBN 978-5-7782-4337-8

В пособии рассмотрены вопросы, связанные с решением задач исследования и прогнозирования временных рядов средствами языка и среды статистических вычислений R. В качестве математического аппарата используются классические параметрические вероятностно-статистические модели и методы анализа временных рядов.

Для каждого метода дано краткое теоретическое описание, позволяющее понять его суть и особенности применения, и приведено описание основных функций языка R, реализующих метод.

Основное внимание в пособии уделено рассмотрению технологии (методики) исследования и прогнозирования временного ряда с помощью среды R. На конкретных примерах рассматриваются вопросы идентификации, анализа адекватности, сравнения и окончательного выбора модели временного ряда.

Предназначено для бакалавров IV курса АВТФ, обучающихся по направлениям 09.03.01 «Информатика и вычислительная техника», 09.03.04 «Программная инженерия» и для магистрантов 1–2-го года обучения, обучающихся по направлениям 09.04.01 «Информатика и вычислительная техника», 09.04.04 «Программная инженерия».

УДК 519.246.8:004(075.8)

ISBN 978-5-7782-4337-8

ВВЕДЕНИЕ

Необходимость решения задач идентификации и прогнозирования временных рядов возникает во многих прикладных областях науки и техники и связана с моделированием естественных и искусственных процессов (объектов). Информация об исследуемом процессе часто представлена в виде дискретного ряда зафиксированных в определенные равноотстоящие моменты времени значений показателя (признака), описывающего процесс. Такой ряд значений называется временным рядом (ВР). Процедура нахождения модели, наиболее адекватно описывающей исследуемый временной ряд, называется идентификацией модели временного ряда. Идентификационная модель исследуемого процесса (объекта) может быть использована для изучения и описания его свойств и особенностей функционирования в различных условиях, а также для определения его прошлых и будущих состояний, т. е. для прогнозирования. Задача прогнозирования будущих показателей процессов на основе их исторических значений является основой для финансового планирования в экономике, прогнозов погоды в метеорологии, для планирования и оптимизации деятельности компаний и производств и т. д. Также прогнозирование является одной из основных задач, которые решаются в рамках динамично развивающегося в настоящее время направления в обработке информации – интеллектуального анализа данных (Data Mining).

В теоретических и прикладных исследованиях рассматривают широкий спектр моделей и методов идентификации и прогнозирования временных рядов. Наиболее распространенные методы анализа временных рядов можно разделить на два основных класса – это параметрические и непараметрические методы. При использовании параметрических методов предполагают, что исследуемый процесс имеет определенную структуру, которую можно описать с помощью аналитической

математической модели, имеющей сравнительно небольшое число параметров, и задача идентификации состоит в том, чтобы определить структуру модели и оценить ее параметры. Параметры модели и модель в целом имеют четкую содержательную интерпретацию в терминах предметной области и описывают закономерности изменения во времени исследуемого процесса. В случае использования непараметрических методов отсутствует свернутое аналитическое параметрическое описание модели, что затрудняет или делает невозможным содержательную интерпретацию модели.

Среди параметрических методов выделяют временные (автокорреляционная и частная автокорреляционная функции, модели авторегрессии и скользящего среднего, модели экспоненциального сглаживания) и частотные (спектральный анализ на основе спектральных функций, гармонический анализ). К непараметрическим методам относятся, например, сингулярный спектральный анализ, нейросетевые модели, генетические алгоритмы, экспертные методы.

В настоящем учебном пособии рассматриваются только параметрические методы исследования временного ряда. Для каждого метода даны его краткое теоретическое описание, идея метода, математическая модель, лежащая в основе метода, условия использования и область применения. Более подробное описание классической теории анализа временных рядов можно найти в работах основоположников этого направления (Дж. Бокс и Г. Дженкинс [12], Т. Андерсон [11], М. Кендэл [18], Д. Бриллинджер [14], Э. Хеннан [27]) и в многочисленной учебной литературе [6, 7, 9, 17, 24, 30, 31].

В настоящем пособии основной упор сделан на практическое применение методов анализа временных рядов. Алгоритмы методов рассмотрены на примере решения конкретной задачи исследования временного ряда, оценены адекватность и точность построенных идентификационных моделей ВР, выполнен сравнительный анализ моделей. В качестве программной среды реализации алгоритмов исследования ВР выбран язык R и среда R, R-Studio.

Язык R – интерпретируемый язык программирования и среда для статистических вычислений и графического анализа с открытым исходным кодом [32], широко используются как статистическое программное обеспечение, поддерживаются большим и активным исследовательским сообществом по всему миру и фактически стали стандартом в области анализа данных. В языке R реализованы все

актуальные методы статистического анализа данных [16, 20, 22, 23, 25, 29], а также множество специфических алгоритмов для решения узкоспециализированных задач из разных предметных областей. К достоинствам R относится возможность создания графиков высокого качества, которые могут быть экспортированы в основные графические форматы и далее использоваться в презентациях и научных публикациях.

Функции языка R объединяются в пакеты – загружаемые модули, которые подключаются к любой программе и предоставляют объединенные в них вычислительные средства. Причем, пакеты для R могут разрабатываться и на других языках программирования. В целом, как язык программирования, R довольно прост и имеет ограниченные изобразительные средства, что компенсируется возможностью неограниченного его расширения с помощью пакетов. В базовую поставку R включен основной набор пакетов, а всего по состоянию на сентябрь 2020 года доступно более 16 200 пакетов [33]. Кроме того, язык R интегрирован в профессиональные статистические пакеты, такие как Statistica, SPSS, SAS, что позволяет запускать код R в оболочке пакета.

1. АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

1.1. Временной ряд. Основные определения и понятия. Разложение временного ряда на составляющие

Под *анализом временных рядов* (ВР) понимают процесс применения математико-статистических методов и методов машинного обучения для выявления закономерностей в поведении ВР, определения структуры ВР и прогнозирования значений ВР на будущие периоды.

Временным рядом называют последовательность наблюдений анализируемого показателя (признака) Y, упорядоченных во времени [7]. Как правило, при решении практических задач рассматривают дискретные (по времени наблюдения) временные ряды, в которых значения показателя фиксируются в равноотстоящие моменты времени, через заданный временной такт (секунда, минута, месяц, квартал, год и т. п.). В этом случае ВР представляется в виде

$$y_1, y_2, \ldots, y_n,$$

где y_t – значение исследуемого показателя, зафиксированное в t-м такте времени ($t = 1, 2, \ldots, n$).

В теории моделирования принято рассматривать временной ряд как упорядоченную последовательность наблюдений анализируемых случайных величин $Y(t_1),\ Y(t_2), \ldots, Y(t_n),\ t_i < t_{i+1}$, произведенных в последовательные моменты времени $t_1,\ t_2, \ldots, t_n:\ y_i = Y(t_i),\ i = 1, 2, \ldots, n$ [7].

Также временной ряд можно интерпретировать как наблюдения над непрерывным случайным процессом $Y(t)$ в моменты времени $t = t_i$ и рассматривать ВР как одну из реализаций случайного процесса.

Принципиальные отличия временного ряда от случайной выборки заключаются в следующем: во-первых, члены ВР не являются статистически независимыми; во-вторых, члены ВР не являются одинаково распределенными, т. е. $P\{y_1 < y\} \neq P\{y_2 < y\}$ при $t_1 \neq t_2$.

Взаимозависимость членов временного ряда позволяет применять специфический математический аппарат для построения прогнозных моделей ВР, основанный на выявлении и описании корреляционных взаимосвязей между членами ВР.

Один из подходов к идентификации ВР заключается в его разложении на детерминированные и случайные составляющие, каждая из которых описывает вклад определенного типа факторов в формирование значений ВР.

В общем виде аддитивная модель разложения ВР на составляющие задается следующим образом [7]:

$$Y(t_i) = F(t_i) + S(t_i) + C(t_i) + \varepsilon(t_i), \qquad (1.1)$$

где $F(t_i)$ – трендовая составляющая (компонента); $S(t_i)$ – сезонная составляющая; $C(t_i)$ – циклическая составляющая; $\varepsilon(t_i)$ – случайная составляющая.

Трендовая составляющая $F(t_i)$ описывает вклад долговременных факторов в формирование значений ВР, определяющих устойчивые закономерности в изменении наблюдаемого процесса в течение длительного интервала времени. Обычно тренд описывается неслучайной функцией, зависящей от времени (аргумент функции), часто монотонного характера.

Сезонная составляющая $S(t_i)$ описывает влияние сезонных факторов, которые обусловливают периодические колебания значений ВР в течение года. Сезонность характерна для многих природных и экономических процессов (например, изменение климатических и метеорологических показателей в течение года, сезонность спроса на товары и услуги, колебания объемов производства, материальных запасов и т. п.). Для описания сезонной составляющей используют тригонометрические функции (гармоники).

Циклическая составляющая $C(t_i)$ описывает влияние длительных (более одного сезона) периодически изменяющихся факторов экономической, астрофизической, демографической природы (например, циклы

солнечной активности, циклы экономического развития и т. п.). Циклические колебания, как и сезонные, математически описываются с помощью тригонометрических функций, отличие только в длине периода колебаний.

Случайная составляющая $\varepsilon(t_i)$ отражает воздействие случайных факторов, которые не поддаются учету и регистрации. Их воздействие как раз и определяет стохастическую природу элементов ВР и необходимость их интерпретации как наблюдений над случайными величинами.

Возможна также мультипликативная модель разложения ВР, в которой ВР представлен как произведение составляющих:

$$Y(t_i) = F(t_i)S(t_i)C(t_i)\varepsilon(t_i). \qquad (1.2)$$

В разложениях (1.1), (1.2) ВР могут присутствовать не все составляющие, обязательным является только наличие случайной компоненты $\varepsilon(t_i)$. Выводы о влиянии того или иного типа факторов на формирование значений ВР, о наличии определенных составляющих в разложении ВР делаются как на основе априорного содержательного анализа изучаемого процесса, так и по результатам статистического анализа исследуемого ВР.

1.2. Статистические характеристики временного ряда

Из определения ВР следует, что в каждый момент времени t_i величина $Y(t_i)$ является случайной и подчиняется некоторому вероятностному закону распределения. Для описания ВР используются те же числовые характеристики, что и для определения случайной величины. Так, математическое ожидание и дисперсия ВР в момент времени t_i определяются выражениями:

$$M(Y(t_i)) = m(t_i), \quad D(Y(t_i)) = D(t_i) = \sigma^2(t_i). \qquad (1.3)$$

Временные ряды классифицируют на два больших класса стационарных и нестационарных ВР.

Ряд называют *строго стационарным* (или стационарным в узком смысле), если совместное распределение вероятностей m наблюдений $y_1, y_2, \ldots, y_m$ такое же, как и для m наблюдений $y_{1+\tau}, y_{2+\tau}, \ldots, y_{m+\tau}$,

при любых m и τ. Другими словами, если для каждого момента времени t_i случайные величины $Y(t_i)$ имеют одинаковое распределение [7].

Ряд называют *стационарным в широком смысле*, если статистические характеристики случайных величин $Y(t_i)$ не зависят от времени. Очевидно, что из стационарности в узком смысле следует стационарность в широком смысле. Обратное в общем случае неверно. В дальнейшем изложении будут рассматриваться только стационарные временные ряды в широком смысле. Статистические характеристики стационарного ВР не меняются во времени, т. е. постоянны математическое ожидание и дисперсия на всем интервале наблюдения:

$$M\left(Y(t_i)\right) = m, \quad D\left(Y(t_i)\right) = D = \sigma^2. \tag{1.4}$$

Выборочные аналоги математического ожидания и дисперсии – соответственно среднее значение (оценка математического ожидания) и выборочная дисперсия (оценка дисперсии) – рассчитываются по формулам:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{1.5}$$

Кроме того, для описания ВР используют специфические характеристики, а именно автокорреляционную и частную автокорреляционную функции (АКФ и ЧАКФ).

Автокорреляционная функция описывает степень взаимосвязи между последовательными наблюдениями ВР: $Y(t_1)$, $Y(t_2)$, …, $Y(t_n)$ сдвинутыми относительно друг друга на l тактов времени (или, как говорят, с лагом l):

$$r(l) = \frac{M\left[\left(Y(t_i) - m\right)\left(Y(t_{i+l}) - m\right)\right]}{\sigma^2}. \tag{1.6}$$

Значение автокорреляционной функции для стационарного ВР зависит только от величины лага l, т. е. $r(l) = r(-l)$. Если $l = 0$, то $r(0) = 1$.

Оценка автокорреляционной функции (выборочная АКФ) рассчитывается по формуле

$$\hat{r}(l) = \frac{\frac{1}{n-l}\sum_{i=1}^{n-l}\left[(y_i - \bar{y})(y_{i+l} - \bar{y})\right]}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{1.7}$$

Фактически формула (1.7) соответствует линейному коэффициенту корреляции Пирсона, который вычисляется для каждого лага l.

Частная автокорреляционная функция измеряет автокорреляцию, существующую между разделенными l тактами времени членами временного ряда $Y(t_i)$ и $Y(t_{i+l})$ при устраненном опосредованном влиянии на эту взаимосвязь всех промежуточных членов этого ВР:

$$\rho(l) = r_{t,t-l(t-1,\,t-2,\,\ldots)} = -\frac{A_{t,t-l}}{\sqrt{A_{tt}\,A_{t-l,t-l}}}, \tag{1.8}$$

где $A_{t,t-l}$, A_{tt}, $A_{t-l,t-l}$ – алгебраические дополнения матрицы R автокорреляционных коэффициентов:

$$R = \begin{pmatrix} 1 & r(1) & \ldots & r(l) \\ r(1) & 1 & \ldots & r(l-1) \\ \ldots & \ldots & \ldots & \ldots \\ r(l) & r(l-1) & \ldots & 1 \end{pmatrix}. \tag{1.9}$$

Оценка частной автокорреляционной функции $\hat{\rho}(l)$ вычисляется по аналогичной формуле с заменой теоретических величин на их соответствующие выборочные значения:

$$\hat{\rho}(l) = \hat{r}_{t,t-l(t-1,t-2,\ldots)} = -\frac{\hat{A}_{t,t-l}}{\sqrt{\hat{A}_{tt}\,\hat{A}_{t-l,t-l}}}. \tag{1.10}$$

Еще одно важное понятие, которое используют в анализе временных рядов, – это понятие *белого шума*. Под белым шумом понимают стационарный временной ряд, у которого математическое ожидание равно нулю, а величины $\varepsilon(t_i)$ некоррелированы, т. е. автокорреляционная и частная автокорреляционная функции нулевые: $r(l) = 0$ и $\rho(l) = 0$ при $l > 0$. Если $l = 0$, то $r(0) = 1$ и $\rho(0) = 1$.

1.3. Подготовка временного ряда для анализа в среде R Исследование структуры ВР

Основные функции среды R, реализующие методы анализа и прогнозирования временных рядов, включены в пакеты tseries и forecast.

Поэтому перед началом работы необходимо установить эти пакеты, используя команды:

```
install.packages("tseries")
install.packages("forecast")
```

Для загрузки пакета в рабочую область используется функция library():

```
library(tseries)
library(forecast)
```

либо функция require():

```
require(tseries)
require(forecast)
```

Основной функцией для импортирования данных в рабочую среду R является read.table(). Функция имеет большое количество управляющих аргументов, из которых, как правило, достаточно использовать минимальный набор:

```
data<-read.table(file="C:\\User\\Desktop\\data.csv", header=TRUE, sep=";")
```

Аргументы функции:

– file – путь к импортируемому файлу;

– header – наличие в загружаемом файле строки с заголовками столбцов, по умолчанию значение FALSE;

– sep – тип разделителя значений переменных, хранящихся в столбцах, по умолчанию предполагается пробел или знак табуляции sep="";

– dec – тип разделителя целой и дробной части, по умолчанию dec=".".

В результате загружаемая в формате .csv таблица сохраняется в виде объекта с именем data.

Полное описание функций и возвращаемых ими значений с примерами реализаций можно найти в справке. Дополнительно можно вызывать справку клавишей F1, когда курсор стоит на имени функции

в тексте скрипта или в консоли. Если необходимо найти какую-либо функцию по ее имени или части имени, то удобно пользоваться функциями из пакета sos:

```
install.packages("sos")
require(sos)
findFn("read.table")
???read.table
```

Чтобы взять только необходимый столбец (например, столбец, хранящий значения временного ряда) из данных, воспользуйтесь конструкцией:

tsData <- data[,c(y)], где y – номер столбца.

В результате в объекте tsData будет храниться одномерный массив (вектор значений ВР).

В ряде функций, реализующих исследование и модельное описание ВР, в качестве аргумента функции используется объект типа временного ряда. Для преобразования числового массива во временной ряд применяется функция ts():

```
tsData <- ts(tsData)
```

Если измерения временного ряда сделаны на регулярных интервалах меньших, чем один год (например, по месяцам или кварталам), можно указать количество измерений, сделанных за год, используя параметр frequency функции ts(). Для временных рядов по месяцам frequency=12, а для данных по кварталам frequency=4. Также можно указать первый год, в который собирались данные, и первый интервал в году с помощью параметра start функции ts(). Например, если первая точка данных соответствует второму кварталу 2000 года, то start=c(2000,2). Для задания ВР месячных изменений значений показателя, начиная с января 2004 года:

```
tsData <- ts(tsData, frequency=12, start=c(2004,1))
tsData
```

Для построения графика временного ряда используется функция

plot.ts(tsData).

Пример 1. На рис. 1 приведен график изменения значений ВР во времени, полученный в результате применения функции plot.ts(tsData). Длина ВР составляет 192 наблюдения, с января 2004 г. по декабрь 2019 г.

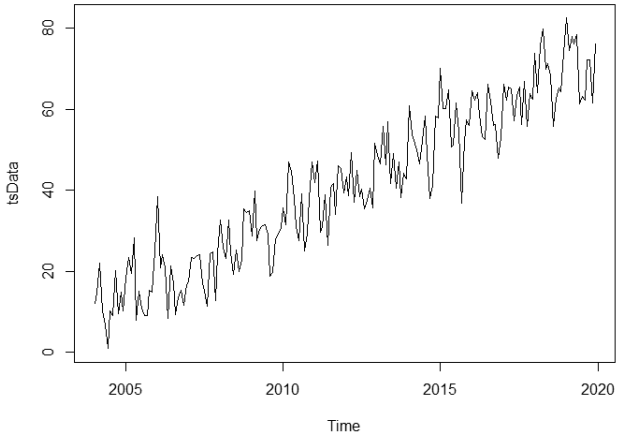


Рис. 1. График значений временного ряда

Функция plot.ts() имеет большое количество настраиваемых параметров для отображения графика (шрифт, цвет и размер символов и линий, заголовок графика, название осей и т. п.), также для построения графика применяется ряд дополнительных функций. Ниже приведен фрагмент кода для построения графика с заданием некоторых параметров и конечный вид графика (рис. 2):

```
par(mar=c(3.5,3.5,0.5,0.5))
plot.ts(tsData, xaxt="n", yaxt="n", ylab="", xlab="", col="blue", main="")
axis(1,at=seq(2004, 2020, by=2), tck=1, col.ticks="light gray")
axis(2,at=seq(0, 80, by=10), tck=1, col.ticks="light gray")
title(ylab = "Значение показателя", cex.lab = 1.2, line = 2.2)
title(xlab = "Год", cex.lab = 1.2, line = 2.2)
```

Функция par() изменяет постоянные графические параметры, т. е. последующие графики будут строиться относительно параметров, указанных пользователем (всего доступно 68 параметров). Вектор mar состоит из четырех числовых значений, которые управляют пространством между осями и границей рисунка, формат c(bottom,left,top,right), значения по умолчанию: c(5.1,4.1,4.1,2.1). В функции plot.ts() используются параметры: xaxt = "n" ось X установлена, но не нарисована; yaxt="n" ось Y установлена, но не нарисована; xlab – название оси X; ylab – название оси Y; col – цвет линий графика; main – название графика. Функция axis() используется для задания параметров оси: 1 – ось X;

2 – ось Y; at – диапазон и шаг изменения значений; tck=1 – нарисовать линии сетки; col.ticks – цвет линий сетки. Функция title() задает параметры заголовков осей: cex.lab – регулирует размер текста относительно значения по умолчанию; line – положение заголовка относительно границы рисунка.

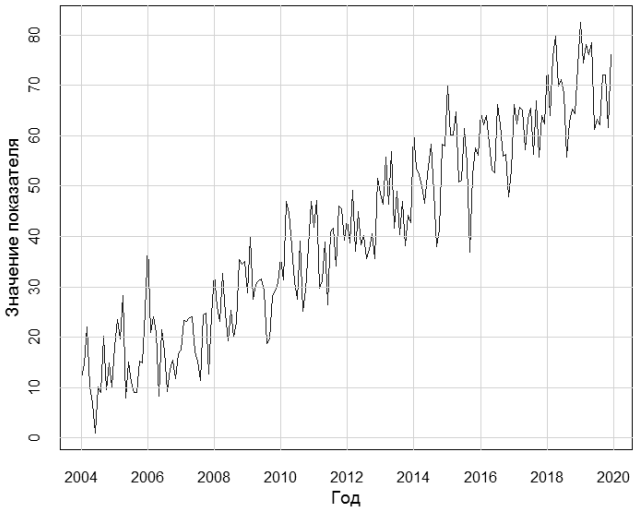


Рис. 2. График значений временного ряда
(с настройкой параметров)

Для разложения ВР на детерминированные и случайные составляющие используется функция decompose():

decompose(x,type = c("additive","multiplicative"),filter = NULL)

Аргументы функции:

– x – временной ряд;

– type – тип модели разложения ВР (аддитивная или мультипликативная), по умолчанию – аддитивная;

– filter – вектор коэффициентов скользящей средней в обратном по времени порядке; если NULL (по умолчанию), то используется скользящая средняя с одинаковыми весами (арифметическая скользящая средняя).

Функция decompose() сначала определяет трендовую составляющую $F(t_i)$ с помощью вычисления скользящей средней. Под скользящей средней понимается функция, значения которой в каждый момент

времени равны среднему значению исходного ВР за предыдущий, текущий и последующие периоды времени, причем вклад каждого значения ВР определяется соответствующим весовым коэффициентом (см. раздел 2.1.1). Скользящая средняя вычитается из ВР. Затем сезонная компонента $S(t_i)$ вычисляется путем усреднения для каждой единицы времени по всем периодам (например, среднее значение за январь 2004–2019 гг.). Затем происходит центрирование сезонной составляющей. На последнем этапе случайная составляющая $\varepsilon(t_i)$ определяется путем удаления трендовой и сезонной компонент из исходного временного ряда. Применение функции decompose() дает хорошие результаты только в случае, если длина ВР равна целому числу периодов.

Применим функцию decompose() для исследуемого временного ряда (см. рис. 2):

```
tsDataComponents <- decompose(tsData)
tsDataComponents
plot(tsDataComponents, col="blue")
```

Функция decompose() возвращает список объектов в качестве результата, где содержатся оценки сезонной составляющей, тренда и случайной компоненты, хранящиеся в именованных элементах этого списка объектов, называемых seasonal, trend и random соответственно. На рис. 3 приведен график декомпозиции исходного ВР, анализ которого позволяет сделать вывод о наличии во ВР монотонного возрастающего тренда, близкого к линейному виду, сезонной компоненты сложной структуры (наличие четко выраженной годовой периодичности и более коротких периодичностей внутри года).

Отметим, что циклическая компонента $C(t_i)$ отдельно не выделяется и в случае ее наличия может быть включена либо в тренд, либо в сезонную составляющую (при задании параметра frequency равным периоду цикла).

Для исследования структуры ряда также используются автокорреляционная и частная автокорреляционная функции. В среде R реализованы функции acf() и Acf(), pacf() и Pacf() соответственно для построения АКФ (*ACF – autocorrelation function*) и ЧАКФ (*PACF – partial autocorrelation function*). Рассмотрим подробнее эти функции.

```
acf(x, lag.max = NULL, type = c("correlation", "covariance", "partial"),
plot = TRUE, na.action = na.fail, ...)
```

Acf(x, lag.max = NULL, type = c("correlation","covariance", "partial"), plot = TRUE, na.action = na.contiguous, ...)

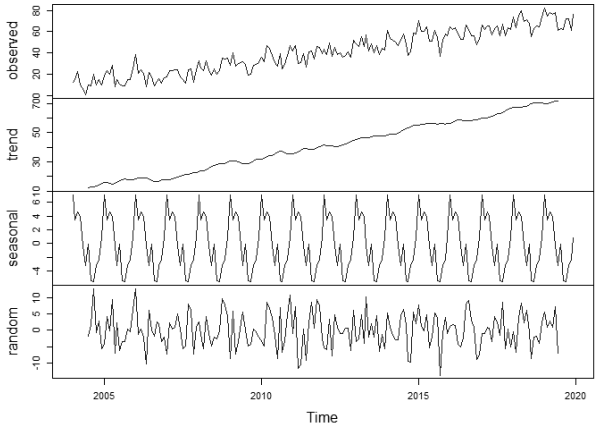


Рис. 3. График декомпозиции временного ряда на аддитивные составляющие

Аргументы функций:

– x – временной ряд или числовой вектор;

– lag.max – максимальный лаг, который будет вычислен; по умолчанию максимальный лаг вычисляется по формуле *10*log10*(*N/m*), где *N* – длина временного ряда, *m* – количество периодов;

– type – тип вычисляемой функции: "correlation"(по умолчанию) – автокорреляционная, "covariance" – автоковариационная, "partial" – частная автокорреляционная;

– plot – построение графика функции (по умолчанию: TRUE);

– na.action – тип обработки пропущенных значений в исходном ВР: если установлено na.fail (по умолчанию в функции acf()), то во ВР не допускаются пропущенные значения и выдается ошибка, в случае их наличия; если задано na.contiguous (по умолчанию в функции Acf()), то выбирается самый длинный фрагмент ВР между пропущенными значениями и по нему рассчитывается функция; na.interp – заполняются пропущенные значения во ВР на основе линейной интерполяции.

Основное отличие функций acf() и Acf() заключается в том, что Acf() не отображает значение коэффициента автокорреляции при лаге 0 (это избыточно, так как значение всегда равно единице) и по горизонтальной оси указаны лаги в единицах времени, а не в сезонных единицах.

Построим *ACF* по временному ряду примера 1 с использованием функций acf()и Acf():

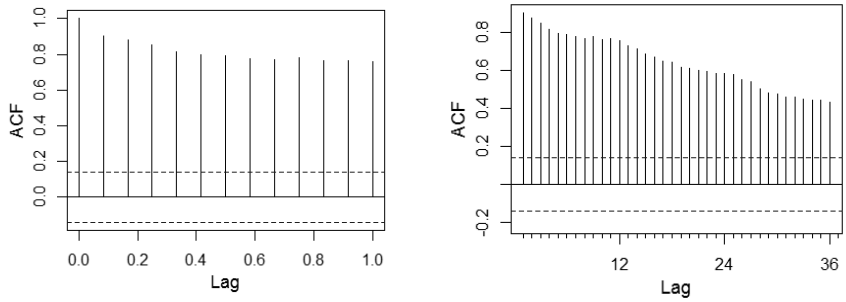
```
tsacf<-acf(tsData, lag.max=12)
tsacf
tsAcf<-Acf(tsData, lag.max=36)
tsAcf
```

Функции возвращают объекты, в которых хранятся значения автокорреляционной функции, вычисленные для каждого лага, ниже приведен фрагмент вывода (первые 10 лагов):

Autocorrelations of series 'tsData', by lag

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
0.902	*0.878*	*0.852*	*0.816*	*0.799*	*0.791*	*0.778*	*0.769*	*0.782*	*0.766*

На рис. 4 показаны построенные графики *ACF*: *а* – первые 12 лагов в сезонных значениях (единица на горизонтальной оси соответствует полному периоду равному 12 месяцев); *b* – первые 36 лагов в единицах времени (в месяцах).



а *б*

Рис. *4*. Автокорреляционная функция исходного ВР (пример 1), построенная с помощью функций:

а – acf(); *б* – Acf()

Для построения частной автокорреляционной функции временного ряда реализованы в R также отдельные функции pacf() и Pacf():

```
pacf(x, lag.max = NULL, plot = TRUE, na.action = na.fail, ...)
Pacf(x, lag.max = NULL, plot = TRUE, na.action = na.contiguous, ...)
```

В функциях используются аналогичные аргументы, как в функциях acf() и Acf(), с такими же установками параметров по умолчанию.

Для построения *PACF* по временному ряду примера 1 используем код:

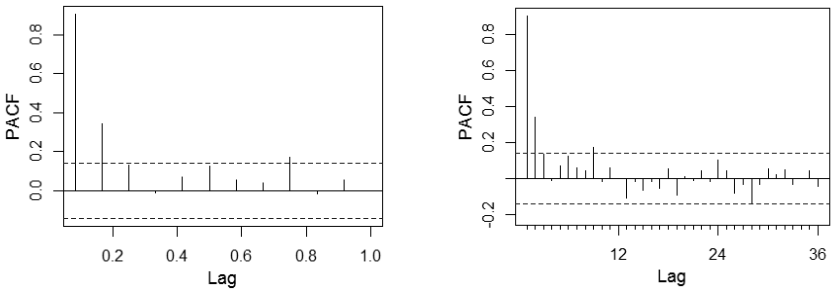
```
tspacf<-pacf(tsData, lag.max=12)
tspacf
tsPacf<-Pacf(tsData, lag.max=36)
tsPacf
```

В результате будут вычислены и выведены значения *PACF* для каждого лага, ниже приведен фрагмент вывода (первые 10 лагов):

Partial autocorrelations of series 'tsData', by lag

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
0.902	*0.34*	*0.13*	*–0.012*	*0.070*	*0.127*	*0.058*	*0.04*	*0.17*	*–0.02*

На рис. 5 представлены построенные графики PACF по аналогии с рис. 4.



а *б*

Рис. 5. Частная автокорреляционная функция исходного ВР (пример 1), построенная с помощью функций:

а – pacf(); *б* – Pacf()

Автокорреляционная функция ВР (см. рис. 4) медленно затухает, и значения *ACF* на всех лагах (1–36) статистически значимы при уровне 0,05 (границы доверительного интервала приведены на рис. 4; если значение коэффициента автокорреляции не входит в интервал, то коэффициент считается статистически значимым). Частная автокорреляционная функция (см. рис. 5) имеет статистически значимые пики при лагах 1–2,

а далее быстро затухает. Такое поведение функций свидетельствует о наличии в исследуемом ВР трендовой составляющей. Выводы о наличии сезонной составляющей можно будет сделать после оценки функции тренда и исключения трендовой компоненты из ВР. Далее по полученному ВР остатка строится *ACF* и *PACF*. Если эти функции изменяются периодически во времени, то в исследуемом ВР присутствует сезонная составляющая с периодом колебаний, равным периоду колебаний *ACF* и *PACF*.

1.4. Основные модели и методы идентификации ВР

Решение задач анализа и прогнозирования временной изменчивости исследуемого процесса (объекта) на основе построенной идентификационной модели играет важнейшую роль в сферах стратегического планирования и оперативного управления в различных областях науки и техники. Поэтому к настоящему моменту разработано множество математических моделей и методов разных классов для анализа и прогнозирования временных рядов. К ним относятся вероятностно-статистические, нейросетевые, экспертные методы, генетические алгоритмы, методы деревьев решений, нечеткой логики и другие.

Далее в учебном пособии рассматриваются основные классические модели и методы анализа и прогнозирования временных рядов, относящиеся к классу вероятностно-статистических параметрических методов исследования. В разделе 2.1 рассмотрен метод последовательной идентификации временного ряда, основанный на последовательном выделении и модельном описании структурных составляющих ВР (трендовая, сезонная, циклическая). Трендовая составляющая описывается с помощью линейных и нелинейных регрессионных моделей. Для идентификации сезонной и циклической компонент ВР применяются спектральный анализ одномерных рядов Фурье и гармонический анализ.

В разделе 2.2 приведено описание и применение моделей экспоненциального сглаживания, которые широко используются в силу их простоты и наглядности. В основу экспоненциального сглаживания заложена идея постоянной корректировки прогнозных значений по мере поступления фактических данных [3, 21]. Модель экспоненциального сглаживания присваивает экспоненциально убывающие веса наблюдениям по мере их старения. Таким образом, последние доступные наблюдения вносят больший вклад в прогнозное значение, чем более отдаленные от текущего момента времени наблюдения.

В разделе 2.3 рассмотрены разные виды моделей авторегрессии и скользящего среднего и методология их построения. Модель авторегрессии и скользящего среднего (АРСС), разработанная Боксом и Дженкинсом [12], считается одной из классических моделей, лежащей в основе многих более сложных методов, позволяет прогнозировать временные ряды и является, с одной стороны, достаточно простой для понимания, а с другой – достаточно гибкой, может быть адаптирована к разным типам стационарных и нестационарных временных рядов. Модель АРСС описывает случайный процесс, представленный временным рядом, как аддитивную композицию двух процессов: процесса авторегрессии и процесса скользящего среднего. Авторегрессия представляет текущее значение процесса через конечную линейную совокупность предыдущих значений процесса, скользящее среднее является по сути дела фильтром низких частот. Модель АРСС применяется для идентификации стационарных ВР, но имеет много модификаций, позволяющих описывать нестационарные ВР, имеющие трендовые и сезонные компоненты.

В разделе 2.4 приведен пример прогнозирования ВР цен на электроэнергию на основе рассмотренных методов и подходов и выполнен сравнительный анализ построенных моделей по точности.

1.5. Оценка качества модели, сравнение и выбор лучшей модели ВР

1.5.1. Статистические характеристики точности модели

Проверка адекватности (соответствия) модели реальному процессу проводится на основе анализа временного ряда остатков, полученного после удаления неслучайной составляющей, описанной моделью. Отметим, что в случае, когда временной ряд остатков обнаруживает некоторые закономерности, необходимо продолжить процесс идентификации и подбирать модель для остатков, и тогда на предмет адекватности проверяется временной ряд финальной остаточной компоненты.

Временной ряд остатков получается как разность наблюдаемых и рассчитанных по модели значений: $\varepsilon_t = y_t - \hat{y}_t$.

На основе ВР остатков рассчитывается ряд статистических характеристик, анализ которых позволяет оценивать точность модели и сравнивать разные модели между собой с целью выбора окончательной,

лучшей по набору статистических характеристик модели ВР. Приведем основные формулы для вычисления характеристик остатков, которые используются на практике:

– средняя ошибка (Mean Error):

$$ME = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t; \tag{1.11}$$

– минимальная ошибка (Minimum Error):

$$Min = \min_{1\le t\le n}\{\varepsilon_t\}; \tag{1.12}$$

– максимальная ошибка (Maximum Error):

$$Max = \max_{1\le t\le n}\{\varepsilon_t\}; \tag{1.13}$$

– средняя абсолютная ошибка (Mean Absolute Error):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|\varepsilon_t\right|; \tag{1.14}$$

– среднее квадратическое отклонение ошибки (Standard Deviation):

$$SD = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(\varepsilon_t - ME)^2}; \tag{1.15}$$

– средняя процентная ошибка (Mean Percentage Error):

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\varepsilon_t}{y_t}100\,\%; \tag{1.16}$$

– средняя абсолютная процентная ошибка (Mean Absolute Percentage Error):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|\varepsilon_t\right|}{y_t}100\,\%; \tag{1.17}$$

– среднеквадратическая ошибка (Mean Square Error):

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2; \tag{1.18}$$

– квадратный корень из среднеквадратической ошибки (Root Mean Square Error):

$$RMSE = \sqrt{MSE}; \tag{1.19}$$

– коэффициент детерминации (Determination Coefficient), первый способ вычисления:

$$R^2 = DC1 = 1 - \frac{\sum_{t=1}^{n} \varepsilon_t^2}{\sum_{t=1}^{n} (y_t - \overline{y_t})^2}; \tag{1.20}$$

– коэффициент детерминации, второй способ вычисления:

$$R^2 = DC2 = \frac{\sum_{t=1}^{n} (\hat{y}_t - \overline{y_t})^2}{\sum_{t=1}^{n} (y_t - \overline{y_t})^2}. \tag{1.21}$$

Средняя ошибка (1.11) и средняя процентная ошибка (1.16) характеризуют степень смещенности модельных результатов. В идеальном случае положительные отклонения модельных значений от наблюдаемых значений ВР должны компенсироваться отрицательными отклонениями и в сумме должны быть равны нулю. Для адекватной модели обе меры (*ME* и *MPE*) стремятся к нулю.

Характеристики *MAE* и *MAPE* используется для сравнения точности разных моделей ВР и выбора наилучшей модели.

Среднеквадратическая ошибка *MSE*, по сути, используется при оценке параметров модели на основе критерия МНК и служит характеристикой точности модели.

Суммарной мерой общего качества построенной модели и ее соответствия исходному ВР является коэффициент детерминации. В случае линейной модели тренда временного ряда коэффициент детерминации равен квадрату коэффициента корреляции Пирсона и квадрату множественного коэффициента корреляции. В общем случае коэффициент детерминации рассчитывается по формулам (1.20) или (1.21). Коэффициент детерминации определяет долю дисперсии ВР y_t, объясненную построенной моделью. Диапазон изменения коэффициента детерминации: $0 \le R^2 \le 1$. Чем ближе значение R^2 к единице, тем теснее связь. В случае нелинейной модели формула (1.21) гарантирует получение неотрицательного значения, но не гарантирует значение меньше единицы.

Напротив, формула (1.20) обеспечивает значение R^2 меньше единицы, но не обеспечивает его неотрицательность. Для линейной модели значения коэффициента детерминации, вычисленные по формулам (1.20) и (1.21), равны между собой; для нелинейной модели в общем случае равенство не выполняется. В литературе можно найти обе формулы, и нет однозначных рекомендаций, какую формулу лучше использовать. Следует отметить, что для моделей ВР коэффициент детерминации, как правило, близок к единице, но это может свидетельствовать не о хорошем качестве модели, а об имеющемся взаимном тренде исходного и модельного ВР. Тем не менее на основе этой характеристики можно сравнивать разные модели ВР по точности.

В ходе исследования следует использовать комплекс статистических характеристик для всесторонней оценки качества и точности построенной модели, так как характеристики взаимно дополняют друг друга.

Временной ряд остатков должен представлять собой реализации значений случайной величины, имеющей нормальное распределение с нулевым математическим ожиданием и конечной дисперсией, и остатки не должны быть линейно зависимы друг от друга (автокоррелированы). Другими словами, ряд остатков должен представлять собой белый шум, имеющий нормальное распределение.

Для проверки гипотезы о нормальном законе распределения остатков используют графический способ:

– построение гистограммы остатков с наложением функции плотности нормального закона распределения, позволяющей визуально оценить симметричность и близость теоретического и эмпирического законов распределения;

– график остатков на нормальной вероятностной бумаге: остатки должны «лежать» на прямой ожидаемого нормального распределения.

Также реализуется тестирование ВР остатков на соответствие нормальному закону распределения с помощью критериев согласия Пирсона, Колмогорова – Смирнова, Шапиро – Уилка и других [8].

1.5.2. Информационные критерии

Информационные критерии измеряют меру относительного качества модели, учитывающую степень соответствия модели исходным данным с корректировкой (штрафом) на используемое количество параметров, т. е. расчет критериев основан на неком компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс. Информационные критерии используются исключительно для сравнения моделей между собой, значения этих критериев содержательной интерпретации не имеют и не связаны с проверкой статистических гипотез относительно качества модели. Обычно чем меньше значения критериев, тем выше относительное качество модели.

Наиболее известны информационные критерии Акаике (*AIC – Akaike information criterion*) и байесовский информационный критерий (*BIC – Bayesian information criterion*) или другое название – байесовский критерий Шварца (*SBC – Schwarz Bayesian criterion*).

Информационные критерии основаны на концепции информационной энтропии, исторически первый критерий – это критерий *AIC*, предложенный Хиротсугу Акаике в 1971 году и описанный в статье [1].

Расчетная формула критерия имеет вид

$$AIC = kp - 2l, \tag{1.22}$$

где $k = 2$ для обычного AIC; p – количество оцененных параметров в построенной модели; l – значение логарифмической функции правдоподобия модели.

В частном случае классической нормальной линейной регрессии логарифмическая функция правдоподобия равна

$$l = -n/2(1 + ln2\pi + ln\hat{\sigma}^2), \tag{1.23}$$

где $\hat{\sigma}^2$ – состоятельная оценка дисперсии (метода максимального правдоподобия) случайной ошибки модели, равная отношению суммы квадратов остатков к объему выборки n.

Чем меньше значение критерия, тем лучше модель. Многие другие критерии являются модификациями *AIC*.

Скорректированный критерий Акаике (*Corrected AIC – AICc*) рекомендуется применять на малых выборках объема n (предложен в 1978 году Sugiura):

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}. \quad (1.24)$$

Байесовский информационный критерий (*BIC*) предложен Шварцем в 1978 году [5]. Критерий разработан исходя из байесовского подхода и является наиболее часто используемой модификацией *AIC*:

$$BIC = SBC = pln(n) - 2l. \quad (1.25)$$

Как следует из формулы (1.25), критерий *BIC* налагает больший штраф на увеличение количества параметров по сравнению с *AIC*, так как $ln(n)$ больше двух уже при объеме выборки, равном восьми наблюдениям.

Применение разных критериев может привести к выбору разных моделей. Во многих работах эти критерии сравниваются, однако нет окончательного вывода о предпочтительности того или иного критерия. Поэтому программные продукты обычно приводят как минимум два критерия: *AIC* и *BIC*. Известно, что для авторегрессионных моделей критерий *AIC* переоценивает порядок модели, то есть оценка порядка модели на основе этого критерия несостоятельна. Состоятельным критерием выбора порядка авторегрессионной модели является *BIC*.

Контрольные вопросы

1. Приведите постановку задачи исследования и прогнозирования временного ряда как одной из задач интеллектуального анализа данных. В каких областях науки и практики решаются задачи, связанные с анализом ВР?

2. Приведите определение временного ряда. Каковы принципиальные отличия временного ряда от случайной выборки? Что понимается под стационарным и нестационарным ВР?

3. Какие типы факторов, влияющих на формирование значений ВР, можно выделить? Какие структурные составляющие присутствуют в разложении ВР?

4. Приведите аддитивную и мультипликативную модель разложения ВР на составляющие.

5. В чем заключается метод декомпозиции ВР на структурные составляющие?

6. Приведите основные статистические характеристики ВР. Что понимается под автокорреляционной и частной автокорреляционной функциями?

7. Приведите и поясните классификацию моделей и методов анализа и прогнозирования ВР.

8. Какие статистические характеристики остатков используются для оценки точности построенной идентификационной модели ВР?

9. Для чего используются информационные критерии? Приведите и поясните основные информационные критерии.

10. Опишите методику исследования ВР в среде R. Какие основные библиотеки и функции языка R используются для анализа ВР?

2. МОДЕЛИ И МЕТОДЫ ИДЕНТИФИКАЦИИ ВРЕМЕННОГО РЯДА

2.1. Метод последовательной идентификации составляющих временного ряда

Временные ряды, которые описывают реальные процессы, как правило, имеют сложную нелинейную структуру и нестационарны по своей природе. Один из методов, который широко используется для построения модели ВР, – метод последовательной идентификации составляющих ВР. В основе метода лежит поэтапная декомпозиция ВР на отдельные интерпретируемые компоненты, их структурная и параметрическая идентификация. На очередном шаге алгоритма выделяются определенные составляющие ВР: трендовые, сезонные, циклические и иные компоненты, которые вносят значимый вклад в поведение ряда. После каждого шага выполняется анализ остатков, полученных вычитанием выделенных компонент из исходного ряда, проверяются адекватность и точность идентификации. Алгоритм метода можно представить в виде последовательности выполнения следующих шагов.

1. Представление временного ряда в виде модели:

$$Y(t_i) = F(t_i) + \varepsilon_1(t_i);$$

структурная и параметрическая идентификация трендовой составляющей ВР $F(t_i)$; нахождение модельных остатков: $\varepsilon_1(t_i) = Y(t_i) - F(t_i)$; исследование точности и адекватности модели.

2. Представление ВР остатков $\varepsilon_1(t_i)$ в виде модели:

$$\varepsilon_1(t_i) = S(t_i) + C(t_i) + \varepsilon_2(t_i);$$

исследование структуры модельных остатков $\varepsilon_1(t_i)$ с помощью построения автокорреляционной, частной автокорреляционной функций ВР

$\varepsilon_1(t_i)$ и периодограммы ВР $\varepsilon_1(t_i)$; структурная и параметрическая идентификация периодической составляющей ВР $S(t_i)+C(t_i)$; нахождение модельных остатков: $\varepsilon_2(t_i)=\varepsilon_1(t_i)-S(t_i)-C(t_i)$; исследование точности и адекватности модели. Здесь сезонная $S(t_i)$ и циклическая $C(t_i)$ составляющие ВР отдельно не рассматриваются, так как математически эти составляющие описывается одинаково (на основе методов гармонического анализа).

3. Представление ВР остатков $\varepsilon_2(t_i)$ в виде модели:

$$\varepsilon_2(t_i)=R(t_i)+\varepsilon_3(t_i),$$

где $R(t_i)$ – авторегрессионная или инерционная составляющая ВР, описывающая зависимость текущего значения ВР от предыдущих значений ВР; исследование структуры модельных остатков $\varepsilon_2(t_i)$ с помощью построения автокорреляционной, частной автокорреляционной функций ВР $\varepsilon_2(t_i)$; структурная и параметрическая идентификация авторегрессионной составляющей ВР $R(t_i)$; нахождение модельных остатков: $\varepsilon_3(t_i)=\varepsilon_2(t_i)-R(t_i)$; исследование точности и адекватности модели. Процесс выделения структурных составляющих ВР можно продолжить и включить, например, составляющую $F(t_i)$, описывающую зависимость значений ВР от величины объясняющих факторов, которые также представлены в виде временных рядов.

4. Построение полной структурной модели ВР с включением всех выделенных на предыдущих шагах составляющих и параметрическая идентификация полной модели. Этот заключительный этап выполняется для более точной оценки коэффициентов модели ВР, так как последовательное выделение составляющих приводит к накоплению погрешности в оценке модельных коэффициентов. Также на этом этапе проводится исследование точности и адекватности полной модели. В случае корректности выполнения всех этапов идентификации итоговые модельные остатки должны в идеале представлять собой белый шум.

Количество этапов, которые необходимо выполнить в процессе идентификации модели исследуемого ВР, зависит от специфики ВР и его структурных особенностей. Как уже ранее отмечалось, в структуре ВР

могут присутствовать не все составляющие, в этом случае будут выполнены не все представленные этапы. Этапы итерационно взаимодействуют, и для наиболее точной идентификации модели ВР может понадобиться в процессе исследования выполнить некоторые этапы многократно или вернуться на предыдущий этап.

Примеры практического применения метода последовательной идентификации составляющих ВР приведены в работах [2, 10].

2.1.1. Модельное описание трендовой составляющей временного ряда

Трендовая составляющая временного ряда отражает влияние долговременных факторов на формирование значений ряда и описывает долговременную тенденцию изменения ВР во времени. Модельное описание тренда позволяет провести ретроспективный анализ закономерностей изменения изучаемого процесса и дать долговременный прогноз развития процесса на будущие моменты времени.

Для выделения и описания тренда используют два основных класса методов, а именно сглаживающие методы и методы линейного и нелинейного регрессионного анализа.

Сглаживающие методы основаны на расчете скользящих средних, в этом случае оценка трендовой составляющей определяется по формуле

$$\widehat{F}_t = \sum_{l=-L}^{L} c_l y_{t+l}, \quad L = 1, 2, \ldots, \tag{2.1}$$

где c_l – весовые множители, удовлетворяющие условию

$$\sum_{l=-L}^{L} c_l = 1.$$

Таким образом, суммируются L значений ВР до и после текущего значения y_t и само значение y_t. Длина интервала сглаживания (или скользящее окно) равна $2L+1$. Наиболее часто используют арифметическое скользящее среднее, когда весовые множители c_l равны между собой.

Сглаживающие методы оценки тренда реализованы в функции decompose() (см. раздел 1.3), длина окна сглаживания зависит от установленного в параметре frequency периода ВР. Если установлен период frequency=5, то $L = 2$ и соответственно длина интервала усреднения

равна 5. Если установлен четный период, например frequency=12, то $L = 6$, длина окна равна 13, но при вычислении скользящей средней первое и последние значение на интервале будет взято с весом 0.5.

Временной ряд скользящей средней содержит меньше наблюдений, чем исходный ВР, за счет потери первых и последних L наблюдений при вычислении значений по формуле (2.1). Так, в результате декомпозиции ВР из примера 1 временной ряд, описывающий сглаженную трендовую составляющую, короче исходного ВР на 12 значений (см. рис. 3, отсутствуют первые и последние шесть наблюдений).

Очевидно, что чем больше величина L, тем меньше колебания ВР $\widehat{F}_t$ и трендовая составляющая более гладкая (сглаженная). Если $D(\varepsilon_t) = \sigma^2$, то дисперсия оценки $\widehat{F}_t$ будет равна $\sigma^2 / (2L+1)$. Следует учитывать, что при возрастании L увеличивается систематическая ошибка.

Если основная тенденция ВР близка к линейному виду, то систематическая ошибка будет небольшой. В случае же нелинейного тренда возможны существенные искажения формы основной тенденции в результате применения скользящего преобразования.

Для параметрического описания трендовой составляющей ВР широко применяют методы парного регрессионного анализа. Модель временного ряда в рамках парного регрессионного анализа имеет вид

$$y_t = F_t + \varepsilon_t = f(t) + \varepsilon_t, \tag{2.2}$$

где $f(t)$ – функция, зависящая от времени или отсчетов времени (1, 2, 3, …).

Как правило, используют одну из базовых моделей тренда [28]. В самом простом случае рассматривают линейную функцию тренда вида

$$f(t) = \alpha_0 + \alpha_1 t, \tag{2.3}$$

где α_0 и α_1 – оцениваемые параметры тренда.

Для описания тренда более сложной формы применяют различные нелинейные функции:

– полиномиальная:

$$f(t) = \alpha_0 + \alpha_1 t + \ldots + \alpha_p t^p, \tag{2.4}$$

где $\alpha_0, \alpha_1, \alpha_p$ – оцениваемые параметры тренда; p – степень полинома (p = 1 соответствует линейной функции);

– логарифмическая:

$$f(t) = \alpha_0 + \alpha_1 \log 10(t); \tag{2.5}$$

– логистическая:

$$f(t) = \frac{k}{1 + \alpha_0 e^{-\alpha_1 t}}, \tag{2.6}$$

где k – горизонтальная асимптота;

– экспоненциальная:

$$f(t) = \alpha_0 \alpha_1^t; \tag{2.7}$$

– модифицированная экспоненциальная:

$$f(t) = k + \alpha_0 \alpha_1^t; \tag{2.8}$$

– кривая Гомпертца:

$$f(t) = k e^{\alpha_0 e^{-\alpha_1 t}}. \tag{2.9}$$

Выбор модели тренда основан на содержательном анализе изучаемого явления или процесса, который представлен в виде временного ряда наблюдаемых значений некоторого показателя (характеристики) процесса. Например, характерен ли рост, спад значений временного ряда, стремление к асимптотам, монотонность, периоды возрастания и убывания значений, наличие экстремумов и т. п.

Тенденции изменения временных рядов весьма многообразны, поэтому и тренды имеют самые разные формы. На рис. 6 приведены графические формы наиболее часто используемых моделей тренда.

Линейный тренд описывает равномерное изменение показателя во времени. Параметр модели α_0 характеризует первоначальный уровень ряда, относительно которого процесс начинает развиваться; параметр α_1 задает среднюю скорость изменения уровня ряда и равен тангенсу угла наклона тренда к оси времени.

В модели, описывающей параболический тренд, вводится дополнительный параметр, отвечающий за ускорение процесса, – α_2. Графиком этой модели является парабола с осью симметрии, параллельной оси ординат. Обычно для моделирования процессов используют одну из ветвей параболического тренда, что позволяет моделировать различные

ситуации: рост с ускорением – правая восходящая ветвь, $\alpha_2 > 0$ (см. рис. 6); снижение с замедлением – левая нисходящая ветвь, $\alpha_2 > 0$; рост с замедлением – левая восходящая ветвь, $\alpha_2 > 0$; снижение с ускорением – правая нисходящая ветвь, $\alpha_2 > 0$. Используя параболу для описания трендовой составляющей ВР, надо учитывать, что такая модель с течением времени может привести к экстремуму, после чего значения ВР начнут изменяться в противоположную сторону.

Также используют полиномиальные модели тренда более высоких степеней. С ростом степени полинома увеличивается точность аппроксимации значений ВР, но при этом точность прогноза может не увеличиваться, а даже снижаться. Известно, что через любые *n* точек можно провести полиномиальную функцию (*n* – 1)-го порядка (через две точки – прямая, через три – парабола, через четыре – полином третьего порядка и т. д.), однако в таком случае не выявляется общая тенденция, а всего лишь осуществляется «подгонка» модели под исходный ряд данных и модель в результате отражает случайные колебания процесса. Кроме того, следует помнить о наличии (*n* – 1) экстремумов в полиномиальной модели *n*-го порядка. На практике полиномиальные модели применяют для описания тренда сложной формы, при этом, как правило, не используют степень полинома выше пятой.

Экспоненциальную модель тренда применяют для описания процессов, имеющих «лавинообразный характер», т. е. когда прирост уровней ряда зависит от достигнутого уровня. В случае если рост или спад значений показателя ограничен предельным значением, используют для описания тенденции модифицированную экспоненциальную модель, один из параметров которой – значение горизонтальной асимптоты.

Кривая Гомпертца и логистическая функция (*S*-образные кривые) описывают тенденцию роста показателя с изменяющимся отношением прироста ко времени. *S*-образная форма тренда подходит для описания такого процесса, при котором изучаемый показатель проходит полный цикл развития, начиная, как правило, от нулевого уровня, сначала медленно, но с ускорением возрастая, затем ускорение становится нулевым в середине цикла, т. е. рост происходит по линейному тренду, затем, в завершающей части цикла, рост замедляется по гиперболе по мере приближения к предельному значению показателя. В сущности, *S*-образные кривые описывают два последовательных лавинообразных процесса:

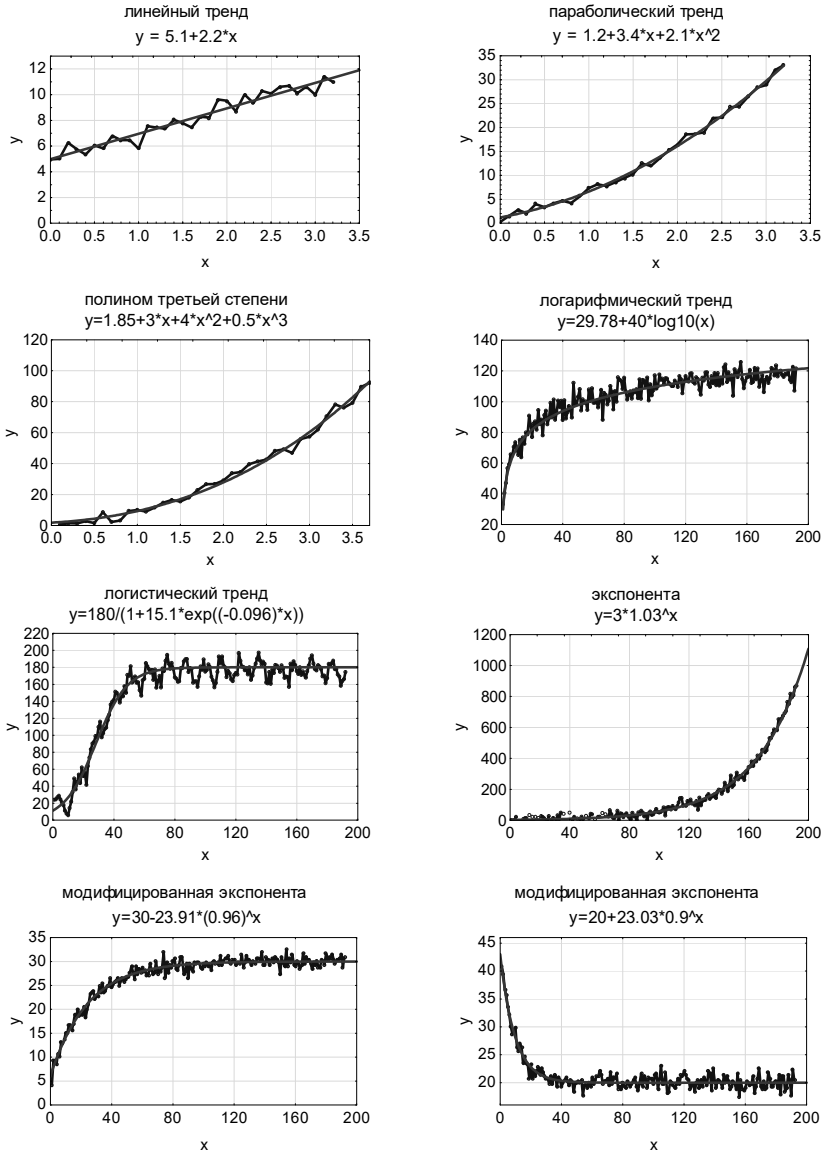


Рис. 6. Виды моделей тренда (см. также с. 34)

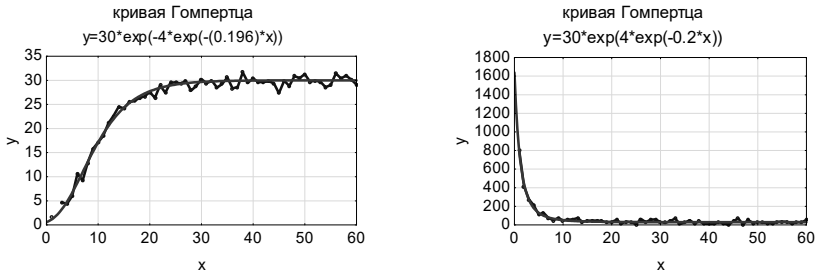


Рис. 6. Окончание

один с ускорением развития, другой – с замедлением, и оба процесса имеют горизонтальные асимптоты. *S*-образные тренды применяют для описания изменения количественных показателей спроса, выпуска продукции, доходности, численности населения и т. д.

На практике исследование временного ряда часто проходит в условиях полной априорной неопределенности, когда нет информации о закономерностях изменения изучаемого процесса во времени и нет возможности содержательно обосновать вид модели тренда. В этом случае используют эмпирические методы для выбора модели трендовой составляющей ВР. Для начала строят линейный график ВР (см. рис. 1 и 2) и на основе графического изображения ряда выдвигают гипотезы о виде функции тренда, при этом может быть выбрано несколько моделей.

Для определения степени полиномиальной функции тренда используют метод последовательных разностей, заключающийся в вычислении разностей:

– первого порядка: $\Delta_t = y_t - y_{t-1}, \quad t = 1, 2, \ldots, n-1;$

– второго порядка: $\Delta_t^2 = \Delta_t - \Delta_{t-1}, \quad t = 1, 2, \ldots, n-2;$

– k-го порядка: $\Delta_t^k = \Delta_t^{k-1} - \Delta_{t-1}^{k-1}, \quad t = 1, 2, \ldots, n-k.$

Расчет разностей ведется до тех пор, пока они не будут примерно равны на всей длине ВР, при этом порядок стабилизации разности соответствует степени полинома. Для примера, на рис. 7 исходный ВР описывается функцией параболы, график ВР первых разностей линейный, график ВР вторых разностей постоянный и не изменяется во времени.

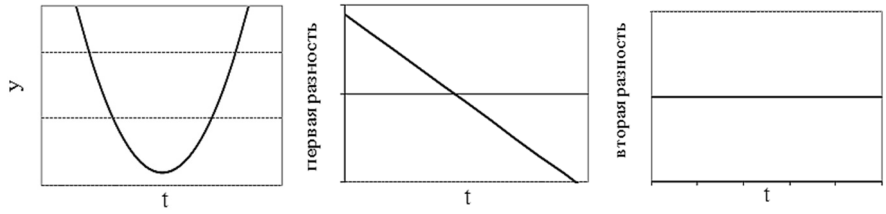


Рис. 7. Определение степени полинома

Для определения других видов тренда разработан метод характеристик прироста, основанный на преобразованиях исходного ВР (определение средних приростов, производных характеристик прироста) и анализе графического изображения полученных ВР [28]. В зависимости от поведения временного ряда характеристики прироста (линейное изменение или постоянный уровень) делаются выводы о виде функции тренда. По сути, выбор модели тренда сводится к визуальному анализу графиков, как и в методе последовательных разностей, и крайне субъективен (полностью зависит от исследователя).

Определение общего вида функции тренда называется *структурной идентификацией модели*. Под *структурой модели* понимают вид функции f, связывающей независимую переменную t со значением ВР y с точностью до ее неизвестных коэффициентов (параметров). Этап выбора структуры модели трудно формализуем, связан с эвристическими решениями, в основном решается только с участием человека.

После выбора структуры модели решается задача *параметрической идентификации*, заключающаяся в оценке параметров функции на основе наблюдений исследуемого ВР. Задача обычно ставится как задача математического программирования, так как при этом минимизируется некоторый функционал, являющийся ущербом, наносимым процессом идентификации.

Для нахождения оценок коэффициентов функции тренда, как правило, используют метод наименьших квадратов (МНК) [9]. В основе МНК лежит критерий: сумма квадратов отклонений экспериментальных наблюдений от построенной линии функции тренда минимальна. В случае линейной функции тренда критерий МНК имеет вид

$$Q = \sum_{t=1}^{n} (y_t - \alpha_0 - \alpha_1 t)^2 = \sum_{t=1}^{n} \varepsilon_t^2 \to \min. \qquad (2.10)$$

По сути, критерий МНК – это обобщенный показатель разброса экспериментальных наблюдений относительно искомой функции тренда. Оценки коэффициентов $\hat{\alpha}_0$ и $\hat{\alpha}_1$ находят из условия равенства частных производных функции Q нулю в точке минимума:

$$\frac{\partial Q}{\partial \alpha_0} = \frac{\partial Q}{\partial \alpha_1} = 0. \tag{2.11}$$

Приведем конечные формулы, полученные в результате применения МНК. Оценки коэффициентов $\hat{\alpha}_0$ и $\hat{\alpha}_1$ вычисляются соответственно по формулам:

$$\hat{\alpha}_0 = \frac{1}{n}\sum_{t=1}^{n} y_t - \frac{\hat{\alpha}_1}{n}\sum_{t=1}^{n} t, \tag{2.12}$$

$$\hat{\alpha}_1 = \frac{n\sum_{t=1}^{n} y_t t - \sum_{t=1}^{n} y_t \sum_{t=1}^{n} t}{n\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2}. \tag{2.13}$$

В случае построения нескольких моделей функции тренда критерий МНК может послужить основой для выбора окончательного вида модели. Лучшей считается модель, для которой критерий МНК минимален. Это один из статистических приемов выбора вида функции тренда.

В случае нелинейной модели используют преобразования для сведения к линейному виду: например, введение фиктивных переменных, логарифмирование и т. д. Далее применяют МНК к преобразованной модели [9].

Бывают случаи, когда невозможно подобрать преобразование для перехода к новой линейной функции и приходится использовать нелинейный метод наименьших квадратов. Минимизирующий функционал МНК определяется выражением

$$Q = \sum_{t=1}^{n} (y_t - \widehat{F}_t)^2 = \sum_{t=1}^{n} \varepsilon_t^2 \to \min. \tag{2.14}$$

Для оценки параметров модели используют алгоритмы Гаусса – Ньютона, Голуба – Перейры, Левенберга – Маквардта и ряд других [15].

После нахождения оценок параметров модели полученное уравнение принимается в качестве оценки для функции тренда и может

быть использовано для дальнейшего исследования ВР или его прогнозирования.

Пример 2. Выполним построение модели тренда для ВР примера 1.

Визуальный анализ исходного графика (см. рис. 1 и 2) и графика декомпозиции ВР на структурные составляющие (см. рис. 3) позволяет предположить линейную функцию тренда.

Для построения линейной модели в среде R используется функция lm():

```
lm(formula, data, subset,...)
```

Аргументы функции:

– formula – модель с указанием зависимой y и независимых переменных $x_1, x_2, \ldots, x_p$ в формате $y \sim x_1 + x_2 + \ldots + x_p$; в случае построения модели ВР $y \sim t$;

– data – объект, в котором хранятся исходные данные (переменные);

– subset – вектор, определяющий подмножество исходных данных для построения модели.

Сначала выделим из исходного массива данных data столбцы, соответствующие отсчетам времени (столбец под номером один) и значениям временного ряда (столбец 5), и переименуем столбцы:

```
data1 <- data[,c(1,5)]
names(data1)[1] <-"t"
names(data1)[2] <-"y"
```

В результате в объекте data1 будет храниться двумерный массив с исходными данными для построения модели (названия столбцов y и t).

Для оценки параметров модели используется конструкция:

```
m1<-lm(y ~ t, data=data1)
m1
```

Функция lm() возвращает объект m1, в котором хранятся следующие компоненты: оценки коэффициентов, модельные остатки, подогнанные (рассчитанные по модели) значения и др.

В результате работы функции выдаются оценки свободного члена модели $\hat{\alpha}_0$ и параметра $\hat{\alpha}_1$:

```
Coefficients:
(Intercept)           t
     8.6055      0.3342
```

К объекту класса lm можно применить ряд функций для получения развернутой статистики по результатам построения модели. Например, функция summary() возвращает:

– оценки параметров модели, стандартные ошибки оценки параметров (*Std. error*);

– результаты оценки статистической значимости коэффициентов модели на основе критерия Стьюдента (расчетное значение критической статистики – *t – value*, *p* – значение);

– характеристики остатков (минимальное значение – *Min*, нижний квартиль – *1Q*, медиана – *Median*, верхний квартиль – *3Q*, максимальное значение – *Max*);

– характеристики точности модели (значение квадрата множественного коэффициента корреляции – *Multiple R-squared*);

– результаты оценки статистической значимости модели по критерию Фишера (расчетное значение критической статистики – *F-statistic*, *p* – значение).

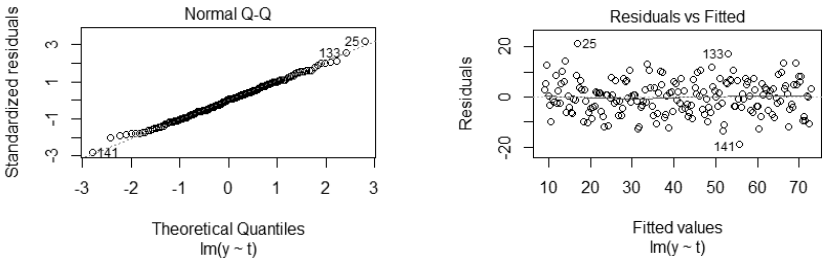
Приведем результаты применения функции summary() для построенной линейной модели:

```
summary(m1)
Call:
lm(formula = y ~ t, data = data1)
Residuals:
     Min       1Q   Median       3Q      Max
–18.8646  –4.5659   0.3248   4.8878  21.4135

Coefficients:
             Estimate   Std. Error   t – value   Pr(>|t|)
(Intercept) 8.605501    0.977368      8.805      8.09e-16 ***
      t     0.334240    0.008783     38.057      < 2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.745 on 190 degrees of freedom
Multiple R-squared:  0.884,  Adjusted R-squared:  0.8834
F-statistic:  1448 on 1 and 190 DF,  p-value: < 2.2e-16
```

В данном случае все параметры модели статистически значимы при уровне 0.0001 ($p < 0.0001$, следовательно гипотеза о статистической незначимости параметров отвергается по критерию Стьюдента [8, 9]).

Построенная модель на 88.4 % объясняет колебания значений исследуемого ВР и является статистически значимой при уровне 0.0001 ($p < 0.0001$, следовательно, гипотеза о статистической незначимости модели отвергается по критерию Фишера [8, 9]). Критерий Фишера позволяет оценить в какой мере замена $\bar{y}$ на $\widehat{F_t}$ позволяет увеличить точность прогнозирования y_t по заданным значениям *t*. Если гипотеза о статистической незначимости не отвергается, то использовать построенную модель нецелесообразно, зависимость y_t от *t* точнее описывается просто средним.

Функция plot(m1) выводит различные графики поведения остатков (рис. 8). В случае адекватности построенной модели исходным данным остатки должны подчиняться нормальному закону распределения и в поведении остатков не должно быть закономерностей. На рис. 8, *а* точки на графике лежат близко к прямой, что свидетельствует о хорошем соответствии распределения остатков нормальному закону; на рис. 8, *б* не наблюдается зависимости остатков от предсказанных модельных значений. Поэтому можно сделать вывод об адекватности построенной модели.



а *б*

Рис. 8. Графики остатков:

а – на нормальной вероятностной бумаге; *b* – зависимость остатков от оцененных по модели значений

На рис. 9 приведен исходный ВР с наложением линейного тренда, для построения которого применяется конструкция

```
plot(data1$y,xlab="t",ylab="y", type="l", col="blue")
lines(fitted(m1), col="red")
```

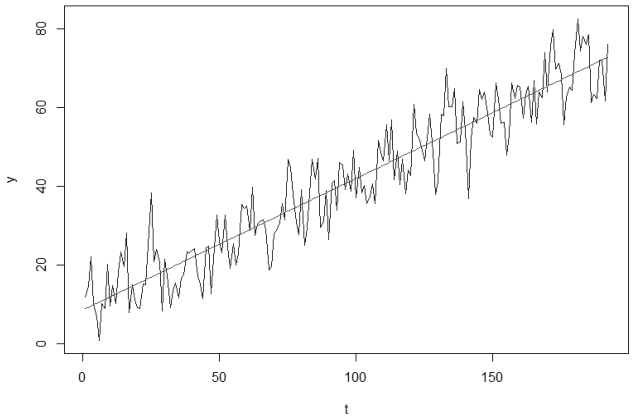


Рис. 9. Временной ряд с наложением линейного тренда

Пример 3. Выполним построение модели тренда для временного ряда, приведенного на рис. 10. Визуальный анализ графика позволяет предположить нелинейную модель тренда, однако конкретный вид модели однозначно определить нельзя, поэтому идентифицируем и сравним по точности несколько возможных моделей, а именно: логистическую (2.6), модифицированную экспоненту (2.8) и параболическую (2.4).

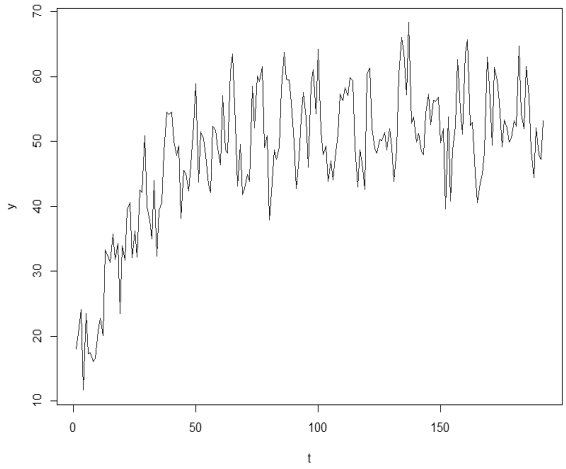


Рис. 10. Временной ряд изменения показателя (пример 3)

Для оценки параметров нелинейной модели на основе нелинейного метода наименьших квадратов используется функция nls():

nls(formula, data, start, algorithm, subset, ...)

Аргументы функции:

– formula – модель в общем виде с указанием зависимой y, независимых переменных $x_1, x_2, \ldots, x_p$ и параметров модели;

– data – объект, в котором хранятся исходные данные (переменные);

– start – вектор начальных оценок параметров модели; в случае если начальные значения не заданы, поиск решения может быть долгим либо решение не будет найдено;

– algorithm – алгоритм оценки параметров модели; по умолчанию используется метод Гаусса – Ньютона, при установке «plinear» – алгоритм Голуба – Перейры, при установке «port» – алгоритм nl2sol, реализующий адаптивный нелинейный метод наименьших квадратов;

– subset – вектор, определяющий подмножество исходных данных для построения модели.

Ниже приведен программный код для идентификации разных вариантов модели тренда:

```
data2<-data[,c(3,11)]
names(data2)[1] <-"t"
names(data2)[2] <-"y"
m2<-nls(y ~ k/(1+a*exp(b*t)),data=data2,start=c(k=55,a=0.1,b=0.1))
m3<-nls(y ~ k-a*b^t,data=data2,start=c(k=55,a=10,b=0.9))
m4<-nls(y ~ a0+a1*t+a2*t^2, data=data2, start=c(a0=1,a1=1,a2=0.3))
summary(m2)
summary(m3)
summary(m4)
```

Исходные данные для построения модели представлены в объекте data2. Функции nls() возвращают соответственно объект m2 (логистический тренд), объект m3 (модель модифицированной экспоненты), объект m4 (параболический тренд), в которых хранятся компоненты модели. В результате работы функции summary() выводятся оценки коэффициентов модели и результаты проверки их статистической значимости на основе критерия Стьюдента. Отметим, что все коэффициенты статистически значимы при уровне значимости 0,0001 для всех построенных моделей тренда:

Formula: *y ~ k/(1 + a * exp(b * t))*
Parameters:
Estimate Std. Error t value Pr(>|t|)
*k 52.64620 0.55750 94.43 < 2e-16 ****
*a 2.44590 0.38612 6.33 1.7e-09 ****
*b -0.07060 0.00767 -9.20 < 2e-16 ****
Signif. codes: *0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: *6.14 on 189 degrees of freedom*
Number of iterations to convergence: *7*
Achieved convergence tolerance: *2.39e-06*

Formula: *y ~ k - a * b^t*
Parameters:
Estimate Std. Error t value Pr(>|t|)
*k 53.17832 0.63204 84.1 <2e-16 ****
*a 42.52929 2.62622 16.2 <2e-16 ****
*b 0.95947 0.00415 230.9 <2e-16 ****
Signif. codes: *0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: *6.19 on 189 degrees of freedom*
Number of iterations to convergence: *7*
Achieved convergence tolerance: *7.74e-06*

Formula: *y ~ a0 + a1 * t + a2 * t^2*
Parameters:
Estimate Std. Error t value Pr(>|t|)
*a0 24.323907 1.513118 16.1 <2e-16 ****
*a1 0.491712 0.036200 13.6 <2e-16 ****
*a2 -0.001925 0.000182 -10.6 <2e-16 ****
Signif. codes: *0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: *6.92 on 189 degrees of freedom*
Number of iterations to convergence: *1*
Achieved convergence tolerance: *2.53e-06*

Нахождение временных рядов остатков модели реализуется с помощью функции residuals():

```
r2<-residuals(m2)
r3<-residuals(m3)
r4<-residuals(m4)
```

Для вычисления статистических характеристик модельных остатков удобно написать собственную функцию. В приведенном ниже коде

(функция mypr()) реализовано вычисление характеристик точности модели на основе формул (1.11)–(1.21). Аргументы функции x – временной ряд остатков; y – исходный временной ряд:

```
mypr <- function(x,y)
{
ME <- mean(x)
n <- length(x)
SD <- sd(x)
Min <- min(x)
Max <- max(x)
MAE <- mean(abs(x))
MPE <- mean(100*x/y)
MAPE <- mean(abs(100*x/y))
RMSE <- sqrt(mean(x*x))
DC1<-1-sum(x^2)/sum((y-mean(y))^2)
DC2<-sum((y-x-mean(y))^2)/sum((y-mean(y))^2)

return(data.frame(n=n, SD=SD, Min=Min, Max=Max, ME=ME, MAE=MAE, MPE=MPE, MAPE=MAPE, RMSE=RMSE, DC1=DC1,DC2=DC2))
}

options(digits=3)
mypr(r2,data2$y)
mypr(r3,data2$y)
mypr(r4,data2$y)
```

Результаты применения функции mypr() к ВР остатков построенных моделей тренда приведены в табл. 1.

Для построения графика исходного временного ряда с наложением построенных моделей тренда (рис. 11) используется код:

```
plot(data2$y,xlab="t",ylab="y", type="l", col="blue")
lines(data2$y-r2, col="green", lwd=2)
lines(data2$y-r3, col="red", lwd=2)

lines(data2$y-r4, col="black", lwd=2)
```

Анализ данных табл. 1 позволяет сделать вывод о хорошей точности построенных моделей тренда: средняя абсолютная процентная ошибка (*MAPE*) составляет 11.3…13.3 %, коэффициент детерминации (*DC1*, *DC2*) равен 0.606…0.689. Для всех моделей тренда получены близкие

Таблица 1

Статистические характеристики точности модели

Статистическая характеристика	Модель 2 (m2)	Модель 3 (m3)	Модель 4 (m4)
n	192	192	192
SD	6.11	6.16	6.88
Min	–14.3	–13.7	–15
Max	15.7	15.3	15.4
ME	–0.00242	2.71e-08	–1.09e-05
MAE	5.02	5.06	5.59
MPE	–1.89	–1.87	–3.2
MAPE	11.3	11.5	13.3
RMSE	6.1	6.14	6.86
DC1	0.689	0.684	0.606
DC2	0.687	0.684	0.606

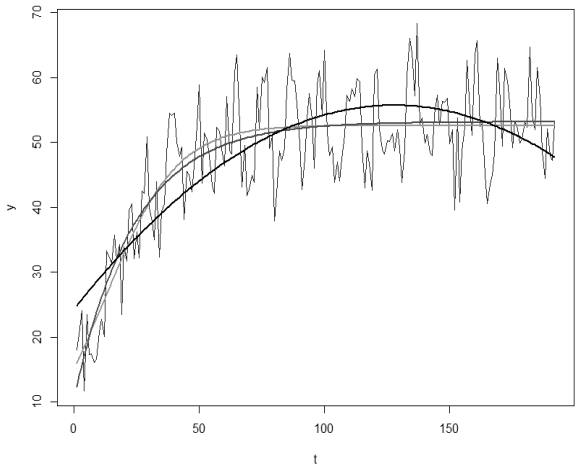


Рис. 11. Временной ряд (пример 3) с наложением моделей тренда

по значениям статистические характеристики остатков. Формально по совокупности характеристик лучшая модель тренда – логистическая. На графике (рис. 11) модель модифицированной экспоненты и логистическая модель расположены очень близко друг к другу и визуально

практически совпадают. Параболическая модель проходит точку экстремума (максимума) на наблюдаемом интервале изменения временного ряда и в дальнейшем прогнозирует уменьшение значений ВР, что визуально плохо согласуется с имеющимися наблюдениями. Поэтому в качестве модели тренда можно рекомендовать либо логистическую кривую, либо модель модифицированной экспоненты.

2.1.2. Модельное описание периодической составляющей временного ряда

К периодической составляющей временного ряда относят:

– сезонную составляющую $S(t_i)$, отражающую периодичность изменения исследуемого процесса (как правило, в течение года);

– циклическую составляющую $C(t_i)$, описывающую долговременные циклы периодичностью более года.

Для анализа структуры периодической составляющей строят автокорреляционную, частную автокорреляционную функции ВР и периодограмму ВР. Для модельного описания периодической составляющей ВР применяют методы гармонического анализа периодических функций.

2.1.2.1. Основные положения гармонического анализа периодических функций

Функцию, заданную в каждой точке изучаемого интервала времени, можно представить бесконечным рядом синусоидальных и косинусоидальных функций. Синусоидальная или косинусоидальная функция с определенным периодом называется *гармоникой*.

Пусть функция $x(t)$ является непрерывной функцией с периодом T. Тогда функцию $x(t)$ можно представить тригонометрическим рядом Фурье:

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left\{ a_k \cos\left(\frac{2\pi k}{T} t \right) + b_k \sin\left(\frac{2\pi k}{T} t \right) \right\}, \quad 0 \le t \le T, \qquad (2.15)$$

где k – номер гармоники. Очевидно, что при увеличении номера гармоники уменьшается период функций $\cos\left(\frac{2\pi k}{T} t \right)$, $\sin\left(\frac{2\pi k}{T} t \right)$. Коэффициенты разложения определяются формулами:

$$a_0 = \frac{2}{T}\int_0^T x(t)\,dt; \tag{2.16}$$

$$a_k = \frac{2}{T}\int_0^T x(t)\cos\left(\frac{2\pi k}{T}t\right)dt; \tag{2.17}$$

$$b_k = \frac{2}{T}\int_0^T x(t)\sin\left(\frac{2\pi k}{T}t\right)dt. \tag{2.18}$$

Аргументы тригонометрических функций cos, sin можно трактовать как частоты ω_k, определяемые соответствующим номером гармоники, т. е.

$$\omega_k = \frac{2\pi}{T}k. \tag{2.19}$$

Также используют другую форму записи тригонометрического ряда Фурье:

$$x(t) = \frac{A_0}{2} + \sum_{k=1}^{\infty} A_k \cos(k\omega_1 t + \varphi_k), \tag{2.20}$$

где амплитуда A_k и фаза φ_k k-й гармонической составляющей связаны с коэффициентами a_k и b_k соотношениями:

$$A_k = \sqrt{a_k^2 + b_k^2}; \quad \varphi_k = -\operatorname{arctg}\frac{a_k}{b_k}, \tag{2.21}$$

или $a_k = A_k\cos\varphi_k$; $b_k = A_k\sin\varphi_k$.

Согласно (2.20) периодическая функция $x(t)$ содержит в себе не зависящую от времени постоянную составляющую $A_0/2$ и бесконечный набор гармоник с частотами $\omega_k = k\omega_1\,(k = 1,2,\ldots)$, кратными основной частоте $\omega_1 = 2\pi/T$ периодической функции. Спектральную составляющую с частотой ω_k называют основной гармоникой, а составляющие с частотами $\omega_k = k\omega_1\,(k > 1)$ – высшими гармониками периодической функции.

Представление периодической функции в виде совокупности постоянной составляющей и суммы гармонических колебаний с кратными частотами называют *спектральным разложением* $x(t)$ в базисе гармонических функций, или *гармоническим анализом периодической функции*.

Амплитуды A_k характеризуют «энергетический вклад» k-й гармоники в функцию $x(t)$. Зависимость амплитуды A_k от номера гармоники k (или от частоты ω_k) характеризует спектральный состав (или спектр) функции $x(t)$ и графически изображается в виде спектральной диаграммы периодической функции. Сравнительно большие величины A_k определяют частоты, на которых сосредоточена основная энергия функции $x(t)$.

Под аппроксимацией функции $x(t)$ рядом Фурье понимают новую функцию $\hat{x}(t)$, полученную суммированием первых членов ряда (2.15), число которых обозначим k_0:

$$\hat{x}(t) = \frac{a_0}{2} + \sum_{k=1}^{k_0} \left\{ a_k \cos\left(\frac{2\pi k}{T} t \right) + b_k \sin\left(\frac{2\pi k}{T} t \right) \right\}. \qquad (2.22)$$

В функции $\hat{x}(t)$ отсутствуют «высокочастотные» гармоники с номерами $k > k_0$, которые присутствовали в исходной функции $x(t)$. Такой способ получения $\hat{x}(t)$ часто называют *низкочастотной фильтрацией* функции $x(t)$.

Для выделения и описания периодической составляющей временного ряда широко используется гармонический анализ, при этом используется разложение $\hat{x}(t)$, в котором содержатся только гармоники, соответствующие наибольшим значениям спектра A_k.

2.1.2.2. Выделение периодической составляющей временного ряда на основе гармонического анализа

При использовании гармонического анализа и разложения Фурье для описания периодической составляющей временного ряда необходимо учесть следующие моменты.

1. Значения временного ряда заданы в дискретные последовательные моменты времени, и чаще всего это равноотстоящие моменты

времени с шагом $\Delta_t = 1$. Тогда в качестве периода принимается величина $T = n\Delta_t$, а условие периодичности ВР имеет вид: $y_{n+t} = y_t, t = 1, 2, \ldots$

2. Временной ряд $Y(t_i)$, кроме периодической составляющей, содержит случайную составляющую $\varepsilon(t_i)$, которую необходимо отделить от периодической составляющей. Модель временного ряда в этом случае имеет вид

$$y_t = S_t + \varepsilon_t = s(t) + \varepsilon_t, \tag{2.23}$$

где $s(t)$ – тригонометрическая функция:

$$s(t) = a_0 + \sum_{t=1}^{n/2} \left\{ a_k \cos\left(\frac{2\pi k}{n} t\right) + b_k \sin\left(\frac{2\pi k}{n} t\right) \right\}. \tag{2.24}$$

Первая гармоника описывает период, равный длине временного ряда или полному периоду повторения *n*. Вторая гармоника имеет период, равный половине основного, третья – одной трети основного, и т. д. Если есть *n* наблюдений, то число гармоник не будет превышать $n/2$. Для периодической функции не всегда требуется определять все $n/2$ гармоник. Как правило, достаточно несколько первых гармоник, которые объясняют большую часть дисперсии временного ряда [19]. Для определения статистически значимых гармоник в разложении ВР строят периодограмму – график зависимости $A_k^2 = a_k^2 + b_k^2$ от периода n/k. Значения периодограммы вычисляются в точках $n/k = 2$ (если *n* четно), $\ldots n/4, n/3, n/2, n$. Также бывает, что периодограмму строят не от периода, а от частоты k/n (в литературе еще встречается название спектрограмма).

Можно определить, какая часть дисперсии исходного временного ряда учитывается *k*-й гармоникой по формуле $\sigma_k^2 = A_k^2/2$. Для последней гармоники $\sigma_{n/2}^2 = A_k^2$. Часть дисперсии, учитываемая одной гармоникой, определяется в виде отношения величины σ_k^2 или $\sigma_{n/2}^2$ к общей дисперсии ВР: σ_y^2.

Так как гармоники не коррелируют между собой, то они не будут учитывать одну и ту же часть общей дисперсии, т. е. дисперсии, учитываемые различными гармониками, складываются.

Коэффициенты разложения (2.24) оцениваются по нелинейному методу наименьших квадратов, в результате применения которого получаются следующие формулы:

$$a_0 = \frac{1}{n}\sum_{t=1}^{n} y_t; \tag{2.25}$$

$$a_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(\frac{2\pi k}{n} t\right); \tag{2.26}$$

$$b_k = \frac{2}{n}\sum_{t=1}^{n} y_t \sin\left(\frac{2\pi k}{n} t\right). \tag{2.27}$$

Формула (2.23) определяет, что a_0 – среднее значение временного ряда за период наблюдения.

Пример 4. Задан временной ряд среднемесячного изменения температуры воздуха с 2000 по 2019 г. (за 20-летний период) в г. Новосибирске (рис. 12). Данные взяты с сайта: http://www.pogodaiklimat.ru/ history/ 29638.htm. Известно, что для ВР изменения температуры воздуха характерна четко выраженная внутригодовая периодичность в течение года. Поэтому полный период повторения значений ВР составляет 12 месяцев, что хорошо видно на рис. 12. Следовательно, периодограмма ВР должна иметь пик, соответствующий периоду, равному 12. Чтобы проверить это предположение, выполним построение периодограммы ВР.

Для построения периодограммы ВР в среде R используется функция spec.pgram():

spec.pgram(x, demean = FALSE, detrend = TRUE, plot = TRUE, log = c("yes", "dB", "no"), ...)

Аргументы функции:

– x – временной ряд или числовой вектор;

– demean – если установлено значение TRUE, то из значений ВР вычитается среднее;

– detrend – если установлено значение TRUE, то вычитается линейный тренд из ВР;

– plot – если установлено значение TRUE, то строится график периодограммы;

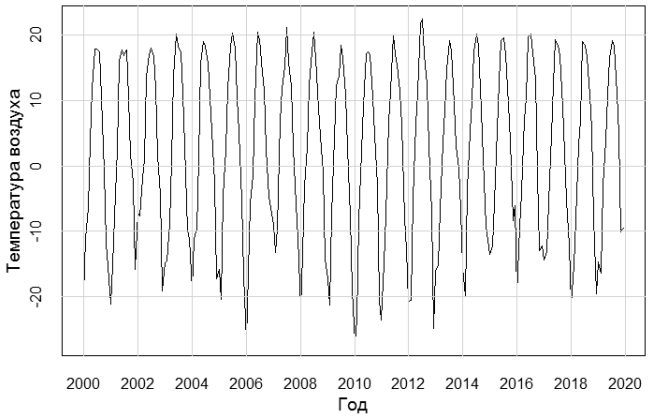


Рис. 12. Временной ряд среднемесячной температуры воздуха в г. Новосибирске за 2000–2019 гг.

– log – задает масштаб построения периодограммы; если установлено "no", то используется линейный масштаб, если "yes" – логарифмический масштаб, "dB" задает шкалу log10.

Для построения периодограммы используем следующий код:

```
data <- read.csv("C://Users//Desktop//температура-Новосибирск.csv",
header = TRUE)
tsData <- data[,1]
sp<-spec.pgram(tsData,detrend = FALSE, log = "no", fast = FALSE,
plot = TRUE, xlab="",ylab="", main="")
sp
title(ylab = "Значение периодограммы", cex.lab = 1.1, line = 2.2)
title(xlab = "Частота", cex.lab = 1.1, line = 2.2)
```

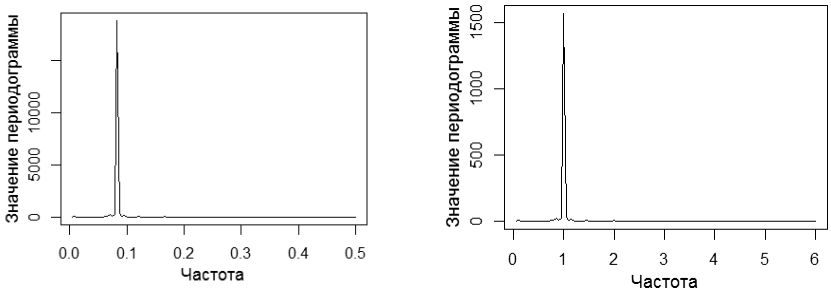
В качестве исходных данных рассматривается вектор tsData, содержащий среднемесячные значения температуры воздуха в г. Новосибирске за 2000–2019 гг. (всего 240 значений). Функция spec.pgram() возвращает объект sp, в котором хранятся значения периодограммы, вычисленные для каждой частоты. Всего в полном разложении ВР используется 120 гармоник (длина ВР/2).

На рис. 13, *а* приведена построенная периодограмма ВР, которая имеет единственный пик, соответствующий 20-й гармонике (k = 20), описывающей период колебаний, равный n/k = 240/20 = 12, с частотой k / n = 20/240 = 0.083(3).

Если перед построением периодограммы выполнить преобразование вектора tsData в объект типа временной ряд с помощью функции ts():

tsData <- ts(tsData, frequency=12, start=c(2000,1)),

то изменится масштаб периодограммы (рис. 13, *б*): значения частоты ВР будут умножены на 12 (основной период ВР), а значения периодограммы будут в 12 раз меньше на соответствующих частотах. При этом общий вид периодограммы не изменится.



а *б*

Рис. 13. Периодограмма ВР температуры воздуха, построенная по числовому вектору (*а*); по объекту типа временной ряд (*б*)

Таким образом, анализ периодограммы подтвердил наличие внутригодовой периодичности изменения ВР температуры воздуха. Для полного описания сезонной составляющей ВР с периодом 12 необходимо использовать 6 гармоник:

1) $a_1 \cos\left(\frac{2\pi}{12}t\right) + b_1 \sin\left(\frac{2\pi}{12}t\right)$, описывает периодичность, равную 12;

2) $a_2 \cos\left(\frac{4\pi}{12}t\right) + b_2 \sin\left(\frac{4\pi}{12}t\right)$, описывает периодичность, равную 6;

3) $a_3 \cos\left(\frac{6\pi}{12}t\right) + b_3 \sin\left(\frac{6\pi}{12}t\right)$, описывает периодичность, равную 4;

4) $a_4 \cos\left(\frac{8\pi}{12}t\right) + b_4 \sin\left(\frac{8\pi}{12}t\right)$, описывает периодичность, равную 3;

5) $a_5 \cos\left(\frac{10\pi}{12}t\right) + b_5 \sin\left(\frac{10\pi}{12}t\right)$, описывает периодичность, равную 2.4;

6) $a_6 \cos(\pi t) + b_6 \sin(\pi t)$, описывает периодичность, равную 2.

Периодограмма ВР температуры воздуха (см. рис. 13) имеет только один статистически значимый пик, соответствующий периоду 12, поэтому для описания сезонной составляющей ВР достаточно использовать только одну гармонику, вклад остальных гармоник статистически незначим, и их учет не приведет к существенному увеличению точности модели:

$$s(t) = a_0 + a_1 \cos\left(\frac{2\pi}{12}t\right) + b_1 \sin\left(\frac{2\pi}{12}t\right). \qquad (2.28)$$

Для оценки параметров модели $s(t)$ на основе нелинейного метода наименьших квадратов используется функция nls(), описанная в разделе 2.1.1.

Ниже приведен программный код для идентификации модели сезонной составляющей $s(t)$:

```
tsData<-data[,c(1,2)]
names(tsData)[1] <-"y"
names(tsData)[2] <-"t"
m5<-nls(y ~ a0+a1*cos(2*pi*t/12)+b1*sin(2*pi*t/12),data=tsData)
summary(m5)
r5<-residuals(m5)
options(digits=3)
mypr(r5,tsData$y)
par(mar=c(3.5,3.5,0.5,0.5))
plot(tsData$y, xaxt="n", yaxt="n", ylab="", xlab="", col="black",
main="", type="l")
axis(1,at=seq(1, 241, by=24), tck=1, col.ticks="light gray", cex.xlab = 5,
labels=c(2000,2002,2004,2006,2008,2010,2012,2014,2016,2018,2020))
axis(2,at=seq(-30, 30, by=10), tck=1, col.ticks="light gray")
title(ylab = "Температура воздуха", cex.lab = 1.2, line = 2.2)
title(xlab = "Год", cex.lab = 1.2, line = 2.2)
lines(tsData$y-r5, col="red", lwd=2)
```

Исходные данные для построения модели представлены в объекте tsData (y – значения исходного временного ряда, t – моменты времени

от 1 до 240). Функция nls() возвращает объект m5 (сезонная модель), в котором хранятся компоненты модели. В результате работы функции summary() выводятся оценки коэффициентов модели и результаты проверки их статистической значимости на основе критерия Стьюдента (все коэффициенты статистически значимы при уровне значимости 0.0001):

Formula: *y ~ a0 + a1 * cos(2 * pi * t/12) + b1 * sin(2 * pi * t/12)*
Parameters:
Estimate Std. Error t value Pr(>|t|)
*a0 1.9654 0.2039 9.637 <2e-16 ****
*a1 -16.0526 0.2884 -55.658 <2e-16 ****
*b1 -8.9135 0.2884 -30.905 <2e-16 ****
Signif. codes: *0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: *3.159 on 237 degrees of freedom*
Number of iterations to convergence: *1*
Achieved convergence tolerance: *5.3e-09*

В результате идентификации построенная сезонная модель временного ряда температуры воздуха имеет вид

$$\hat{s}(t) = 1.97 - 16.05\cos\left(\frac{2\pi}{12}t\right) - 8.91\sin\left(\frac{2\pi}{12}t\right). \qquad (2.29)$$

Нахождение временных рядов остатков модели реализуется с помощью функции residuals(). Для вычисления статистических характеристик модельных остатков используется функция mypr(), описанная в разделе 2.1.1.

В результате применения функции mypr() к ВР остатков сезонной модели $\hat{s}(t)$ выводятся характеристики точности модели:

n	*SD*	*Min*	*Max*	*ME*	*MAE*	*MPE*	*MAPE*	*RMSE*	*DC1*	*DC2*
240	3.15	-10.8	9.29	-3.14e-09	2.31	-13.8	43.8	3.14	0.945	0.945

Далее строится график исходного временного ряда с наложением сезонной модели с помощью функции plot( ) (рис. 14).

Построенная сезонная модель адекватна исходным данным (*ME* – средняя ошибка равна нулю), визуально хорошо согласуется с ВР температуры воздуха, и коэффициент детерминации модели *DC1* близок к единице (0.945). Однако средняя абсолютная процентная ошибка (*MAPE*) составляет 43.8 %, что свидетельствует о невысокой точности построенной модели и плохой прогнозной способности модели в целом.

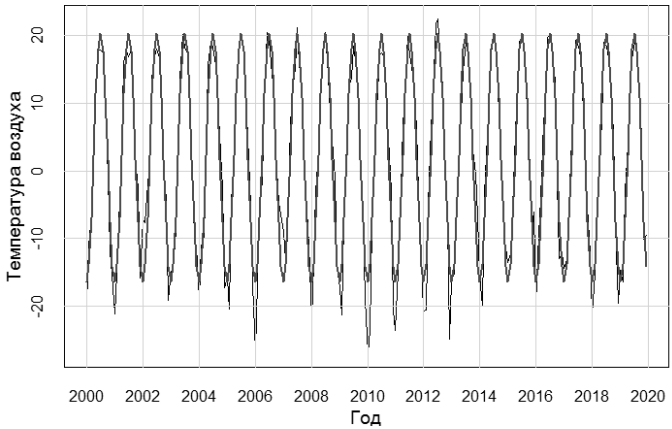


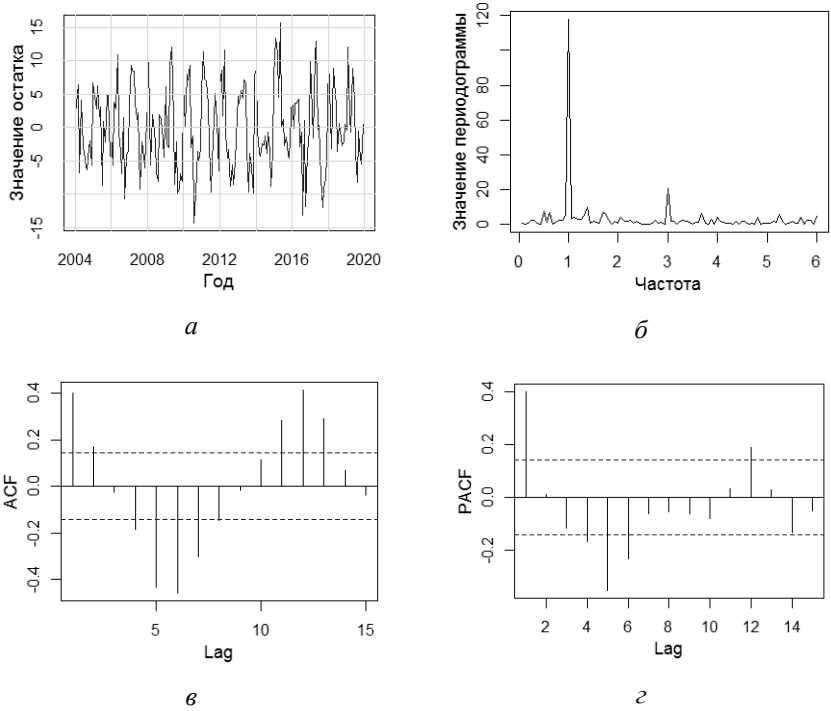
Рис. 14. Временной ряд среднемесячной температуры воздуха в г. Новосибирске с наложением сезонной модели за 2000–2019 гг.

Пример 5. В примере 3 был рассмотрен ВР, имеющий сложную нелинейную структуру, и описано исследование трендовой составляющей ВР, в результате которого идентифицирована логистическая модель тренда. Теперь выполним этап исследования сезонной составляющей ВР.

Для нахождения модельных остатков используется функция residuals(). Временной ряд остатков, полученный после вычитания из исходного ВР логистического тренда, приведен на рис. 15, *а*. Далее для исследования структуры остатков строятся периодограмма ВР остатка (рис. 15, *б*), *ACF* (рис. 15, *в*) и *PACF* (рис. 15, *г*) остатков:

```
r2<-residuals(m2)
r2<-ts(r2, frequency=12, start=c(2004,1))
par(mar=c(3.5,3.5,0.5,0.5))
plot.ts(r2, xaxt="n", yaxt="n", ylab="", xlab="", col="blue", main="")
axis(1,at=seq(2004, 2020, by=2), tck=1, col.ticks="light gray")
axis(2,at=seq(-15, 15, by=5), tck=1, col.ticks="light gray")
title(ylab = "Значение остатка", cex.lab = 1.2, line=2)
title(xlab = "Год", cex.lab = 1.2, line = 2.2)

spec.pgram(r2,detrend = FALSE, log = "no", plot = TRUE,
xlab="",ylab="", main="")
title(ylab = "Значение периодограммы", cex.lab = 1.1, line = 2.2)
title(xlab = "Частота", cex.lab = 1.1, line = 2.2)
```

```
Acf(r2,lag.max=15, xlab="",ylab="", main="")
title(ylab = "ACF", cex.lab = 1.1, line = 2.2)
title(xlab = "Lag", cex.lab = 1.1, line = 2.2)

Pacf(r2,lag.max=15, xlab="",ylab="", main="")
title(ylab = "PACF", cex.lab = 1.1, line = 2.2)
title(xlab = "Lag", cex.lab = 1.1, line = 2.2)
```



а *б* *в* *г*

Рис. 15. Модельные остатки:

а – ВР модельных остатков; *б* – периодограмма ВР модельных остатков; *в* – автокорреляционная функция остатков; *г* – частная автокорреляционная функция остатков

Периодограмма ВР остатка имеет два пика: первый пик соответствует полному периоду повторения ВР, равному 12 (номер гармоники $k = 16$, $n / k = 192 / 16 = 12$, частота на графике (*12): $12k / n = 1$), второй пик соответствует периоду 4 (номер гармоники $k = 48$, $n / k = 192/48 = 4$, частота на графике (*12): $12k / n = 3$). *ACF* и *PACF* описывают колебательный

процесс с периодом 12. В результате анализа графиков (рис. 15) определена структурная модель сезонной составляющей с периодом 12, включающая две гармоники (первую и третью):

$$s(t) = a_0 + a_1 \cos\left(\frac{2\pi}{12}t\right) + b_1 \sin\left(\frac{2\pi}{12}t\right) + \\ + a_2 \cos\left(\frac{6\pi}{12}t\right) + b_2 \sin\left(\frac{6\pi}{12}t\right). \qquad (2.30)$$

Для построения модели исследуемого ВР с включением трендовой и сезонной составляющих используется функция nls(), в качестве начальных параметров тренда можно задать оценки коэффициентов, полученные на предыдущем шаге (см. пример 3):

```
m3<-nls(y ~
k/(1+a*exp(b*t))+a1*cos(2*pi*t/12)+b1*sin(2*pi*t/12)+a2*cos(6*pi*t/12)
+b2*sin(6*pi*t/12),data=data2,start=c(k=53,a=2.45,b=-0.07,a1=0.1,b1=0.1,
a2=0.1,b2=0.1))
m3
```

Оценки коэффициентов модели хранятся в объекте m3:

```
Nonlinear regression model
  model: y ~ k/(1 + a * exp(b * t)) + a1 * cos(2 * pi * t/12) + b1 *
 sin(2 *    pi * t/12) + a2 * cos(6 * pi * t/12) + b2 * sin(6 * pi *    t/12)
data: data2
      k         a          b         a1        b1         a2        b2
52.67219  2.67110 -0.07397 -0.42263  5.51381 -0.75674  2.21795
residual sum-of-squares: 3682
Number of iterations to convergence: 4
Achieved convergence tolerance: 8.623e-06
```

Результирующая модель имеет вид

$$\hat{y}(t) = 52.67 / (1 + 2.67\exp(-0.07t) - 0.42\cos\left(\frac{\pi}{6}t\right) + \\ + 5.51\sin\left(\frac{\pi}{6}t\right) - 0.76\cos\left(\frac{\pi}{2}t\right) + 2.22\sin\left(\frac{\pi}{2}t\right). \qquad (2.31)$$

Заметим, что оценки коэффициентов тренда немного отличаются от оценок, полученных в разделе 2.1.1 (пример 3), из-за учета в модели сезонной составляющей, помимо трендовой компоненты.

Характеристики точности модели получены с помощью функции mypr():

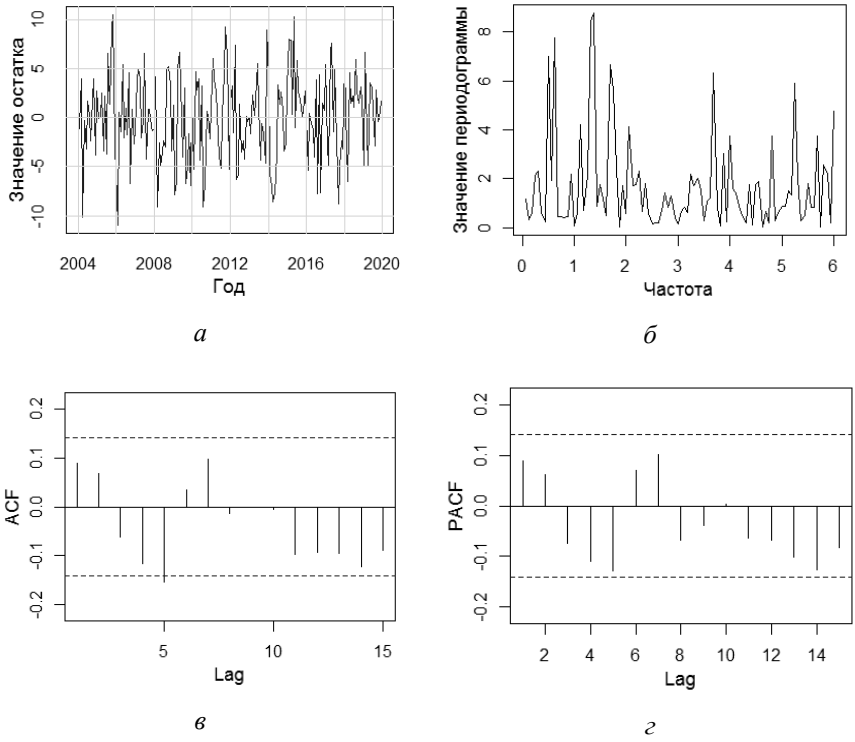
```
r3<-residuals(m3)
mypr(r3,data2$y)
```

n	*SD*	*Min*	*Max*	*ME*	*MAE*	*MPE*	*MAPE*	*RMSE*	*DC1*	*DC2*
192	4.39	-10.98	10.47	-0.01	3.55	-1.13	8.06	4.38	0.84	0.83

Учет сезонной составляющей позволил увеличить точность модели ВР: средняя абсолютная ошибка составила *MAE* = 3.55; средняя абсолютная процентная ошибка *MAPE* = 8.06; коэффициент детерминации *DC* равен 0.83…0.84. На следующем этапе продолжим исследование ВР для выявления оставшихся детерминированных составляющих, построим периодограмму ВР модельных остатков, *ACF* и *PACF* остатков (рис. 16):

```
plot(data2$y,xlab="t",ylab="y", type="l", col="blue")
lines(data2$y-r3, col="red")

r3<-ts(r3, frequency=12, start=c(2004,1))
plot.ts(r3, xaxt="n", yaxt="n", ylab="", xlab="", col="blue", main="")
axis(1,at=seq(2004, 2020, by=2), tck=1, col.ticks="light gray")
axis(2,at=seq(-15, 15, by=5), tck=1, col.ticks="light gray")
title(ylab = "Значение остатка", cex.lab = 1.2, line=2)
title(xlab = "Год", cex.lab = 1.2, line = 2.2)

Acf(r3,lag.max=15, xlab="",ylab="", main="")
title(ylab = "ACF", cex.lab = 1.1, line = 2.2)
title(xlab = "Lag", cex.lab = 1.1, line = 2.2)

Pacf(r3,lag.max=15, xlab="",ylab="", main="")
title(ylab = "PACF", cex.lab = 1.1, line = 2.2)
title(xlab = "Lag", cex.lab = 1.1, line = 2.2)

spec.pgram(r3,detrend = FALSE, log = "no", plot = TRUE,
 xlab="",ylab="", main="")
title(ylab = "Значение периодограммы", cex.lab = 1.1, line = 2.2)
title(xlab = "Частота", cex.lab = 1.1, line = 2.2)
```

Периодограмма остатков не имеет статистически значимых пиков, значения функций *ACF* и *PACF* также статистически не значимы на всех лагах («нулевые») и попадают в 95 % доверительный интервал (границы интервала обозначены на рис. 16, *в* и *г* пунктирной линией). Анализ графиков позволяет сделать вывод об отсутствии во ВР остатков детерминированных структурных составляющих. Следовательно, процесс идентификации модели завершен. На рис. 17 представлен график ВР с наложением модельной кривой. Итоговая модель ВР включает две детерминированные компоненты: логистический тренд и периодическую составляющую с периодами 12 и 4.



а *б* *в* *г*

Рис. 16. Модельные остатки:

а – ВР модельных остатков; *б* – периодограмма ВР модельных остатков; *в* – автокорреляционная функция остатков; *г* – частная автокорреляционная функция остатков

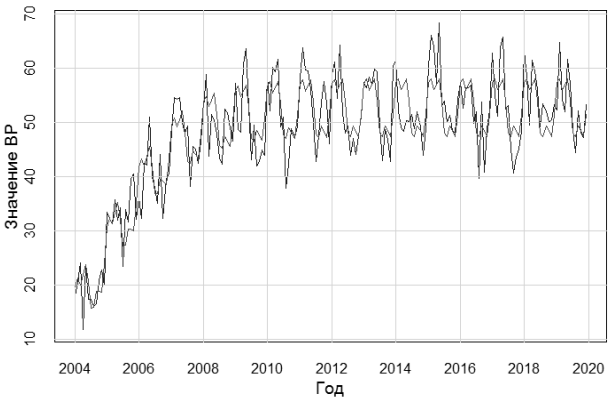


Рис. 17. Временной ряд с наложением полной модели за 2004–2019 гг.

Возможна ситуация, когда в остатках выделяется еще и инерционная составляющая, в этом случае для ее описания используются модели авторегрессии и скользящего среднего, которые подробно обсуждаются в разделе 2.3.

2.2. Модели экспоненциального сглаживания

Класс методов и моделей экспоненциального сглаживания (ЭС) применяется для описания и прогнозирования значений стационарных и нестационарных временных рядов [3, 21]. В основе построения моделей ЭС лежит предположение об инерционности изменения ВР, описывающего процесс, в течение времени.

Наиболее часто на практике используются следующие модели (методы) ЭС: простое экспоненциальное сглаживание (модель Брауна); двойное экспоненциальное сглаживание (модель Хольта); тройное экспоненциальное сглаживание (модель Хольта – Винтерса) и модель Тейла – Вейджа.

Суть методов экспоненциального сглаживания сводится к учету предыдущих значений временного ряда с помощью взвешенной скользящей средней, веса которой подчиняются экспоненциальному закону распределения. Взвешенная скользящая средняя с экспоненциально распределенными весами характеризует значение рассматриваемого

процесса на конце интервала сглаживания, т. е. является средней характеристикой последних наблюдений ВР:

$$A_t = \alpha y_{t-1} + (1-\alpha) A_{t-1}, \quad y_t = A_t + \varepsilon_t, . \tag{2.32}$$

где y_t – значение ВР в момент времени t; A_t, A_{t-1} – значение экспоненциальной средней соответственно в моменты времени t и t – 1; $\alpha \in [0,1]$ – параметр сглаживания; ε_t – значение случайной составляющей в момент времени t.

Экспоненциальность заключена в рекурсивности формулы: каждый раз выполняется умножение $(1-\alpha)$ на предыдущее модельное значение, которое, в свою очередь, также содержит в себе $(1-\alpha)$, и так до первого наблюдения ВР.

Модель Брауна задается формулой

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t, \tag{2.33}$$

где $\hat{y}_t$, $\hat{y}_{t+1}$ – прогнозное значение ВР соответственно в моменты времени t и $t + 1$.

В модели Брауна прогнозное значение определяется через предыдущее спрогнозированное значение, но скорректированное на величину отклонения факта от прогноза ($y_t - \hat{y}_t$). В учете ошибки прогноза заключается суть адаптации модели. Параметр сглаживания определяет вклад составляющих в модель. Чем больше значение параметра, тем больше вклад последнего значения ВР и тем меньше влияние на прогнозные оценки предшествующих значений ряда, меньше сглаживающее влияние экспоненциальной средней.

Модель Брауна не позволяет учесть трендовую и сезонную составляющие ВР и используется только при краткосрочном прогнозировании.

Для учета линейного тренда разработана модель Хольта:

$$\begin{gathered} A_t = \alpha y_t + (1-\alpha)(A_{t-1} + B_{t-1}), \\ B_t = \beta(A_t - A_{t-1}) + (1-\beta)B_{t-1}, \quad \hat{y}_{t+d} = A_t + dB_t, \end{gathered} \tag{2.34}$$

где α, $\beta \in [0,1]$ – параметры сглаживания; d – интервал прогнозирования; B_t, B_{t-1} – значение трендовой составляющей в моменты времени t и t – 1.

В модели Хольта A_t (экспоненциальное среднее) описывает значение ВР в момент времени t, которое зависит от текущего значения ряда

(первое слагаемое) и от предыдущего сглаженного значения ВР и тренда (второе слагаемое). Составляющая B_t описывает тренд, который зависит от изменения сглаженных значений ВР на текущем шаге и от предыдущего значения тренда. Таким образом, для описания тренда также применяется экспоненциальное сглаживание с параметром β. Прогноз значений ВР на момент времени $t + d$ представляет собой сумму экспоненциального среднего и трендовой составляющей $\hat{y}_{t+d}$.

Модель Хольта – Винтерса позволяет учесть трендовую составляющую и мультипликативную сезонность:

$$\begin{aligned} A_t &= \alpha(y_t / S_{t-L}) + (1-\alpha)(A_{t-1} + B_{t-1}), \\ B_t &= \beta(A_t - A_{t-1}) + (1-\beta)B_{t-1}, \\ S_t &= \gamma(y_t / A_t) + (1-\gamma)S_{t-L}, \\ \hat{y}_{t+d} &= (A_t + dB_t)S_{t-L+1+(d-1)modL}, \end{aligned} \tag{2.35}$$

где $\alpha, \beta, \gamma \in [0,1]$ – параметры сглаживания; L – период сезонности; S_t, S_{t-L} – значение сезонной составляющей в моменты времени t и $t – L$.

В модели индекс $t-L+1+(d-1)modL$ сезонной составляющей S означает, что используется соответствующий коэффициент сезонности из прошлого периода (например, для прогноза на январь используется значение сезонной составляющей января прошлого года).

Модель Тейла – Вейджа основана на экспоненциальном сглаживании с учетом тренда и аддитивной сезонности. Основное отличие от предыдущей модели заключается в аддитивном характере взаимодействия тренда и сезонной составляющей (сезонность является не коэффициентом, на который умножается полученный прогноз, а целым числом, которое прибавляется или вычитается из прогноза).

Основные соотношения, описывающие модель Тейла – Вейджа, следующие:

$$\begin{aligned} A_t &= \alpha(y_t - S_{t-L}) + (1-\alpha)(A_{t-1} + B_{t-1}), \\ B_t &= \beta(A_t - A_{t-1}) + (1-\beta)B_{t-1}, \\ S_t &= \gamma(y_t - A_t) + (1-\gamma)S_{t-L}, \\ \hat{y}_{t+d} &= A_t + dB_t + S_{t-L+1+(d-1)modL}. \end{aligned} \tag{2.36}$$

Для построения разных типов моделей экспоненциального сглаживания в среде R используется функция HoltWinters() из пакета stats.

HoltWinters(x, alpha = NULL, beta = NULL, gamma = NULL, seasonal = c("additive", "multiplicative"), start.periods = 2, l.start = NULL, b.start = NULL, s.start = NULL, optim.start = c(alpha = 0.3, beta = 0.1, gamma = 0.1), optim.control = list())

Аргументы функции:

– x – временной ряд;

– alpha – альфа-параметр экспоненциальной модели;

– beta – бета-параметр; если установлено значение FALSE, то функция будет выполнять простое экспоненциальное сглаживание (модель Брауна);

– gamma – гамма-параметр, учитывающий сезонную компоненту; если установлено значение FALSE, то используется несезонная модель;

– seasonal – тип сезонной компоненты: "additive" (аддитивная – по умолчанию) или "multiplicative" (мультипликативная), действует, только если задан параметр gamma;

– start.periods – начальные периоды, используемые при автоопределении начальных значений модели (должно быть не менее 2);

– l.start – начальное значение для уровня ($a[0] = A_0$);

– b.start – начальное значение для трендовой составляющей ($b[0] = B_0$);

– s.start – вектор начальных значений для сезонной составляющей ($s1[0] = S_0, ..., sL[0] = S_L$);

– optim.start – вектор с именованными компонентами alpha, beta и gamma, содержащий начальные значения для поиска оптимальных значений параметров модели, игнорируется в случае с одним параметром;

– optim.control – необязательный список с дополнительными параметрами управления, передаваемыми в optim, игнорируется в случае с одним параметром.

Неизвестные параметры модели оцениваются на основе метода наименьших квадратов. Если все параметры модели равны NULL (по умолчанию), то строится полная экспоненциальная модель ВР, учитывающая трендовую и сезонную составляющие. Начальные значения параметров вычисляются на основе декомпозиции ВР на тренд и сезонную составляющую с использованием скользящих средних (см. функцию decompose()).

Пример 6. В качестве исходных данных рассматривается вектор tsData, содержащий среднемесячные значения показателя за 2004–2019 гг. (всего 192 значения). Функция ts() преобразует вектор tsData в объект типа временной ряд с периодичностью, равной 12 месяцам. Функция HoltWinters() с заданными по умолчанию параметрами строит сезонную модель Тейла – Вейджа, учитывающую тренд и аддитивную сезонность:

```
tsData <- ts(tsData, frequency=12, start=c(2004,1))
plot(tsData, xlim=c(2004,2020), xaxt="n", yaxt="n", ylab="значение
 показателя", xlab="время", col="blue")
axis(1,at=seq(2004, 2020, by=2))
axis(2,at=seq(0, 200, by=25))
tsDataHW <- HoltWinters(tsData)
tsDataHW
```

В результате построения модели оптимальные значения параметров и коэффициентов равны:

Smoothing parameters:
alpha: *0.2335344*
beta : *0.2181513*
gamma: *0.3370024*
Coefficients:
a *183.4454472*
b *0.3855616*
s1 *6.9285586*
s2 *13.0344721*
s3 *9.5739253*
s4 *6.9376295*
s5 *8.0712655*
s6 *-1.7944235*
s7 *–6.4786586*
s8 *–9.6874527*
s9 *–11.2790269*
s10 *–13.1991559*
s11 *–8.9686294*
s12 *–3.7559162*

Далее выполняем прогнозирование значений ВР с помощью построенной модели на год (12 значений):

```
holtWintersModelForecast <- forecast(tsDataHW, h = 12, col="red")
holtWintersModelForecast
plot(holtWintersModelForecast, xaxt="n", yaxt="n", ylab="", xlab="", col="black", main="")
axis(1,at=seq(2004, 2020, by=2), tck=1, col.ticks="light gray", cex.xlab = 5)
axis(2,at=seq(0, 200, by=25), tck=1, col.ticks="light gray")
title(ylab = "Значение показателя", cex.lab = 1.2, line = 2.2)
title(xlab = "Год", cex.lab = 1.2, line = 2.2)
lines(fitted(tsDataHW)[,1], col="blue")
```

На графике рис. 18 приведены исходный и модельный ВР, прогноз значений временного ряда с учетом 80 % и 95 %-го доверительных интервалов. Визуально модельный ВР хорошо согласуется с исходным ВР.

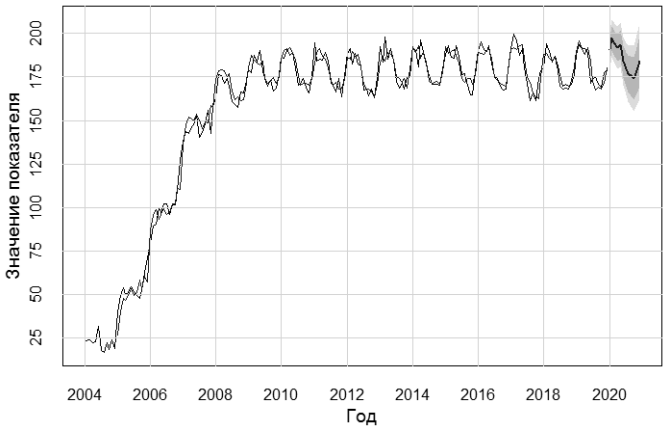
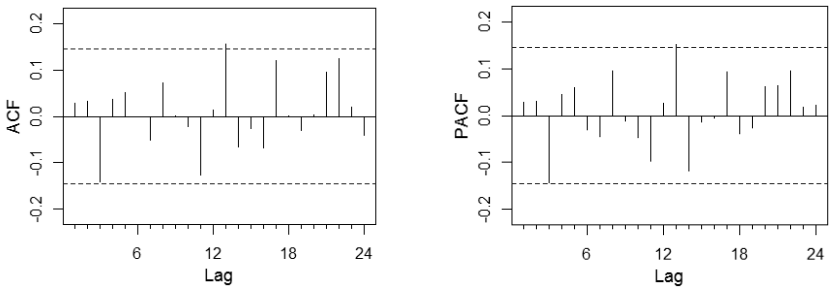


Рис. 18. Прогнозирование значений ВР на основе модели Тейла – Вейджа

Для проверки адекватности модели необходимо построить *ACF* и *PACF* остатков:

```
Acf(holtWintersModelForecast$residuals)
Pacf(holtWintersModelForecast$residuals)
```

На рис. 19 приведены графики рассчитанных *ACF* и *PACF* модельных остатков.



а *б*

Рис. 19. Остатки модели Тейла – Вейджа:

а – автокорреляционная функция остатков; *б* – частная автокорреляционная функция остатков

Как видно из рис. 19, *ACF* и *PACF* остатков нулевые, отсутствуют статистически значимые корреляции, что свидетельствует об адекватности построенной модели.

Для вычисления статистических характеристик точности модели Тейла – Вейджа можно использовать функцию accuracy():

```
accuracy(holtWintersModelForecast)
```

```
ME      RMSE     MAE       MPE      MAPE     MASE       ACF1
4.962542 24.51547 18.89714 0.8096548 7.960331 0.1729135 0.05035711
```

В итоге средняя абсолютная ошибка модели составила *MAE* = 18.9, средняя абсолютная процентная ошибка *MAPE* = 7.96, что является неплохим результатом.

2.3. Модели авторегрессии и скользящего среднего

Модели и методология *ARMA* (*AutoRegressive Moving Average Model* – модель авторегрессии и скользящего среднего – АРСС) предложены Боксом и Дженкинсом в 1974 г. для исследования и прогнозирования временных рядов [12]. Согласно этой методологии временной ряд представляет собой реализацию некоторого случайного процесса, а задача моделирования заключается в восстановлении характеристик случайного процесса по одной его реализации. Модель *ARMA* включает,

как частные случаи, модели авторегрессии *AR* и скользящего среднего *MA* и основана на предположении о наличии линейных корреляционных взаимосвязей между последовательными наблюдениями временного ряда. Модель *ARMA* может интерпретироваться как линейная модель множественной регрессии, в которой в качестве объясняющих переменных выступают прошлые значения самой зависимой переменной, а в качестве регрессионного остатка – скользящие средние из элементов белого шума. Модели *ARMA* позволяют описывать широкий спектр стационарных процессов. Для моделирования нестационарных процессов используется класс моделей *ARIMA* (*AutoRegressive Integrated Moving Average Model* – интегрированная модель авторегрессии и скользящего среднего), который является надстройкой над классом моделей *ARMA*. Модели *ARIMA* работают с нестационарными рядами, которые могут быть сведены к стационарным взятием разностей некоторого порядка от исходного временного ряда. Такие ряды называют интегрированными или разностно-стационарными. Для учета периодичности ВР разработана модель *SARIMA* (*Seasonal ARIMA*), включающая сезонную компоненту и стандартную модель *ARIMA*. Для учета влияния внешних факторов разработаны соответственно несезонные и сезонные модели: *ARIMAX* (*ARIMA eXtended*) и *SARIMAX*.

Рассмотрим подробнее разные варианты моделей.

Модель авторегрессии и скользящего среднего *ARMA*(*p*, *q*) в общем виде выглядит следующим образом:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \\ + \beta + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}, \qquad (2.37)$$

где $\beta, \alpha_i,\ i = \overline{1, p};\ \theta_j, j = \overline{1, q}$ – параметры модели; p – порядок авторегрессионной составляющей; q – порядок скользящей средней; ε_t – белый шум, обладающий свойствами: математическое ожидание $m_1 = 0$; дисперсия $\gamma_0 = \sigma^2$; автоковариация на k-х лагах $\gamma_k = 0,\ k > 0$.

Модель *ARIMA* (*p*, *d*, *q*) включает разностный оператор, который применяется *d* раз для сведения нестационарного ВР к стационарному виду и его описанию в рамках модели *ARMA* (*p*, *q*). Значения ВР первой разности рассчитываются по формуле: $\Delta y_t = y_t - y_{t-1}$. Вторая разность задается формулой: $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$.

Последующие разности вычисляются аналогично: путем последовательного *d*-кратного применения разностного оператора.
Модель *ARIMA*(*p*, *d*, *q*) имеет вид

$$\Delta^d y_t = \beta + \sum_{i=1}^{p} \alpha_i \Delta^d y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \qquad (2.38)$$

где $\Delta^d y_t$ – стационарный ВР.

Также применяется запись модели *ARIMA*(*p*, *d*, *q*) через лаговый оператор:

$$\alpha_p(L)\Delta^d y_t = \beta + \theta_q(L)\varepsilon_t, \qquad (2.39)$$

где

$$Ly_t = y_{t-1}, \quad L^k y_t = y_{t-k}, \quad \alpha_p(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p,$$
$$\theta_q(L) = 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q.$$

Модель *SARIMA* (*p*, *d*, *q*) (P_s, D_s, Q_s) разработана для учета сезонной составляющей ВР и требует задания дополнительных параметров:

$$\alpha_p(L)\alpha_{Ps}(L)\Delta^d \Delta_s^D y_t = \beta + \theta_q(L)\theta_{Qs}(L)\varepsilon_t, \qquad (2.40)$$

где P_s – сезонный порядок авторегрессии; Q_s – сезонный порядок скользящей средней; D_s – порядок сезонной разности;

$$\alpha_{Ps}(L) = 1 - \alpha_{s1} L - \alpha_{s2} L^2 - \ldots - \alpha_{sP} L^{Ps},$$

$$\theta_{Qs}(L) = 1 + \theta_{s1} L + \theta_{s2} L^2 + \ldots + \theta_{sQ} L^{Qs}.$$

Модели *ARIMAX* и *SARIMAX* дополнительно учитывают влияние внешнего фактора по сравнению с моделями *ARIMA* и *SARIMA*. Пусть z_t – ВР внешнего фактора, влияющего на формирование значений y_t. В модели *ARIMA* (2.36) и *SARIMA* (2.38) добавляется слагаемое $\sum_{r=1}^{w} \gamma_r z_{t-r}$, описывающее авторегрессию *w*-го порядка ВР внешнего фактора.

В авторегрессионных моделях прогнозирования можно учесть влияние более одного внешнего фактора, но, как правило, в модель включают не больше трех факторов. Это связано с тем, что в модели обычно используются будущие значения внешних факторов, которые, в свою очередь, спрогнозированы. Таким образом, каждый внешний фактор несет в себе ошибку собственного прогноза, которая может быть значительной, что в результате отрицательно скажется на точности прогнозирования исследуемого временного ряда.

В основе прогнозирования с использованием моделей классов *ARIMA*, *SARIMA* лежит методология Бокса – Дженкинса, которая состоит из трех основных этапов: 1) структурная идентификация модели, 2) оценивание и проверка адекватности модели, 3) прогнозирование.

На этапе идентификации модели необходимо протестировать исследуемый ВР на стационарность и в случае отклонения гипотезы стационарности свести ряд к стационарному виду путем преобразований (взятие последовательных разностей, логарифмирование). Далее задается структура модели на основе определения порядков разности d, авторегрессионной составляющей p, скользящей средней q и аналогичных сезонных параметров D_s, P_s, Q_s с помощью построения и интерпретации выборочных автокорреляционной и частной автокорреляционной функций.

На втором этапе параметры модели оцениваются методом максимального правдоподобия и выполняется проверка адекватности модели (статистическая значимость коэффициентов модели, остатки модели должны иметь свойства белого шума). В случае, когда несколько моделей адекватны, выбирается модель с наименьшим количеством параметров и наилучшими статистическими характеристиками качества подгонки модели (наименьшие значения характеристик остатков, наибольший коэффициент детерминации, наилучшие значения информационных критериев Акаике (*AIC*) и Шварца (*BIC*)).

На третьем этапе реализуется прогнозирование значений ВР с помощью выбранной модели на будущие периоды с оценкой доверительных интервалов для прогнозных значений.

Для определения стационарности временного ряда существует несколько способов. Во-первых, выполняется графический анализ: наличие тренда, сезонной составляющей, изменение амплитуды колебаний значений на разных участках ВР свидетельствует о нестационарности ряда.

Во-вторых, анализ поведения *ACF* и *PACF*. В случае стационарности ВР значения функций статистически значимы только на нескольких первых лагах, а далее функции быстро убывают. В случае нестационарности ВР значение на первом лаге автокорреляционной функции, т. е. *ACF*(*1*), близко к единице, а затем коррелограмма медленно убывает по угасающей экспоненте (синусоиде); значение на первом лаге *PACF*(*1*) = *ACF*(*1*) близко к единице, остальные значения коэффициентов корреляции статистически незначимы, т. е. значения функции не выходят за пределы доверительного интервала.

В третьих, использование статистических тестов на стационарность, например, расширенного *ADF*-теста Дики – Фуллера (*Augmented Dickey – Fuller*), *PP*-теста Филлипса – Перрона (*Phillips – Perron*), *KPSS*-теста Квятковского – Филлипса – Шмидта – Шина (*Kwiatkowski – Phillips–Schmidt– Shin*) [30].

Для определения структуры (параметров) моделей класса *ARMA* с помощью выборочных автокорреляционной (*ACF*) и частной автокорреляционной (*PACF*) функций разработаны следующие правила [12, 13]:

– один параметр авторегрессии (модель *AR*(*1*)): *ACF* экспоненциально затухает; *PACF* имеет статистически значимый выброс на лаге 1 (нет корреляций для других задержек);

– два параметра авторегрессии (модель *AR*(*2*)): *ACF* имеет форму затухающей синусоиды или экспоненциально затухает; значения *PACF* статистически значимы только для сдвигов 1 и 2 (значения для других задержек нулевые);

– один параметр скользящего среднего (модель *MA*(*1*)): *ACF* имеет выброс на лаге 1 (остальные значения нулевые); *PACF* экспоненциально затухает монотонно либо осциллируя, т. е. меняя знак;

– два параметра скользящего среднего (модель *MA*(*2*)): *ACF* имеет выбросы на лаге 1 и 2 (остальные значения нулевые); *PACF* экспоненциально затухает или изменяется синусоидально;

– один параметр авторегрессии и один параметр скользящего среднего (модель *ARMA*(*1*,*1*)): *ACF* и *PACF* экспоненциально затухают, начиная с первого лага (первое значение ненулевое), затухание может быть монотонное или колебательное.

Для идентификации сезонных моделей авторегрессии и скользящего среднего используются критерии, приведенные выше, только рассматривается поведение *ACF* и *PACF* на лагах, кратных периоду сезонности ВР.

Пример 7. На рис. 19 приведен ВР среднемесячного изменения показателя за 2004–2019 гг. Выполним идентификацию и исследование

модели авторегрессии и скользящего среднего, описывающей поведение ВР, с помощью методологии Бокса – Дженкинса.

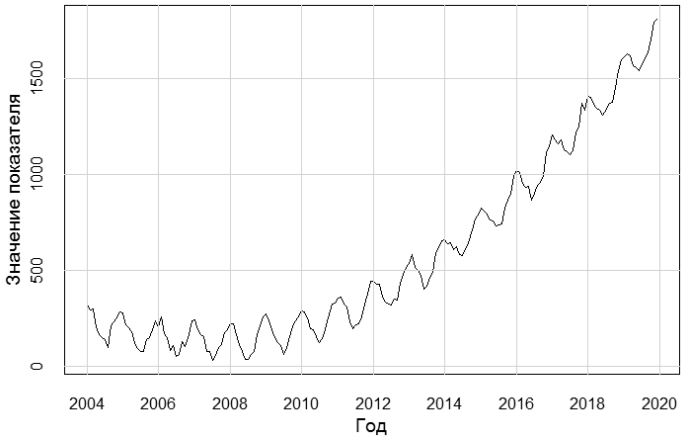
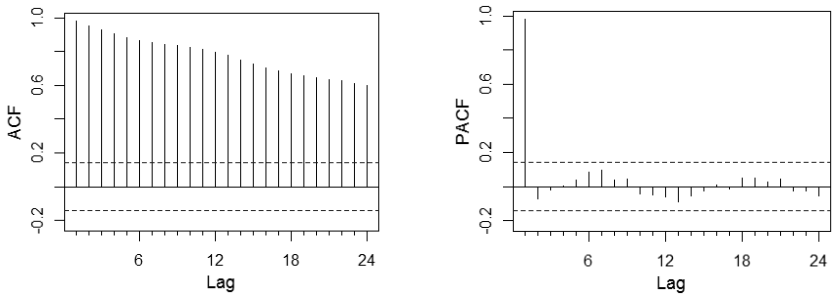


Рис. 19. ВР изменения значения показателя за 2004–2019 гг.

Временной ряд не является стационарным, на графике (см. рис. 19) хорошо видно наличие монотонного возрастающего тренда и периодических колебаний одинаковой амплитуды относительно тренда на всем диапазоне изменения значений ВР. Вывод о нестационарности ВР также подтверждается по результатам анализа графиков *ACF* и *PACF* (рис. 20). Значение автокорреляционной функции ВР на первом лаге близко к единице (0.978), далее автокорреляционная функция медленно экспоненциально затухает. Значение частной автокорреляционной функции ВР на первом лаге также близко к единице (0.978), статистически значимых корреляций на других лагах не наблюдается.

В результате проверки гипотезы о наличии единичного корня в исходном ВР (о нестационарности ВР) с помощью расширенного теста Дики – Фуллера гипотеза не отвергается при уровне значимости 0.05 (*p-value* = 0.77>0.05):

```
adf.test(tsData, k=0)

Augmented Dickey-Fuller Test
data:  tsData
Dickey-Fuller = -1.5318, Lag order = 0, p-value = 0.7721
alternative hypothesis: stationary
```



а *б*

Рис. 20. Автокорреляционная (*а*) и частная автокорреляционная (*б*) функции исходного ВР

Также не отвергается гипотеза о наличии единичных корней в исходном ВР для первых пяти лагов при уровне значимости 0.05 (*p-value* = 0.96>0.05):

```
adf.test(tsData, k=5)
```

```
    Augmented Dickey-Fuller Test
data:  tsData
Dickey-Fuller = -0.78444, Lag order = 5, p-value = 0.9614
alternative hypothesis: stationary
```

Далее выполним преобразования временного ряда для его сведения к стационарному виду с помощью разностных операторов:

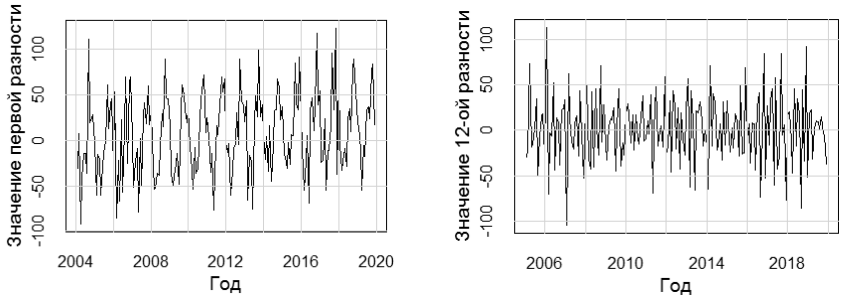
```
tsdiff1 <- diff(tsData, lag=1, differences=1)
plot.ts(tsdiff1)
tsdiff1
adf.test(tsdiff1)
```

```
    Augmented Dickey-Fuller Test
data:  tsdiff1
Dickey-Fuller = -15.116, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

На рис. 21, *а* приведен ряд первых разностей исходного ВР: визуально наблюдается ярко выраженный колебательный процесс и небольшой монотонный рост значений ВР, однако в результате применения

теста Дики – Фуллера к ВР первой разности отвергается гипотеза о наличии единичных корней (*p-value* = 0.01<0.05):

```
spec.pgram(tsdiff1,detrend = FALSE, log = "no", fast = FALSE, plot = TRUE, xlab="",ylab="", main="")
title(ylab = "Значение периодограммы", cex.lab = 1.1, line = 2.2)
title(xlab = "Частота", cex.lab = 1.1, line = 2.2)
```



а *б*

Рис. 21. ВР первой разности (*а*) и ВР 12-й разности (*b*) исходного ВР

Следовательно, ВР первой разности не имеет статистически значимого тренда. Для определения периода колебаний (сезонности) ВР была построена периодограмма по ВР первой разности (рис. 22).

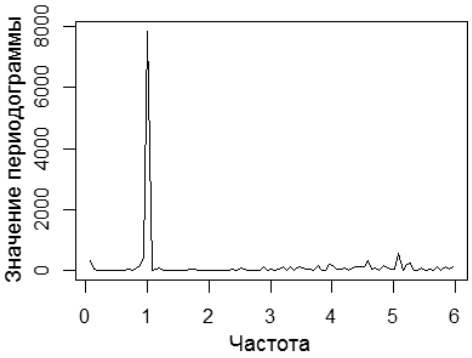
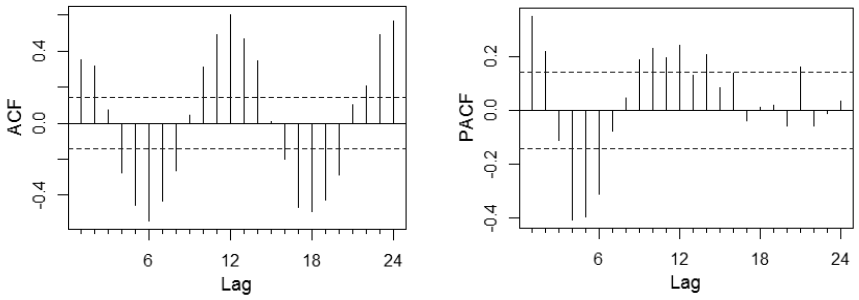


Рис. 22. Периодограмма ВР первой разности

Периодограмма имеет один статистически значимый пик, соответствующий годовой периодичности (12 месяцев) изменения наблюдаемого

показателя. Также сезонность ВР хорошо видна на графиках *ACF* и *PACF* ВР первой разности, значения изменяются по синусоиде с периодом колебаний, равным 12 (рис. 23).



а *б*

Рис. 23. Автокорреляционная (*а*) и частная автокорреляционная (*б*) функции ВР первой разности

Для исключения сезонной составляющей был применен разностный оператор с лагом 12 к ВР первой разности:

```
tsdiff12 <- diff(tsdiff1, lag=12, differences=1)
plot.ts(tsdiff12)
tsdiff12
```

Результат преобразования показан на рис. 21, *б*. Последовательное применение разностных операторов соответственно с лагом 1 (рис. 21, *а*) и лагом 12 (рис. 21, *б*) позволило привести исследуемый ВР к стационарному типу. Теперь можно закончить структурную идентификацию модели *SARIMA*, а именно: определить порядки несезонной и сезонной авторегрессионной составляющей и скользящей средней на основе анализа *ACF* и *PACF* преобразованного ВР (рис. 24).

Автокорреляционная функция имеет статистически значимые пики на лагах 1,12 и 13, частная автокорреляционная функция – на лагах 1, 2, 12 и 13. Согласно правилам определения структуры модели (см. выше) такая ситуация соответствует модели *SARIMA*(*2*,*1*,*1*)(*2*,*1*,*2*). Отметим, что идентификация модели на основе *ACF* и *PACF* является достаточно грубой процедурой, в которой получают прикидочные значения порядка модели. Идентификация модели выполняется с помощью анализа поведения выборочных *ACF* и *PACF*, которые более или менее точно

оценивают неизвестные теоретические *ACF* и *PACF*. Кроме того, не всегда удается однозначно определить порядки модели, визуально анализируя графики *ACF* и *PACF*. Например, на рис. 24, *а* автокорреляционная функция также имеет небольшой статистически значимый пик на лаге 11. На графике частной автокорреляционной функции пик (рис. 24, *б*), соответствующий лагу 11, выражен еще более явно. Наличие этих пиков можно объяснить не идеально четкой 12-месячной периодичностью исходного ВР и влиянием случайной составляющей на изменение значений ВР. Также *PACF* имеет пики, выходящие за 95 % доверительный интервал, на лагах 4, 6, 18. Как их интерпретировать и стоит ли учитывать в структуре модели, зависит от исследователя. Довольно типично на этапе идентификации получить несколько приемлемых моделей, которые с достаточной степенью точности подходят к наблюдаемому ВР и в дальнейшем оцениваются [13].

Для идентификации моделей класса авторегрессии и скользящего среднего в среде R используется функция Arima():

```
tsDataA <- Arima(tsData, order=c(2,1,1), seasonal=c(2,1,2))
tsDataA
```

В результате применения функции выдаются оценки коэффициентов модели, значения стандартной ошибки коэффициентов модели, значения информационных критериев Акаике и Шварца.

```
Series: tsData
ARIMA(2,1,1)(2,1,2)[12]

Coefficients:
      ar1      ar2      ma1     sar1     sar2     sma1     sma2
   -0.1555  -0.0512  -0.6043   0.9849   6e-04  -1.7511   0.7903
s.e.  0.0015   0.0014      NaN      NaN      NaN   0.0037      NaN

sigma^2 estimated as 544:  log likelihood=-826.83
AIC=1669.66   AICc=1670.51   BIC=1695.16
Warning message:
In sqrt(diag(x$var.coef)) : NaNs produced
```

Обратите внимание, стандартные ошибки коэффициентов модели рассчитаны не для всех параметров, о чем выдано соответствующее предупреждение (*NaNs produced*). Подобная ситуация возникает, если исследуемый ВР после применения разностных преобразований находится на границе стационарности. Можно попробовать изменить

структуру модели и дополнительно использовать вторую разность (на рис. 21, *а* после взятия первой разности исходного ВР визуально наблюдался небольшой линейный рост). Результаты построения *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*) следующие:

```
tsDataA <- Arima(tsData,order=c(2,2,1), seasonal=c(2,1,2))
tsDataA

Series: tsData
ARIMA(2,2,1)(2,1,2)[12]
Coefficients:
      ar1      ar2      ma1     sar1     sar2    sma1     sma2
  -0.6967  -0.3673  -0.9998  -0.5213  0.0178  -0.3586  -0.4428
s.e.  0.0715  0.0740  0.0207  1.2619  0.1172  1.2656  1.0672

sigma^2 estimated as 530:  log likelihood=-823.44
AIC=1662.88   AICc=1663.73   BIC=1688.34
```

Теперь стандартные ошибки коэффициентов модели рассчитаны для всех параметров и уменьшились значения всех информационных критериев, что свидетельствует о более высокой точности модели.

Ниже приведены статистические характеристики модельных остатков:

```
accuracy(tsDataA)

ME          RMSE     MAE       MPE      MAPE     MASE       ACF1
0.2019814 21.72616 16.49109 –1.166893 7.188686 0.1508975 -0.06106652
```

Для проверки адекватности модели *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*) были построены периодограмма (рис. 25), *ACF* и *PACF* временного ряда остатков (рис. 26).

Периодограмма (рис. 25) не имеет статистически значимых пиков (максимальные значения не превосходят ста) и соответствует белому шуму. Формально графики *ACF* и *PACF* (рис. 26), построенные по модельным остаткам, имеют небольшие (значение по модулю около 0.2) статистически значимые пики, и можно продолжить исследование, пытаться строить модели более высоких порядков. Но необходимо помнить, что модель должна быть содержательно объяснена и интерпретирована, поэтому на практике, как правило, не используют модели выше второго-третьего порядков. Модель с большим количеством параметров может с высокой точностью описывать наблюдаемые данные, но при

прогнозе оказаться несостоятельной и обеспечить меньшую точность по сравнению с моделью более простой структуры. Подобный эффект наблюдается в результате переобучения модели: модель описывает уже не закономерности изменения реального процесса во времени, а его случайные колебания, обусловленные влиянием стохастических факторов.

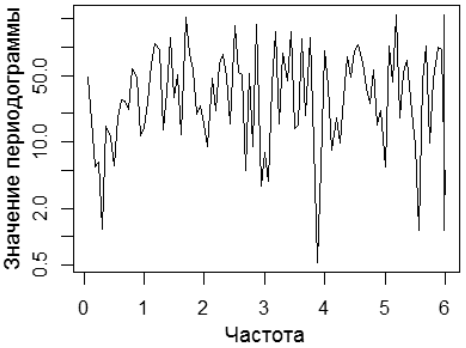
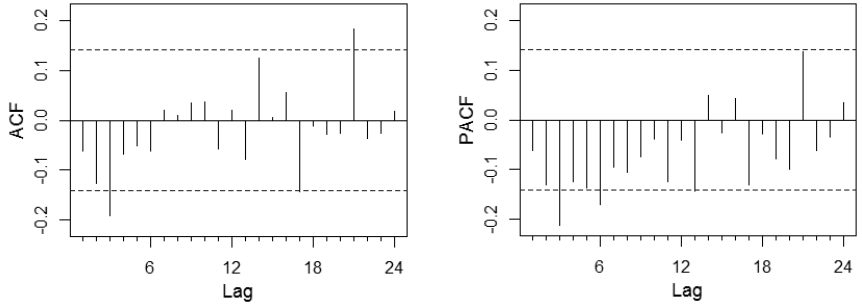


Рис. 25. Периодограмма ВР остатков модели *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*)



а *б*

Рис. 26. Автокорреляционная (*а*) и частная автокорреляционная (*б*) функции ВР остатков модели *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*)

Также в среде R реализована возможность построения моделей класса *SARIMA* с автоматическим подбором структуры модели с помощью функции auto.arima (). Функция возвращает лучшую модель *SARIMA* в соответствии со значением выбранного информационного

критерия: *AIC*, *AICc* или *BIC*. Функция перебирает все варианты структуры модели в пределах заданных максимальных порядков:

```
tsDataAutoA<- auto.arima(tsData,ic = "aic",max.p = 5,max.q = 5,
max.P = 5,max.Q = 5,max.d = 3,max.D = 3)
tsDataAutoA
accuracy(tsDataAutoA)
```

В результате применения функции auto.arima() построена более простая по структуре модель класса *SARIMA*, чем при «ручном» подборе.

Series: *tsData*
ARIMA(*1*,*1*,*3*)(*1*,*1*,*1*)*[12]*

Coefficients:

	ar1	*ma1*	*ma2*	*ma3*	*sar1*	*sma1*
	-0.7625	*0.1143*	*-0.4398*	*0.1139*	*-0.1654*	*-0.4731*
s.e.	*0.5071*	*0.5072*	*0.3154*	*0.0728*	*0.1222*	*0.1285*

sigma^2 estimated as 638.8: *log likelihood=-831.94*
AIC=1677.89 AICc=1678.54 BIC=1700.2

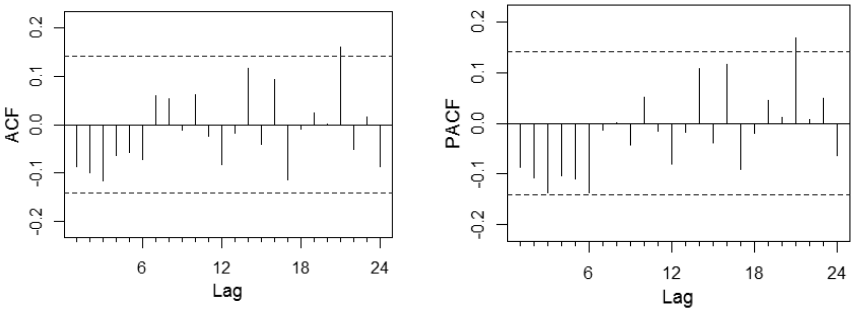
ME	*RMSE*	*MAE*	*MPE*	*MAPE*	*MASE*	*ACF1*
6.942812	23.99203	18.42788	1.482427	7.58042	0.1686197	–0.08571279

Однако для этой модели значения информационных критериев сравнительно хуже и больше по абсолютным значениям статистические характеристики остатков. На рис. 27 приведены графики *ACF* и *PACF* остатков модели *SARIMA*(*1*,*1*,*3*)(*1*,*1*,*1*).

На рис. 28 приведен исходный ВР с наложением построенной модели *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*) и прогноз значений ВР на 2020 г. с указанием доверительного интервала.

Заметим, что окончательный выбор вида модели не может опираться только на статистические характеристики точности построенных моделей и требует содержательного обоснования модели, анализа изучаемого процесса с привлечением специалиста предметной области.

Подход Бокса – Дженкинса к анализу временных рядов является весьма мощным инструментом для построения точных прогнозов с малой дальностью прогнозирования. Классы моделей авторегрессии и скользящего среднего достаточно гибкие и могут описывать широкий



а *б*

Рис. 27. Автокорреляционная (*а*) и частная автокорреляционная (*б*) функции ВР остатков модели *SARIMA*(*1*,*1*,*3*)(*1*,*1*,*1*)

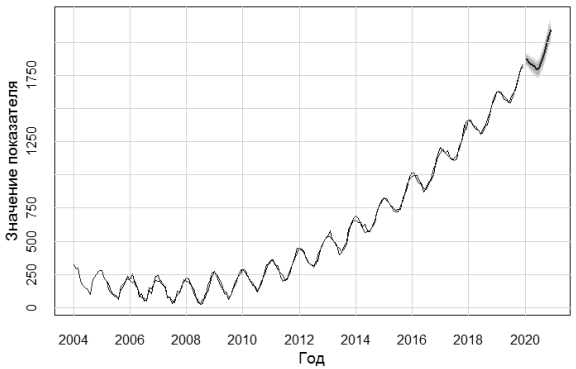


Рис. 28. Прогнозирование значений исходного ВР на основе модели *SARIMA*(*2*,*2*,*1*)(*2*,*1*,*2*)

спектр характеристик временных рядов, которые встречаются на практике. Недостатки моделей связаны, прежде всего, с неоднородностью временных рядов и особенностями практической реализации метода: необходимость относительно большого количества исходных данных; не существует простого способа корректировки параметров моделей авторегрессии и скользящего среднего (когда привлекаются новые данные, модель приходится почти полностью перестраивать, а иногда требуется выбор абсолютно новой модели); для оценок используется та или иная модель, что означает наличие модельного риска в расчетах.

Как следствие, необходима периодическая проверка адекватности применяемой модели. К достоинствам моделей класса авторегрессии и скользящего среднего относятся «прозрачность» моделирования и наличие четкой содержательной интерпретации модели и ее параметров, а также возможность оценки статистической значимости параметров модели и построения доверительных интервалов для прогнозных значений. Модели класса авторегрессии и скользящего среднего широко используются на практике и имеют множество примеров успешного применения.

2.4. Пример исследования временного ряда цен на электроэнергию

В заключение обзора классических параметрических моделей и методов идентификации и прогнозирования временных рядов приведем пример исследования ВР свободной цены на электроэнергию «рынка на сутки вперед» (РСВ) на основе рассмотренных методов. Для исследования был выбран фрагмент ВР почасовых цен РСВ за январь 2017 г., всего 744 наблюдения. Исходный ВР был разделен на две части: обучающая (720 наблюдений) и тестовая (последние 24 наблюдения).

ВР цен РСВ приведен на рис. 29, ряд имеет нелинейную структуру и не стационарен. Визуальный анализ ряда позволяет предположить наличие тренда сложной формы, периодической составляющей и случайной составляющей. Для исследования и прогнозирования ВР использовались метод последовательной идентификации составляющих ВР, экспоненциальное сглаживание и была построена сезонная модель авторегрессии и скользящего среднего.

Методика анализа ВР с помощью этих методов была подробно рассмотрена в предыдущих разделах, поэтому здесь остановимся только на основных моментах и полученных результатах.

Метод последовательной идентификации позволил выделить во ВР три детерминированные составляющие: трендовую, периодическую и инерционную. Для описания трендовой составляющей использовался полином пятой степени, учитывающий сложную форму тренда. В результате построения и анализа *ACF*, *PACF* и периодограммы ВР была выявлена четко выраженная периодичность в изменении значений ВР – 24 часа (сутки), также небольшой пик соответствует периоду 12 часов. Поэтому для описания периодической составляющей ВР было использовано две гармоники, соответствующие периодичности 24 и 12. Кроме

того, в итоговую модель была включена инерционная составляющая, описывающая зависимость текущего значения ВР от прошлого значения, – модель $AR(1)$.

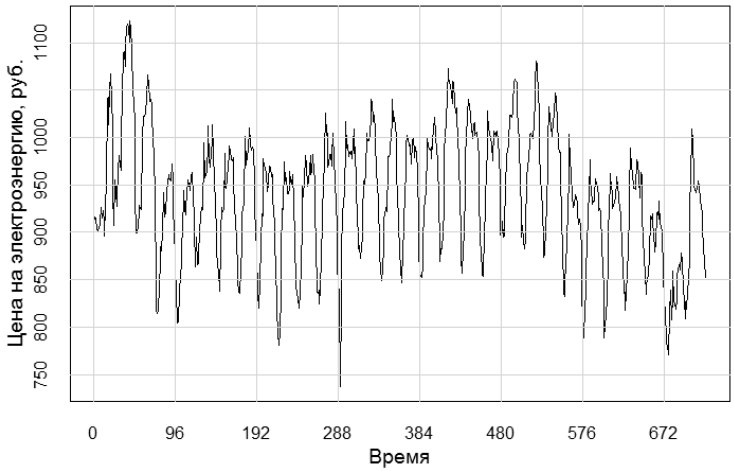


Рис. 29. ВР цен на электроэнергию

В результате итоговая модель (модель 1) имеет вид

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 + \alpha_5 t^5 +$$

$$+ b_1 \sin\left(\frac{2\pi}{24} t\right) + c_1 \cos\left(\frac{2\pi}{24} t\right) + b_2 \sin\left(\frac{4\pi}{24} t\right) +$$

$$+ c_2 \cos\left(\frac{4\pi}{24} t\right) + f y_{t-1} + \varepsilon_t, \qquad (2.41)$$

где α_i, $i = \overline{1,5}$; b_1, b_2, c_1, c_2, f – оцениваемые параметры модели.

Модель экспоненциального сглаживания (модель 2) была построена с помощью функции HoltWinters() и соответствует модели Тейла – Вейджа, учитывающей тренд и аддитивную сезонность (см. раздел 2.2).

Для построения сезонной модели авторегрессии и скользящего среднего (модель 3) использовались функции auto.arima() и Arima(), рассмотренные в разделе 2.3. Результирующая модель соответствует

SARIMA(*1*,*0*,*3*)(*3*,*1*,*0*) и включает следующие компоненты: авторегрессию первого порядка, скользящее среднее третьего порядка, сезонную авторегрессию третьего порядка и сезонную разность первого порядка.

В табл. 2 приведены статистические характеристики точности моделей на обучающей и тестовой частях ВР. Все построенные модели имеют высокую точность на обучающей части ВР: коэффициент детерминации *DC1* равен соответственно 0.93; 0.9 и 0.94; средняя абсолютная процентная ошибка *MAPE* составила 1.39; 1.81 и 1.24. На тестовой части лучшая по точности – сезонная модель авторегрессии и скользящего среднего (*DC1* = 0.98; *MAPE* = 1.15); наименее точна модель, построенная с помощью метода последовательной идентификации (*DC1* = 0.57; *MAPE* = 6.21).

Таблица 2

Статистические характеристики точности модели

Статистическая характеристика	Модель 1		Модель 2		Модель 3	
	обучение	тест	обучение	тест	обучение	тест
n	720	24	720	24	720	24
SD	17.41	45.61	22.25	33.84	16.31	9.39
Min	–90.1	–25.1	–124.2	–20.1	–87.6	–24.5
Max	73.64	131.8	90.03	87.68	68.86	6.14
ME	0	54.67	–2.24	28.18	–0.33	–9.52
MAE	13.02	59.39	16.72	34.95	11.51	10.56
MPE	–0.03	5.66	–0.26	3.01	–0.05	–1.04
MAPE	1.39	6.21	1.81	3.73	1.24	1.15
RMSE	17.4	70.58	22.4	43.5	16.3	13.2
DC1	0.93	0.57	0.90	0.8	0.94	0.98

На рис. 30 приведен исходный ВР с наложением сезонной модели авторегрессии и скользящего среднего и прогноз на 24 часа вперед. Визуально исходный и модельный ВР хорошо согласуются.

Отметим, что для окончательного выбора модели необходимо провести экспериментальные исследования на разных фрагментах ВР цен на электроэнергию. Кроме того, можно проверить эффективность использования для прогнозирования ВР цен на электроэнергию методы других классов (нейросетевые, генетические алгоритмы, деревья решений и т. д.). Например, в работе [4] приведены результаты исследования ВР с помощью алгоритма адаптивной резонансной теории *ART-2*.

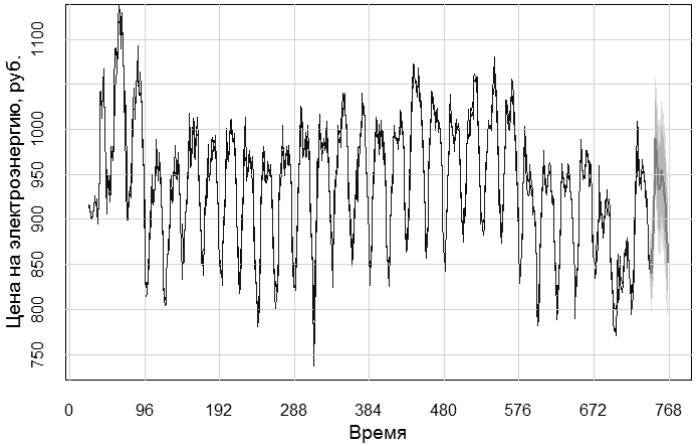


Рис. *30*. Прогнозирование ВР цен на электроэнергию на основе модели *SARIMA*(*1*,*0*,*3*)(*3*,*1*,*0*)

Также перспективным подходом представляется использование ансамблевых методов, когда прогнозы, полученные по разным моделям, объединяются с учетом весовых коэффициентов с целью получения более точных прогнозов [26].

Контрольные вопросы

1. В чем заключается метод последовательной идентификации составляющих ВР? Опишите этапы метода.

2. Какие методы используются для определения наличия трендовой составляющей в исследуемом ВР? Поясните на примере.

3. Какие методы используются для структурной идентификации трендовой составляющей ВР? Поясните на примере.

4. Приведите наиболее часто используемые модели тренда на примерах. Для описания каких процессов используются эти модели?

5. В чем заключается суть метода наименьших квадратов?

6. Как определить структуру периодической гармонической функции, описывающей сезонную составляющую временного ряда?

7. Известно, что в формировании значений временного ряда участвуют колебания двух периодов: 36 и 12. Напишите структурную модель периодической гармонической функции.

8. Какие виды моделей авторегрессии и скользящего среднего используются для описания ВР? В чем заключаются достоинства и недостатки подхода Бокса – Дженкинса?

9. В каких случаях выделяют авторегрессионую составляющую временного ряда? Как определить порядок модели авторегрессии и скользящего среднего, сезонной модели авторегрессии и скользящего среднего?

10. Приведите понятия стационарного и нестационарного ВР. Как привести ВР к стационарному виду?

11. Приведите правила интерпретации поведения автокорреляционной и частной автокорреляционной функции ВР на примерах.

12. Опишите методику построения моделей авторегрессии и скользящего среднего для исследования и прогнозирования ВР на примере.

13. Какие виды моделей экспоненциального сглаживания используются для описания ВР? В чем заключаются достоинства и недостатки метода экспоненциального сглаживания?

14. Опишите методику построения моделей экспоненциального сглаживания на примере.

15. Как выполнить проверку адекватности построенной идентификационной модели ВР данным наблюдения?

16. Как выполнить сравнительный анализ построенных моделей ВР и выбрать окончательный вид модели?

17. Какие библиотеки и функции языка R используются для построения и исследования моделей временных рядов?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Akaike*, *Hirotugu*. A new look at the statistical model identification / *Akaike*, *Hirotugu* // IEEE Transactions on Automatic Control. – 1974. – Vol. 19. – № 6. – Pp. 716–723.

2. *Alsova O. K*. Rotavirus seasonality: an application of singular spectrum analysis and polyharmonic modeling / O. K. Alsova, V. B. Loktev, E. N. Naumova // International Journal of Environmental Research and Public Health. – 2019. – Vol. 16, iss.22. – Art. 4309 – DOI: 10.3390/ijerph16224309.

3. *Brown R. G*. Smoothing forecasting and prediction of discrete time series / *R.G. Brown*. – N.Y. – 1963. – 468 p.

4. *Gavrilov A. V*. Time series prediction using the adaptive resonance theory algorithm ART-2 / A. V. Gavrilov, O. K. Alsova // Journal of Physics: Conference Series. – 2019. – Vol.1333: Information Technologies in Business and Industry. – Art. 032004 (6 p.). – DOI: 10.1088/1742-6596/1333/3/032004.

5. *Schwarz G*. Estimating the Dimension of a Model / *Schwarz G*. // Annals of Statistics. – 1978 – № 6. – Pp. 461–464.

6. *Авдеенко Т. В*. Компьютерные методы анализа временных рядов и прогнозирования: учеб. пособие / Т. В. Авдеенко. – Новосибирск: Изд-во НГТУ, 2008. – 272 с.

7. *Айвазян С. А*. Прикладная статистика. Основы эконометрики. Том 2 / С. А. Айвазян. – М.: Юнити-Дана, 2001. – 432 с.

8. *Айвазян С. А*. Прикладная статистика: Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика, 1983. – 471 с.

9. *Айвазян С. А*. Прикладная статистика и основы эконометрики: учебник для вузов / С. А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ, 1998. – 1022 с.

10. *Альсова О. К*. Решение задач управления Новосибирским водохранилищем на основе прогнозирования притока воды в створ ГЭС / О. К. Альсова, В. В. Губарев // 12 Всероссийское совещание по проблемам управления (ВСПУ-2014) : труды, Москва, 16-19 июня 2014 г. – М.: ИПУ РАН, 2014. – С. 3148–3158.

11. *Андерсон Т.* Статистический анализ временных рядов / Т. Андерсон. – М.: Мир, 1976. – 755 с.

12. *Бокс Дж.* Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, вып. 1, 1974. – 406 с.; вып. 2. – 197 с.

13. *Боровиков В. П.* Прогнозирование в системе Statistica в среде Windows. Основы теории и интенсивная практика / В. П. Боровиков, Г. И. Ивченко. – М.: Финансы и статистика, 2000. – 380 с.

14. *Бриллинджер Д. Р.* Временные ряды. Обработка данных и теория / Д. Р. Бриллинджер; пер. с англ. А. В. Булинского, И. Г. Журбенко; под ред. А. Н. Колмогорова. – М.: Мир, 1980. – 536 с.

15. *Гончаров В. А.* Методы оптимизации : учебное пособие для вузов / В. А. Гончаров. – М.: Юрайт, 2020. – 191 с. – (Высшее образование). – Текст: электронный // ЭБС Юрайт [сайт]. – URL: http://biblio-online.ru/ bcode/463500.

16. *Кабаков Р. И.* R в действии. Анализ и визуализация данных в программе R / Р. И. Кабаков. – М.: ДМК Пресс, 2014. – 588 с.

17. *Каштанов В. А.* Случайные процессы: учебник и практикум для вузов / В. А. Каштанов, Н. Ю. Энатская. – М.: Юрайт, 2020. – 156 с. – (Высшее образование). – Текст: электронный//ЭБС Юрайт [сайт]. – URL: http://biblio-online.ru/bcode/452779.

18. *Кендэл М.* Временные ряды / М. Кендэл. – М.: Финансы и статистика, 1981. – 199 с.

19. *Кильдишев Г. С.* Анализ временных рядов и прогнозирование / Г. С. Кильдишев, А. А. Френкель. – М.: Статистика, 1973. – 103 с.

20. *Лонг Дж. Д.* Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации данных / Дж. Д. Лонг, П. Титор; пер. с англ. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 510 с.

21. *Лукашин Ю. П.* Адаптивные методы краткосрочного прогнозирования временных рядов: учеб. пособие / Ю. П. Лукашин. – М.: Финансы и статистика, 2003. – 416 с.

22. *Мастицкий С. Э.* Анализ временных рядов с помощью R / С. Э. Мастицкий. – 2020 – Текст: электронный. – URL: https://ranalytics.github.io/tsa-with-r.

23. *Мастицкий С. Э.* Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. – М.: ДМК Пресс., 2015. – 496 с.

24. *Подкорытова О. А.* Анализ временных рядов: учебное пособие для вузов / О. А. Подкорытова, М. В. Соколов. – 2-е изд., перераб. и доп. – М.: Юрайт, 2020. – 267 с. – (Высшее образование). – Текст: электронный // ЭБС Юрайт [сайт]. – URL: https://urait.ru/bcode/450587.

25. *Уикем Х.* Язык R в задачах науки о данных. Импорт, подготовка, обработка, визуализация и моделирование данных / Х. Уикем, Г. Гроулмунд; пер. с англ. А. Г. Гузикевича. – М.: Вильямс, 2018. – 592 с.

26. *Френкель А. А.* Методологические подходы к улучшению точности прогнозирования путем объединения прогнозов / А. А. Френкель, А. А. Сурков // Вопросы статистики. – 2015. – № 8. – С. 17–36.

27. *Хеннан Э.* Многомерные временные ряды / Э. Хеннан. – М.: Мир, 1974. – 576 с.

28. *Четыркин Е. М.* Статистические методы прогнозирования / Е. М. Четыркин. – М.: Статистика, 1977. – 199 с.

29. *Шитиков В.К. Э.* Классификация, регрессия, алгоритмы Data Mining с использованием R / В. К. Шитиков, С. Э. Мастицкий. – 2017 – Текст: электронный – URL: https://github.com/ranalytics/data-mining.

30. Эконометрика / под общ. ред. В. С. Мхитаряна. – М.: Проспект, 2014. – 384 с.

31. Эконометрика : учебник для вузов / И. И. Елисеева [и др.]. – М.: Юрайт, 2020. – 449 с. – (Высшее образование). – Текст : электронный // ЭБС Юрайт [сайт]. – URL: http://biblio-online.ru/bcode/449677.

32. Официальный сайт R [Электронный ресурс]. – URL: https://www.r-project.org/ (дата обращения: 15.09.2020).

33. Репозиторий пакетов R [Электронный ресурс]. – URL: https://cran.r-project.org/web/packages (дата обращения: 15.09.2020).

ОГЛАВЛЕНИЕ

Альсова Ольга Константиновна

ИССЛЕДОВАНИЕ ВРЕМЕННЫХ РЯДОВ В СРЕДЕ R

Учебное пособие

Выпускающий редактор *И.П. Брованова*
Корректор *Л.Н. Киншт*
Дизайн обложки *А.В. Ладыжская*
Компьютерная верстка *Л.А. Веселовская*

Подписано в печать 11.01.2021. Формат 60 × 84 1/16. Бумага офсетная. Тираж 100 экз.
Уч.-изд. л. 5,11. Печ. л. 5,5. Изд. № 175/2020. Заказ № 107. Цена договорная

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20