САРАТОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИ ВЕРСИТЕТ

Факультет нелинейных процессов

Кафедра электроники, колебаний и волн

САРАТОВСКОЕ ОТДЕЛЕНИЕ ИНСТИТУТА РАДИОТЕХНИКИ И ЭЛЕКТРОНИКИ РАН

Учебно-научная лаборатория

«Нелинейная динамика (физический эксперимент)»

Поддержано ФЦП «Интеграция» (проект А0057/99) и грантом «Ведущие научные школы» (проект РФФИ (96-15-96536)

Б.П. БЕЗРУЧКО, Д.А. СМИРНОВ

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПО ВРЕМЕННЫМ РЯДАМ

Учебно-методическое пособие

Государственный учебно-научный центр «Колледж»

Саратов, 2000

УДК 530.18
Б 39

Безручко Б.П., Смирнов Д.А.
Б 39 Статистическое моделирование по временным рядам. Учебно-методическое пособие, – Саратов: Издательство ГосУНЦ “Колледж”, 2000 – 23 с.

Рассматриваются подходы к использованию дискретных последовательностей экспериментальных данных (временных рядов) для конструирования статистических моделей, предназначенных для прогноза поведения объекта. Представлены: экстраполяция временной зависимости, а также линейные модели авторегрессии и проинтегрированного скользящего среднего. Предлагается, пользуясь готовыми программами, по экспериментальным временным рядам сконструировать прогностические модели и оценить их качество.

Работа предназначена для практических занятий по курсу “Математическое моделирование” для студентов факультета нелинейных процессов и физического факультета Саратовского госуниверситета.

Рецензент: старший научный сотрудник Саратовского отделения института радиотехники и электроники РАН, к.ф.-м.н. Селезнев Е.П.

Содержание

1. Введение

(динамический и статистический подходы к моделированию)

Математическая модель – описание какого-либо класса явлений внешнего мира, выраженное с помощью математической символики. Формальная математическая конструкция становится моделью после «наполнения» ее физическим содержанием, указанием связи символов с характеристиками объекта. Поэтому при моделировании очень важно выбрать математический аппарат, наиболее соответствующий целям моделирования, и структуру формул, наиболее приспособленную к упомянутому «наполнению». Этот выбор делается на начальном этапе моделирования при исходном рассмотрении объекта или информации о нем и определяется целями моделирования. Так если требуется однозначный прогноз и имеется возможность точного задания величин, характеризующих состояние объекта, конструируют *динамические* модели. Для этого обычно используют аппарат дифференциальных уравнений и однозначные отображения. Если от претензий на точное описание объекта отказываются и объявляют наблюдаемые процессы случайными (непредсказуемыми), строят *статистические* (*вероятностные*) модели. Обычно это делают с помощью математического аппарата статистики и теории вероятностей, если по условиям задачи достаточно указать вероятность того или иного из возможных состояний системы или устраивает приближенное описание с помощью усредненных величин. Более того, в ряде случаев динамическое описание даже не представляется возможным из-за сложности моделируемой системы или ее поведения.

Следует добавить, что хаотические решения простых (маломерных) нелинейных динамических систем могут представлять собой весьма нерегулярные, беспорядочные зависимости от времени, так что для их описания тоже весьма уместны статистические характеристики. С другой стороны для описания непредсказуемых однозначно явлений в системах с малыми шумами используют стохастические дифференциальные уравнения (с малыми случайны-

ми добавками). Эти представители вероятностных моделей в некотором смысле находятся на стыке с динамическими.

В данной работе мы ограничимся статистическим рассмотрением процессов эволюции: определением зависимостей величин, характеризующих объект, от времени с целью прогноза их дальнейшего поведения. Кроме того, разговор пойдет лишь об *эмпирических моделях,* которые конструируют непосредственно из экспериментальных данных, представленных в виде временных рядов (последовательностей чисел) [1]. Впервые задача построения модели по временному ряду была поставлена в рамках статистики в связи с проблемой прогноза. В самом деле, естественным представляется следующий вопрос: если известно поведение объекта до настоящего момента времени, то возможно ли предсказать его будущее, и насколько далеко? Сначала задача прогноза наблюдаемого процесса формулировалась как одна из наиболее распространенных задач статистического анализа — изучение связи между переменными. До 1920-х гг. она решалась методом экстраполяции наблюдаемой временной зависимости, затем появились и получили развитие другие методы, в которых, главным образом, ограничиваются линейными приближениями [1-3]. Несмотря на активное повсеместное продвижение нелинейных представлений, эти ставшие классическими подходы остаются актуальными, а знакомство с ними, реализуемое в данной работе, — необходимым.

---

[1] Выделяют неструктурные (непараметрические) и структурные (параметрические) методы анализа временного ряда. К первым относят оценивание по данным спектра Фурье, автокорреляционной функции, гистограммы и т.д. Эти методы характеризуются тем, что по временному ряду оценивается очень большое число параметров, “данные говорят сами за себя” [2] (так, например, значение автокорреляционной функции для каждого времени задержки — это отдельный параметр). Ко вторым относят методы, ориентированные на оценку по данным небольшого числа параметров при некоторых дополнительных предположениях о свойствах наблюдаемых величин. Например, построение гистограммы по наблюдаемым значениям — это неструктурный подход к оцениванию плотности распределения вероятностей случайной величины. А выбор явного вида функции, например, $p(x) = \exp(-\frac{x^2}{2\sigma^2}) \Big/ \sqrt{2\pi\sigma^2}$, и оценка ее параметра σ^2 — подход структурный . К структурным относятся и подходы. о которых пойдет речь дальше.

2. Временные ряды

Временными рядами называют дискретные последовательности чисел, характеризующих состояние объекта наблюдения в отдельные моменты времени[2]. Такой способ представления информации о явлениях, эволюционных процессах, движениях весьма распространен. В одних случаях дискретное задание наблюдаемой естественно или даже единственно возможно. Например, курс валют устанавливается с дневным интервалом, статистические данные о состоянии производства собираются по годам и кварталам, подсчеты численности биологических популяций ведутся по сезонам, данные о погоде представляются метеостанциями также дискретно. В других случаях дискретность является результатом приближения или связана с выбором способа наблюдения. Так непрерывное во времени изменение положения движущегося тела становится дискретным при его наблюдении с использованием стробоскопа, а непрерывный сигнал – после преобразования с помощью аналого-цифрового преобразователя.

Далее мы будем рассматривать лишь примеры эволюции во времени и временные ряды с отсчетами, сделанными с постоянным шагом – через равные интервалы времени Δt (*интервалы выборки*). Члены ряда v_i – значения наблюдаемой величины в дискретные моменты – будем называть *точками, i* – порядковым номером точки или *дискретным временем;* количество точек в ряде N – *длиной ряда* (или *длиной временной реализации)*. Для обозначения самого ряда будем использовать фигурные скобки: $\{v_i\} = v_1, v_2, ..., v_N$.

Скалярным называют ряд, сформированный из отсчетов скалярной наблюдаемой – когда величина, характеризующая моделируемый объект или явление единственна и ей соответствует точка на числовой оси. В случаях, когда

[2] "Дискретный" – раздельный, состоящий из отдельных частей. В отличие от "непрерывного" (способного принимать любое значение, без промежутков, "континуального") дискретный набор значений содержит, например, только 0 или 1, только целые, только рациональные числа и т.п. Здесь речь идет о дискретности значений только времени, а значения величины, характеризующей объект наблюдения, (будем ее называть просто "наблюдаемой" или "переменной") могут быть любыми.

состояние объекта или явление в данный момент времени характеризуется двумя и более скалярными величинами, говорят о *векторном* ряде.

3. Экстраполяция временной зависимости

Пусть имеется скалярный временной ряд $\{v_i\}_{i=1}^{N}$. Необходимо найти (предсказать) значения величины v в моменты времени $t > t_N$. Одним из возможных подходов является *экстраполяция*[3] наблюдаемой временной зависимости. Для реализации этого подхода предполагают явный вид функциональной зависимости v от t и оценивают неизвестные параметры по временному ряду. Затем полученная функция используется для прогноза дальнейшего поведения. Обычно ищется зависимость среднего значения v от t – *регрессия*[4]. Например, если экспериментальные точки на плоскости v-t (рис.1а) располагаются вдоль некоторой прямой, целесообразно использовать линейную функцию $v(t)$ (линейный регрессионный анализ). Модель строится в виде

$$v_i = b_0 + b_1 t_i + a_i, \qquad (1)$$

где a_i — независимые одинаково распределенные случайные величины: значения некоторого случайного процесса в моменты времени t_i. Параметры b_0 и b_1 оцениваются методом максимального правдоподобия (см. Приложение). Если предположить, что величины a_i распределены по нормальному закону, то этот

[3] Экстраполяцией называется продолжение некоторой функции $v(t)$ за пределы области ее определения. В физике под экстраполяцией часто понимается распространение полученной экспериментально временной зависимости $v(t)$ на другие промежутки времени. В зависимости от того, насколько значения, полученные путем непосредственных измерений на новом промежутке, совпадают со значениями функции $v(t)$ на нем говорят об успехе или неуспехе экстраполяции.

[4] Регрессией называется зависимость среднего значения некоторой величины v от другой величины параметра t. Иногда этим же термином обозначают и процесс отыскания рассматриваемой зависимости. Необходимость отыскания функции среднего возникает в двух основных случаях: первый — когда по каким-либо причинам невозможно точно измерить величину v; второй случай заключается в том, что природа измеряемой величины такова, что она в действительности v принимает различные значения при одном и том же t в разных экспериментах (то есть, существуют некоторые факторы, которые не удается учесть в процессе измерения вследствие незнания, либо эти факторы намеренно игнорируются).

метод сводится к методу наименьших квадратов. Выбираются такие значения параметров модели, которые минимизируют средний квадрат «ошибки» a:

$$\frac{1}{N}\sum_{i=1}^{N}(v_i - b_0 - b_1 t_i)^2 = \min. \qquad (2)$$

В данном случае задача сводится к решению линейной системы двух алгебраических уравнений. После того, как значения параметров b_0 и b_1 определены, наиболее вероятное значение величины v в любой момент времени t можно вычислять по формуле $v(t) = b_0 + b_1 t$. Можно также определить границы доверительного интервала, в который с заданной вероятностью попадет значение $v(t)$. Экстраполяция и состоит в использовании полученной формулы для прогноза значений v при $t > t_N$.

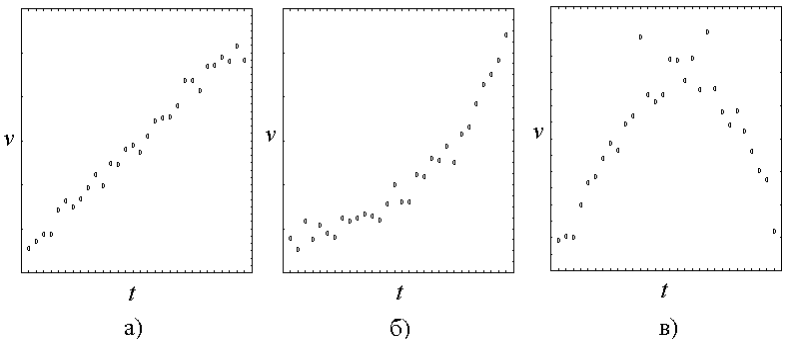


Рис.1. Возможные варианты изменения наблюдаемой величины v, для которых достаточно просто подобрать явный вид зависимости $v(t)$: а) линейная, б) экспоненциальная, в) полиномиальная (3-го порядка).

Для проверки адекватности полученной модели экспериментальным данным полезно провести анализ остатков (остаточных ошибок) модели[5]. Модель (1) можно считать удовлетворительной, если остатки некоррелированы и распределены (приблизительно) по нормальному закону. Для проверки этих утверждений необходимо рассчитать автокорреляционную функцию остатков и плотность их распределения (построить гистограмму).

В общем случае могут потребоваться другие аппроксимирующие функции (см. например, рис.1б,в, где точки располагаются явно нелинейно). Как при выборе линейного вида зависимости $v(t)$, так и при использовании нелинейных функций (экспоненты, полинома и т.д.), описанный метод прогноза эффективен

[5] Поясним, что такое остатки или остаточные ошибки модели. Пересчитаем наиболее вероятные значения переменной v, исходя из построенной модели (т.е. $v_i = b_0 + b_1 t_i$), для моментов времени t_i. Эти значения называются предсказанными значениями. Разность между исходными и предсказанными значениями называется остатками.

лишь для узкого класса процессов, соответствующих выбранному виду временной зависимости. Однако во многих случаях удовлетворительно подобрать явный вид $v(t)$ не удается.

4. Модель в виде случайного процесса

В тех случаях, когда в распределении экспериментальных точек на плоскости *v*-*t* не просматривается какой-либо закономерности (разброс точек очень велик), может оказаться более эффективным моделирование временного ряда *случайным процессом*. Случайный процесс — это функция $v = v(t_i, \omega)$, где ω — случайное событие (т.е. значение *v* в любой момент времени является случайной величиной).

В качестве базовой модели при таком подходе обычно принимают *нормальный белый шум* $a(t)$. Это случайный процесс, значения которого в различные моменты времени статистически независимы, а значения в любой фиксированный момент времени распределены одинаково по нормальному закону $p(a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{a^2}{2\sigma^2}}$, где σ^2 — дисперсия[6] (среднее значение принято равным нулю).

Однако свойства наблюдаемого процесса могут противоречить гипотезе о том, что это нормальный белый шум (например, автокорреляции могут быть существенно отличны от нуля для ненулевых задержек). В этом случае полезным для многих практических ситуаций оказался следующий подход. Предпо-

[6] В общем случае, чтобы полностью описать случайный процесс, необходимо определить всевозможные конечномерные распределения этого процесса. Однако можно показать, что для полного описания нормального случайного процесса (т.е. процесса, для которого одномерное распределение в любой момент времени является нормальным) достаточно задать только его среднее значение, дисперсию и автокорреляционную функцию. Таким образом, для построения модели в виде нормального белого шума по данным нужно оценить только его дисперсию и среднее значение (значения автокорреляционной функции при любой ненулевой задержке равны нулю). Нормальный белый шум описывает так называемое случайное блуждание, можно сказать, что это наиболее непредсказуемый процесс. Кроме того, выделенное положение нормального закона распределения обусловлено тем, что суммарный

лагалось, что наблюдаемый процесс — это нормальный белый шум, преобразованный *линейным фильтром*[7].

Значение нормального белого шума, преобразованного линейным фильтром, для любого момента времени t_n определяется выражением

$$v_n = a_n + \sum_{i=1}^{\infty} \psi_i a_{n-i}, \qquad (3)$$

где веса ψ_i должны удовлетворять условию $\sum_{i=1}^{\infty} \psi_i^2 \le const$, чтобы процесс v_n был стационарным[8].

Таким образом, вид модели выбран на основе двух предположений: к наблюдаемому временному ряду имеют отношение нормальный белый шум и линейный фильтр. Для построения модели вида (3) необходимо оценить по данным дисперсию белого шума и веса ψ_i.[9]

5. Модель скользящего среднего

Разумеется, вычислить бесконечное количество весов не представляется возможным, однако, как правило, значения ψ_i быстро убывают с ростом номера i и на практике достаточно ограничиться моделью с конечным числом весов q. Таким образом, можно получить модель скользящего среднего порядка q — MA(q) (от английского «Moving Average» — скользящее среднее):

$$v_n = a_n - \sum_{i=1}^{q} \theta_i a_{n-i}, \qquad (4)$$

вклад очень большого числа случайных факторов распределен асимптотически нормально (согласно центральной предельной теореме).

[7] Как известно, линейный фильтр обладает тем свойством, что если на его вход подан нормальный белый шум, то выходной сигнал может не быть δ-коррелированным. Кроме того, линейные фильтры были хорошо изучены и широко использовались на практике.

[8] Процесс называется *стационарным в узком смысле*, если его всевозможные конечномерные распределения не меняются при сдвиге по времени. Процесс называется *стационарным в широком смысле*, если его среднее значение и дисперсия не зависят от времени (причем дисперсия конечна), а автокорреляционная функция зависит только от модуля разности аргументов. Для рассматриваемого нормального случайного процесса оба понятия совпадают.

[9] В принципе, нужно также оценить среднее значение шума. Но мы (без потери общности) будем считать, что наблюдаемый процесс предварительно приведен к нулевому среднему.

Эта модель содержит q+1 параметров (θ_1, θ_2, …, θ_q и σ^2), значения которых нужно оценить по временному ряду.

6. Модель авторегрессии

Заметим теперь, что общее выражение (3) можно эквивалентно переписать в виде:

$$v_n = a_n + \sum_{i=1}^{\infty} \pi_i v_{n-i}, \tag{5}$$

где веса π_i выражаются[10] через ψ_i.

Аналогично, для построения модели вида (5) нужно было бы оценить значения бесконечного количества весов π_i, но практически во многих случаях вполне достаточно ограничиться некоторым конечным числом. Таким образом, приходим к модели, которая называется процессом авторегрессии порядка p — AR(p) (от английского «AutoRegressive» — авторегрессионная):

$$v_n = a_n + \sum_{i=1}^{p} \phi_i v_{n-i}, \tag{6}$$

Эта модель содержит p+1 параметров (ϕ_1, ϕ_2, …, ϕ_p и σ^2), значения которых нужно оценить по временному ряду (причем значения параметров должны удовлетворять определенным соотношениям [2], чтобы процесс был стационарным).

[10] Переход от (3) к (5) можно осуществить следующим образом: нужно последовательно исключать из выражения (3) величины a_{n-1}, a_{n-2} и т.д. Для этого сначала нужно выразить значение шума a_{n-1} через v_{n-1} и предыдущие значения a с помощью формулы $a_{n-1} = v_{n-1} - \sum_{i=1}^{\infty} \psi_i a_{n-1-i}$, затем подставить это выражение в формулу (3), исключив из нее, таким образом, a_{n-1}. Далее аналогично исключается a_{n-2}, и т.д.)

7. ARIMA- модель

Практически наиболее эффективный подход состоит в том, чтобы объединить модели (4) и (6)[11]. Таким образом, получаем модель авторегрессии и скользящего среднего порядка (p, q) — ARMA(p, q):

$$v_n = a_n + \sum_{i=1}^{p} \phi_i v_{n-i} - \sum_{i=1}^{q} \theta_i a_{n-i}, \qquad (7)$$

которая содержит $p+q+1$ параметров.

Если наблюдаемый ряд $v(t)$ имеет признаки нестационарности (например, какие-либо детерминированные тренды — линейный, полиномиальный и т.п.), то стационарный процесс (7) не может быть адекватной моделью. Однако в таком случае может оказаться стационарной некоторая разность наблюдаемого процесса порядка d: $w_n = \nabla^d v_n$, где $\nabla v_n = v_n - v_{n-1}$ — первая разность (аналог дифференцирования), а ∇^d означает последовательное применение d раз оператора ∇. Для описания процесса w_n уже может быть эффективной ARMA-модель. Таким образом, приходим к модели авторегрессии и проинтегрированного скользящего среднего порядка (p, d, q) — ARIMA(p, d, q) (от английского — «AutoRegressive Integrated Moving Average»):

$$\begin{aligned} w_n &= a_n + \mu + \sum_{i=1}^{p} \phi_i w_{n-i} - \sum_{i=1}^{q} \theta_i a_{n-i}, \\ w_n &= \nabla^d v_n. \end{aligned} \qquad (8)$$

Чтобы выразить значения наблюдаемого процесса v_n через значения процесса w_n, описываемого ARMA-моделью (в которую при необходимости может быть включен и постоянный член μ – как в (8)), нужно использовать оператор сум-

[11] Эта необходимость вызвана следующими обстоятельствами. Предположим, что наблюдаемый временной ряд генерируется процессом авторегрессии порядка 1. Если попытаться описать его процессом скользящего среднего, то потребуется модель (4) с бесконечным числом параметров θ_i (по крайней мере, с очень большим). Оценки значений большого числа параметров менее надежны, и в данном случае это обязательно приведет к существенному снижению эффективности модели. И обратно, если ряд генерируется процессом скользящего среднего порядка 1, то для его описания потребовался бы процесс авторегрессии очень высокого порядка. Поэтому целесообразно объединить в модели выражения (4) и (6), чтобы

мирования (аналог интегрирования), обратный оператору ∇. Этим объясняется наличие слова «проинтегрированного» в названии модели.

8. Методика построения ARIMA-модели по временному ряду

Общий подход предложенный Боксом и Дженкинсом [2], представлен схемой на рис.2.

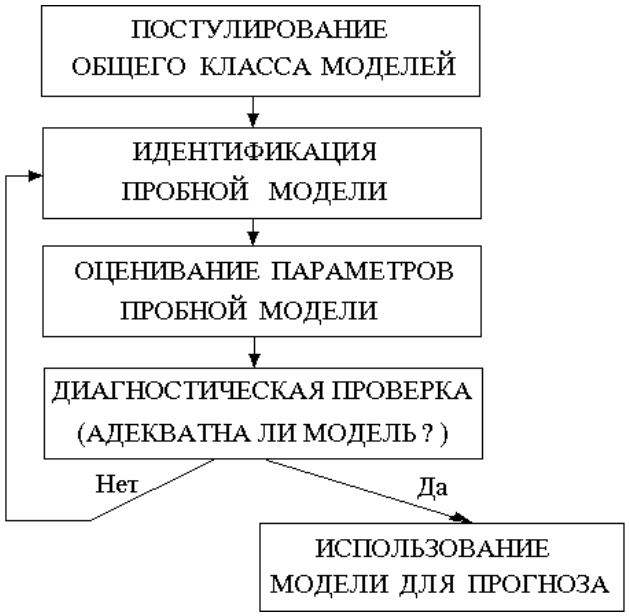


Рис.2. Схема построения модели по временному ряду (приводится по [2]).

Первый этап моделирования — *выбор (постулирование) общего класса моделей*. В данном случае это класс ARIMA-моделей.

Второй этап — *идентификация* пробной модели. Под идентификацией модели понимают определение подкласса моделей, наиболее подходящего для описания процесса. В данном случае выбор подкласса — это выбор конкретных значений величин p, d, q (на этом этапе делаются также грубые оценки параметров θ_1, θ_2, …, θ_q, ϕ_1, ϕ_2, …, ϕ_p, σ^2 и μ). Выбор подкласса осуществляется, главным образом, на основе анализа автокорреляционной функции наблюдаемого ряда[12]. Обычно редко используются модели, для которых хотя бы одна из величин p, d, q больше 2. Не следует без необходимости выбирать слишком большое значение d, т.к. это приводит к менее стабильным оценкам параметров модели.

Третий этап — *оценивание* (*подгонка*) параметров пробной модели. На этом этапе с помощью специальных численных процедур по наблюдаемым

можно было экономично (при помощи небольшого числа параметров) описать наблюдаемый процесс и вида (4), и вида (6), и смешанный.

[12] Так, если она спадает экспоненциально, то следует выбрать $p = 1$, $q = 0$; если она имеет большое значение при задержке, равной 1, и равна нулю для других значений, то следует использовать значения $p = 0$, $q = 1$ и т.д.

данным оцениваются значения параметров θ_1, θ_2, …, θ_q, ϕ_1, ϕ_2, …, ϕ_p, σ^2 и μ. Значения параметров вычисляются на основе принципа максимального правдоподобия, который приводит в данном случае к нелинейному методу наименьших квадратов.

Четвертый этап — *диагностическая проверка* модели. На этом этапе выясняется, не имеет ли наблюдаемый ряд каких-либо свойств, противоречащих построенной модели. В данном случае исследуются следующие свойства данных: автокорреляционная функция остаточных ошибок модели (в данном случае остаточные ошибки — это предполагаемые значения a_n) и их спектр Фурье. Если отличие этих характеристик от характеристик нормального белого шума статистически значимо, то модель следует признать неадекватной.

Если диагностическая проверка показала, что модель не адекватна наблюдаемым данным, то, возможно, нужно вернуться к этапу оценивания и использовать другую численную процедуру. Но в [2] показано, что процедура оценивания, применяемая для ARIMA-моделей, эффективно использует данные, поэтому более вероятно, что модель не верно идентифицирована (второй этап) и следует опробовать другой подкласс моделей. Если же и выбор других подклассов моделей не приводит к успеху, то выбранный общий класс моделей (ARIMA-модели) не подходит для описания наблюдаемого процесса[13].

Данный подход интенсивно развивался с 1927 года (когда впервые была предложена авторегрессионная модель для описания временного ряда годовых чисел солнечных пятен [1]) в течение следующих 50 лет. Были детально разработаны и обоснованы [2,3] методы выбора оптимальных значений *p*, *d*, *q*, вычисления оптимальных значений параметров $\theta_i, \phi_i, \sigma^2$ и проверки адекватности

[13] Отметим, что перед построением модели иногда проводят и некоторые другие преобразования временного ряда (кроме взятия первой разности). Во-первых, при наличии сезонных изменений в ряде используют разность со сдвигом *s*: $\nabla_s v_i = v_i - v_{i-s}$. Во-вторых, если размах колебаний величины *v* возрастает с течением времени, то часто применяется операция логарифмирования наблюдаемого ряда с целью стабилизации дисперсии. Это преобразование можно интерпретировать следующим образом: чем выше абсолютное значение переменной,

построенной модели наблюдаемым данным. ARIMA-модели оказались достаточно эффективны для описания различных процессов. Они применялись и применяются в технике, метеорологии, экономике, социологии для решения задач прогнозирования и/или управления. В настоящее время эти методы уже реализованы в прикладных программах с богатыми графическими возможностями, в частности, в системе **Statistica**.

9. Практическое задание

В данной работе предлагается

- построить аппроксимацию предложенной временной зависимости,
- познакомиться со способами предварительной обработки ряда,
- потренироваться в выборе параметров и использовании ARIMA-модели,
- проверить адекватность модели с помощью анализа остаточных ошибок.

Одной из целей работы является также знакомство с системой **Statistica**[14] (фирма-производитель StatSoft Inc., USA) — мощным средством статистического анализа и обработки данных (см., например, [4]).

Задания:

Дважды щелкните мышью на ярлыке программы **Statistica**. На экране появится переключатель модулей. Выберите модуль **Множественная регрессия — Multiple Regression**. После запуска модуля на экране откроется основное окно системы **Statistica**. При запуске системы в нее автоматически загружается последний файл, с которым вы работали в ней. Одновременно с этим появляется *Стартовая панель модуля*, содержащая основные операции, которые доступны в данном модуле, и позволяющая определить различные параметры анализа.

тем выше и уровень случайных ошибок. При логарифмировании все ошибки становятся примерно одинаковыми.

[14] По вопросам приобретения программы можно обращаться: корпорация СофтЛайн, 17036, Москва, ул. Шверника, дом 4, тел. (095) 232-0023, 126-9969, 126-9065, e-mail: root@softline.msk.su, или: StatSoft Inc., 2325 Easr 13th Street, Tulsa, OK 74104, USA, тел. (918) 583-4149.

Задание 1. Построить аппроксимацию зависимости оптовой цены на марочные вина от их «возраста».

1) Исходные данные в системе **Statistica** организованы в виде таблицы. Столбцы таблицы называются *Variables — Переменные*, строки *Cases — Случаи*. В качестве переменных выступают исследуемые величины, случаи — это значения, которые принимают переменные и которые меняются в процессе наблюдения. Откройте файл с таблицей данных (файл prices.sta): *оптовые цены на марочные вина в зависимости от года закладки* [4]. В таблице имеется две переменных, зависимость между которыми требуется найти, — в первом столбце переменная *Год* (год закладки вина), во втором — *Цена* (цена бутылки в долларах). Перед применением статистических процедур может потребоваться преобразование данных. Так, перейдем от переменных *Год* и *Цена* к новым переменным *Возраст* и *Цена_Лог*, которые связаны с исходными формулами:

$$Возраст = 1972 - Год,\ Цена\_Лог = \ln(Цена).$$

После этого таблица будет содержать четыре переменные. Формулы для преобразования переменных задаются в диалоговом окне спецификаций переменной. Для его вызова достаточно дважды щелкнуть мышью на имени переменной в таблице с данными. Теперь имеет смысл отобразить данные на графике. Воспользуйтесь пунктом меню Graphics и выберите нужный тип графика (2D Scatterplots).

2) Будем искать зависимость вида (1) между переменными *Цена_Лог* и *Возраст* (фактически мы располагаем временным рядом значений цены на вино, только выборка произведена неравномерно). Оценим параметры модели (1) и проверим адекватность построенной модели исходным данным.

Необходимо вызвать стартовую панель модуля (пункт меню **Analysis — Startup Panel**). Далее нажмите кнопку **Variables**. В открывшемся окне нужно выбрать переменные для анализа — зависимую и независимые (*Цена_Лог* и *Возраст*). Нажмите кнопку **OK**. Выберите и дополнительные опции и параметры анализа (например, расчет с расширенной точностью и выбор метода по

умолчанию). В стартовой панели нажмите кнопку **OK**. Система произведет вычисления и через секунду появится окно результатов. Верхняя часть окна — информационная, нижняя содержит функциональные кнопки, позволяющие подробно просмотреть результаты анализа.

В информационной части смотрим прежде всего на значение коэффициента детерминации RI (он показывает долю общего разброса значений зависимой переменной, которая объясняется построенной регрессией). В данном случае оно равно 92.9 %, что является хорошим результатом. Щелкните кнопку **Regression Summary**, чтобы просмотреть таблицу с результатами анализа.

3) Важным моментом при проверке адекватности модели является анализ остатков. В окне **Результаты множественной регрессии** нажмите кнопку **Residual Analysis — Анализ остатков**. Вы откроете окно **Анализ остатков**. Для оценки адекватности модели построим график остатков на нормальной вероятностной бумаге (**Normal Probability Plot of Residuals**). Если они достаточно хорошо ложатся на прямую, соответствующую нормальному закону, то предположение о нормальном распределении ошибок a_i выполнено.

4) Возможен вывод результатов анализа в файл с отчетом (для этого выберите пункты меню File—Page Output/Setup и установите флажок напротив опции Window – печать в окно). Для него на рабочем пространстве системы откроется специальное окно. В него можно распечатать любой документ (таблицу или график). Создайте такой файл с результатами вашей работы.

Задание 2. Осуществить прогноз ряда количества перевозок авиапассажиров с помощью ARIMA-модели.

1) Перейдем теперь к построению ARIMA-моделей (8). Откройте модуль **Анализ временных рядов/Прогнозирование**. Откроется стартовая панель модуля. С помощью кнопки **Open Data** откройте файл с данными для анализа (файл series_g.sta, в котором содержится одна переменная: количество общих перевозок авиапассажиров в данном месяце; все месяцы пронумерованы от 1 до

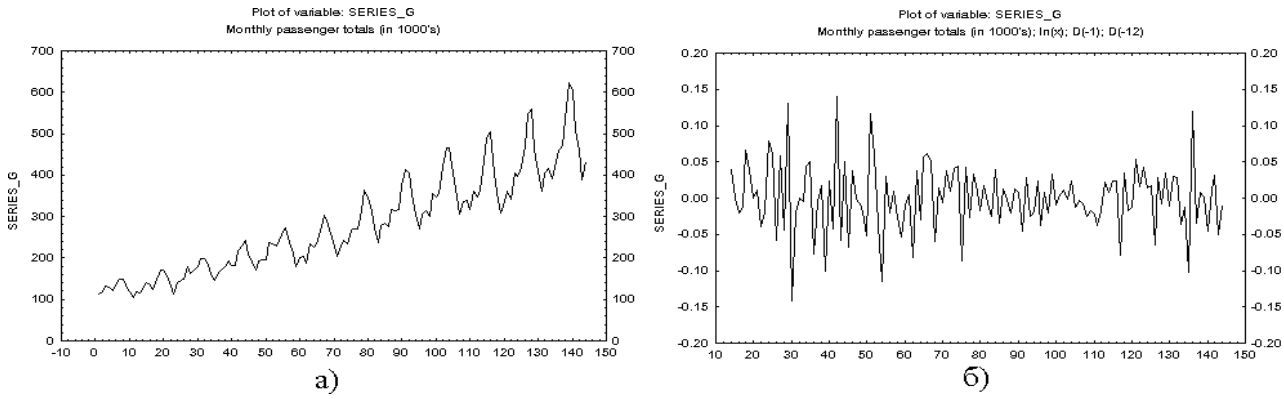


Рис.3. Наблюдаемый график зависимости количества перевозок авиапассажиров от времени (а) и график после преобразований временного ряда: ln(v), разность со сдвигом 1 и разность со сдвигом s = 12 (б).

144, с января 1949 по декабрь 1962 года). С помощью кнопки **Variables** выберите переменную для анализа. В данном случае имеется только одна переменная. На стартовой панели модуля инициируйте кнопку **ARIMA & autocorrelation functions**. На экране появится стартовая панель **Single Series ARIMA**.

2) Прежде всего, просмотрите данные графически. Для этого инициируйте кнопку **Plot**. На экране появится график временного ряда из открытого файла данных (рис.3а). В данном графике имеются отчетливые годовые периоды, присутствуют резко выраженные пики, амплитуда колебаний возрастает. В ряде имеется отчетливый тренд: средние значения перевозок постепенно увеличиваются. Имеется также и сезонность: как и следовало ожидать, из года в год пик перевозок приходится на одни и те же месяцы — июль либо август. Характер перевозок также очень похож со сдвигом на год. Просмотрев ряд на графике, вернитесь в стартовое окно **ARIMA**, нажав на кнопку **Continue**.

3) Прежде, чем подогнать к временному ряду авторегрессионную модель, его следует «сделать стационарным». Мы будем последовательно преобразовывать ряд, делая его раз за разом все более похожим на стационарный. Инициируйте кнопку **Other transformations & plots**. Откроется окно **Преобразования переменных**. Выберите опцию **Plot variables (series) after each transformation**. Теперь система будет автоматически показывать графики преобразованных данных после каждого преобразования ряда. Инициируйте кнопку

OK(Transform highlighted variables). Откроется окно **Преобразования временного ряда**. Следует выбрать какое-либо преобразование и нажать кнопку **OK(Transform)**. На каждом шаге можно выполнить только одно преобразование значений переменной.

Прежде всего, применим преобразование **Natural log**. Далее возьмем первую разность ряда (устраняем линейный тренд), смотрите опции в правом нижнем углу окна. Вспоминая, что ряд имеет сезонную составляющую, возьмите сезонную разность: выберите преобразование **Differencing** со сдвигом *lag* = 12 (см. рис.3б). Нажмите кнопку **Exit** для выхода в окно **ARIMA**.

9) Вернемся в окно **Single Series ARIMA** (нажмите кнопку **Continue**, а затем **Exit**). Нужно задать значения следующих параметров ARIMA-модели: **p — Autoregressive, P — Seasonal, q — Mouving average, Q — Seasonal.**[15] По меньшей мере, один из этих параметров должен быть отличен от нуля. Выберем значения: p = P = 0, q = Q = 1 (здесь сезонные параметры P и Q должны соответствовать сдвигу *lag* = 12).

Пометьте нужные опции, показывая системе, какие преобразования производились с исходным рядом. Выберите метод оценки параметров, например, **Exact — Точный**. В подокне **Variables** не забудьте высветить первую строку — исходный временной ряд. Запустите процедуру оценивания щелчком кнопки **OK(Begin parameter estimation)**. Откроется окно **Оценивание параметров**. Если находите оценки приемлемыми, щелкните **OK** и просмотрите всесторонне результаты.

4) Откроется окно **Результаты ARIMA**. Просмотрите численные оценки, щелкнув по кнопке **Parameter estimates**. Обязательно обратите внимание на остатки и просмотрите их графики. Щелкните, например, кнопку **Normal** (вы построите график на нормальной вероятностной бумаге: он должен быть близок к прямой, тогда модель можно считать адекватной исходным данным). Щелкнув кнопку **Hist**, вы увидите гистограмму значений остатков с наложен-

ной нормальной плотностью. Полезно также просмотреть автокорреляции остатков (**Autocorrelations of residuals**).

5) В левой части окна результатов имеется группа кнопка **Forecasting — Прогнозирование**. Установите опции — прогноз на 24 случая вперед, уровень доверия 0.9. Щелкните кнопку **Plot series & forecasts**. Вы увидите прогноз, который дает построенная модель.

Задание 3. Осуществить прогноз того же ряда (что и в задании 2) и сравнить его с экспериментальными данными.

Теперь аналогичным образом подгоните модель ARIMA к тому же ряду, используя случаи с 1 по 120 (воспользуйтесь кнопкой **Select cases** на стартовой панели модуля). Осуществите прогноз на 24 месяца вперед и сравните его с экспериментальными данными.

Задание 4. Провести анализ временного ряда курса акций IBM.

Проведите аналогичный анализ временного ряда курса акций IBM (файл ibm1.sta). Сохраните наиболее существенные результаты своей работы в файле отчета.

Приложение.

О методах максимального правдоподобия и наименьших квадратов

В статистике одним из часто используемых методов оценки параметров некоторого распределения по экспериментальным данным является метод максимального правдоподобия. Пусть $p(a)$ — функция плотности распределения вероятностей случайной величины a в некоторый момент времени t (для любого t одинакова). Пусть эта функция зависит от параметров, значения которых следует оценить по экспериментальным данным. Совместная N-мерная плотность распределения вероятностей величин $a(t_i)$ в моменты времени t_i (i = 1,

[15] Сезонные параметры P и Q отвечают за добавку в правую часть модели (7) еще двух сумм,

…, *N*) называется «функцией правдоподобия» **L** и в случае независимых $a(t_i)$ выражается как:

$$\mathbf{L} = p_N(a_1, a_2, \ldots, a_N) = p(a_1)p(a_2)\ldots p(a_N). \qquad (\text{П}.1)$$

Для рассмотренного примера (1) функция правдоподобия зависит от параметров b_0 и b_1, т.к. $a_i = v_i - b_0 - b_1 t_i$. Метод максимального правдоподобия состоит в том, что значения параметров b_0 и b_1 выбираются таким образом, чтобы максимизировать **L**. Их можно определить из системы уравнений ($\partial\mathbf{L}/\partial b_0 = 0$ и $\partial\mathbf{L}/\partial b_1 = 0$) или ($\partial \ln\mathbf{L}/\partial b_0 = 0$ и $\partial \ln\mathbf{L}/\partial b_1 = 0$).

Другим методом оценки является метод наименьших квадратов. Пусть случайные величины $a(t_i)$ — это отклонения значений наблюдаемой величины от значений аппроксимирующей функции. Для примера (1) $a_i = v_i - b_0 - b_1 t_i$ — аппроксимирующей функцией является линейная функция. Значения параметров определяются из условия, чтобы сумма квадратов отклонений $a(t_i)$ была минимальна:

$$\sum_{i=1}^{N} a^2(t_i) = \min. \qquad (\text{П}.2)$$

Можно показать, что в том случае, когда предполагается нормальное распределение случайных величин $a(t_i)$, метод максимального правдоподобия и метод наименьших квадратов совпадают.

в которых значения *v* и *a* берутся с некоторым сдвигом *lag*, не равным 1.

Литература

1. Yule G.U. «On a method of investigating periodicities in disturbed series with special reference to wolfer's sunspot numbers», Phil.Trans.R.Soc.London A, 1927, Vol. 226, P. 267-298.
2. Бокс Дж., Дженкинс Т. «Анализ временных рядов. Прогноз и управление», М.: Мир, 1974, 242 с.
3. Льюнг Л. «Идентификация систем. Теория для пользователя», М.: Наука, 1991, 432 с.
4. Боровиков В.П., Боровиков И.П. «Statistica. Статистический анализ и обработка данных в среде Windows», М.: Информационно-издательский дом "Филинъ", 1997, 608 с.

Контрольные вопросы

1. Что такое временной ряд? Каким образом могут быть получены временные ряды на практике?
2. В чем заключается метод экстраполяции наблюдаемой временной зависимости для прогноза?
3. Что такое модели скользящего среднего и авторегрессии? Каковы предпосылки для использования этих математических конструкций?
4. Что такое ARIMA-модель? Когда возникает необходимость в использовании такой модели?
5. Какие преобразования наблюдаемого временного ряда применяются перед построением модели? Какова их цель?
6. Из каких основных этапов состоит процедура построения ARIMA-модели по временному ряду? Каково содержание каждого из этих этапов (какие операции выполняются)?

Учебно-методическое пособие

БЕЗРУЧКО Борис Петрович

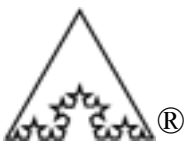
СМИРНОВ Дмитрий Алексеевич

Статистическое моделирование по временным рядам

ГосУНЦ «Колледж», Лицензия ЛР №020773 от 15.05.98

Подписано к печати 31.03.2000. Формат 60x84 1/16.
Бумага Papirus Slim. Гарнитура Times.
Усл. печ. л. 1,39 (1,5). Уч.-изд. л. 1.1. Тираж 100 экз. Заказ 161.

Издательство ГосУНЦ «Колледж»
410026, Саратов, ул. Астраханская, 83.
Тел. (845-2) 523864



Отпечатано на ризографе издательства ГосУНЦ «Колледж»