МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра «Прикладная математика и информатика»
(наименование кафедры)

02.03.03 Математическое обеспечение и администрирование информационных систем
(код и наименование направления подготовки, специальности)

Технология программирования
(направленность (профиль)/специализация)

БАКАЛАВРСКАЯ РАБОТА

на тему «Реализация алгоритма прогнозирования на основе авторегрессионной модели ARIMA»

Студент Н. М. Шогунова _____
(И.О. Фамилия) (личная подпись)

Руководитель Г. А. Тырыгина _____
(И.О. Фамилия) (личная подпись)

Консультанты К. А. Селиверстова _____
(И.О. Фамилия) (личная подпись)

Допустить к защите

Заведующий кафедрой к. т. н., доцент А. В. Очеповский _____
(ученая степень, звание, И.О. Фамилия) (личная подпись)

«_____» _____ 20___ г.

Тольятти 2019

АННОТАЦИЯ

Тема Бакалаврской работы: «Реализация алгоритма прогнозирования на основе авторегрессионной модели ARIMA».

В данной выпускной квалификационной работе исследуется статистическая модель для анализа дискретных временных рядов. Рассматривается авторегрессионная интегрированная скользящая средняя (ARIMA) модель. В работе изучены свойства модели и представлены статистические методы для спецификации модели, оценки параметров и проверки модели. Описано приведение нестационарного временного ряда к стационарному ряду. В результате, на основании рассмотренной модели, была реализована программа для решения задачи прогнозирования случайных данных. В качестве исходных данных был выбран временной ряд, отображающий колебания.

Были выполнены следующие задачи:

- выявлены основные виды авторегрессионных моделей временных рядов;
- обосновано использование модели ARIMA;
- осуществлена программная реализация алгоритма прогнозирования на основе модели ARIMA.

Для реализации в качестве языка программирования использовался Python с использованием библиотек: matplotlib, pandas, statsmodels. Было проведено исследование с использованием данной программы и получены остатки отображающие качество прогноза предложенной модели.

Цель состояла в том, чтобы предоставить работу, которая является практичным и полезным для людей, занимающихся разработкой систем для осуществления прогнозирования на основе применения моделей временных рядов.

Данная бакалаврская работа состоит из пояснительной записки на 40 стр., включая 12 рисунков и 3 приложений.

ABSTRACT

The title of the graduation work: "Implementation of the prediction algorithm based on the ARIMA autoregressive model."

In this final qualifying work, a statistical model is investigated for analyzing discrete time series. An autoregressive integrated moving average (ARIMA) model is considered. The work studies the properties of the model and presents statistical methods for model specification, parameter estimation, and model validation. The reduction of a non-stationary time series to a stationary series is described. As a result, based on the considered model, a program was implemented for solving the problem of predicting random data. As a source of data was selected time series, displaying fluctuations.

The following tasks were performed:

- the main types of autoregressive time series models were identified;
- the use of the ARIMA model is justified;
- implemented software implementation of the algorithm.

For implementation, programming was used as a language by Python using libraries: matplotlib, pandas, statsmodels. A study was conducted using this program and obtained residues reflecting the forecast quality of the proposed model.

The goal was to provide work that is practical and useful for people developing systems to implement forecasting using time series models.

This undergraduate work consists of an explanatory note on 40 p., including 12 pictures, and 3 applications.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Временной ряд - это последовательность наблюдений, сделанных последовательно во времени. Многие наборы данных отображаются в виде временных рядов: еженедельная последовательность количества товаров отправленных с завода, еженедельная серия данных о дорожно-транспортных происшествиях, суточных количествах осадков, ежечасных наблюдениях за выходом химического процесса и т.д. Примеры временных рядов можно найти в таких областях как экономика, бизнес, инженерия, естественные науки (особенно геофизика и метеорология) и социальные науки. Внутренняя особенность временного ряда заключается в том что, как правило, соседние наблюдения являются зависимыми. Характер этой зависимости среди наблюдений временного ряда представляет значительный практический интерес. Это требует разработки стохастических динамических моделей для временных рядов и использования таких моделей в различных областях для анализа данных. Постановка задачи состоит в том, чтобы для данных находящихся в файле data.csv осуществить адекватный прогноз. Этим определяется актуальность работы. Объект исследования – временной ряд. Предмет исследования – авторегрессионная модель ARIMA. Цель – реализация алгоритма прогнозирование на основе авторегрессионой модели ARIMA. Для реализации цели следует выполнить задачи:

1) выявить основные виды авторегрессионых моделей временных рядов;

2) обосновать использование модели ARIMA;

3) осуществить программную реализацию алгоритма этой модели.

Работа состоит из введения, трех глав, заключения, список используемой литературы и три приложения.

В первой главе рассматривается основные понятия теории временного ряда, их основные компоненты, стационарность и тесты для проверки ряда на стационарность. ([1]-[3])

В главе 2 рассматриваются конкретные стохастические процессы, используемые для моделирования временных рядов: авторегрессия (AR),

скользящая средняя (MA), смешанная авторегрессия скользящего среднего (ARMA) и авторегрессионная интегрированная скользящая средняя (ARIMA).([18]-[23]) Далее, в третьей главе реализован алгоритм прогнозирования на основе модели ARIMA с использованием языка программирования Python. ([7])

Глава 1 ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВРЕМЕННЫХ РЯДОВ

1.1 Основные компоненты временных рядов

Временной ряд - это последовательность числовых данных, в которой каждый элемент связан с конкретным моментом времени. Можно привести многочисленные примеры: месячная безработица, еженедельные показатели денежной массы, ежедневные цены закрытия фондовых индексов, ежедневные процентные ставки, ежечасный индекс процентной ставки и цены акции за минуту (или даже секунду) и так далее.

Анализ одной последовательности данных называется *одномерным анализом* временных рядов. Анализ нескольких наборов данных для одной и той же последовательности временных периодов называется *многомерным анализом временных рядов* или анализом нескольких временных рядов (например, анализ на основе ежемесячных данных, взаимосвязи между безработицей, уровнем цен, денежная масса и т.д. подпадает под анализ нескольких временных рядов). Целью анализа временных рядов является изучение динамики или временной структуры данных.

Анализ временных рядов включает в себя понимание различных аспектов о присущей природе рядов, чтобы создавать значимые и точные прогнозы. Для визуализации временного ряда представленного на рисунке 1 использовался пакет matplotlib. Рисунок показывает динамику временного ряда (Приложение А).

Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

- факторы, формирующие тенденцию временного ряда;
- факторы, формирующие циклическое колебание ряда;
- случайные факторы.

Рисунок 1.1 - График временного ряда ежемесячного противодиабетических препаратов с 2001 по 2018 год

Тенденция наблюдается при увеличении или уменьшении наклона, наблюдаемого во временном ряде. Тенденция можно наблюдать на рисунке 1.2 (Приложение А).

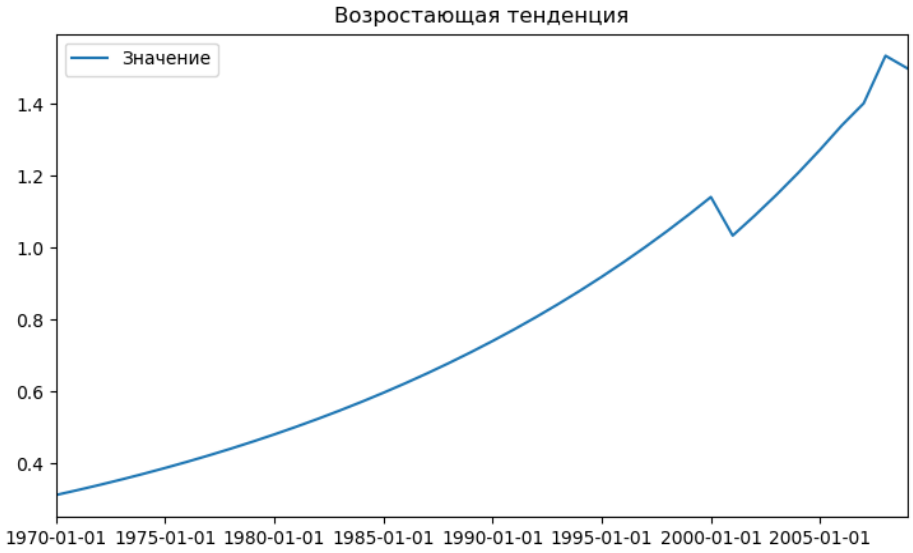

Рисунок 1.2 - Тенденция временного ряда

Факторы, формирующие циклическое колебание, могут носить сезонный характер, когда наблюдается отчетливая повторяющаяся картина, наблюдаемая между регулярными интервалами из-за сезонных факторов. Это может быть из-за месяца года, дня месяца, будних дней или даже времени суток. На рисунке

1.3 продемонстрирован пример графика сезонности временного ряда (Приложение А).

Рисунок 1.3 - График временного ряда, содержащего только сезонную компоненту

Временные ряды не содержат только тенденцию или циклической компоненту, а каждый их уровень образуется как комбинация тренда, сезонности и ошибки. На рисунке 1.4 можно наблюдать данный вид временного ряда (Приложение А).

Однако необязательно, чтобы все временные ряды имели тренд или сезонность. Временной ряд может не иметь четкой тенденции, но иметь сезонность. Верно и обратное.

1.2 Стационарность временных рядов

С теоретической точки зрения временной ряд представляет собой набор величин $\{X_t\}$. Такой набор случайных величин, упорядоченных по времени, называется случайным процессом. Слово «стохастик» имеет греческое происхождение и означает «относящимся к случайности». Если это непрерывная переменная, то случайные переменные принято обозначать через $X(t)$, а если t - дискретная переменная, то принято обозначать их через X_t.

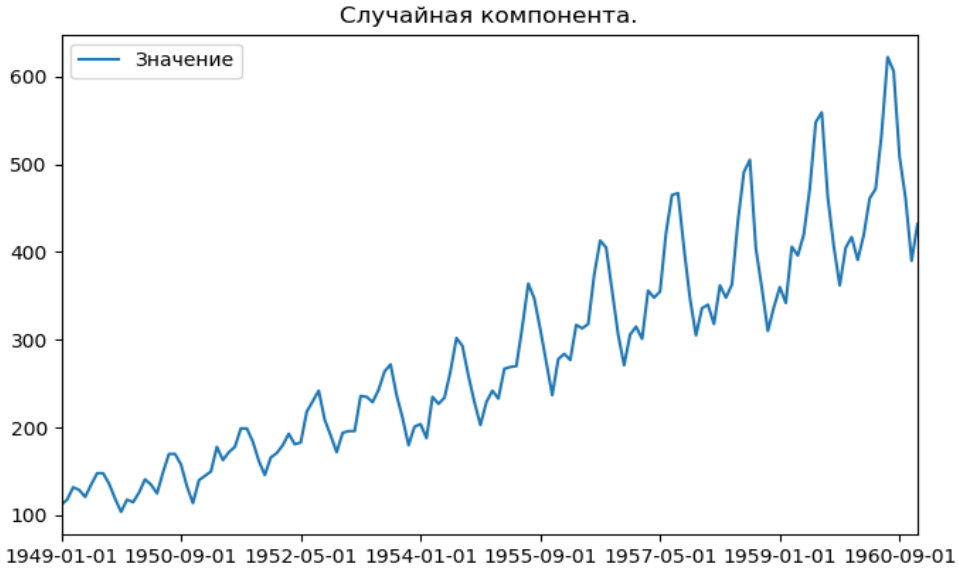

Рисунок 1.4 - График временного ряда, содержащего случайную компоненту

Примером непрерывной переменной $X(t)$ является запись электрокардиограммы. Примером дискретной случайной величины X_t является данные о безработице, денежной массе, ценах закрытия акции и т.д.

Случайные величины $\{X_t\}$, как правило, не являются независимыми. Кроме того, у нас есть только выборка размера 1 по каждой из случайных величин (например, если мы говорим, что уровень безработицы в конце этой недели является случайной величиной, у нас есть только одно наблюдение по этой конкретной случайной переменной). Нет другого способа получить наблюдение, поэтому мы имеем то, что называется *одиночной реализацией*.

Один из способов описания стохастического процесса – указать совместное распределение переменных X_t. Это довольно сложно и обычно не делается на практике. Вместо этого обычно делается то, что мы определяем первый и второй момент переменных X_t. Это:

1) среднее значение: $\mu\ t\ = E\ X_t\ ;$

2) дисперсия $\sigma^2\ t\ = var(X_t);$

3) автоковариации $\gamma\ t_1, t_2\ = cov\ X_{t_1}, X_{t_2}\ .$

Когда $t_1 = t_2 = t$, автоковариация равна всего $\sigma^2(t)$. Одним из важных случайных процессов является случай стационарных случайных процессов. Временной ряд называется строго стационарным, если совместное распределение любого набора n из наблюдений $X(t_1), X(t_2),..., X(t_n)$ такое же, как совместное распределение $X(t_1 + k), X(t_2 + k),..., X(t_n + k)$ для всех n и k.

Вышеприведенное определение строгой стационарности справедливо для всех значений n. Подставляя $n = 1$, мы получаем для всех t $\mu(t) = \mu$ и $\sigma^2(t) = \sigma^2$. Кроме того, если подставить $n = 2$, мы получим результат, что совместное распределение $X(t_1)$ и $X(t_2)$ такое же, как у $X(t_1 + k)$ и $X(t_2 + k)$. Записывая $t_1 + k = t_2$, мы видим, что это то же самое, что и распределение $X(t_2)$ и $X(t_2 + k)$.

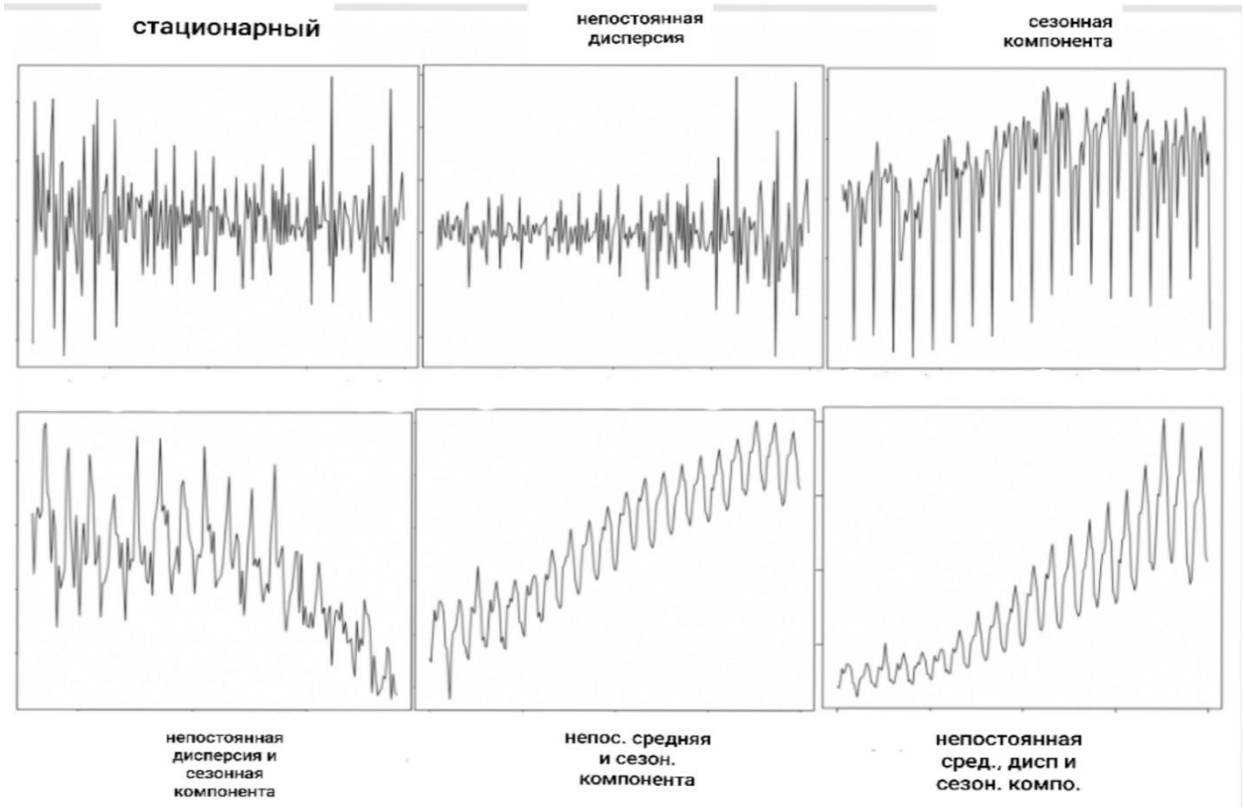

Рисунок 1.5 - Примеры стационарных и нестационарных временных рядов

Таким образом, это зависит только от разности $(t_2 - t_1)$, которая называется *лагом.* Следовательно, мы можем записать автоковариантную функцию $\gamma(t_1, t_2)$ как $\gamma(k)$, где $k = t_2 - t_1$ лаг. Таким образом, $\gamma(k) = \mathrm{cov}[X(t), X(t + k)]$ - коэффициент автоковариации при лаге k. $\gamma(k)$

называется функцией автоковариации и будет сокращенно обозначаться как АКВФ. $\gamma(0)$ - это конечно, дисперсия σ^2.

Поскольку коэффициенты автоковариации зависят от единиц измерения $X(t)$, удобно рассмотреть автокорреляции, свободные от единиц измерения. Поскольку $\text{var}\, X(t) = \text{var}\, X(t+k) = \sigma^2 = \gamma(0)$, мы имеем коэффициент автокорреляции $p(k)$ при лаге k как

$$p(k) = \frac{\gamma(k)}{\gamma(0)}$$

$p(k)$ называется функцией автокорреляции и будет обозначаться как АКФ. График $p(k)$ называется коррелограммой.

Для строго стационарных временных рядов распределение $X(t)$ не зависит от t. Таким образом, не только среднее значение и дисперсия являются постоянными. Все моменты высшего порядка не зависят от значения t, так же, как и все моменты высшего порядка совместного распределения любых комбинаций переменных $X(t_1), X(t_2), X(t_3), \ldots$. На практике это очень сильное предположение, и полезно определить стационарность менее ограничительным образом. Это определение касается только первого и второго моментов.

Временной ряд называется слабостационарным, если его среднее значение является постоянным, а его АКВФ зависит только от запаздывания, то есть

$$E[X(t)] = \mu \text{ и } \text{cov}[X(t), X(t+k)] = \gamma(k)$$

Не делается предположений о моментах высшего порядка. Альтернативными терминами, используемыми для слабостационарного временного ряда, являются стационарными в широком смысле, стационарными ковариационными или стационарными второго порядка.

Если $X(t_1), X(t_2), \ldots, X(t_n)$ имеют многомерное нормальное распределение, понятия строгой стационарности и слабой стационарности эквивалентны, так

как многомерное нормальное распределение полностью характеризуется первым и вторым моментами. Для других распределений это не так.

Прогнозировать стационарные ряды относительно легко и прогнозы более надежны. Причина заключается в том, что модели авторегрессионного прогнозирования по существу являются моделями линейной регрессии, которые используют лаг самого ряда в качестве предикторов.

Мы знаем, что линейная регрессия работает лучше всего, если предикторы (X переменные) не коррелируют друг с другом. Таким образом, стационаризация ряда решает эту проблему, поскольку устраняет любую устойчивую автокорреляцию, тем самым делая предикторы (лаги ряда) в моделях прогнозирования почти независимыми. Этим определяется важность приведения ряда к стационарному виду.

При выполнении прогноза необходимо учитывать правомерность допущения о стационарности ряда. Проверка на стационарность осуществляется специальными тестами – тестами на стационарность.

Идентификация временного ряда (как стационарного так и нестационарного), основанная на проверке постоянного значения математического ожидания, дисперсии и ковариации, невозможна, так как структура временного ряда изначально неизвестна. График временного ряда и его коррелограммы могут помочь в субъективном определении временного ряда y_t как стационарного.

Методика проверки временных рядов на стационарность включает в себя целый ряд тестов, направленных на выявление «единичного корня» (Unit Root Tests), базовым из которых является тест Дики–Фуллера (DF -test). Этот тест рекомендовано использовать при условии гомоскедастичности и некоррелированности случайных отклонений тестируемой модели. Понятие «единичного корня» используется в анализе временных рядов как свойство, характеризующее нестационарный временной ряд.

Существует несколько реализаций тестов Unit Root, таких как:

1) Расширенный Дики Фуллер тест (ADF Test);

2) Тест Квятковского-Филлипса-Шмидта-Шина-KPSS (тренд стационарный);

3) Филипс Перрон тест (тест PP);

Наиболее часто используется тест ADF, где нулевая гипотеза рассматривает наличие единичного корня (unit root, один из корней характеристического уравнения полинома лежит на единичной окружности), т.е. не стационарность ряда. Малые значения p-value теста свидетельствуют о стационарности временного ряда. То есть, если p-значение в тесте ADF меньше уровня значимости (0.05), отклоняется нулевая гипотеза.

Глава 2 АВТОРЕГРЕССИОННЫЕ МОДЕЛИ ВРЕМЕННЫХ РЯДОВ

2.1 Модели авторегрессии и скользящей средней

Об автокорреляционной функции АКВФ следует отметить два важных момента.

1) АКВФ является четной функцией лага k [т.е. $p(k) = p(-k)$]. Это следует из результата $\gamma(k) = \text{cov}(X_t, X_{t+k}) = \text{cov}(X_{t-k}, X_t)$ из за стационарности $\gamma(-k)$.

2) Для данного АКВФ будет только один нормальный процесс. Но можно найти несколько ненормальных процессов, которые имеют одинаковую активность.

Случайный процесс это дискретный процесс $\{X_t\}$, состоящий из последовательности взаимно независимых одинаково распределенных случайных величин. Он имеет постоянное среднее значение и постоянную дисперсию и АКВФ:

$$\gamma(k) = \text{cov}(X_t, X_{t+k}) = 0 \text{ для } k \neq 0$$

АКВФ определяется как

$p(k) = 1$ для $k = 0$

$p(k) = 0$ для $k \neq 0$.

Случайный процесс также называется *белым шумом*.

Этот процесс часто используется для описания поведения цен на акции (хотя есть некоторые исследователи, которые не согласны с этой теорией случайного блуждания). Предположим, что $\{\varepsilon_t\}$ - чисто случайный ряд со средним μ и дисперсией σ^2. Тогда процесс $\{X_t\}$ называется случайным блужданием, если

$$X_t = X_{t-1} + \varepsilon_t$$

Предположим, что X_0 равно нулю. Затем процесс развивается следующим образом:

$X_1 = \varepsilon_1$

$X_2 = X_1 + \varepsilon_2 = \varepsilon_1 + \varepsilon_2$ и т.д.

Получится следующее путем последовательной замены

$X_t \sum_{t=1}^{t} \varepsilon_t$

Следовательно, $E(X_t) = t\mu$ и $\text{var}(X_t) = t\sigma^2$. Поскольку среднее значение и дисперсия изменяются с t, процесс не стационарен, но его первое отличие является стационарным. В отношении цены акции, это говорит о том, что изменения в цене акций будут чисто случайным процессом.

Предположим, что $\{\varepsilon_t\}$ - чисто случайный процесс со средним нулем и дисперсией σ^2. Тогда процесс $\{X_t\}$ определяется как

$$X = \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + ... + \beta_m\varepsilon_{t-m}$$

Этот процесс называется процессом скользящего среднего порядка m и обозначается MA(m). Поскольку ε являются ненаблюдаемыми, мы масштабируем их так, чтобы $\beta_0 = 1$. Поскольку $E\ \varepsilon_t = 0$ для всех t, $E(X_t) = 0$. Кроме того, $\text{var}(X_t) = (\sum_{t=0}^{m} \beta_t^2)\sigma^2$, так как ε_t независимы с общей дисперсией σ^2. Далее, выписываем выражения для X_t и X_{t-k}, и с учетом независимости ε получаем

$$\gamma\ k = cov\ X_t, X_{t-k} = \begin{cases} \sigma^2 \sum_{t=0}^{m-k} \beta_t\beta_{t+k}, & k = 0,1,2,\dots,m \\ 0, & k > m \end{cases}$$

Также принимая во внимание $\text{cov}(X_t, X_{t+k})$, мы получаем то же выражение, что и для $\gamma(k)$. Следовательно, $\gamma(-k) = \gamma(k)$.

АКФ можно получить, разделив $\gamma(k)$ на $\text{var}(X_t)$.

Для процесса MA $p(k)=0$ для $k>m$, то есть они равны нулю для лагов, превышающих порядок процесса. Поскольку $\gamma(k)$ не зависит от t, процесс MA(m) слабо стационарен по отношению к процессу MA.

Для упрощения обозначений используется оператор запаздывания L. Он определяется как $L^j X_t = X_{t-j}$ для всех j. Таким образом, $LX_t = X_{t-1}, L^2 X_t = X_{t-2}, L^{-1} X_t = X_{t+1}$ и так далее.

С этим обозначением процесс MA(m) может быть записан как (так как $\beta_0 = 1$)

$$X_t = (1 + \beta_1 L + \beta_2 L^2 + ... + \beta_m L^m)\varepsilon_t = \beta(L)\varepsilon_t$$

Многочлен от L имеет m корней, и мы можем записать

$$X_t = (1 - \pi_1 L)(1 - \pi_2 L)...(1 - \pi_m L)\varepsilon_t,$$

где $\pi_1, \pi_2, ..., \pi_m$ - корни уравнения.

$$Y^m + \beta_1 Y^{m-1} + ... + \beta_m = 0$$

После оценки модели мы можем вычислить остаток из $\varepsilon_t = [\beta(L)]^{-1} x_t$ при условии, что $[\beta(L)]^{-1}$ существует. Это условие называется условием обратимости. Условием обратимости является то, что $|\pi_i| < 1$ для всех i. Это означает, что процесс MA(m) можно записать как процесс AR(∞). Например, для процесса MA(2)

$$X_t = (1 + \beta_1 L + \beta_2 L^2)\varepsilon_t,$$

π_1 и π_2 - корни квадратного уравнения $Y^2 + \beta_1 Y + \beta_2 = 0$.

Условие $|\pi_i| < 1$ дает

$$\left|\frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_2}}{2}\right| < 1$$

Это дает результат, что β_1 и β_2 должны удовлетворять:

$$\begin{aligned} &\beta_1 + \beta_2 > -1 \\ &\beta_2 - \beta_1 > -1. \\ &|\beta_2| < 1 \end{aligned} \qquad (1)$$

Последнее условие вытекает из того факта, что $\beta_2 = \pi_1 \pi_2$ произведение корней. Первые два условия вытекают из того факта, что если $\beta_1^2 - 4\beta_2 > 0$, то $(\beta_1^2 - 4\beta_2) < (2 + \beta_1)^2$ или $(\beta_1^2 - 4\beta_2) < (2 - \beta_1)^2$.

Процессы скользящего среднего возникают при решении задач устранения тренда. Одной из процедур, часто используемых для устранения тренда, является процедура последовательного дифференцирования временного ряда X_t. Пусть у нас есть

$$X_t = a_0 + a_1 t + a_2 t^2 + \varepsilon_t,$$

где ε_t - чисто случайный процесс, последовательное различие X_t устранит тенденцию, но получающийся ряд представляет собой процесс со скользящим средним, который может показать цикл. Таким образом, серия с устранением тренда может показывать цикл, даже если в исходной серии его не было. Этот феномен случайных циклов известен как эффект Слуцкого.

Предположим, что $\{\varepsilon_t\}$ – чисто случайный процесс со средним нулем и дисперсией σ^2. Тогда процесс $\{X_t\}$ будет выглядит следующим образом

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + ... + a_r X_{t-r} + \varepsilon_t \quad (2)$$

Формула (2) называется авторегрессионным процессом порядка r и обозначается AR(r). Поскольку выражение похоже на уравнение множественной регрессии, поэтому, оно называется «регрессивным». Тем не

менее, регрессия X_t заключается на его собственных прошлых значениях. Следовательно, это называется авторегрессия.

В терминах оператора запаздывания L процесс AR (2) можно записать в виде

$$X_t = (a_1 L + a_2 L^2 + ... + a_r L^r) X_t + \varepsilon_t$$

или

$$(1 - a_1 L - a_2 L^2 ... - a_r L^r) X_t = \varepsilon_t \qquad (3)$$

или

$$X_t = \frac{1}{1 - a_1 L - a_2 L^2 ... a_r L^r} \varepsilon_t = \frac{1}{(1 - \pi_1 L)(1 - \pi_2 L)...(1 - \pi_r L)} \varepsilon_t ,$$

где $\pi_1, \pi_2, ..., \pi_r$ - корни уравнения:

$$Y^r - a_1 Y^{r-1} - a_2 Y^{r-2} ... a_r = 0 .$$

Условие того, что разложение (3) верно и дисперсия X_t конечна, состоит в том, что $|\pi_i| < 1$ для всех i.

Чтобы найти АКВФ, предполагается, что процесс является стационарным и найдется p(k). Для этого, уравнение (2) умножается на X_{t-k}, вычисляется математическое ожидание и делится на $\text{var}(X)$, которое предполагается конечным.

Получается:

$$p(k) = a_1 p(k-1) + ... + a_k p(k-r)$$

(подставляя $k = 1,2,...,r$ и учитывая $p(k) = p(-k)$, мы получаем уравнения для определения r параметров $a_1, a_2, ..., a_r$). Эти уравнения известны как уравнения Юла-Уокера.

Чтобы проиллюстрировать эти процедуры, рассмотрим процесс AR(2)

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \varepsilon_t,$$

π_1 и π_2 - корни уравнения:

$$Y^2 - a_1 Y - a_2 = 0.$$

Таким образом, $|\pi_i| < 1$ означает, что:

$$\left| \frac{a_1 \pm \sqrt{a_1^2 + 4a_2}}{2} \right| < 1.$$

Это дает:

$$\begin{aligned} &a_1 + a_2 < 1; \\ &a_1 - a_2 > 1; \qquad (4) \\ &|a_2| < 1. \end{aligned}$$

(Условия аналогичны условием (1), полученным для обратимости процесса MA(2).)

В случае процесса AR(2) можно получить получается $p(k)$ рекурсивно, используя уравнения Юла-Уокера. Известно:

$$p(0) = 1 \text{ и } p(1) = a_1 p(0) + a_2 p(-1) = a_1 p(0) + a_2 p(1) \text{ или } p(1) = \frac{a_1}{1 - a_2}.$$

Таким образом:

$$p(2) = a_1 p(1) + a_2 p(0) = \frac{a_1^2}{1 - a_2} + a_2$$

$$p(3) = a_1 p(2) + a_2 p(1) = \frac{a_1 (a_1^2 + a_2)}{1 - a_2} + a_1 a_2$$

и так далее.

Этот метод можно использовать независимо от того, является ли корни действительными числами или комплексными.

2.2 Процесс авторегрессии ARMA и ARIMA

Рассмотрим модели, которые являются комбинациями моделей AR и MA. Они называются моделями ARMA. Модель ARMA(p,q) определяется как:

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q} \ ,$$

где $\{\varepsilon_e\}$ – чисто случайный процесс со средним нулем и дисперсией σ^2. Преимущество этих методов заключается в том, что они приводят к экономным представлениям AR(p) или MA(q) более высокого порядка.

Используя оператор запаздывания L, мы можем записать это как

$\varphi\ L\ X_t = \theta(L)\varepsilon_t$,

где $\varphi\ L$ и $\theta\ L$ - полиномы порядков p и q соответственно, определяемые как

$$\varphi\ L\ = 1 - a_1 L - a_2 L^2 \ldots a_p L^p,$$

$$\theta\ L\ = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q.$$

Для стационарных временных рядов мы требуем, чтобы корни $\varphi\ L\ = 0$ лежали вне единичной окружности. Для обратимости компонента MA требуется, чтобы корни $\theta(L)$ лежали вне единичной окружности. Например, для процесса ARMA(2,2) эти условия задаются уравнениями (1) и (4). АКВФ и АКФ модели ARMA сложнее, чем модели AR и MA.

Получим АКФ для простейшего случая: процесса ARMA(1,1)

$$X_t = a_1 X_{t-1} + \varepsilon_t + \beta_1 \varepsilon_{t-1}.$$

В терминах оператора запаздывания L это можно записать как

$$X_t - a_1 X_{t-1} = \varepsilon_t + \beta_1 \varepsilon_{t-1}$$

или же

$$1 - a_1 L\ X_t = (1 + \beta_1 L)\varepsilon_t$$

или можно записать как

$$X_t = \frac{1+\beta L}{1-aL}\varepsilon_t = \ 1+\beta L \ \ 1+aL+a^2L^2+\cdots \ \varepsilon_t = [1+\ a+\beta\ L+a\ a+\beta\ L^2+ a^2\ a+\beta\ L^3+\cdots]\varepsilon_t.$$

Поскольку ε_t – чисто случайный процесс с дисперсией σ^2, мы получим

$$var\ X_t \ = \ 1+\ a+\beta\ ^2+a^2\ a+\beta\ ^2+\cdots\ \sigma^2 = \ 1+\frac{(a+\beta)^2}{1-a^2}\ \sigma^2 = \frac{1+\beta^2+2a\beta}{1-a^2}\sigma^2,$$

также

$$cov\ X_t, X_{t-1} \ = [\ a+\beta\ +a\ a+\beta\ ^2+a^3\ a+\beta)^2+\cdots\ \sigma^2 = \ a+\beta+ \frac{a+\beta\ ^2a}{1-a^2}\ \sigma^2 = \frac{(a+\beta)(1+a\beta)}{1-a^2}\sigma^2.$$

Следовательно,

$$p\ 1\ = \frac{cov(X_t,X_{t-1})}{var(X_t)} = \frac{(a+\beta)(1+a\beta)}{1+\beta^2+2a\beta}.$$

Последовательные значения $p(k)$ могут быть получены из рекуррентного отношения $p\ k\ = ap(k-1)$ для $k \geq 2$. Для процесса AR(1) $p\ 1\ = a$. Можно проверить, что $p(1)$ для процесса ARMA(1,1) имеет значение > или $< a$ в зависимости от того, $\beta > 0$ или < 0, соответственно.

Многие эмпирические временные ряды (например, цены на бирже) ведут себя так, как будто они не имеют фиксированного среднего значения. Но даже при этом они выглядят однородными в том смысле, что если не учитывать локальный уровень или, возможно, локальный уровень и тренд, любая часть временного ряда по своему поведению во многом подобна любой другой. Модели, описывающие такое однородное нестационарное поведение, можно получить, предположив, что некая подходящая разность процесса стационарна. Рассмотрим теперь свойства важнейшего класса моделей, в которых d-я разность есть стационарный смешанный процесс авторегрессии – скользящего среднего. Эти модели называются процессами авторегрессии – проинтегрированного скользящего среднего (ARIMA).

На практике большинство временных рядов являются нестационарными. Одной из процедур, которая часто используется для преобразования нестационарного ряда в стационарный ряд, является последовательность дифференцировании. Определим оператор $\Delta = 1 - L$, так что $\Delta X_t = X_t - X_{t-1}$,

$\Delta^2 X_t = X_t - X_{t-1} - (X_{t-1} - X_{t-2})$ и так далее.

Предположим, что $\Delta^d X_t$ является стационарным рядом, который может быть представлен моделью ARMA(p, q). Затем мы говорим, что X_t может быть представлена моделью ARIMA(p, d, q,). Модель называется интегрированной моделью, потому что стационарная модель ARMA, которая подгоняется к разностным данным, должна суммироваться или «интегрироваться», чтобы обеспечить модель для нестационарных данных. Фактически, даже если нет необходимости в компоненте скользящего среднего при моделировании X_t процедура дифференцирования X_t приведет к процессу скользящего среднего (эффект Слуцкого).

Характерная запись модели ARIMA (p,d,q) имеет след. вид:

$$\Delta^d Y_t = \sum_{i=1}^{p} \varphi_i \, \Delta^d Y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} Q_j \, \Delta^d Y_{t-j}, \; \varepsilon_{t \sim N(0,\sigma_t^2)}.$$

Когда AR, MA или ARMA модель подобрана для данного временного ряда, ее рекомендуется проверить (тестировать), что она действительно дает адекватное описание данных. Часто используются два критерия, которые отражают близость приспособленности и количество оцениваемых параметров. Одним из них является информационный критерий Akaike (AIC), а другой – критерий Байеса Шварца (SBC). Последнее также называется байесовским информационным критерием (BIC). Если p – общее число, то, оценивая параметры, мы имеем.

$$AIC\ p = n \log \sigma_p^2 + 2p,$$

$$BIC\ p = n \log \sigma_p^2 + p \log n,$$

здесь n – размер выборки. Если RSS-остаточная сумма квадратов ε_t^2, то $\sigma_p^2 = RSS/(n-p)$. Если мы рассматриваем несколько моделей ARMA, то выбирается одна. Выбирается та модель, у которой самая низкая AIC или BIC. (Два критерия могут привести к различным результатам). Следует проверить серийную картину корреляции из остатков.

Box и Pierce предлагают вычислять $Q = N \sum_{k-1}^{m} r_k^2$, где r_k – автокорреляция лага k, а N – число наблюдения в серии. Если модель подходит, они утверждают, что Q имеет асимптотическое распределение X^2 с $m-p-q$ степенями свободы, где p и q – соответственно порядки компонентов AR и MA.

2.3 Подход Бокс-Дженкинса и прогнозирование

Основными шагами в методологии Бокса-Дженкинса является: (1) дифференцирование серии для того чтобы достигнуть стационарности, (2) идентификация предварительной модели, (3) оценка модели, (4) диагностическая проверка (если модель находится в неадекватном состоянии, то возврат к шагу 2), и (5) использование модели для прогнозирования и управления.

Теперь разберем эти шаги более подробно.

1. Дифференцирование для достижения стационарности: как сделать вывод о там, что является временной ряд стационарным или нет? Мы можем сделать это, изучая граф коррелограмма ряда. Коррелограмма стационарного ряда постепенно уменьшается согласно k, число лагов становятся длиннее, но это не обычный случай для нестационарного ряда. Таким образом, общая процедура состоит в том, чтобы построить коррелограмму данного ряда y_t и последовательных разностей Δy, $\Delta^2 y$, и так далее, и посмотреть на коррелограммы на каждом этапе.

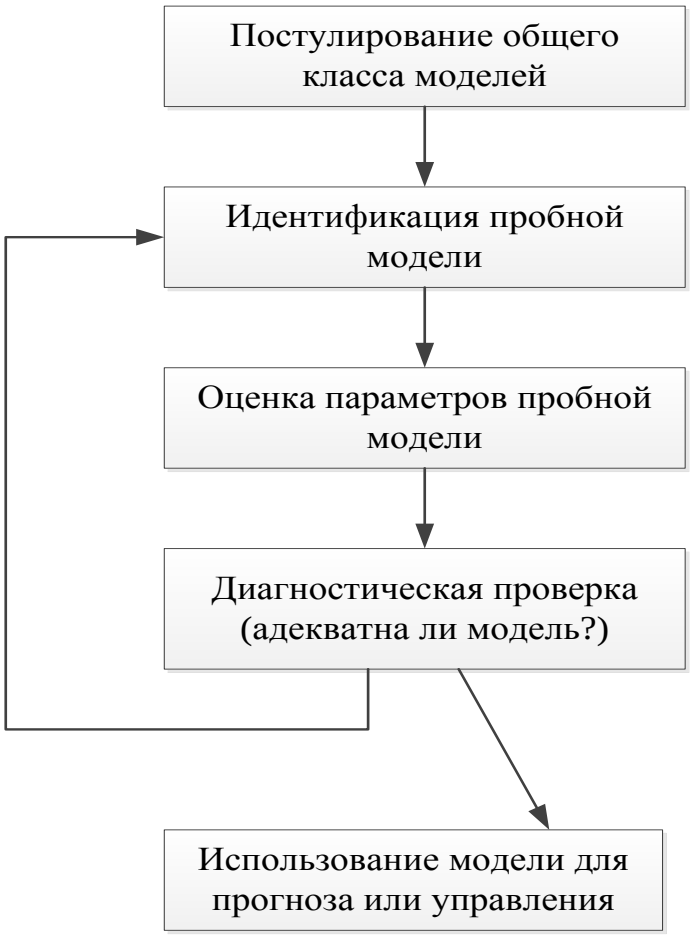

Рисунок 2.1 - Этапы итеративного подхода к построению моделей

Дифференцирование продолжается до тех пор, пока коррелограмма не ослабится. коррелограмму данного ряда y_t и последовательных разностей Δy, $\Delta^2 y$, и так далее, и посмотреть на коррелограммы на каждом этапе. Дифференцирование продолжается до тех пор, пока коррелограмма не ослабится.

2. После того, как мы использовали процедуру дифференцирования, чтобы получить стационарный временной ряд, далее мы изучаем коррелограмму для того, чтобы принять решение о соответствующих порядках компонентов AR и MA. Коррелограмма процесса MA равна нулю после точки. Процесс AR уменьшается геометрически. Коррелограмма ARMA процесс показывает различные паттерны (но все ослабляются через некоторое время). Основываясь на них, мы приходим к предварительной модели ARMA. Это шаг

включает в себя больше осуждающей процедуры, чем использование каких-либо других четких правил.

3. Следующим шагом является оценка предварительной модели ARMA в шаге 2. В предыдущем главе обсуждалось об оценке модели ARMA.

4. Следующим шагом является диагностическая проверка, чтобы проверить адекватность предварительной модели.

5. Последний шаг – прогнозирование, Далее мы рассмотрим эту проблему.

Для закрепления идей мы рассмотрим прогнозирование по модели ARMA(2, 2). Предположим, что мы оценили модель с помощью n наблюдений. Мы хотим прогнозировать x_{n+k}. Это называется k – прогноз на период вперед. Это обозначается как $x_{n,k}$. Первый индекс дает период времени, когда прогноз сделан, а второй нижний индекс обозначает период времени, на которой прогноз сделан. Начнем с $k = 1$, чтобы получить прогноз x_{n+1} в период времени n. Мы имеем

$$x_{n+1} = a_1 x_n + a_2 x_{n-1} + \varepsilon_{n+1} + \beta_1 \varepsilon_n + \beta_2 \varepsilon_{n-1}$$

Мы наблюдали x_n и x_{n-1}. Мы можем заменить ε_n и ε_{n-1} на предсказанные остатки. Неизвестно только ε_{n+1}. Заменяется на его ожидаемое значение ноль. Следовательно,

$$x_{n,1} = a_1 x_n + a x_{n-1} + \beta_1 \varepsilon_n + \beta_2 \varepsilon_{n-1}.$$

Перейдем к $k = 2$. Имеем

$$x_{n+2} = a_1 x_{n+1} + a_2 x_n + \varepsilon_{n+2} + \beta_1 \varepsilon_{n+1} + \beta_2 \varepsilon_n.$$

Заменим ε_{n+2} и ε_{n+1} на ноль, их ожидаемое значение. x_{n+1} не известно, но есть прогноз $x_{n,1}$. Таким образом, получаем

$$x_{n,2} = a_1 x_{n,1} + a_2 x_n + \beta_2 \varepsilon_n.$$

Далее продолжаем:

1) записать выражения для x_{n+k};

2) заменить все будущие значение x_{n+k} $j > 0, j < k$ на их прогноз;

3) заменить все ε_{n+j} $(j > 0)$ на ноль;

4) заменить все ε_{n-j} $(j \leq 0)$ на предсказанные остатки.

Альтернативной процедурой является запись x_t в терминах всех лагов x. Для этого используется ранее описание процедура. Имеем

$$1 + \gamma_1 L + \gamma_2 L^2 + \gamma_3 L^3 + \cdots \; x_t = \varepsilon_t,$$

γ получены, как описано в предыдущем разделе. Теперь имеем

$$x_{t+1} = -\gamma_1 x_t + \gamma_2 x_{t-1} + x_3 x_{t-2} + \cdots \; + \varepsilon_{t+1}.$$

Чтобы получить $x_{t,1}$ заменим ε_{t+1} на ноль, его ожидаемое значение и γ по их оценкам. Процедура такая же, как и раньше за исключением того, что не придется иметь дело с предсказанными остатками.

Глава 3 РЕАЛИЗАЦИЯ АЛГОРИТМА ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ ARIMA

3.1 Модель ARIMA(p, d, q)

Любой "несезонный" временной ряд, который показывает закономерности и не является случайным белым шумом, может быть смоделирован с помощью моделей ARIMA.

Модель ARIMA характеризуется тремя терминами: p, d, q где:

- p-порядок члена AR;
- q порядок термина MA;
- d-число разностей, необходимых для стационарного временного ряда.

Первый шаг для построения модели ARIMA – сделать временной ряд стационарным. Потому что термин "Авторегрессия" в ARIMA означает, что это линейная регрессионная модель, которая использует свои собственные лаги в качестве предикторов. Модели линейной регрессии, как известно, работают лучше всего, когда предикторы не коррелированы и независимы друг от друга.

Самый распространенный подход сделать ряд стационарным, это отличить его. То есть вычесть предыдущее значение из текущего значения. Иногда, в зависимости от сложности ряда, может потребоваться несколько разностей.

Значение d, следовательно, является минимальным числом разностей, необходимых для того, чтобы сделать ряд стационарным. А если временной ряд уже стационарный, то d = 0.

"p" - порядок термина "Авторегрессия" (AR). Это относится к числу лагов Y, которые будут использоваться в качестве предикторов. А " q "- это порядок "Скользящий средний" (MA). Это относится к числу запаздывающих ошибок прогноза, которые должны войти в модель ARIMA.

Авторегрессионная модель (только AR) - это модель, в которой Y_t зависит только от собственных лагов. То есть Y_t является функцией "лагов Y_t".

$$Y_t = a + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_1,$$

где, Y_{t-1} – лаг 1 ряда, β_1- коэффициент лаг 1, который оценивает модель, и a - термин перехвата, также оцениваемый моделью.

Аналогично, модель Скользящего Среднего (только MA) - это модель, в которой Y_t зависит только от запаздывающих ошибок прогноза.

$$Y_t = a + \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q},$$

где членами ошибки являются ошибки авторегрессионных моделей соответствующих лагов. Ошибки ε_t и ε_{t-1} являются ошибками из следующих уравнений:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_0 Y_0 + \varepsilon_t;$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \cdots + \beta_0 Y_0 + \varepsilon_{t-1}.$$

Модель ARIMA - это модель, где временные ряды дифференцировались хотя бы один раз, чтобы сделать ее стационарным, и объединились AR и MA. Таким образом, уравнение становится:

$$Y_t = a + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \varepsilon_t + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \cdots + \varphi_q \varepsilon_{t-q}.$$

Цель дифференцирования - сделать временные ряды стационарными.

Правильный порядок разности - это минимальная разность, необходимая для получения почти стационарного ряда, который бродит вокруг определенного среднего, и график АКФ достигает нуля довольно быстро.

Если автокорреляции положительны для многих чисел лагов (10 или более), то ряд нуждается в дальнейшем дифференцировании.

Для проверки ряда на наличие стационарности, используется, критерия Дики Фуллера (adfuller()), из пакета statsmodels.

Дифференцирование необходимо только в том случае, если ряд нестационарный. В противном случае различие не требуется, то есть d=0.

Как уже говорилось во второй главе, нулевая гипотеза теста АКФ заключается в том, что временной ряд нестационарен. Итак, если p-значение теста меньше уровня значимости (0,05), то откланяется нулевая гипотеза и делается вывод, что временной ряд действительно стационарен.

На рисунке 3.1 представлен тест Дики Фуллера, который показывает что p=0.124419 (Приложении Б). Итак, в нашем случае, p > 0,05, значит, мы продолжаем поиск порядка дифференцирования.

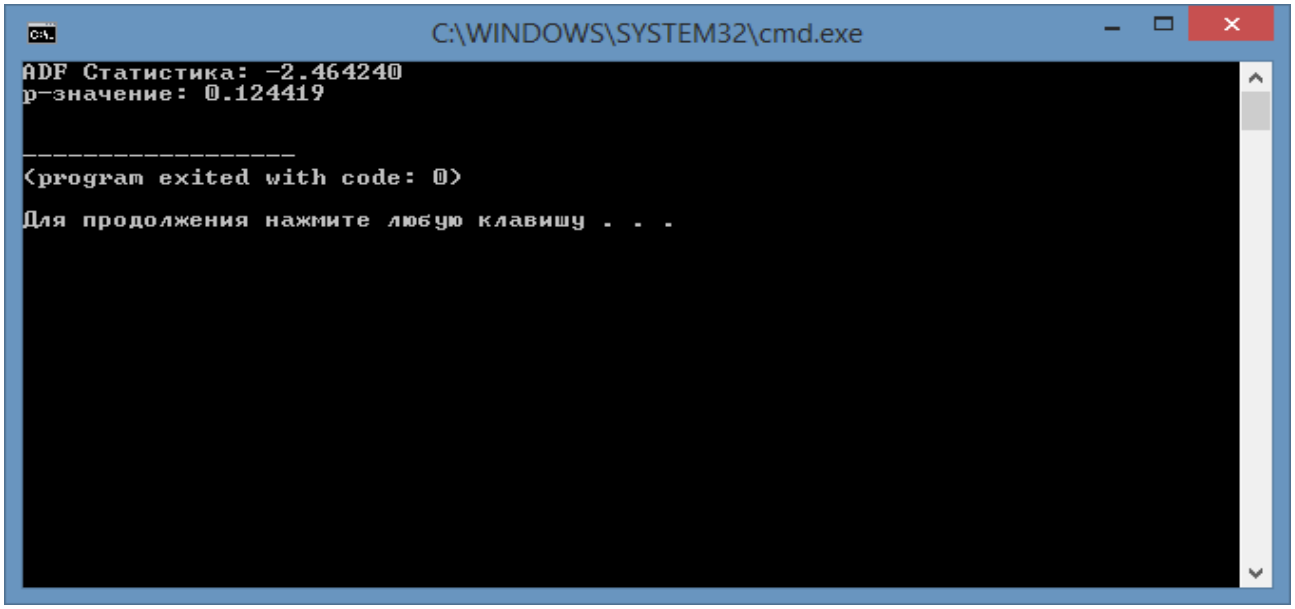

Рисунок 3.1 - Тест Дики Фуллера

На рисунке 3.2 можно наблюдать, что для вышеприведенных рядов временные ряды достигают стационарности с двумя порядками дифференцирования. Но при взгляде на график автокорреляции для 2-го дифференцирования отставание довольно быстро переходит в дальнюю отрицательную зону, что указывает на то, что ряды могли быть более дифференцированы. Итак, можно брать порядок дифференцирования как 1, даже если ряд будет слабо-стационарна (Приложение Б).

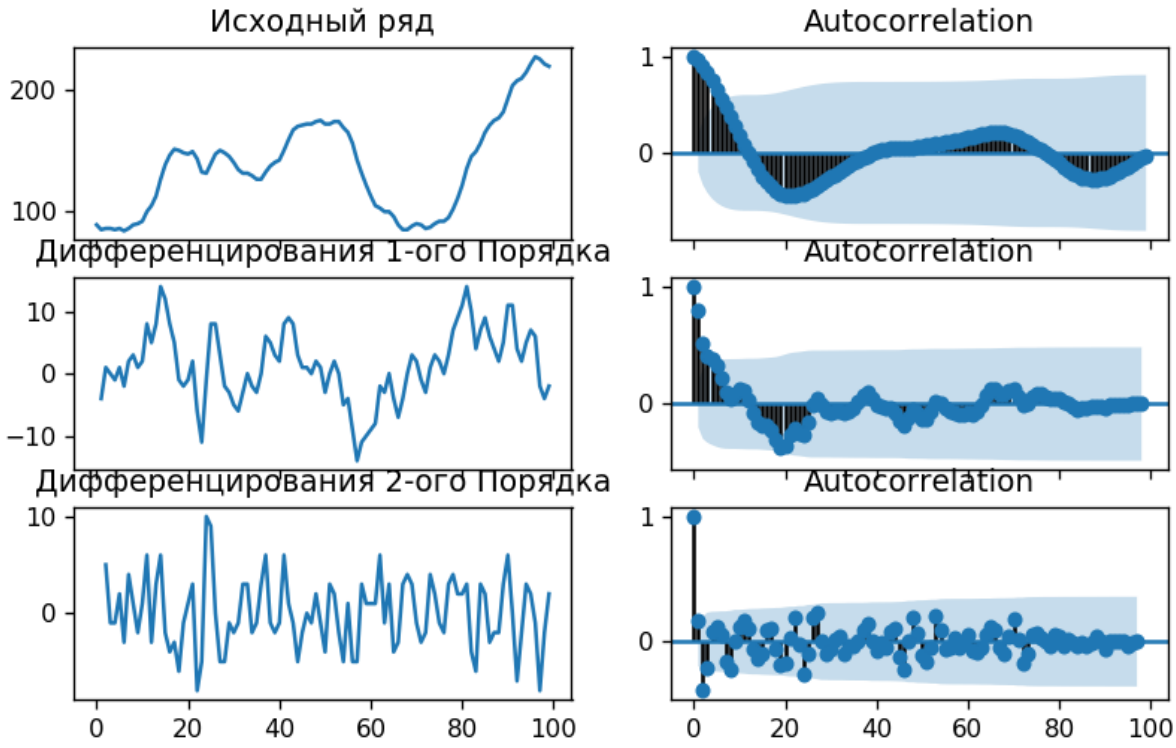

Рисунок 3.2 - график Автокорреляции

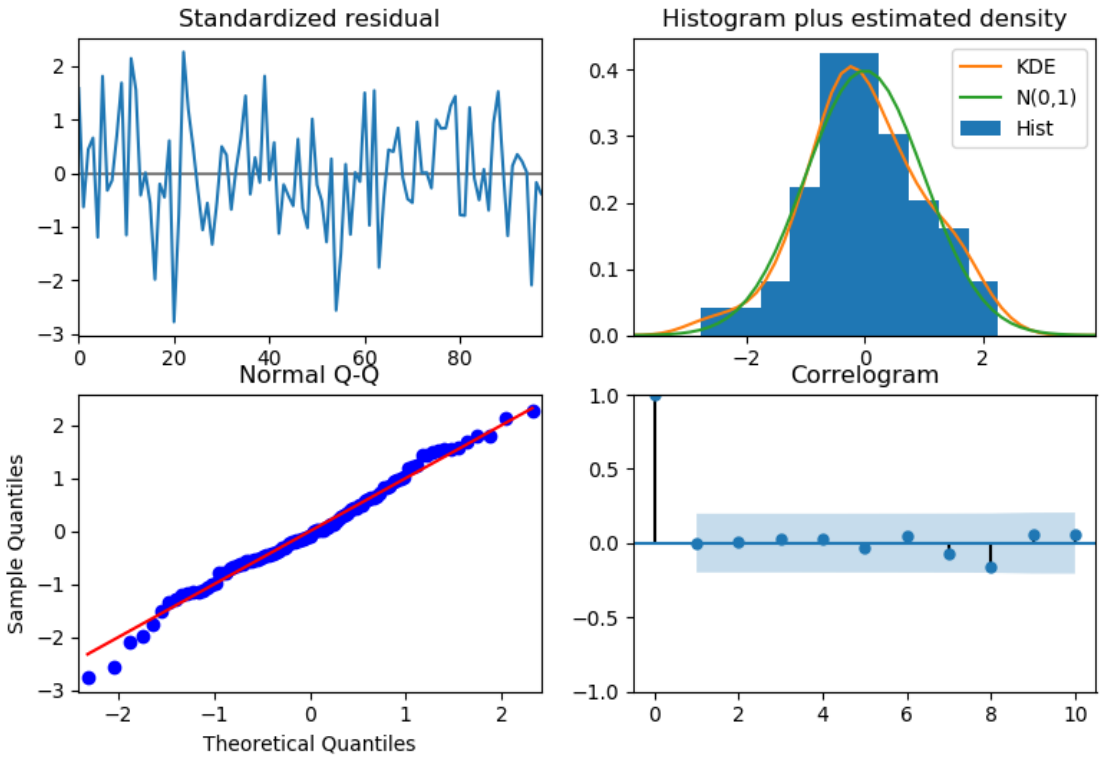

Рисунок 3.3 - Коррелограмма и остатки

На рисунке 3.3 представлено:

1) слева сверху: остаточные ошибки, похоже, колеблются вокруг среднего значения нуля и имеют равномерную дисперсию;

2) вверху справа: график плотности предполагает нормальное распределение со средним нулем;

3) внизу слева: все точки должны идеально совпадать с красной линией. Любые значительные отклонения означают, что распределение искажено;

4) внизу справа: Коррелограмма, график АКФ показывает, что остаточные ошибки не автокоррелированы. Любая автокорреляция будет означать, что в остаточных ошибках есть информация, то есть они не случайные они как то связанны друг с другом и это сигнал о том, что модель можно улучшить. В целом, рассматриваемая модель хорошо подходит для прогнозирования.

При всех лагах автокорреляция маленькая и колеблется около нуля. Квантифицировать то, насколько она далеко стоит от нуля, то есть достаточна ли она большая, чтобы можно было считать, что она не нулевая можно с помощью статистических критериев. Значимость автокорреляции при каком то фиксированном конкретном лаге можно проверить при помощи критерия Стьюдента. На рисунке 3.4 синяя линия это исходный ряд, а красная это прогноз. Прогноз подразумевает собой подобранное значение модели ARIMA (3,2,1). По сути это просто значение в предыдущие моменты времени. Из рисунка видно, что абсолютная ошибка, т.е. разница между фактическим значением и прогнозным значением минимальна, для модели ARIMA(3,2,1). При установке dynamic=False в выборке для прогнозирования используются запаздывающие значения. То есть модель обучается до предыдущего значения, чтобы сделать следующий прогноз. Это может сделать подходящий прогноз и сделать фактические данные искусственно хорошими.

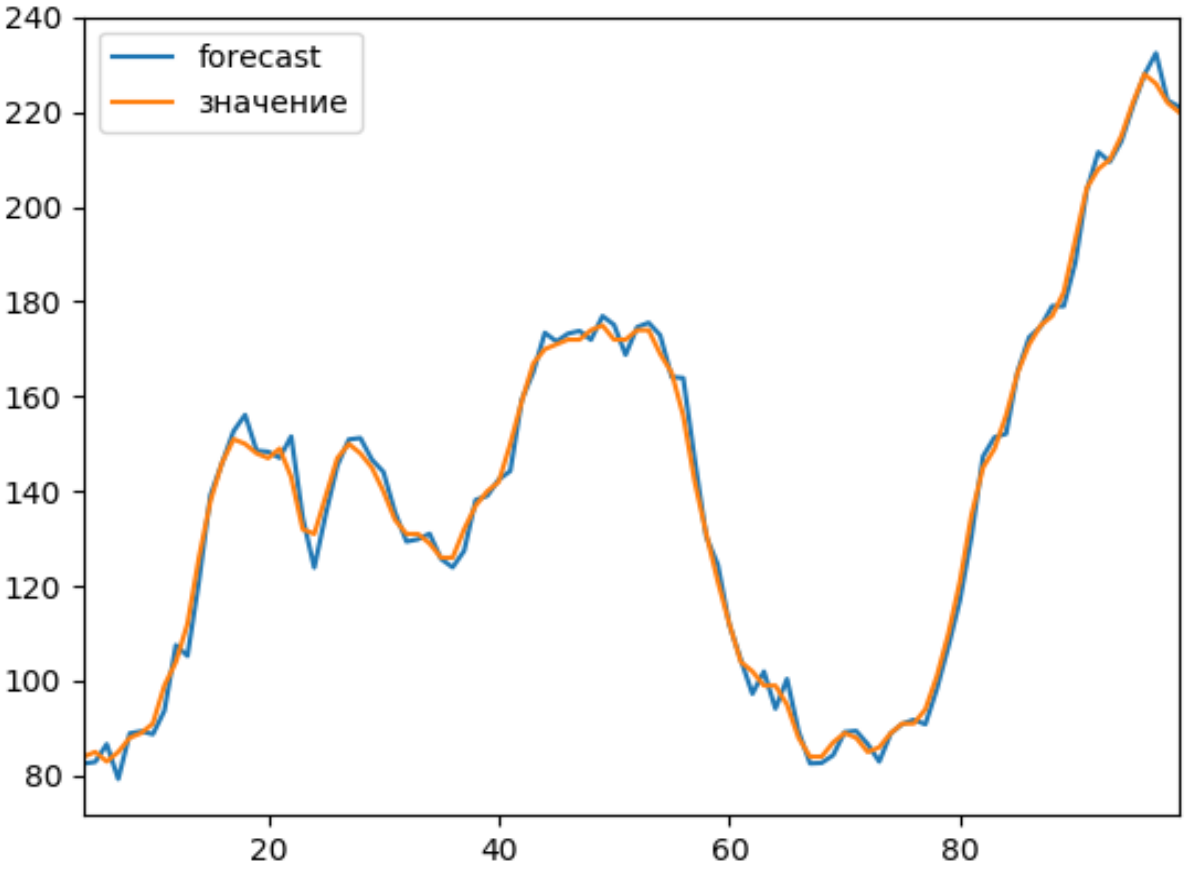

Рисунок 3.4 - График исходных значений и фактических значений

На рисунке 3.5 продемонстрирована работа программы, в которой черная линия показывает окончательный прогноз на 10 шагов вперед.

Для выбора подходящей модели ARIMA(p,d,q) указываются следующие параметры:

1) используется adftest для нахождения оптимального 'd' значения;

2) максимум параметры p и q указываются до 3;

3) частота серии равной 1;

4) пусть модель определит «d»;

5) нет сезонности.

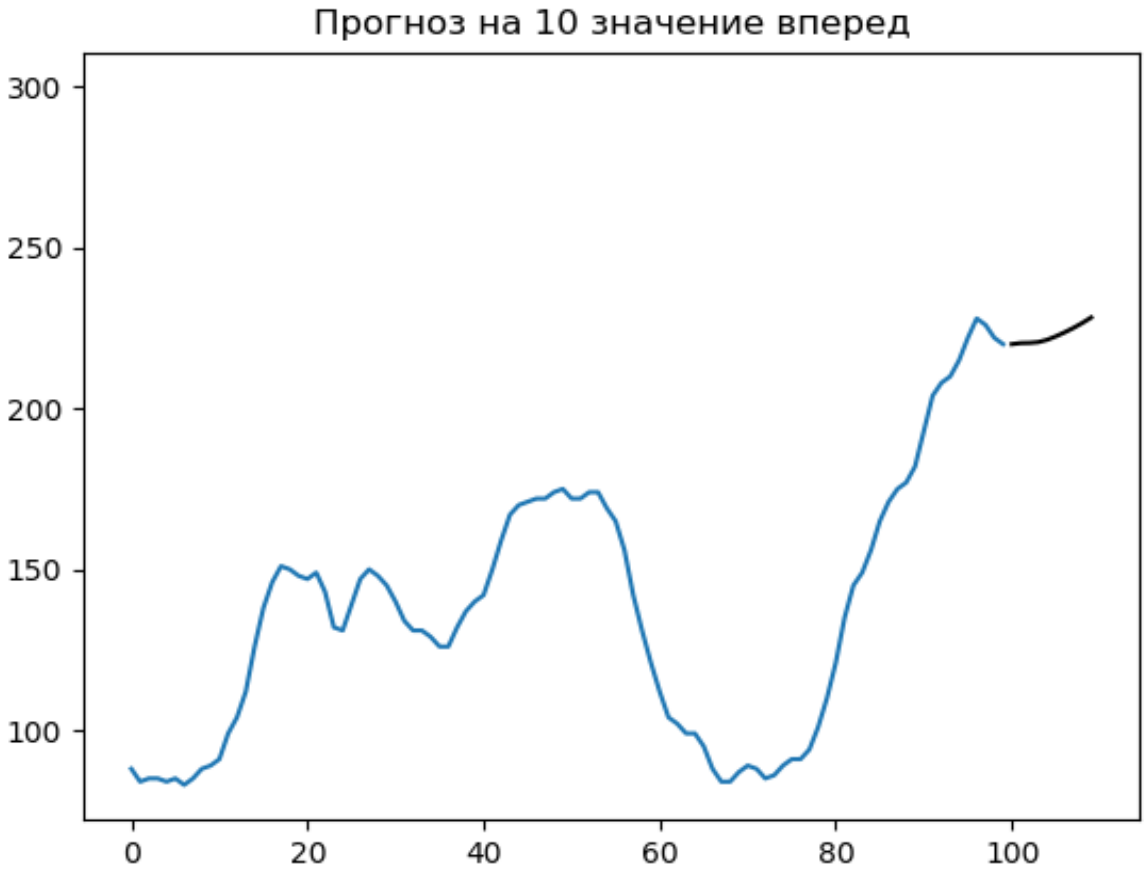

Рисунок 3.5 - График окончательного прогноза на 10 значений вперед

На рисунке 3.6 вывод работы autoARIMA. Даная программа автоматически выбирает модель с наименьшим информационным критерием Акаике (AIC). Далее, если определились порядки p, d и q, строится модель ARIMA. Модель показывает много информации. Таблица в середине - это таблица коэффициентов, где значения в разделе "coef" являются весами соответствующих членов. Значение p должен быть менее 0,05 для того, чтобы соответствующий Y был значительным. Условие стационарности: все корни характеристического уравнения по модулю больше единицы $z_j > 1$, что и наблюдается на таблице (Приложение В).

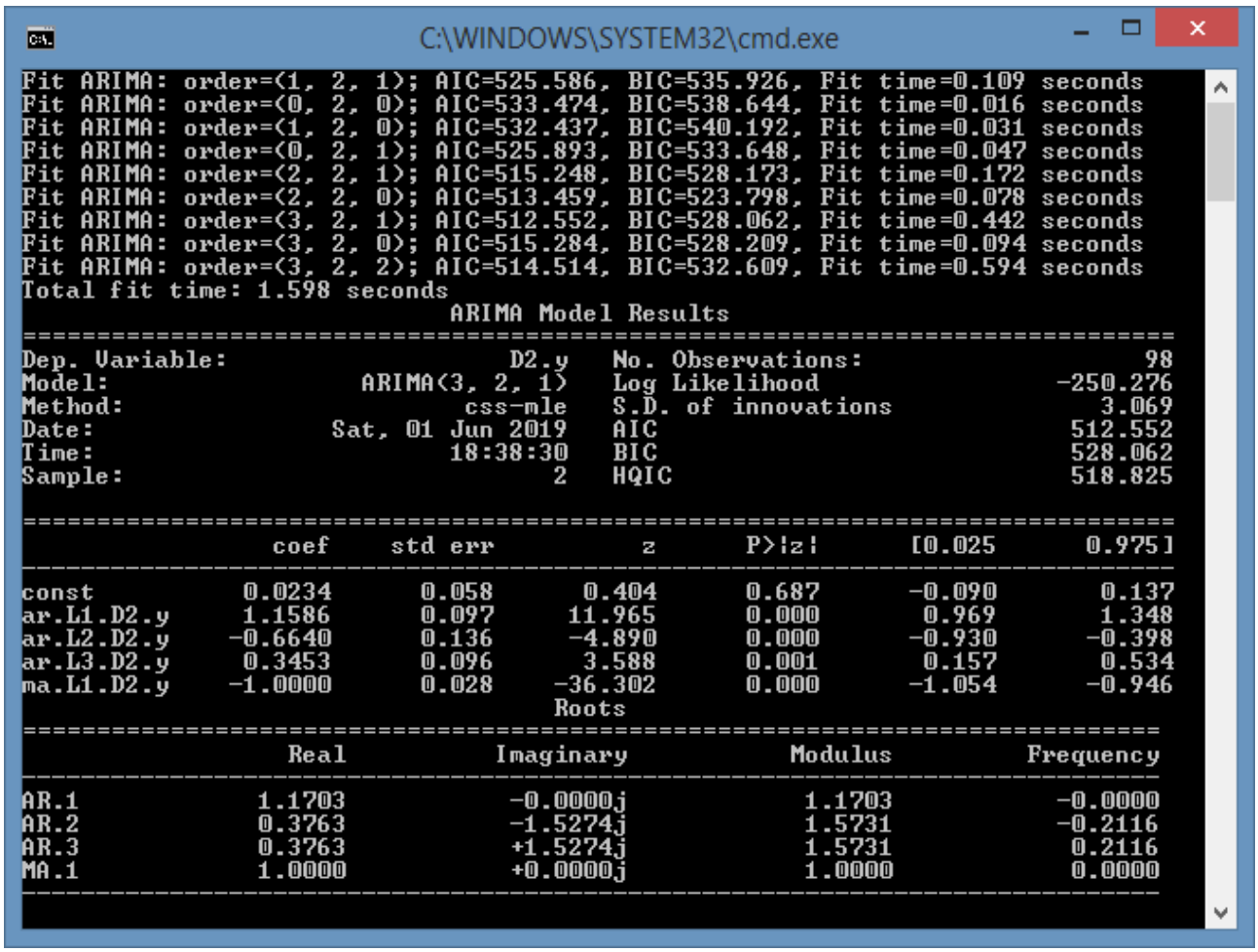

Рисунок 3.6 - Результат работы программы

ЗАКЛЮЧЕНИЕ

Цель данной выпускной квалификационной работы состояло в реализации алгоритма прогнозирования на основе авторегрессионной модели ARIMA. Для достижения этой цели:

1) выявлены основные виды авторегрессионных моделей временных рядов. К авторегрессионным моделям временных рядов относятся процессы авторегрессии (AR), скользящей средней (MA), смешанной авторегрессии скользящего среднего (ARMA) и авторегрессии проинтегрированного скользящего среднего (ARIMA), их свойства, представлены статистические методы для спецификации модели, оценки параметров и проверки модели на стационарность;

2) обосновано использование модели ARIMA. Выявлялась стационарность и нестационарность временных рядов для их анализа. Для работы с нестационарным временным рядом используется так называемая модель авторегрессионного интегрированного скользящего среднего (ARIMA);

3) осуществлена программная реализация алгоритма прогнозирования на основе модели ARIMA. Программа находит наилучший порядок p,d,q модели ARIMA и на основе выбранной модели строит прогноз. Для оценки адекватности модели оцениваются остатки модели. Если остатки несмещённые, стационарные и не автокоррелированные, то модель адекватная. В качестве исходных данных были выбраны временные ряды, которые находятся в файле data.csv, отображающие колебание данных.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

Научная и методическая литература

1. Айвазян, С. А. Эконометрика. — М.:; ИНФРА-М, 2014.

2. Айвазян, С. А. Методы эконометрики : учебник / С. А. Айвазян, В. С. Мхитарян. — М.: Магистр; ИПФРА-М, 2010.

3. Андерсон Т. Статистический анализ временных рядов. — М.: Мир, 1976. — 757 с.

4. Гуриков С. Р. Основы алгоритмизации и программирования на Python : учеб. пособие — М. : ФОРУМ : ИНФРА-М, 2018. — 343 с.

5. Дайитбегов Д.М. Компьютерные технологии анализа данных в эконометрике: Монография / 3-е изд., испр. и доп. - М.: Вузовский учебник: НИЦ Инфра-М, 2013.

6. Доугерти К. Введение в эконометрику: Пер. с англ. – М.: ИНФРА-М, 2003. – 402 с.

7. Доусон М. Программируем на Python. – СПб.: Питер, 2014. – 416 с.

8. Елисеева, ИИ Практикум по эконометрике: учеб. пособие / И И. Елисеева, СВ. Курышева, Н.М.Гордиенко [и др.]; под ред. И.И.Елисеевой. - 2-е изд., перераб. и доп. -М.: Финансы и статистика, 2008. - 344 с.

9. Елисеева И.И. Эконометрика: учебник / И. И. Елисеева, С. В. Курышева, Т. В. Костеева и др.; под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М.: Финансы и статистика. 2007.

10. Коэльо, Луис Педро Построение систем машинного обучения на языке Python / Луис Педро Коэльо, Вилли Ричарт; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва: ДМК Пресс, 2016. - 302 с.

11. Магнус Я.Р. Эконометрика. Начальный курс: учебник./ Магнус Я.Р., Катышев П.К., Пересецкий А.А.. - 7-е изд., испр. - М.:Дело, 2005. - С.253-275.

12. Маккинли, У. Python и анализ данных / Уэс Маккинли ; пер. с англ. А.А. Слинкина. - Москва : ДМК Пресс, 2015. - 482 с.

13. Новиков А. И. Эконометрика: Учебное пособие / 3-е изд., перераб. и доп. - М.: НИЦ ИНФРА-М, 2014. - 272 с.: 60х88 1/16.

14. Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения / С. Рашка; пер. с англ. А.В. Логунова. - Москва : ДМК Пресс, 2017. - 418 с.

15. Саммерфилд, М. Программирование на Python 3. Подробное руководство - М.: Символ-Плюс, 2011. - 608 с.

16. Тихомиров Н. П. Эконометрика: учебник для вузов / Н. П. Тихомиров, Е. Ю. Дорохина. - М. : Экзамен, 2003. - 512 с.

17. Уткин В.Б./Эконометрика, - 2-е изд. - М.:Дашков и К, 2017. - 564 с.

Литература на иностранном языке

18. Box, George; Jenkins, Gwilym M.; Reinsel, Gregory C. (2016). Time Series Analysis: Forecasting and Control (Fifth ed.). ISBN 978-1-118-67502-1.

19. Jagannathan, R., Skoulakis, G. & Wang, Z. (2010), The analysis of the cross section of security returns, in Y. Aït-Sahalia & L. P. Hansen, eds, 'Handbook of financial econometrics', Vol. 2, Elsevier B.V., pp. 73–134.

20. Bollerslev, T. & Wooldridge, J. M. (1992), 'Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances', Econometric Reviews 11(2), 143–172.

21. Hamilton, James D. (1994) Time Series Analysis, Princeton University Press. Description and preview.

22. Downey A., Elkner J., Meyers Ch. How to Think Like a Computer Scientist: Learning with Python. - Wellesley, Massachusetts: Green Tea Press, 2002.

23. Judge, G.G., R.C. Hill, W.E. Griffiths, H. L¨utkepohl and T.C. Lee (1985), The Theory and Practice of Econometrics (John Wiley and Sons: New York).

ПРИЛОЖЕНИЕ А

Листинг программы на Python визуализация временных рядов.

```
from pandas import read_csv

from matplotlib import pyplot

series = read_csv('data.csv', header=0)

print(series.head())

series.plot()

pyplot.show()
```

Листинг программы на Python виды временных рядов.

```
trend = read_csv('guinearice.csv', names=['Значение'], header=0)
print(trend.head())
trend.plot(title='Возростающая тенденция')
pyplot.show()


season= read_csv('sunspotarea.csv', names=['Значение'], header=0)
print(season.head())
season.plot(title='Сезонная компонента.')
pyplot.show()


trandsea = read_csv('AirPassengers.csv', names=['Значение'], header=0)
print(trandsea.head())
trandsea.plot(title='Случайная компонента.')
pyplot.show()
```

ПРИЛОЖЕНИЕ Б

Листинг приложения тест Дики-Фуллера.

```
import numpy as np, pandas as pd

from statsmodels.tsa.stattools import adfuller

from numpy import log

df = pd.read_csv('data.csv', names=['value'], header=0)

result = adfuller(df.value.dropna())

print('ADF Статистика: %f' % result[0])

print('p-значение: %f' % result[1])
```

Листинг приложения графики автокорреляционной функции.

```
import numpy as np, pandas as pd

from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

import matplotlib.pyplot as plt

plt.rcParams.update({'figure.figsize':(9,7), 'figure.dpi':120})

# Импортировать данные из data.csv

df = pd.read_csv('data.csv', names=['value'], header=0)

# Исходный ряд

fig, axes = plt.subplots(3, 2, sharex=True)

axes[0, 0].plot(df.value); axes[0, 0].set_title('Исходный ряд')

plot_acf(df.value, ax=axes[0, 1])

# 1-ое Дифференцирование

axes[1, 0].plot(df.value.diff()); axes[1, 0].set_title('Дифференцирования 1-ого Порядка')

plot_acf(df.value.diff().dropna(), ax=axes[1, 1])

# 2-ое Дифференцирование

axes[2, 0].plot(df.value.diff().diff()); axes[2, 0].set_title('Дифференцирования 2-ого Порядка')

plot_acf(df.value.diff().diff().dropna(), ax=axes[2, 1])

plt.show()
```

ПРИЛОЖЕНИЕ В

Листинг Файла Auto_ARIMA.py

```
import numpy as np, pandas as pd
from statsmodels.tsa.arima_model import ARIMA
import matplotlib.pyplot as plt
import pmdarima as pm
df = pd.read_csv('data.csv', names=['value'], header=0)
model = pm.auto_arima(df.value, start_p=1, start_q=1,
                      test='adf',       # используется adftest, чтобы найти оптимальное 'd'
                      max_p=3, max_q=3, # максимум p и q
                      m=1,              # частота серий
                      d=None,           # пусть модель определит 'd'
                      seasonal=False,   # Нет сезонность
                      start_P=0,
                      D=0,
                      trace=True,
                      error_action='ignore',
                      suppress_warnings=True,
                      stepwise=True)
model.plot_diagnostics(figsize=(7,5))
plt.show()
# Прогнозирование
n_periods = 10
fc, confint = model.predict(n_periods=n_periods, return_conf_int=True)
index_of_fc = np.arange(len(df.value), len(df.value)+n_periods)
# ряд для построения графика
fc_series = pd.Series(fc, index=index_of_fc)
lower_series = pd.Series(confint[:, 0], index=index_of_fc)
```

```
upper_series = pd.Series(confint[:, 1], index=index_of_fc)
# График
plt.plot(df.value)
plt.plot(fc_series, color='black')
plt.fill_between(lower_series.index,
                 lower_series,
                 upper_series,
                 color='w', alpha=0.15)
plt.title("Прогноз на 10 значение вперед")
plt.show()
print(model.summary())
```

Листинг программы файла graphic.py

```
from pandas import read_csv
import numpy as np, pandas as pd
import matplotlib.pyplot as plt
from matplotlib import pyplot
from statsmodels.tsa.arima_model import ARIMA
df = pd.read_csv('data.csv', names=['значение'], header=0)
model = ARIMA(df.value, order=(3,2,1))
model_fit = model.fit(disp=0)
# текущие значение и прогнозируемые
model_fit.plot_predict(dynamic=False)
plt.show()
```