МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра компьютерных технологий и систем

СИВАКОВ Алексей Сергеевич

МЕТОДЫ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Дипломная работа

Научный руководитель:
Зав. кафедрой, профессор,
доктор педагогических наук
В.В. Казачёнок,
Ассистент кафедры КТС
Е.О. Гиро

Допущена к защите
«____»_____20___г.
Заведующий кафедрой компьютерных технологий и систем
доктор педагогических наук, кандидат физико-
математических наук, профессор В.В. Казачёнок

Минск, 2025

ОГЛАВЛЕНИЕ

РЕФЕРАТ

Дипломная работа: 54 с., 13 рис., 4 табл., 3 приложения, 10 источников.

Ключевые слова: ВРЕМЕННОЙ РЯД, АНАЛИЗ, ПРОГНОЗИРОВАВНИЕ, МОДЕЛЬ.

Цель дипломной работы: исследовать существующие методы анализа и прогнозирования временных рядов и применение их на практике.

Методы исследования: анализ и исследование методов, математическое описание моделей, прогнозирование.

В результате работы были рассмотрены существующие методы анализа и прогнозирования временных рядов, рассмотрены основные классы моделей, произведена оптимальная стратегия анализа и прогнозирования, применение её на практике, изучены преимущества и недостатки каждой модели.

Область возможного применения: экономика, астрономия, метеорология, маркетинг, исследовательская деятельность и другое.

Полученные результаты и их новизна: предложена оптимальная стратегия для анализа и формирования прогноза временны рядов. Результаты экспериментов демонстрируют преимущества и недостатки того или иного класса моделей для прогнозирования и важность анализа.

РЭФЕРАТ

Дыпломная праца: 54 с., 13 рыс., 4 табл., 3 дадаткі, 10 крыніц.

Ключавыя словы: ЧАСОВЫ ШЭРАГ, АНАЛІЗ, ПРАГНАЗІРАВАННЕ, МАДЭЛЬ.

Мэта дыпломнай працы: даследаваць існуючыя метады аналізу і прагназіравання часовых шэрагаў і іх прымяненне на практыцы.

Метады даследавання: аналіз і вывучэнне метадаў, матэматычнае апісанне мадэляў, прагназіраванне.

У выніку працы былі разгледжаны існуючыя метады аналізу і прагназавання часавых шэрагаў, разгледжаны асноўныя класы мадэляў, праведзена аптымальная стратэгія аналізу і прагназавання, прымяненне яе на практыцы, вывучаны перавагі і недахопы кожнай мадэлі.

Вобласць магчымага прымянення: эканоміка, астраномія, метэаралогія, маркетынг, даследчая дзейнасць і іншае.

Атрыманыя вынікі і іх навізна: прапанавана аптымальная стратэгія для аналізу і фарміравання прагнозу часовых шэрагаў. Вынікі эксперыментаў дэманструюць перавагі і недахопы таго ці іншага класа мадэляў для прагназіравання і важнасць аналізу.

ANNOTATION

Thesis: 54 p., 13 fig., 4 tables, 3 appendices, 10 sources.

Keywords: TIME SERIES, ANALYSIS, FORECASTING, MODEL.

The purpose of the thesis: investigate the possibilities of digital twin technology to create control systems for mobile robot.

Research methods: Analysis and study of methods, mathematical description of models, forecasting.

As a result of the work, the work examined existing methods of time series analysis and forecasting, reviewed main model classes, developed an optimal strategy for analysis and forecasting with practical implementation, and studied advantages and disadvantages of each model.

The field of possible practical application: economics, astronomy, meteorology, marketing, research activities, and other fields.

Obtained results and their novelty: an optimal strategy for time series analysis and forecasting was proposed. Experimental results demonstrate the advantages and disadvantages of different model classes for forecasting and highlight the importance of analysis.

ВВЕДЕНИЕ

В современном мире способность принимать решения и разрабатывать стратегию развития, сформированную на основе анализа данных, играет важную роль в достижении успеха. По этой причине различные компании и организации создают долгосрочные прогнозы, разрабатывают алгоритмы для определения предпочтений потребителя. Ввиду современных тенденций анализ и прогнозирование временных рядов являются важной сферой исследований во многих областях.

Изучение различного рода процессов играет важнейшую роль. Врач по текущим симптомам пытается определить дальнейшее развитие болезни, инженер изучает, насколько прочным должно быть крыло самолета, чтобы выдержать соответствующую нагрузку, владелец магазина выясняет, в каких объемах проводить закупку на следующую неделю. Примеров можно привести множество, однако все они имеют одну общую деталь: предсказать точный результат невозможно. Например, по текущему состоянию колонии бактерий нельзя сказать, какова в точности будет популяция этой колонии через некоторый промежуток времени. Для выполнения поставленных задач как раз и используются временные ряды.

Временные ряды имеют широкое применение в различных областях науки: астрономии (солнечная активность), экономике, медицине (результаты электрокардиограммы) и т. д.

Главная задача анализа временного ряда заключается в создании прогноза его будущих значений с использованием наиболее подходящей математической структуры. Для достижения этой цели необходимо решить две ключевые задачи: определить, какие факторы влияют на формирование значений временного ряда, и построить математическую модель для каждой из этих составляющих или для их комплексного отражения.

Любой временной ряд можно представить, как сумму тренда, сезонной, циклической и случайной составляющих. Три первых компонента формируют неслучайную часть временного ряда. Случайная составляющая свойственна любому временному ряду. Однако наличие в структуре временного ряда компонентов неслучайной составляющей не является обязательным условием.

Существуют два основных подхода к моделированию временных рядов: одновременное моделирование неслучайной составляющей во всей совокупности и разложение временного ряда на отдельные компоненты, с последующим моделированием значений каждой из этих компонент по отдельности.

Для того, чтобы раскрыть все аспекты работы с временными рядами, а также предложить наиболее эффективный подход к их анализу и моделированию, сформулированы следующие задачи:

- Исследовать существующие подходы к анализу и моделированию рядов;
- Математически обосновать рассмотренные методы и модели;
- Провести собственный анализ данных с применением теоретических знаний;
- Выявить закономерности, недостатки и преимущества той или иной модели;
- Сделать выводы о проделанной работе.

ГЛАВА 1
АНАЛИЗ ВРЕМЕННОГО РЯДА

1.1 Теоретические сведения о временных рядах

Данные временного ряда - это серия наблюдений, записанных через разные или регулярные промежутки времени. В общем случае временной ряд - это последовательность точек данных, снятых через равные промежутки времени. Частота записи точек данных может быть ежечасной, ежедневной, еженедельной, ежемесячной, ежеквартальной или ежегодной. На рисунке 1.1 приведён временной ряд, построенный на основе набора данных по товарообороту.

Рисунок 1.1 – Временной ряд

Важными понятиями при анализе и прогнозировании временных рядов являются лаг и уровень. Уровень представляет собой среднее значение за определённый промежуток времени (например, средняя температура за год), лаг – временное смещение в ряду, или некоторое предыдущее значение ряда.

При исследовании временных рядов выделяют следующие компоненты:

- Тренд: показывает общую тенденцию данных к увеличению или уменьшению в течение длительного периода времени. Тренд - это плавная, общая, долгосрочная, усредненная тенденция. Не обязательно, чтобы увеличение или уменьшение происходило в одном направлении на протяжении определённого промежутка времени;

- Сезонность: данный компонент показывает повторяющиеся тенденции по времени, направлению и величине. Например, пик розничных продаж приходится на декабрь месяц;
- Цикличность: это повторяющиеся закономерности или периодические колебания, которые происходят в течение определенного интервала времени. Это может быть связано с различными факторами, такими как сезонность (ежедневная, еженедельная, ежемесячная, годовая), тренд;
- Нерегулярные колебания: это внезапные изменения, происходящие во временном ряду, которые вряд ли повторятся. Их нельзя объяснить тенденциями, сезонными или циклическими изменениями. Эти вариации иногда называют остаточными или случайными компонентами.

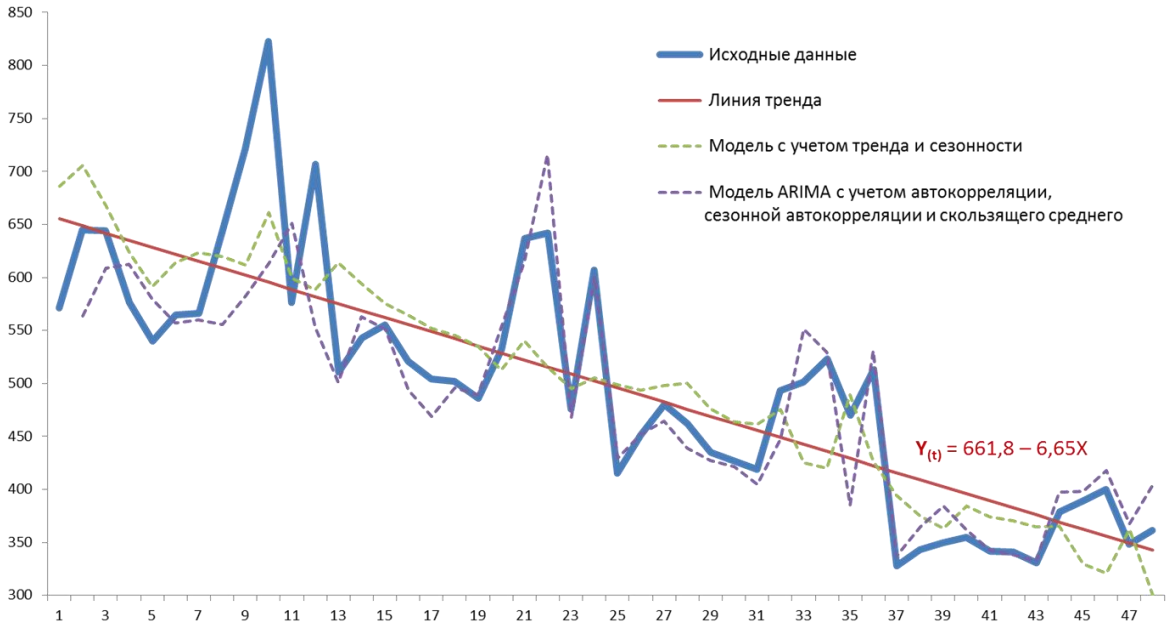

Рисунок 1.2 – Временной ряд с учётом компонент

На рисунке 1.2 показано моделирование временного ряда с учётом компонент. Проанализировав результаты видно, что при прогнозировании учитывается тенденция к уменьшению значений параметра, сезонные перепады, благодаря чему при применении модели ARIMA отличие практически нет, особенно при резких изменениях.

1.1.1 Классификация временных рядов.

Временной ряд в общем случае состоит из пары обязательных параметров: время (отсюда и название данного фундаментального понятия) и,

собственно, значений анализируемых данных в конкретные моменты (интервалы) времени.

Ввиду многообразия возможных вариантов второго показателя и неоднозначности поведения ряда можно проследить основные моменты и закономерности путём анализа. Для упрощения временные ряды введена их классификация в зависимости от характера данных и структуры (Таблица 1).

Таблица 1.1 – Классификация рядов

1. По виду анализируемого показателя	1. Абсолютных величин 2. Относительных величин 3. Средних величин
2. По характеру измерения показателей во времени	1. Интервальные 2. Моментные
3. По виду интервалов времени	1. С равностоящими уровнями 2. С неравностоящими уровнями
4. По поведению статистических свойств	1. Стационарные ряды 2. Нестационарные ряды

Временные ряды можно разделить на три основные группы в зависимости от вида анализируемых показателей:

- Ряды абсолютных величин;
- Ряды относительных величин;
- Ряды средних величин.

К первому типу относятся временные ряды, содержащие фактические количественные значения (в натуральных или стоимостных единицах). К таким величинам можно отнести, например, объем продаж в штуках, ВВП страны в денежном выражении, численность населения города, количество выпущенных автомобилей. Особенность таких рядов заключается в следующем: возможность отождествлять значения для однородных явлений, однако присутствует зависимость от масштаба (например, нельзя напрямую сравнивать ВВП США и маленькой страны) и чувствительность к изменению единиц измерения. Для таких временных рядов удобно применять следующие методы анализа и прогнозирования:

- Изучение динамики (рост/падение);
- Выявление трендов и сезонности;
- Прогнозирование (ARIMA, регрессионные модели);

Второй тип подразумевает отражение соотношения между разными показателями (доли, проценты, индексы, коэффициенты). Примером можно привести уровень инфляции или доля рынка (%), коэффициент рождаемости (на тысячу человек). Особенность данных рядов заключается в независимости

масштаба, возможности сравнения разнородных данных. Подходящие методы анализа:

- Анализ динамики изменений (%);
- Сравнение с benchmark (например, средним по отрасли);
- Построение индексов (Laspeyres, Paasche);

Временные ряды средних величин, в свою очередь, характеризуют усредненное значение показателя за интервал времени (средняя зарплата по стране, среднемесячная температура). Отличительной чертой является сглаживание данных, т.е. позволяют убрать "шум" и выделить общую тенденцию. Также есть зависимость от метода усреднения (медиана, среднее арифметическое, взвешенное среднее). К таким рядам применяют метод скользящего среднего, сравнение средних за разные периоды, а также построение доверительных интервалов.

В зависимости от состояний явлений по времени временные ряды можно разделить на два принципиально разных типа по характеру измерения показателей во времени: интервальные и моментные. Это разделение критически важно для корректного выбора методов анализа и интерпретации результатов, поскольку при моделировании и прогнозировании необходима информация про структуру ряда.

Интервальные временные ряды показывают, как меняется значение какого-либо показателя за определенные периоды времени, такие как год, месяц или квартал. Пример: ежемесячные данные о средней температуре. В отличие от них, существуют ряды, где значения привязаны к конкретным моментам времени, например, к определенному дню недели или дате (1 января).

Временные ряды можно классифицировать по тому, насколько регулярно расположены интервалы времени между измерениями: равномерно или неравномерно.

В случае временных рядов с равностоящими уровнями по времени, наблюдения фиксируются на одинаковых интервалах времени. Такие временные ряды характеризуются фиксированным шагом наблюдений (час, день, неделя, месяц и т.д.), регулярной структурой и стандартным форматом для большинства аналитических методов. К примерам можно отнести ежедневные котировки акций или ежемесячные показатели инфляции. Отсюда такие ряды при пропусках данных требуют:

- Интерполяцию;
- Замена средним/медианным значением;

В случае временных рядов с неравностоящими уровнями по времени, интервалы между наблюдениями могут быть неодинаковыми. К характеристике данных временных рядов приписывают переменный шаг

измерений, сложность анализа, а именно: невозможность прямого применения стандартных методов и необходимость предобработки (например, нормализация, т.е. приведение к более узкому интервалу возможных значений, чаще всего от 0 до 1). Временные ряды с неравностоящими уровнями по времени имеют ряд подходящих для них методов обработки:

- Приведение к равномерному виду: агрегация по временным окнам, сплайны, линейная интерполяция;
- Специальные алгоритмы: неравномерные аналоги ARIMA, точечные процессы;
- Методы машинного обучения: рекуррентные нейронные сети с временными метками и т.д.

Подытожив, выбор между подходами зависит от конкретной задачи, качества данных и требуемой точности анализа. Современные методы глубокого обучения (как, например, InFormer) постепенно стирают границы между обработкой разных типов рядов, но понимание фундаментальных различий остается критически важным для корректного анализа.

Несмотря на важность описанных выше свойств и типов временных рядов, наиболее значимой характеристикой является стационарность – свойство, основная концепция которого заключается в том, что статистические свойства (например, дисперсия или последовательная корреляция), постоянны с течением времени. По наличию данного свойства временные ряды подразделяются на стационарные и нестационарные. В таблице 2 приведены основные отличия по разным статистическим критериям.

Таблица 1.2 – Сравнение по критериям с учётом стационарности

Критерий	Стационарные ряды	Нестационарные ряды
Среднее	Постоянное	Изменяется
Дисперсия	Одинаковая	Меняется (гетероскедастичность)
Автокорреляция	Зависит только от лага	Зависит от времени
Типовые модели	AR, MA, LSTM	ARIMA, SARIMA
Примеры	Белый шум, остатки	Биржевые котировки, ВВП

Для проверки стационарности существует немало способов: от визуальной оценки графика и его важнейший компонент до проверки гипотез. В работе будет посвящено немало времени для оценки данного свойства.

1.2 Методы анализа

Анализ временных рядов является важным инструментом для прогнозирования будущих значений и понимания тенденций в данных. Он позволяет выявлять закономерности и зависимости между данными во времени, что может быть полезным для принятия решений в различных областях, таких как экономика, финансы, маркетинг и т. д. Без предварительного анализа трудно представить любое исследование данных.

Характеристики отдельных компонентов временного ряда определяют группы математических подходов, используемых для их изучения. Для выявления и анализа тренда применяют такие методы, как тест Дики-Фуллера (анализ стационарности), тест Грейнджера (связь между временными рядами), корреляционный анализ и многие другие.

Методы анализа сезонности и цикличности во временных рядах сильно зависят от целей исследования. Если эти аспекты являются важными, то их следует выделять из общего ряда и анализировать с применением соответствующих моделей. В случае, если колебания представляют собой нежелательный шум, то проводится анализ в рамках случайной составляющей.

Отклонения от общего тренда можно выявить при помощи спектрального анализа. Для исследования таких процессов часто используют гармонические модели с применением ряда Фурье. Преимуществом методов спектрального анализа является возможность оценить влияние колебательной компоненты, не учитывая другие компоненты ряда. Описание в частотной области часто применяется в научной деятельности.

Специальный класс стохастических моделей разработан для анализа реакции технических систем на импульсные воздействия. Такие модели позволяют не только оценивать текущее состояние системы, но и предсказывать ее поведение при внешних возмущениях. Для выявления взаимосвязей между различными временными рядами и учета их в прогнозах применяются методы кросс-корреляционного анализа и модели передаточных функций, которые помогают определить степень влияния одних процессов на другие.

Современные вычислительные технологии и усовершенствованные методы анализа нестационарных процессов открывают новые возможности в диагностике технических систем. Благодаря развитию алгоритмов обработки сигналов и применению адаптивных математических методов удается

повысить точность и информативность анализа, а также расширить диапазон обнаруживаемых аномалий.

Развитие методов количественных характеристик степени нерегулярности процессов и возросшие возможности современной вычислительной техники позволяют за счет использования новых подходов к сущности сигналов и новых видов их специальной математической обработки существенно расширить диапазон и повысить информативность применяемых диагностических методов исследования поведения технических систем.

Классический подход к прогнозированию временных рядов включает последовательное выделение и исключение структурных компонент: тренда, сезонных колебаний и эффектов интервенций. После устранения этих детерминированных составляющих оставшийся сигнал должен соответствовать случайному процессу (белому шуму), что подтверждает адекватность модели. На завершающем этапе прогноз строится отдельно для каждой компоненты с последующим синтезом итогового результата.

1.2.1 Спектральный анализ

Непрерывный или дискретный временной ряд может быть проанализирован с точки зрения описаний во временной области и описаний в частотной области. Последний также называется спектральным анализом и выявляет некоторые характеристики временного ряда, которые нелегко увидеть при анализе описания во временной области. Спектральный анализ используется для решения широкого спектра практических задач в технике и науке, например, при изучении вибраций, межфазных волн и анализе устойчивости.

Любой временной ряд можно разложить на гармонические составляющие с помощью преобразования Фурье. Однако на практике обычно требуется выделить не все компоненты, а только основные – те, которые вносят наибольший вклад в дисперсию сигнала.

При работе с дискретными данными (когда общий период наблюдения T разделен на N интервалов, $T = N \cdot \Delta t$) временной ряд x(t) представляется N отсчетами. Дискретное преобразование Фурье (ДПФ) позволяет получить значения спектральной плотности на определенных частотных интервалах, что дает возможность выявить наиболее значимые колебательные составляющие процесса.

С помощью преобразования Фурье любой ряд динамики можно представить в виде суммы конечного числа гармоник. Но часто исследователю нужны не все гармоники, а лишь основные, порождающие основную часть дисперсии. При дискретном временном параметре такая реализация

представлена N (т.к. T=N·Δt) значениями временного ряда x(t). Дискретное преобразование Фурье дает значения спектральной плотности на частотах:

$$f_k = \frac{k}{T} = \frac{k}{\Delta t \times N}, k = \overline{0, N-1} \quad (1.1)$$

При этом коэффициенты Фурье определяются по следующей формуле:

$$X(f_k) = \Delta t \sum_{i=0}^{N-1} x(i) \times [\cos\frac{2\pi ki}{N} - sin\frac{2\pi ki}{N}] \quad (1.2)$$

1.2.2 Корреляционный анализ

При детальном рассмотрении временного ряда и его тенденций придают значение корреляционной зависимости каждого уровня ряда от предыдущих - автокорреляции, анализ которой даст возможность определить структуру ряда т. е. выявить наличие той или иной компоненты. Для исследований вводят автокорреляционную функцию – последовательность коэффициентов автокорреляции уровней первого и последующих порядков, характеризующих тесноту линейной связи между последующими и предыдущими членами временного ряда. Графиком такой функции называют коррелограммой. Пример коррелограммы показан на рисунке 1.3.

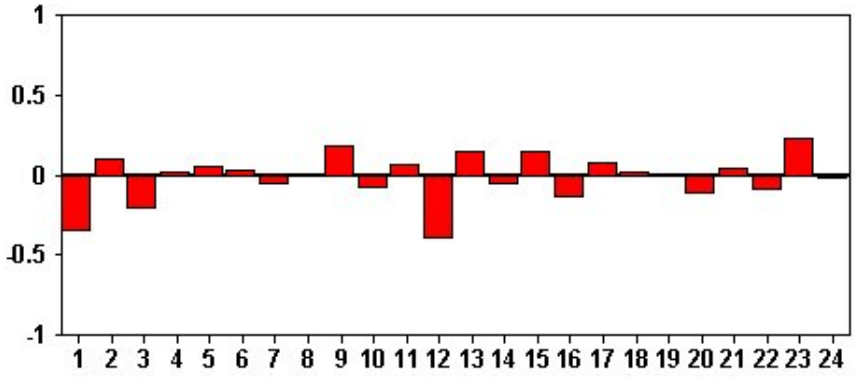

Рисунок 1.3 – График функции

Ось x графика автокорреляционной функции показывает лаг, при котором вычисляется автокорреляция; ось y показывает значение корреляции (от −1 до 1). Например, пик при лаге 1 на коррелограмме показывает сильную корреляцию между значениям ряда и предыдущим значением, пик при лаге 2

показывает сильную корреляцию между каждым значением и значением в более ранний момент на расстоянии 2 от данного и так далее.

На основе графика можно сделать некоторые выводы:

- Положительная корреляция при некотором лаге показывает, что большому текущему значению отвечают большие значения с этим лагом, а отрицательная корреляция показывает обратное;
- Абсолютная величина корреляции служит мерой связанности, причем чем больше абсолютное значение, тем сильнее взаимосвязь.

Исследование автокорреляционной функции и коррелограммы помогает определить структуру временного ряда, выявляя наличие определенных компонент. Когда наибольшее значение имеет коэффициент автокорреляции первого порядка, это свидетельствует о наличии в ряде только трендовой составляющей. В случае, когда максимальным оказывается коэффициент автокорреляции порядка m, можно сделать вывод о присутствии циклических колебаний с периодом m временных интервалов. Если же ни один из коэффициентов автокорреляции не достигает значимого уровня, это означает отсутствие в ряде как тренда, так и циклических компонент.

1.2.3 Анализ стационарности

Первым шагом в любом процессе создания временных рядов является определение стационарности ряда. Пересказывая определение из классификации, переменная называется стационарной, если ее статистические свойства, такие как среднее значение, дисперсия и последовательная корреляция, постоянны с течением времени. Это свойство делает анализ более простым, поскольку можно предсказать, что статистические свойства останутся такими же, какими они были в прошлом.

Стационарность временных рядов играет ключевую роль в анализе данных по нескольким причинам. Во-первых, стационарные ряды значительно упрощают процесс прогнозирования благодаря постоянству своих статистических свойств во времени. Во-вторых, большинство современных методов анализа, включая популярные модели ARIMA, изначально разработаны для работы со стационарными рядами либо предполагают возможность их преобразования к стационарному виду.

Особую важность это требование приобретает в связи с тем, что использование нестационарных данных может привести к серьезным ошибкам: искажению результатов статистического анализа, снижению точности прогнозов и ошибочным выводам. Именно поэтому проверка на стационарность и соответствующие преобразования данных становятся

обязательным этапом предварительного анализа временных рядов перед применением сложных статистических методов.

С данными реального мира часто бывает так, что ряды нестационарны, и поэтому был разработан набор математических методов для преобразования их в стационарные:

- Данные модифицируют, учитывая ряд x(t) и формируя новый ряд y(t):

$$y_i(t) = x_i(t) - x_{i-1}(t) \quad (1.3)$$

Измененные данные будут содержать на одну точку меньше, чем исходные данные. Хотя вы можете изменять данные более одного раза, обычно достаточно одного различия;

- Для непостоянной дисперсии использование логарифма или квадратного корня из ряда может стабилизировать дисперсию. Для отрицательных данных добавляется подходящая константа, чтобы сделать все данные положительными перед применением преобразования. Затем используемая константа может быть вычтена из модели для получения прогнозируемых значений и прогнозов для будущих точек.

Наивным и в то же время простым способом определения стационарности является отображение ряда и проверка на тренд и сезонность. Хотя и можно заметить некоторые свойства ряда, однако разработаны более эффективные подходы для определения стационарности:

- Расширенный тест Дики-Фуллера: данный тест также называют тестом единичного корня. В теории вероятностей и математической статистике единичный корень является характеристикой некоторых стохастических процессов, которые могут вызывать проблемы при статистическом выводе с использованием моделей временных рядов. Проще говоря, единичный корень нестационарен, но не всегда имеет трендовую составляющую. Тест на стационарность строится на проверке двух конкурирующих гипотез: нулевая гипотеза H0 утверждает о наличии единичного корня, что свидетельствует о нестационарности ряда, тогда как альтернативная гипотеза H1 предполагает стационарность временного ряда. В ходе тестирования определяются критические значения для стандартных уровней значимости (1%, 5% и 10%), которые отражают вероятность ошибочного отклонения нулевой гипотезы (Рисунок 1.4). Также рассчитывается p-значение – минимальный уровень значимости, при котором H_0 может быть отвергнута. Например, p-значение 0.05 означает 5% вероятность наличия единичного корня в ряде.

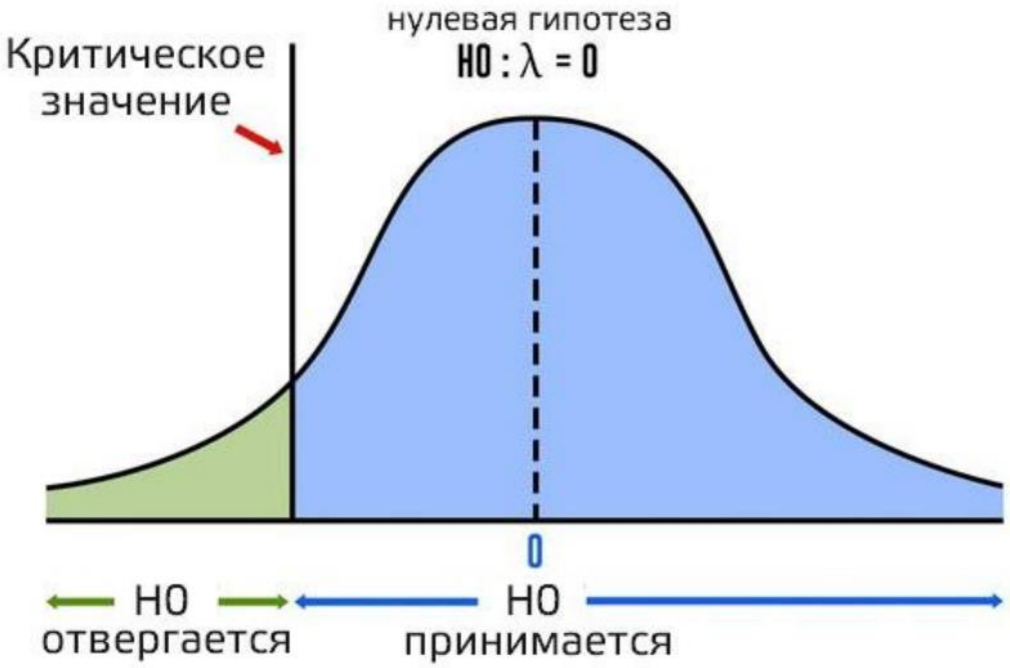

Рисунок 1.4 – Интерпретация результатов теста

- Тест Филлипса-Перрона является расширенной версией теста Дики-Фуллера, основное применение которого работа с временными рядами с выраженной сезонностью и изменениями в структуре сдвигов. Область его применения ограничена условиями, что случайные отклонения имеют различную дисперсию и отсутствует нормальное распределение. Нулевая гипотеза заключается также в наличии единичного корня, тогда ряд нестационарен, в то время как альтернатива говорит об обратном. Как и в тесте Дики-Фуллера, нулевая гипотеза отвергается при p-значении меньше 0.05, либо ряд является стационарным, если полученная статистика меньше заданных критических значений. Данный метод используется в основном для коротких временных рядов (менее 100 наблюдений), при сильной автокорреляции в остатках и при непостоянной дисперсии. Главным недостатком метода является его уязвимость к сильным отклонениям и взрывным трендам, что может привести к искажению результата.

- Тест Квятковского-Филлипса-Шмидта-Шина, который часто используется вместе с тестом Дики-Фуллера для перекрёстной проверки ввиду его специфики. Ключевое отличие состоит в том, что в качестве нулевой гипотезы принимается информация о том, что ряд стационарен (вокруг детерминированного тренда или уровня), в то время как в ADF – наличие

единичного корня (ряд нестационарен). Математическая основа анализа данным методом подразумевает разложение ряда на три компоненты: детерминированный ряд (если есть), случайное блуждание (стохастический тренд) и стационарная ошибка. Статистика KPSS оценивает кумулятивную ошибку от регрессии на тренд:

$$KPSS = \frac{\sum_{t=1}^{T} {S_t}^2}{T^2 * \sigma^2} \quad (1.4)$$

где St – частичная сумма остатков,
σ – корень оценки дисперсии остатков.

Интерпретация результатов следующая: полученная по формуле (1.4) статистика сравнивается с критическими значениями по следующим правилам: если значение статистики больше критического, то гипотеза отвергается, что означает нестационарность рассматриваемого временного ряда, или p – значение меньше заданного порога (по умолчанию берётся 0.05, иначе требуется указать причины, почему оно должно быть другое). Несмотря на свои преимущества, данный метод имеет ряд ограничений: не различает тип нестационарности, чувствителен к выбору лага для оценки дисперсии, и, что самое главное, существует вероятность искажения результата в случае структурных изменениях ряда, поэтому тест имеет смысл применять в комплексе с другими (в частности с тестом Дики-Фуллера).

1.2.4 Тест Грейнджера

Тест Грэйнджера – это статистический тест, предложенный в 1969 году Клайвом Грейнджером (в 2003 году учёному за данную работу получил Нобелевскую премию), который используется для определения причинно-следственных связей между временными рядами. Суть теста заключается в проверке гипотезы о том, что значения одного временного ряда могут быть полезны для прогнозирования другого. Например, в экономической теории существует предположение об обратной зависимости между инфляцией и безработицей (кривая Филипса) – тест Грэйнджера позволяет статистически подтвердить или опровергнуть наличие такой связи, а также определить ее направленность.

Математические описание данного статистического теста имеет следующий вид: пусть на вход принимается два временных ряда X и Y, причём их длина одинаковая (т.е. одинаковое количество наблюдений). Также на длину накладывается условие:

$$L \geq 3 * m + 3 \quad (1.5)$$

где L – длина (количество точек ряда),
m – лаг.

Основная концепция заключается в следующем: если наблюдение x имеет влияние на наблюдение y второго временного ряда, то преобразования x идут раньше преобразований y, исключая обратную ситуацию, т.е. x влияет на прогноз y, но y не участвует в прогнозе x.

Для проверки на данное условие оценивают регрессию на значения лагов для X и Y, используя следующую модель и нулевую гипотезу:

$$y_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j y_{t-j} \quad + \sum_{j=1}^{m} \beta_j x_{t-j} + \varepsilon_t \quad (1.6)$$

Нулевая гипотеза: βi = 0, для i от 1 до m. В таком случае говорят, что x не является причиной Грейнджера для Y.

Аналогично формулируется для y, после чего проверяются результаты в обоих случаях, причём в одном гипотеза должна быть отвергнута, а в другом – принята.

Тест проделывают для нескольких значений m для оценки степени влияния длины лага на результат.

1.3 Выводы по первой главе

Данная глава посвящена глубокому разбору временных рядов как опции для эффективного статистического анализа данных. Рассмотрены наиболее важные методы анализа, широко применяющиеся на практике. Особое внимание получил разбор стационарности как ключевого свойства. В дальнейшем рассмотренная в этой главе теория станет подспорьем для построения моделей и правильной оценки параметров, что позволит существенно повысить точность вычислений и, как итог, прогноза.

ГЛАВА 2
ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ

2.1 Основные положения

Прогнозирование временных рядов – это процесс использования статистических методов и математических моделей для предсказания будущих значений на основе исторических данных, упорядоченных во времени. Многие компании изучают прогнозирование временных рядов как способ принятия более эффективных бизнес-решений. Возьмем в качестве примера отель. Если у менеджера есть хорошее представление о том, сколько гостей ожидается следующим летом, он может использовать эти данные для планирования управления персоналом, бюджета или даже расширения объекта. Аналогичным образом, уверенное понимание будущих событий может принести пользу широкому спектру отраслей и проблем, от традиционного сельского хозяйства до транспорта по требованию и многого другого. На рисунке 2.1 приведены виды методов и моделей прогнозирования.

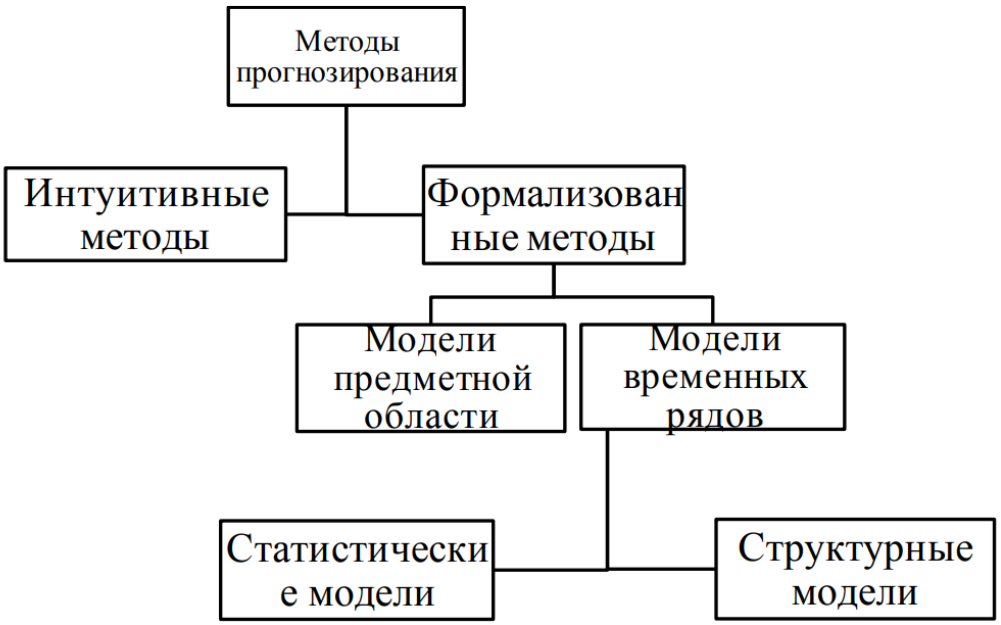

Рисунок 2.1 – Виды методов и моделей прогнозирования

Чтобы дать необходимую теоретическую основу для исследования, дана краткая характеристика каждого из представленных видов прогнозов.

Интуитивные методы прогнозирования основываются на профессиональном опыте и логическом анализе экспертов. Формализованные

методы, в свою очередь, используют математический аппарат, что существенно повышает точность прогнозов, сокращает сроки их разработки и минимизирует субъективный фактор. Особое место занимают предметно-ориентированные модели, которые строятся на основе закономерностей конкретной предметной области и учитывают её специфику. Отдельного внимания заслуживают модели временных рядов, анализирующие внутренние зависимости исторических данных для предсказания будущих значений. Эти модели обладают универсальностью применения и подразделяются на статистические (например, регрессионные уравнения) и структурные (включая нейросетевые архитектуры). Представленный обзор демонстрирует эволюцию методов прогнозирования от субъективных экспертных оценок к строгим математическим моделям, что значительно расширяет возможности точного предсказания в различных сферах деятельности.

В следующем разделе будут рассмотрены конкретные модели прогнозирования, их достоинства и недостатки.

2.2 Модели прогнозирования рядов

2.2.1 Регрессионные модели

Регрессионное моделирование временных рядов основывается на принципе линейной зависимости между прогнозируемым показателем Y и объясняющими переменными X. Суть подхода заключается в выявлении и количественной оценке взаимосвязи между целевым временным рядом и влияющими на него факторами.

В простейшем случае регрессионная модель допускает линейную зависимость между прогнозируемой переменной y и внешним фактором x. Модель прогнозирования в таком случае описывается следующим уравнением:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \varepsilon_t \qquad (2.1)$$

где β0 и β1 – коэффициенты регрессии,
ε – ошибка модели, отражающая любые факторы, влияющие на y.

В случае, если на y влияет множество внешних факторов, рассматривают множественную регрессионную модель, уравнение которой имеет вид:

$$y_t = \beta_0 + \beta_1 \cdot x_{1,t} + \beta_1 \cdot x_{1,t} + \cdots + \beta_k \cdot x_{k,t} + \varepsilon_t \qquad (2.2)$$

Регрессионные модели показывают наилучший результат в случае, если можно без проблем выявить вид функциональной зависимости. Кроме того, модель крайне гибкая и прозрачная для исследования промежуточных вычислений.

При нелинейной зависимости между рядом y и внешним фактором x вступает в силу предположение, что существует известная функция F, описывающая данную зависимость:

$$y_t = F(x(t), A) \tag{2.3}$$

где A – вектор варьируемых параметров функции F.

2.2.2 Авторегрессионные модели

В основу авторегрессионных моделей заложено предположение о том, что значение временного ряда y(t) линейно зависит от некоторого количества предыдущих значений y(t-1), …, y(t-p). К наиболее эффективным и используемым моделям относят:

Авторегрессионная (AR) модель: в данной модели текущее значение ряда выражается как конечная линейная совокупность предыдущих значений ряда и импульса, который называется «белым шумом». Формула (2.4) описывает AR порядка p:

$$(t) = C + \varphi_1 \cdot y(t-1) + \cdots + \varphi_p \cdot y(t-p) + \varepsilon \tag{2.4}$$

где C - вещественная константа,
φi - коэффициенты,
ε - ошибка модели.

Для определения φi и C используют метод наименьших квадратов или метод максимального правдоподобия.

Модель скользящего среднего (MA) представляет собой метод сглаживания временных рядов, который по своей сути функционирует как фильтр низких частот, устраняя высокочастотные колебания и шумы. Следует подчеркнуть, что данный подход реализуется в нескольких модификациях: экспоненциальное, кумулятивное взвешенное и простое скользящее среднее. Описывается следующим уравнением:

$$y(t) = \frac{1}{q}(y(t-1) + \cdots + y(t-q)) + \varepsilon \tag{2.5}$$

где q – порядок,
ε – ошибка модели.

Для повышения эффективности моделирования временных рядов на практике часто используют комбинированные модели, объединяющие подходы авторегрессии и скользящего среднего. Модель ARMA(p,q) сочетает два ключевых компонента: авторегрессионную часть порядка p, которая отражает линейную зависимость текущего значения от предыдущих наблюдений, и компоненту скользящего среднего порядка q, отвечающую за фильтрацию случайных колебаний. В случаях работы с нестационарными данными применяется расширенная версия - модель ARIMA(p,d,q), где параметр d показывает порядок дифференцирования, необходимый для приведения ряда к стационарному виду (как правило, достаточно первого или второго порядка).

Процесс построения такой модели включает несколько этапов: сначала исходный ряд преобразуется в стационарный через дифференцирование, затем подбираются оптимальные параметры p и q, после чего оценивается статистическая значимость коэффициентов. Главные преимущества этих комбинированных моделей заключаются в их универсальности для различных типов временных рядов, способности одновременно учитывать внутренние зависимости и случайные колебания, а также гибкости настройки под конкретные аналитические задачи.

Благодаря этим характеристикам модели ARMA/ARIMA нашли широкое применение в таких областях как эконометрическое прогнозирование, анализ финансовых показателей и обработка технических сигналов, где требуется точное моделирование сложных временных зависимостей.

Развитием модели ARIMA(p,d,q) является модель ARIMAX(p,d,q), которая описывается уравнением:

$$y(t) = AR(p) + \alpha_1 \cdot x_1(t) + \cdots + \alpha_s \cdot x_s(t) \qquad (2.6)$$

где αi — коэффициенты внешних факторов Xi.

В данной модели чаще всего y(t) является результатом модели MA(q), то есть отфильтрованными значениями исходного ряда. Далее для прогнозирования y(t) используется модель авторегрессии, в которой введены дополнительные регрессоры внешних факторов Xi.

Для анализа временных рядов с выраженной сезонной компонентой часто применяют расширенную версию ARIMA-модели - сезонную модель SARIMA. В отличие от базового варианта, SARIMA включает дополнительные сезонные параметры: P (порядок сезонной авторегрессии), D

(порядок сезонного дифференцирования), Q (порядок сезонного скользящего среднего) и s (период сезонности, например, 12 для месячных данных с годовым циклом). Эти параметры позволяют эффективно учитывать регулярные циклические колебания в данных. Однако, несмотря на свою популярность и широкое применение в задачах прогнозирования сезонных процессов, модель SARIMA имеет существенные ограничения в практическом использовании. Основные сложности связаны с трудоемкостью первоначальной настройки - требуется тщательный подбор как несезонных (p,d,q), так и сезонных (P,D,Q,s) параметров, что значительно увеличивает временные затраты на подготовку модели. Кроме того, модель требует значительных вычислительных ресурсов и регулярного переобучения при поступлении новых данных, что делает ее не всегда оптимальным выбором для оперативного прогнозирования или проектов с ограниченными техническими возможностями.

2.2.3 Модели экспоненциального сглаживания

Метод экспоненциального сглаживания, разработанный в конце 1950-х годов, стал фундаментом для многих эффективных прогнозных методик. Суть этого подхода заключается в вычислении взвешенного среднего исторических данных, где веса наблюдений уменьшаются по экспоненциальному закону по мере их удаления от текущего момента. Таким образом, наибольшее влияние на прогноз оказывают самые последние данные, в то время как более ранние наблюдения учитываются с постепенно снижающимся весом. Функция модели ES имеет вид:

$$\begin{cases} X(t) = S(t) + \varepsilon_t, \\ S(t) = \alpha \cdot \mathrm{X(t-1)} + (1-\alpha) \cdot \mathrm{S(t-1)}. \end{cases} \quad (2.7)$$

где α — коэффициент сглаживания в диапазоне от 0 до 1,
X(t) и S(t) – временные ряды.

Начальные условия определяются как S(1) = X(0). В данной модели каждое последующее сглаженное значение S(t) является взвешенным средним между предыдущим значением временного ряда X(t) и предыдущего сглаженного значения S(t-1).

Эффективность метода экспоненциального сглаживания во многом определяется значением коэффициента сглаживания, который играет ключевую роль в формировании прогноза. Когда коэффициент приближается к 1, модель придает максимальный вес последним наблюдениям и практически

игнорирует более ранние данные, что делает прогноз чрезвычайно чувствительным к новейшим изменениям.

Для работы с временными рядами, содержащими трендовую составляющую, был разработан усовершенствованный вариант метода - двойное экспоненциальное сглаживание. Эта модификация учитывает две независимые компоненты: базовый уровень ряда и его тренд, применяя к каждой из них отдельные процедуры сглаживания. Такой подход позволяет более точно моделировать динамику изменяющихся процессов, сохраняя при этом все преимущества оригинального метода, включая простоту вычислений и оперативность получения результатов. Двойное сглаживание особенно эффективно для рядов с устойчивой тенденцией роста или снижения, где традиционное экспоненциальное сглаживание может давать значительные погрешности. При этом сохраняется принцип экспоненциального уменьшения весов для более старых наблюдений, но теперь он применяется отдельно как к уровню ряда, так и к его трендовой составляющей, что существенно повышает точность прогнозов для динамически изменяющихся процессов.

Для моделирования временных рядов, имеющих тренд, была предложена усовершенствованная модель, дающая более точный результат: двойное экспоненциальное сглаживание. Сглаживание уровня и тренда проходит независимо друг от друга по следующим формулам:

$$\begin{cases} X(t) = S(t) + \varepsilon_t \\ S(t) = \alpha \cdot \mathrm{X(t-1)} + (1-\alpha) \cdot (\mathrm{S(t-1)} - B(t-1)) \\ B(t) = \gamma \cdot \big(\mathrm{S(t-1)} - \mathrm{S(t-2)}\big) + (1-\gamma) B(\mathrm{t-1}) \end{cases} \quad (2.8)$$

где α — коэффициент сглаживания,
γ — коэффициент сглаживания тренда.

2.2.4 Нейросетевые модели

Нейронные являются мощным инструментом во многих областях науки: решение задачи классификации, кластеризации, распознавания образов, анализа паттернов, принятия решений и прогнозирования. Для формирования прогноза применяют следующие нейросетевые модели:

1. Рекуррентные нейронные сети (RNN) представляют собой особый класс нейронных сетей, расширяющий возможности традиционных сетей прямого распространения за счет введения механизма внутренней памяти. Их ключевая особенность заключается в способности обрабатывать последовательные данные, сохраняя информацию о предыдущих вычислениях. Принцип работы RNN основан на циклической организации

вычислений – для каждого элемента входной последовательности сеть выполняет одинаковую операцию, но результат текущего вычисления зависит не только от входящих данных, но и от состояния сети, полученного при обработке предыдущих элементов. После генерации выходного значения состояние копируется и передается обратно в сеть для использования при следующем вычислении;

Принцип работы рекуррентных нейронных сетей завязан на понятии петли – нейрон связан сам с собой. Сама структура основывается на блочной архитектуре состояний: на вход принимается последовательность, которая разбивается на части, каждая часть передаётся в соответствующее состояние с результатом предшествующего блока. На рисунке 2.2 представлен вид простейшего элемента архитектуры рекуррентной нейронной сети, где происходит обновление состояния выходного вектора.

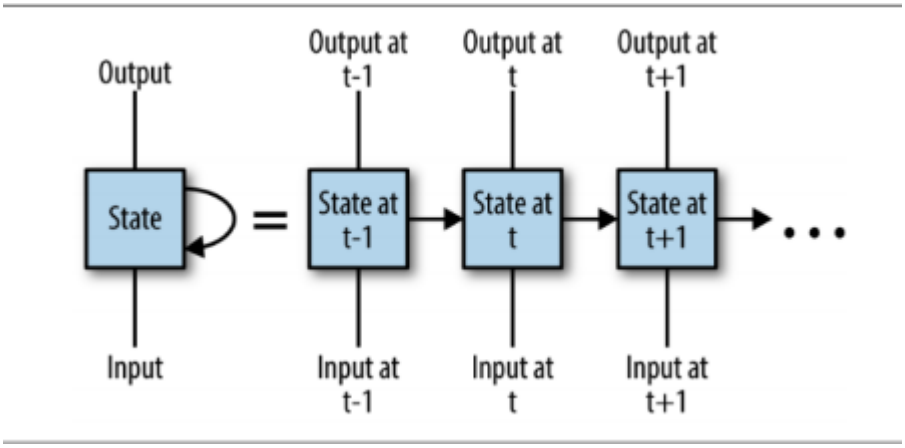

Рисунок 2.2 – Клетка рекуррентной нейронной сети

RNN имеет следующее математическое обоснование: пусть W – матрица весов (необходима для перехода между слоями), V – матрица весов для выходов, U – для входов, f – функция активации, h – функция для получения результата st – некоторое состояние в момент времени t, c и b – отклонения, xt– входящая последовательность значений. Тогда задача блока at формулируется следующим образом:

$$\begin{cases} a_t = b + W \cdot s_{t-1} + U \cdot x_t \\ o_t = c + V s_t \\ s_t = f(a_t) \\ y_t = h(o_t) \end{cases} \quad (2.9)$$

Принцип работы следующий: на вход приходит значение xt, умножается на U – матрицу весов для входов, состояние нейронной сети st умножается на матрицу весов для выходов – W, и, наконец, к полученным двум слагаемым добавляется свободный член – отклонение (для выравнивания значения). Далее полученное значение передаётся в функции активации, выбор которой крайне важен для дальнейшего обучения и точности выходного набора. Чаще всего использует следующие 4 вида функции:

- Сигмоида, определяется по следующей формуле:

$$\sigma = \frac{1}{1 + e^{x}} \tag{2.10}$$

Свойства функции: при убывающем к минус бесконечности x стремится к 0, стремится к 1 при возрастающем к плюс бесконечности, на всей области значений дифференцируема; основной недостаток – высокая скорость сходимости к крайним точкам.

- Гиперболический тангенс:

$$tanh = \frac{e^{x} - e^{-x}}{e^{-x} + e^{x}} \tag{2.11}$$

Как и сигмоида, дифференцируемый на всей области значений, при возрастании x также стремится к 1, при убывании – к -1. Скорость стремления даже больше, чем у сигмоиды, но отсутствие 0 гарантирует, что градиент при обучении не обнулится.

- Функция Хевисайда: принимает значение 0 при отрицательном x, 1 – при положительном; не определена в 0, однако есть возможность доопределить, тогда, как и у сигмоиды, допускается обнуление градиента при обучении.
- Функция ReLU: принимает значение 0 при отрицательном x, x – иначе; преимуществом можно считать простоту в дифференцировании, что помогает при вычислениях и ускоряет время обучения.

Несмотря на свои очевидные преимущества (учет временных зависимостей, работа с разномасштабными данными, поддержка многомерных рядов), при использовании данного типа прогнозирования ряда есть свои недочёты: основной алгоритм при обучении есть метод градиентного спуска, что может привести к возникновению известных проблем, таких как проблема «взрывающегося градиента» или «затухающего градиента».

2. Долгая краткосрочная память (LSTM): LSTM являются разновидностью RNN, специально разработанной для решения проблемы исчезающего градиента. Они обладают способностью запоминать информацию на длительные периоды времени, что делает их очень эффективными для прогнозирования временных рядов с долгосрочными зависимостями. Кроме того, данная модель прогнозирования временных рядов не требует дополнительной проверки стационарности ряда. Принцип работы LSTM основан на контроле состояния каждой ячейки: нейронная сеть уменьшает или увеличивает количество информации в состоянии ячейки, в зависимости от потребностей. Для этого используются тщательно настраиваемые структуры, называемые гейтами (рисунок 2.2).

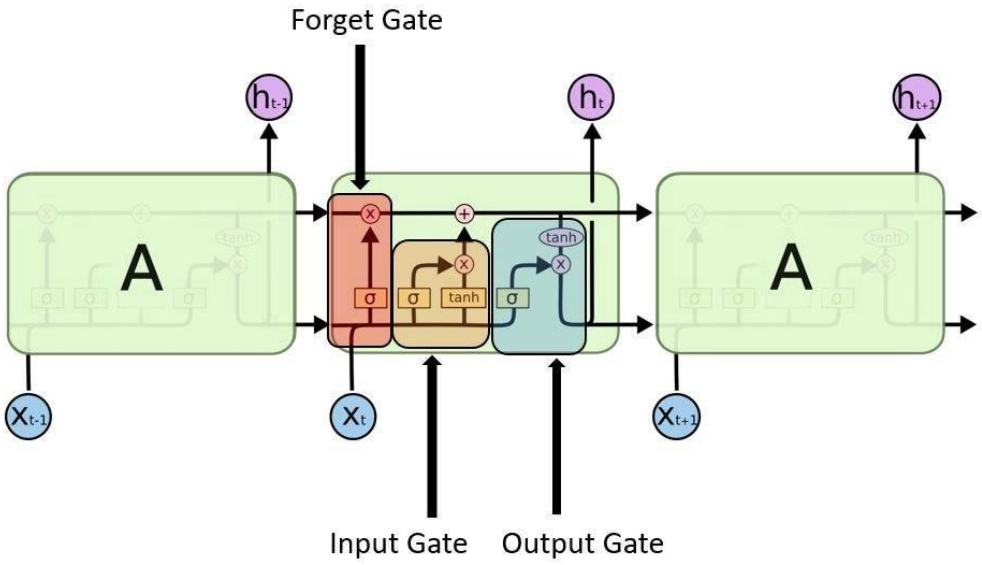

Рисунок 2.3 – Составляющие LSTM ячейки

Forget Gate – простейший из гейтов, поскольку представляет собой один полносвязный слой с сигмоидной функцией активации. Используется для того, чтобы обновить данные в векторе состояния.

Input Gate – данный слой, в отличие от предыдущего, состоит не только из рекуррентного слоя, но и полносвязного слоя с сигмоидной функцией активации, задача которого заключается в определении момента добавления информации из RNN в вектор.

Output Gate – полносвязный слой с сигмоидной функцией активации, который решает, какая часть долгосрочного вектора состояний должна быть передана на следующий шаг.

Такая архитектура довольно громоздкая и трудная для вычислений, что приводит к более медлительному обучению, и делает LSTM более тяжёлой в сравнении с другими RNN и CNN. Также крайне трудно подобрать гиперпараметры. Однако адаптивный механизм запоминания, высокая эффективность для последовательных данных, гибкость архитектурных решений, универсальность применения, возможность эффективно работать с долгосрочными зависимостями и решать проблему исчезающего градиента делает LSTM одним из приоритетных выборов при прогнозировании.

Кроме рассмотренных моделей есть некоторые более сложные, например, свёрточные нейронные сети (CNN), однако область их применения в данной сфере крайне ограничена.

2.3 Выводы по второй главе

В данной главе рассмотрена теоретическая основа для прогноза временных рядов. Классические статистические методы, такие как скользящее среднее, экспоненциальное сглаживание и модели ARIMA, хорошо работают при наличии устойчивых трендов и сезонностей. Они просты в реализации и интерпретации, но могут уступать современным методам при работе со сложными, нелинейными зависимостями.

Современные подходы, использующие машинное обучение и нейронные сети (например, LSTM), позволяют учитывать более сложные закономерности и взаимодействия в данных. Эти методы требуют больших объемов данных и вычислительных ресурсов, но часто обеспечивают более высокую точность прогнозов.

Таким образом, выбор метода прогнозирования временных рядов должен основываться на характеристиках данных, цели анализа и доступных ресурсах. Комплексный подход, сочетающий разные методы, может привести к более точным и надёжным прогнозам.

ГЛАВА 3
ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

3.1 Набор данных

Для решения поставленной задачи был взят набор данных с информацией о среднемесячном количестве солнечных пятен в период с 31.01.1749 по 31.01.2021 (3265 записей). В файле отражается солнечная активность за определённый промежуток времени, что является полезным в астрономии, поскольку анализ таких данных помогает выявить тенденции и аномалии физических явлений.

Исходный временной ряд представлен на графике 3.1.

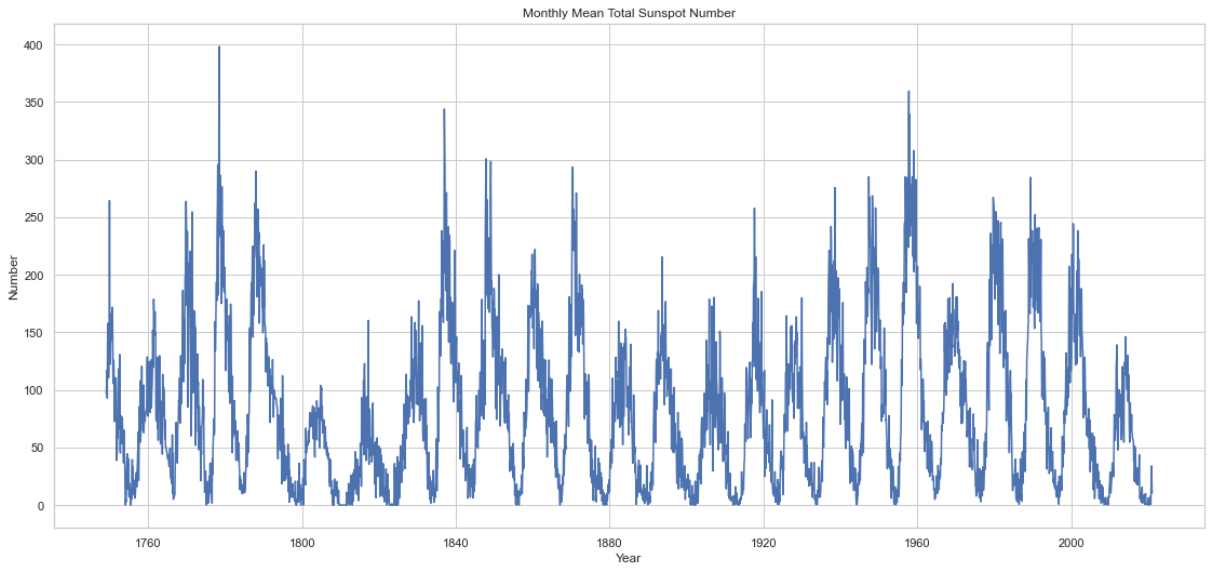

Рисунок 3.1 - Временной ряд

3.2 Модели LSTM

Для прогноза солнечной активности модель LSTM (долгая краткосрочная память) является одни из приоритетных вариантов для решения поставленной задачи ввиду следующих факторов:

- Для большинства моделей прежде, чем перейти к реализации прогноза необходима информация о том, является ли ряд стационарным. Однако данная модель не требует предварительного анализа стационарности;
- Поскольку для данного ряда характерна долгосрочная зависимость, выбор LSTM модели выглядит приоритетным ввиду того, что сеть имеет механизм, называемый "ячейкой памяти", который может сохранять информацию в течение длительного периода времени;

- LSTM также обладает способностью выборочно забывать информацию, что полезно для предотвращения переобучения и повышения точности прогнозов. Эта функция делает данную модель привлекательным выбором для задач прогнозирования временных рядов, особенно при работе с данными, имеющими сложные закономерности.

Реализация прогноза значений данного временного ряда произведена на языке Python с использованием открытых библиотек pandas, которая используется для обработки и анализа данных, и TensorFlow, использующаяся для решения задач построения и обучения нейронной сети. Исходный код представлен в конце работы (ПРИЛОЖЕНИЕ А).

3.2.1 Подготовка данных

Прежде, чем перейти к реализации прогноза солнечной активности, необходимо провести предварительный анализ и обработку данных. Для начала были произведены манипуляции с данными: Важным шагом перед обучением является масштабирование данных к диапазону от 0 до 1 для того, поскольку повышает скорость нахождения оптимального решения и эффективность обучения. Финальной частью подготовки является преобразование набора в формат, подходящий для контролируемого обучения, т. е. создать временные интервалы фиксированной длины, на основе которых формируются будущие значения временного ряда. Для того, чтобы определить длину интервала, необходимо исследовать график автокорреляционной функции (рисунок 3.2).

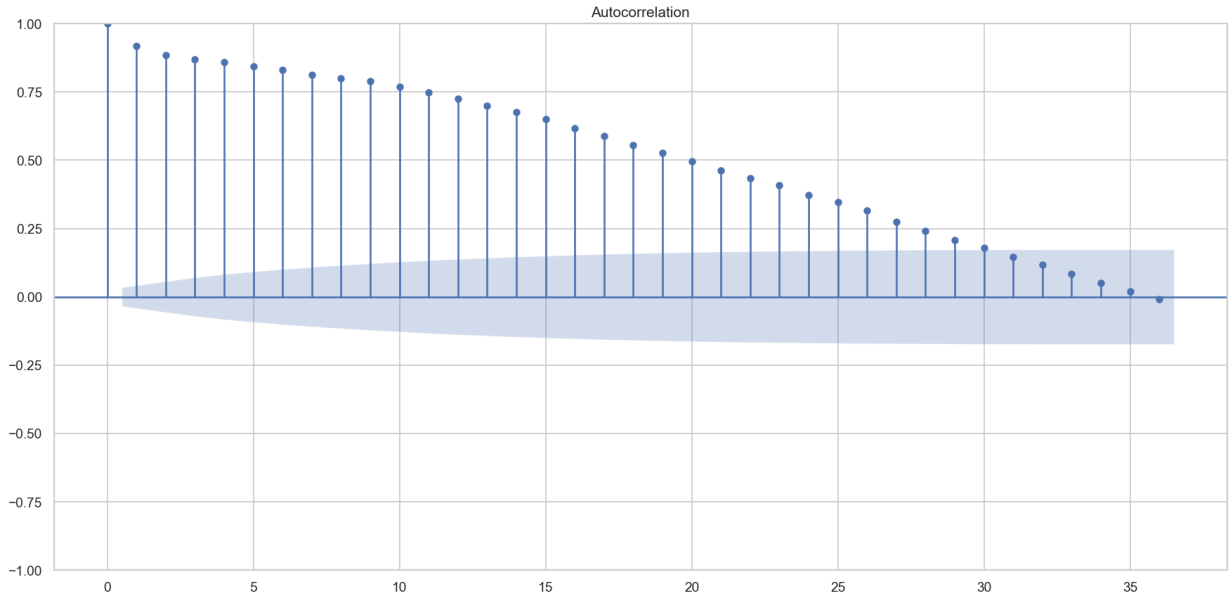

Рисунок 3.2 – График автокорреляционной функции

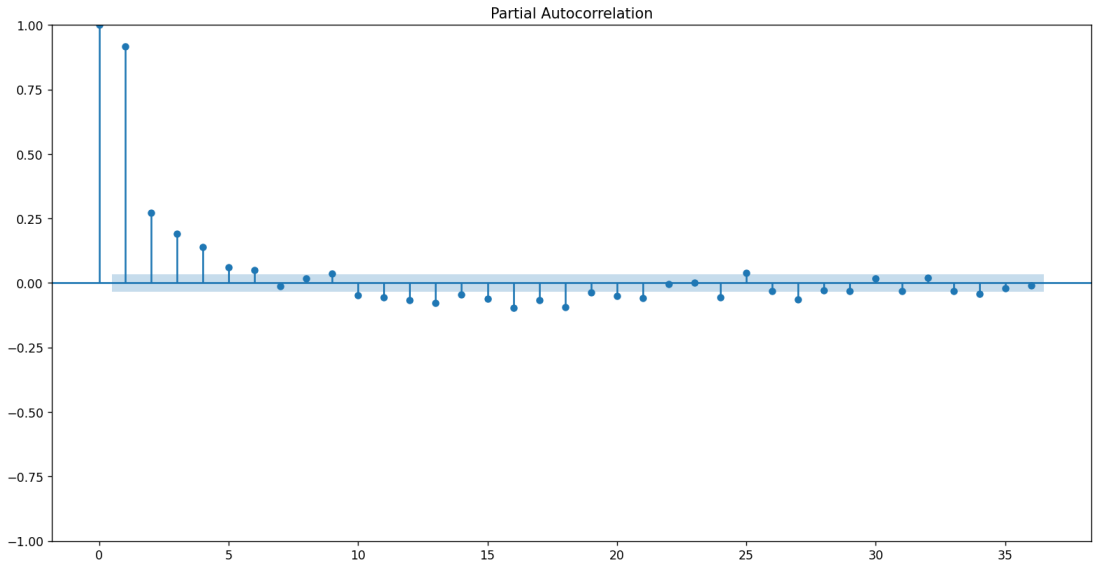

Рисунок 3.3 – График частной автокорреляционной функции

Анализ коррелограммы выявит важные тенденции: в данном случае интересна зависимость прогнозируемого значения от предыдущих. временного ряда показывает, что конкретное значение зависит примерно от 30 предыдущих значений ряда (исходя из того, что значения в синей зоне не являются статистически важными, а значение 30 является пиковым).

Набор данных был разбит на данные для обучения и данные для тестирования (90% и 10% от исходного набора соответственно).

Для проверки правильности загрузки и утверждения формата данных вывел первые несколько строк нашего набора данных: поскольку чтение из файла происходит с использованием функции из библиотеки pandas для чтения большинства типов файлов, то такая на первый взгляд простая проверка позволяет исключить ситуацию, при которой с самого начала есть ошибка, которая будет тянуться всю дальнейшую разработку.

Следующим шагом проверил пропущенные значения: отсутствующие значения могут вызывать проблемы при анализе и искажать результаты прогнозирования. Для проверки пропущенных значений использовал метод isnull() из библиотеки pandas в комбинации с функцией sum. Если результатом вывода будет 0, то данные хранятся в нужном виде, в противном случае все дальнейшие исследования приведут к ошибкам. Для данного набора проблем не обнаружено.

Каждый столбец в датафрейме имеет определенный тип данных. В динамических рядах особенно важен тип данных DataTime, который специально предназначен для хранения дат и времени. Датафрейм в pandas по умолчанию считывает информацию как текст. Даже если в столбце содержатся даты, pandas будет воспринимать их как обычные строки. В данном случае необходимо преобразование столбца в DataTime, чтобы была возможность

работать с временными данными. Для этого использована функция to_datetime.

В pandas каждая строка данных имеет свой уникальный индекс (подобно номеру строки). Однако, иногда бывает удобнее использовать в качестве индекса указатель на конкретный столбец из ваших данных. В случае работы с временными рядами, наиболее удобным выбором для индекса является столбец, содержащий дату или время.

Это позволит легко выбирать и анализировать данные за определенные периоды времени. В рамках данной задачи выбран столбец Date. Использована функция set_index из той же библиотеки pandas.

Например, если количество пассажиров растет каждый год, это указывает на восходящий тренд. Если количество пассажиров растет в одни и те же месяца каждого года, это указывает на годовую сезонность. Для визуализации используется множество функций: figure(size) задаёт форму с указанным размером (в данном случае это прямоугольник), plot(arr) отображает сам ряд, остальные функции нужны для наглядности и указания подписей полей, значений (title, label).

Еще один важный шаг в предварительной обработке данных — это проверка необходимости изменения масштаба данных. Если размах ваших данных слишком велик (например, некоторое значение колеблется от тысяч до миллионов), может потребоваться преобразование данных. Как было сказано ранее, набор был преобразован таким образом, что значения варьировались от 0 до 1. Для этого задействован класс MinMaxScaler и его функция fit_transform.

Данный этап может включать в себя и другие действия, такие как обнаружение и обработка аномалий или выбросов, создание новых переменных или признаков, и разделение данных на подгруппы или категории.

3.2.2 Результаты

После обучения нейронной сети были получены следующие результаты (рисунок 3.4):

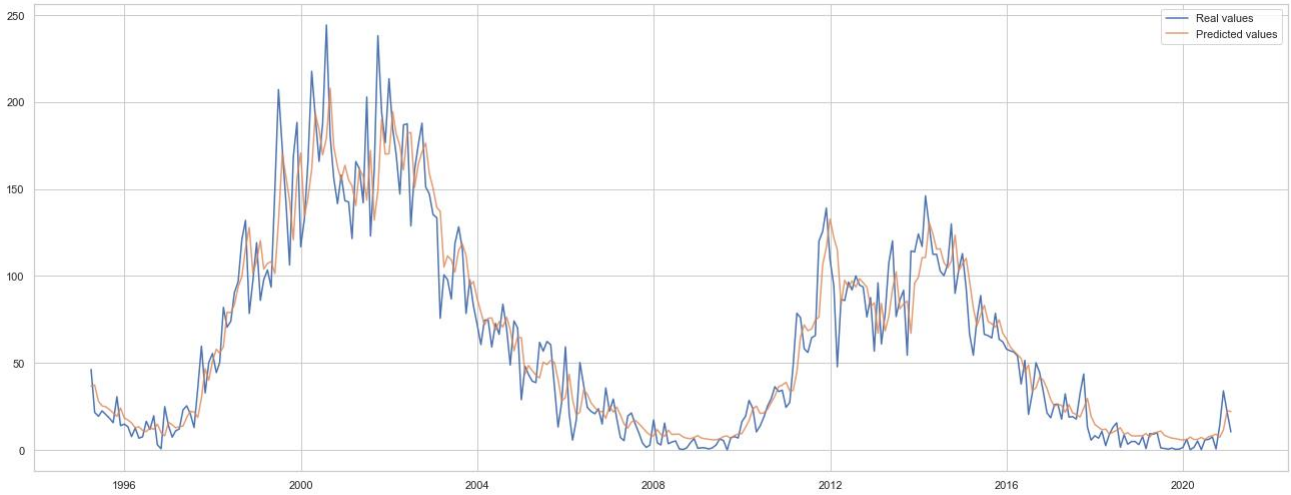

Рисунок 3.4 – Сравнение прогноза и реальных значений

Приведенный выше график сопоставляет прогнозируемые значения с фактическими. Модель эффективно фиксирует временные изменения, о чем свидетельствует близкое соответствие прогнозируемых и реальных точек данных.

Также для оценки результата используется метрика MAE (средняя абсолютная ошибка), которая измеряет среднюю сумму абсолютной разницы между фактическим и прогнозируемым значением:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |real\_value_t - predicted\_value_t| \quad (3)$$

Для данной модели средняя абсолютная разница между прогнозируемыми и реальными значениями составила 12.9912, что подтверждает эффективность применения выбранной модели.

3.3 Модель ARIMA

Модель ARIMA является широко используемым инструментом для прогноза и имеет несколько преимуществ для прогнозирования временных рядов:

1) Универсальность: ARIMA модель является гибкой и может быть применена к широкому спектру временных рядов. Она может обрабатывать данные с различными трендами, сезонностью и шумом.

2) Учет зависимостей во времени: ARIMA модель учитывает зависимости между текущим значением временного ряда и предыдущими

значениями. Она автоматически учитывает лаги и автокорреляцию в данных, что позволяет уловить тенденции и паттерны во временном ряду.

3) Интерпретируемость: ARIMA модель основана на простой и интерпретируемой математической структуре. Она состоит из трех основных компонентов: авторегрессии (AR), интегрирования (I) и скользящего среднего (MA). Каждая из этих компонентов имеет свое значение и может быть интерпретирована в контексте временного ряда.

4) Относительная простота использования: широко доступна в различных программных пакетах и языках программирования, таких как Python и R. Также существуют множество ресурсов и материалов для изучения и понимания ARIMA модели.

Однако стоит отметить, что ARIMA модель имеет и некоторые ограничения. Например, она предполагает стационарность данных и может быть менее эффективной в прогнозировании сложных временных рядов с нетривиальной структурой. Кроме того, ARIMA модель не учитывает внешние факторы или дополнительные переменные, которые могут влиять на временной ряд. В таких случаях могут быть предпочтительны другие модели, такие как ARIMA-X или SARIMA, которые учитывают внешние факторы или сезонность соответственно.

Исходный код на языке Python представлен в конце работы (ПРИЛОЖЕНИЕ Б).

3.3.1 Предварительный анализ

Для наглядности и прозрачности сравнения данных моделей в предварительном анализе проделан похожий порядок шагов в подготовке данных: проверка формата и пропущенных значений, установка индекса, наглядное отображение ряда.

Как было сказано ранее, данная модель требует стационарности ряда. В главе 1 был рассмотрен эффективный способ определения данного признака: тест расширенный тест Дики-Фуллера. На языке Python существует реализация данного метода в библиотеке statsmodels. Для определения стационарности важен такой параметр, как p-значение – минимальная величина уровня значимости, при которой нулевая гипотеза (предположение, что ряд нестационарен или имеет единичный корень) отвергается. Предположение, что ряд стационарен можно выдвинуть, если данный параметр принимает значение менее, чем 0.05, и тестовая статистика меньше любого из критических значений.

После выполнения теста были получены следующие результаты:

Таблица 3.1 – Результаты теста

p-значение	1.108552e-18
Статистика теста	-10.497052
Критическое значение (1%)	-3.432372402612
Критическое значение (5%)	-2.862433576091
Критическое значение (10%)	-2.567245669977

Исходя из полученной информации, можно сделать вывод, что ряд стационарен.

Следующий шаг подразумевает определение трёх параметров, используемых в модели: p - порядок авторегрессии, q - порядок скользящего среднего, d - порядок интегрирования. Поскольку ряд стационарен, то параметр d принимается равным 0.

Для определения параметров p и q изучаются графики автокорреляционной и частной автокорреляционной функций (рисунки 3.2 и 3.5 соответственно). В данном случае, значения данных параметров принимаются равными 2.

Для оптимального подбора параметров ARIMA-модели применяется метод grid search (поиск по сетке), который последовательно перебирает все возможные комбинации гиперпараметров из заданного диапазона значений, обучая модель для каждой комбинации. Этот исчерпывающий подход гарантирует нахождение наиболее подходящей конфигурации модели для конкретного временного ряда.

Соответствие этим критериям указывает на хорошую модель. Для проверки используется функция summary (), в которую в качестве параметра передаются спрогнозированные значения. Она выводит таблицей всю информацию по модели, по которой формируется вывод.

3.3.2 Результаты

Результат работы модели ARIMA приведён на рисунке 3.5.

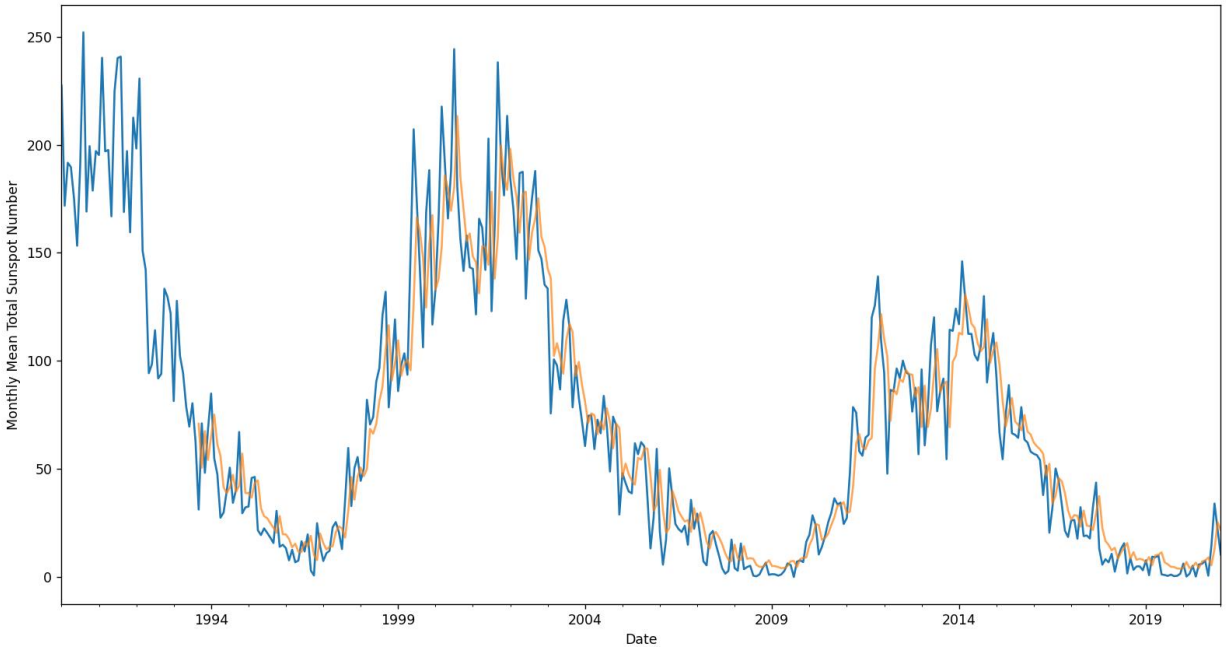

Рисунок 3.5 – Прогноз модели ARIMA

Оранжевым показан прогноз, синим – исходный временной ряд. Тут наглядно показано, что модель удачно улавливает тенденции роста и падения среднего числа солнечных пятен в разные периоды времени.

Исходя из того факта, что средняя абсолютная разница между прогнозируемыми и реальными значениями составила 13.3484, модель показала себя результативной, результат приближен к полученному LSTM моделью.

3.4 Модель экспоненциального сглаживания

Модель экспоненциального сглаживания – довольно мощный инструмент при прогнозировании значений временного ряда, к его преимуществам можно отнести:

- Простота и интерпретируемость модели позволяет быстро получить нужный результат (при моделировании учитываются только влияние предыдущих значений, тренд и сезонность);
- Не требует больших вычислительных ресурсов, в сравнении с нейросетевыми и авторегрессионными моделями (следует из первого пункта); работает даже на небольших временных рядах;
- Хорошая точность для краткосрочных прогнозов;

Однако никакая модель не идеальна, недостатки также весомы:

- Насколько удачно работает с краткосрочными зависимостями, настолько плохо работает с долгосрочными;

- Нестабильная сезонность (например, изменение сезонности ввиду кризиса) может привести к искажению результата;

Для более точного прогноза (с учётом тренда и/или сезонности) существуют более сложные структуры, такие как модели двойного и тройного экспоненциального сглаживания. Первый вариант учитывает только тренд, в то время как вторая опция включает в себя такой параметр, как сезонность. В таблице 4 представлена краткая сравнительная характеристика.

Таблица 3.2 – Сравнение по критериям с учётом стационарности

Критерий	ЭС	ДЭС	ТЭС
Учитывает тренд	нет	да	да
Учитывает сезонность	нет	нет	да
Точность прогноза	средняя	средняя	высокая
Примеры использования	Стабильные продажи	Рост ВВП	Прогноз температуры

Исходя из предварительного анализа рассмотренных ранее моделей, в котором благодаря корреляционному анализу удалось выявить зависимость текущего значения от 30 предыдущих (долгосрочная зависимость), данная модель не покажет себя должным образом. Учитывая полученные данные (наличие стационарности), применение более сложных по структуре моделей не имеет смысла (тренд отсутствует, сезонность крайне слабо влияет на последующие значения). Подтвердим данную гипотезу с помощью вычислительного эксперимента.

3.4.1 Предварительный анализ

Первоначально провёл те же манипуляции с набором данных вплоть до разбиения на тестовый и тренировочный наборы (проверка формата и пропущенных значений, установка индекса). В библиотеке statsmodels существует реализация данной модели (само устройство и формулы рассмотрены во второй главе), в неё можно передать следующие важны параметры:

- Непосредственно значения временного ряда (единственный обязательный параметр);

- Тип тренда: аддитивный (линейный рост или падение) или мультипликативный (изменение в %); ввиду стационарности ряда, тренд отсутствует;
- Тип сезонности;
- Коэффициент сглаживания: по умолчанию для определения используется метод максимального правдоподобия, что является лучшим выбором в большинстве случаев;

В реализации воспользуемся реализованным методом определения коэффициента сглаживания. Метод максимального правдоподобия работает следующим образом: строится предположение о том, как распределены ошибки прогноза (разность между реальным и спрогнозированным значением) – самой распространённой практикой является предположение о нормальном распределении; цель метода – максимизировать логарифм правдоподобия по параметрам: используется градиентный спуск; ищется такой коэффициент, при котором сумма квадратов ошибок минимальна, что эквивалентно методу наименьших квадратов, если ошибки нормальны. Преимущества такого подхода:

- Учитывает вероятностную природу модели;
- Позволяет добавить другие параметры (тренд, сезонность);
- Гибкость: работает для сложных моделей (например, с затухающим трендом);
- Даёт обоснованные оценки и доверительные интервалы;

Корректность метода можно проверить с помощью оценки остатков на нормальность (тест Шапиро-Уилка).

3.4.2 Результаты

Результат работы модели экспоненциального сглаживания приведён на рисунке 3.6. Оценка качества работы модели также определена с помощью формулы MAE.

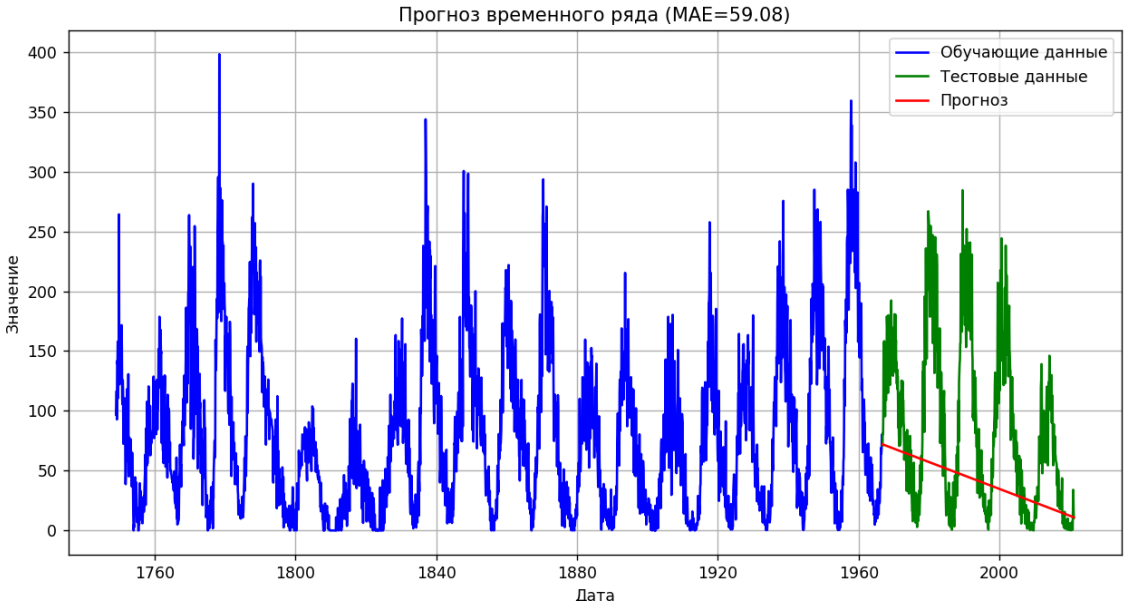

Рисунок 3.6 – Прогноз модели экспоненциального сглаживания

Синим выделена обучающая часть данных, зелёным – тесовая, красным – прогноз. Только из визуальной составляющей сделанная ранее гипотеза о том, что данная модель слабо проявит себя ввиду долгосрочной зависимости. Двойное и тройное сглаживание также не дали бы результата. Статистически также имеется подтверждение: средняя абсолютная разница между прогнозируемыми и реальными значениями составила всего 59.08, что явно хуже полученных ранее результатов для ARIMA и LSTM.

3.5 Выводы по третьей главе

Модель LSTM (с долгой краткосрочной памятью) выдала приличный результат, основываясь на такой метрике, как средняя абсолютная ошибка. Такой высокий показатель качества прогноза был достигнут с помощью грамотного выбора временного шага, т.е. предварительного анализа, а также верного выбора значений гаперпараметров вроде числа слоёв, количества эпох обучения, размера мини-выборки и т.д.

Модель ARIMA, в свою очередь, показала себя не хуже, чем предыдущая. Для достижения данного итога немалую роль сыграл тот факт, что временной ряд оказался стационарным. Кроме того, подбор параметров p, d, и q основан на анализе графиков и нельзя точно сказать, что их выбор является оптимальным. Для этого существуют более точные технологии и средства (например, проход через все возможные комбинации параметров из заранее определенной сетки значений и обучение на каждой из этих комбинаций).

Обобщая вышесказанное, LSTM и ARIMA модели показали себя как эффективные модели прогнозирования, поскольку полученные значения были приближены к реальным. Однако можно добиться результата лучше, экспериментируя с различными конфигурациями архитектуры, гиперпараметрами или включая внешние источники данных, относящиеся к активности солнечных пятен. Не стоит забывать, что существует огромное количество методов прогнозирования временных рядов, каждый из которых может выдать другой по точности результат. Конкретно сравнивая данные рассмотренные модели, то каждая имеет свои преимущества и недостатки, поэтому их применение зависит от внешних условий и входных данных.

ЗАКЛЮЧЕНИЕ

Анализ и прогнозирование временных рядов является важным инструментом для многих областей деятельности, таких как экономика, финансы, маркетинг, производство и т.д. Он позволяет предсказывать будущие значения временных рядов на основе анализа прошлых данных, что помогает принимать правильные решения и планировать действия в будущем.

Были выполнены следующие задачи:

- Проведен обзор на научные статьи, касающиеся темы анализа и прогнозирования временных рядов, рассмотрены основные методы и классы моделей, среди методов анализа выделены наиболее эффективные и информативные (к ним относятся анализ стационарности, спектральный и корреляционный анализ, рассмотрен широко распространённый на практике метод анализа – тест Грейнджера, рассмотренный только с теоретической точки зрения, поскольку требует наличие второго временного ряда со связанными данными);
- Проведена предобработка исходных данных, формирование тренировочных и тестовых наборов; рассмотрены различные библиотеки на языке программирования Python: pandas, для обработки и нормализации данных, matplotlib, для визуализации исходного временного ряда и результатов работы моделей прогнозирования, statsmodels, для корреляционного анализа (отображения графика), анализа стационарности с помощью теста Дики-Фуллера, а также для задействования модели ARIMA, TensorFlow, основная библиотека, в которой находится весь функционал модели LSTM; также рассмотрена модель экспоненциального сглаживания, на примере которой показана важность выбора модели при прогнозировании и анализа (в данном случае стационарность, а значит отсутствие тренда, неопределённая сезонность, долгосрочная зависимость, определённая с помощью короллелограммы). Можно сделать вывод, что недостатки стоит учитывать также, как и преимущества. Для других моделей в качестве минусов были упомянуты зависимости от определения стационарности и параметров у ARIMA (и класса регрессионных и авторегрессионных моделей в целом), затраченное время на обучение, громоздкость структуры нейронной сети, а значит и возросшая вычислительная сложность у LSTM (как и у класса всех нейросетевых подходов к прогнозированию).
- Произведено обучение нейронных сетей и выполнен прогноз солнечной активности;

На данный момент существует большое количество самых разнообразных подходов к анализу и прогнозированию временных рядов,

улучшению предсказательной способности, и данная работа является хорошей основой для дальнейших исследований.

СПИСОК ИСПОЛЬЗОВАНЫХ ИСТОЧНИКОВ

1) Бокс, Дж. Анализ временных рядов прогноз и управление. Выпуск 1 / Дж. Бокс, Г. Дженкинс. - М.: Мир, 1974. - 408 с.

2) Иванов, А.А. Применение ARIMA и SARIMA для прогнозирования финансовых временных рядов / А.А. Иванов // Экономика и математические методы. – 2020. – Т. 56, № 3. – С. 45–58.

3) Лукьяненко, А.В. Прогнозирование экономических временных рядов: методы и модели / А.В. Лукьяненко, И.Г. Персианов. – М.: Финансы и статистика, 2018. – 304 с.

4) Официальный сайт Python-библиотеки statsmodels [Электронный ресурс]. – Режим доступа: https://www.statsmodels.org/stable/index.html – Дата доступа: 15.05.2025.

5) Садовникова, Н. А. Анализ временных рядов и прогнозирование / Н.А. Садовникова, Р.А. Шмойлова. – М.: Изд. центр ЕАОИ, 2016. - 152 с.

6) A Complete Tutorial on Time Series Modeling [Электронный ресурс]. – Режим доступа: https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/. – Дата доступа: 10.05.2025.

7) Cryer, J. Time Series Analysis / J. Cryer, K. Chan. – New York: Springer, 2008. - 491 p.

8) Forecasting: Principles and Practice [Электронный ресурс]. – Режим доступа: https://otexts.com/fpp2/. – Дата доступа: 11.04.2025.

9) Introduction to time series analysis [Электронный ресурс]. – Режим доступа: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm. – Дата доступа: 20.04.2025.

10) LSTM — нейронная сеть с долгой краткосрочной памятью [Электронный ресурс]. – Режим доступа: https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/. – Дата доступа: 11.05.2025.

ПРИЛОЖЕНИЕ А

Код программы для LSTM модели

```
def adf_test(series, max_lag=None):
    n = len(series)
    if n <= 2:
        raise ValueError("Series must have more than 2 observations.")


    if max_lag is None:
        max_lag = int(n ** 0.5)
    lags = range(1, max_lag + 1)
    results = []
    for lag in lags:
        X = []
        Y = []
        for i in range(lag, n):
            X.append([1, series[i - lag]])  # Константа и лаг
            Y.append(series[i])  # Текущая наблюдаемая величина
        beta = estimate_coefficients(X, Y)
        residuals = calculate_residuals(X, Y, beta)
        tau_statistic = calculate_tau_statistic(residuals, lag)
        results.append((lag, tau_statistic))
    return results
def estimate_coefficients(X, Y):
    import numpy as np
    X = np.array(X)
    Y = np.array(Y)
```

```
    beta = np.linalg.inv(X.T @ X) @ (X.T @ Y)
    return beta
def calculate_residuals(X, Y, beta):
    import numpy as np
    X = np.array(X)
    Y = np.array(Y)
    predictions = X @ beta
    residuals = Y - predictions
    return residuals
def calculate_tau_statistic(residuals, lag):
    n = len(residuals)
    residual_variance = sum(residuals**2) / n
    tau_statistic = (sum(residuals) / n) / (residual_variance / lag)
    return tau_statistic
df = pd.read_csv("Sunspots.csv")
df.drop(columns="Unnamed: 0", inplace=True)
df["Date"] = pd.to_datetime(df["Date"])
df.set_index(df["Date"], inplace=True)
df.drop(columns="Date", inplace=True)
df.head()
plt.figure(figsize=(17,8))
plt.plot(df.index, df["Monthly Mean Total Sunspot Number"])
plt.xlabel("Year")
plt.ylabel("Number")
plt.title("Monthly Mean Total Sunspot Number")
fig, ax = plt.subplots(figsize=(15, 6))
plot_acf(df["Monthly Mean Total Sunspot Number"], ax=ax)
def create_dataset(X, y, time_steps=1):
```

```
    Xs, ys = [], []
    for i in range(len(X) - time_steps):
        v = X[i:(i + time_steps), :]
        Xs.append(v)
        ys.append(y[i + time_steps])
    return np.array(Xs), np.array(ys)
split = int(len(df) * 0.9)
stc = MinMaxScaler()
df_scaled = stc.fit_transform(df)
df_train = df_scaled[:split]
df_test = df_scaled[split:]
time_steps = 30
X_train, y_train = create_dataset(df_train, df_train, time_steps)
X_test, y_test = create_dataset(df_test, df_test, time_steps)
model = Sequential()
model.add(LSTM(64, return_sequences=True, input_shape=(time_steps, 1)))
model.add(LSTM(64, return_sequences=True))
model.add(LSTM(32))
model.add(Dense(32, activation="relu"))
model.add(Dense(1))
model.compile(optimizer=Nadam(), loss='mse', metrics="mae")
history = model.fit(
    X_train, y_train,
    validation_data=(X_test, y_test),
    epochs=50,
    batch_size=32,
    verbose=0,
)
y_pred = model.predict(X_test)
```

```
y_pred = [x[0] for x in y_pred]
y_pred = [x[0] for x in stc.inverse_transform(pd.DataFrame(y_pred))]
y_test_real = [x[0] for x in stc.inverse_transform(pd.DataFrame(y_test))]
print("The MAE is: ", mean_absolute_error(y_test_real, y_pred))
plt.figure(figsize=(20,8))
plt.plot(df.index[split+time_steps:], y_test_real, label="Real values")
plt.plot(df.index[split+time_steps:], y_pred, label="Predicted values", alpha=0.8)
plt.show()
```

Код программы для ARIMA модели

```
def ADF_test(timeseries):
    print('Results of Dickey-Fuller Test:')
    dftest = adfuller(timeseries,autolag='AIC')
    dfoutput = pd.Series(dftest[0:4],index=['Test Statistic','p-value','#Lags Used','NUmber of Observations Used'])
    for key,value in dftest[4].items():
        dfoutput['Critical Value(%s)'%key]="{:.12f}".format(value)
    print(dfoutput)


data = pd.read_csv('Sunspots.csv')
data['Date'] = pd.to_datetime(data['Date'])
data.set_index('Date', inplace=True)
y = data['Monthly Mean Total Sunspot Number']
print(ADF_test(y))


plot_acf(y)
plt.show()
plot_pacf(y)
plt.show()


y_to_train = y[:'1993-08-31']
y_to_test = y['1993-09-30':]
predict_date = len(y)-len(y[:'1993-09-30'])
mod = ARIMA(y, order=(2, 0, 2))
results = mod.fit()
print(results.summary())
```

```
pred = results.get_prediction(start = pd.to_datetime('1993-09-30'), dynamic = False)
pred_array = pred.predicted_mean.values
plt.figure(figsize=(20,8))
pred_ci = pred.conf_int()
ax = y['1990':].plot(label='observed')
pred.predicted_mean.plot(ax=ax, label='One-step ahead forecast', alpha=.7)
print("The MAE is: ", mean_absolute_error(y_to_test, pred_array))
ax.set_xlabel('Date')
ax.set_ylabel('Monthly Mean Total Sunspot Number')
plt.show()
```

Код программы для модели экспоненциального сглаживания

```
data = pd.read_csv('Sunspots.csv')
data['Date'] = pd.to_datetime(data['Date'])
data.set_index('Date', inplace=True)
y = data['Monthly Mean Total Sunspot Number']

plot_acf(y)
plt.show()
plot_pacf(y)
plt.show()

train_size = int(len(data) * 0.8)
train, test = data.iloc[:train_size], data.iloc[train_size:]



model = ExponentialSmoothing(
    train['Monthly Mean Total Sunspot Number'],
    trend='add',

)
fitted_model = model.fit()

forecast = fitted_model.forecast(steps=len(test))

mae = mean_absolute_error(test['Monthly Mean Total Sunspot Number'], forecast)
print(f'MAE: {mae:.2f}')
```

```
plt.figure(figsize=(12, 6))

plt.plot(train.index, train['Monthly Mean Total Sunspot Number'], label='Обучающие данные', color='blue')

plt.plot(test.index, test['Monthly Mean Total Sunspot Number'], label='Тестовые данные', color='green')

plt.plot(test.index, forecast, label='Прогноз', color='red')

plt.title(f'Прогноз временного ряда (MAE={mae:.2f})')

plt.xlabel('Дата')

plt.ylabel('Значение')

plt.legend()

plt.grid()

plt.show()
```