

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

ДОПУСТИТЬ К ЗАЩИТЕ

Заведующий кафедрой №52

д.т.н., профессор

должность, уч. степень, звание

А.М. Тюрликов

подпись, дата

инициалы, фамилия

БАКАЛАВРСКАЯ РАБОТА

на тему Прогнозирования трафика инфокоммуникационных сетей

выполнена

Урусовым Кириллом Андреевичем

фамилия, имя, отчество студента в творительном падеже

по направлению подготовки

11.03.02

Инфокоммуникационные технологии

код

наименование направления

и системы связи

наименование направления

направленности

03

Программно-защищенные

код

наименование направленности

инфокоммуникации

наименование направленности

Студент группы №

5823

К.А. Урусов

подпись, дата

инициалы, фамилия

Руководитель

д.т.н., профессор 52 кафедры

должность, уч. степень, звание

Т.М. Татарникова

подпись, дата

инициалы, фамилия

Санкт-Петербург 2022

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

УТВЕРЖДАЮ

Заведующий кафедрой №52

д.т.н., профессор

А.М. Тюрликов

должность, уч. степень, звание

подпись, дата

инициалы, фамилия

ЗАДАНИЕ НА ВЫПОЛНЕНИЕ БАКАЛАВРСКОЙ РАБОТЫ

студенту группы

5823

Урусову Кириллу Андреевичу

номер

фамилия, имя, отчество

на тему

Прогнозирования трафика инфокоммуникационных сетей

утвержденную приказом ГУАП от

31.03.2022

№ 11-352/22

Цель работы:

Задачи, подлежащие решению:

Содержание работы (основные разделы):

Срок сдачи работы « 02 » июня 2022

Руководитель

д.т.н., профессор 52 кафедры

Т.М. Татарникова

должность, уч. степень, звание

подпись, дата

инициалы, фамилия

Задание принял(а) к исполнению

студент группы №

5823

К.А. Урусов

подпись, дата

инициалы, фамилия

Оглавление

Введение	4
1. Методов прогнозирования, использующихся для анализа трафика	6
1.1. Анализ существующих моделей прогнозирования	6
1.2. Описание выбранной модели прогнозирования	8
1.3. Оценка качества прогнозирования	13
2. Этапы подбора данных для анализа	14
2.1. Оценка качества данных	14
2.2. Предобработка данных	16
2.3. Представление данных	18
3. Выбор инструментов для построения модели	19
4. Примеры работы модели прогнозирования трафика инфокоммуникационных сетей	20
4.1. Получение данных для экспериментов	20
4.2. Прогнозирование на основе дампа трафика за 24 часа	20
4.3. Прогнозирование на основе дампа трафика за 96 часов	29
4.4. Прогнозирование на основе дампа трафика за 48 часов	36
Заключение	44
Библиографический список	47

Введение

За последнее десятилетие произошло сильное изменение объемов сетевого трафика. С каждым годом и к интернету присоединяются новые смартфоны, планшеты, устройства интернет вещей. Так же доступность видеохостингов и стриминговых сервисов значительно увеличивает количество передаваемого трафика. По данным сайта datareportal.com число интернет-пользователей увеличивается в среднем на 8,2% в год. По сравнению с этим, население мира увеличивается в среднем всего на 1,05% в год.

Такой непрерывный рост трафика приводит к возникновению перегрузке каналов передачи данных. Перегрузки вызывают снижению качества звука и изображения, увеличивается количество потерянных пакетов данных. Постоянные инвестиции денежных средств в новое оборудование и коммуникации, с целью увеличить пропускную способность не всегда могут обеспечить необходимый запас производительности для качественной работы приложения.

В 2020 году один из крупнейших стриминговых сервисов в мире Netflix выложил отчетность, в которой указал на рост пользователей на 16 млн. в связи с пандемией Covid-19. Такой рост пользователей, а следовательно и трафика, вынудил компанию снизить максимально разрешенное качество трансляций для жителей Европы, чтобы не перегружать инфокоммуникационные сети.

Из всего вышесказанного можно сделать вывод, что основной задачей сетевых компаний становится поддержка сбалансированной нагрузки на сеть. Снижение пропускной способности на одном из промежуточных узлов приводит к ухудшению общей работоспособности сети в том числе потере пакетов, перегрузкой буферов коммутаторов. Решение проблемы перегрузок на сетевом оборудовании за счет управления трафиком позволит сократить расходы на обслуживание существующих сегментов сети.

Управление сетевым трафиком осуществляется конфигурацией коммутационного оборудования с помощью различных политик, фильтров и

ограничения максимальной нагрузки. Избыток трафика, который может вызвать перегрузку можно обработать тремя способами:

- Заблокировать, то есть удалить пакет, что обычно приводит к повторной передаче данного пакета, а значит только усугубляет ситуацию;
- Временно запретить источнику генерировать пакеты определенного типа;
- Доставлять пакеты адресату с худшим показателем качества, за большее время или с большей долей потерей пакетов;

Третий вариант выглядит предпочтительней, так как он позволяет динамически управлять трафиком и ограничивать поток данных только в кратковременные промежутки времени. Это более эффективный вариант, но он требует преждевременного решения о таком воздействии, так как этот процесс требует времени для осуществления ограничений.

Таким образом, решению задачи по управлению трафика инфокоммуникационных сетей могут способствовать данные, полученные с помощью прогноза по загрузке линий связи. Для прогноза используются статистические данные, которые собираются в процессе эксплуатации коммуникационного оборудования. Хорошо подобранная модель способна с заданной точностью прогнозировать трафик как в краткосрочном, так и в долгосрочном промежутке времени.

Целью выпускной квалификационной работы является разработка модели прогнозирования, которая позволит на основании имеющихся данных спрогнозировать объем трафика в будущем.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Произвести анализ существующих методов прогнозирования, использующихся для анализа трафика.
2. Осуществить подборку данных для дальнейшего анализа.
3. Разработать модель прогнозирования трафика, которая позволит произвести все необходимые этапы анализа данных.
4. Проверить работоспособность модели на известных данных.

5. Произвести анализ полученных результатов прогнозирования.

Результатом выпускной квалификационной работы является разработка программы, основанной на статистических методах анализа данных, которая позволит:

- Обработать данные трафика для приведения их к систематическому виду.
- Построить модель прогноза трафика инфокоммуникационных сетей.
- Оценить качество прогнозирования полученной модели.

Полученные на выходе алгоритма данные будут наглядно представлены в отчетах с помощью средств визуализации, которые позволят оценить точность модели и эффективность ее использования.

1. Методов прогнозирования, использующихся для анализа трафика

1.1. Анализ существующих моделей прогнозирования

На протяжении многих лет люди занимаются прогнозами экономически и политических событий, погодных условий, спортивных результатов. Раньше для этого использовались интуиция, экспертные мнения, использование прошлых результатов для сравнения с традиционной статистикой. Самым современным методом является прогнозирование временных рядов. Такой способ имеет обширную область применения. В данной работе мы будем рассматривать этот метод для прогнозирования трафика инфокоммуникационных сетей.

Временной ряд - собранный в разные моменты времени статистический материал о значении каких-либо параметров исследуемого процесса. Каждая единица статистического материала называется измерением или отсчётом, также допустимо называть его уровнем на указанный с ним момент времени. Во временном ряде каждому отчету должно быть указано время измерения или номер измерения по порядку. Временной ряд существенно отличается от простой выборки данных, так как при анализе учитывается взаимосвязь

измерений со временем, а не только статистическое разнообразие и статистические характеристики выборки.

Стационарный временной ряд - такой ряд, который не имеет циклической компоненты, каждый следующий уровень равен сумме среднего ряда и случайной компоненты.

Модели временных рядов - математические модели прогнозирования, которые стремятся обнаружить зависимость будущих значений от прошлого и на этой зависимости вычислить прогноз. Такие модели универсальны для любого типа временных рядов, так как их общий вид не изменяется для различных предметных областей. Мы можем использовать нейронные сети для прогнозирования стоимости акций, а затем применить эту же модель для прогнозирования температуры воздуха.

Существует два основных вида моделей прогнозирования: статистические и структурные. Для статистических моделей функциональная зависимость задается аналитически. К таким моделям относятся: регрессионные, авторегрессионные и экспоненциального сглаживания.

Структурные модели основаны на зависимости структур. К ним относятся нейронные сети, модели на базе цепей Маркова и деревья принятия решений. В таблице ниже приведены достоинства и недостатки выше перечисленных методов.

Таблица 1 - Модели и методы прогнозирования

Методы и модели	Достоинства	Недостатки
Регрессионные	Простота в моделировании и проектировании	Сложность нахождения оптимальных коэффициентов и функциональной зависимости

Методы и модели	Достоинства	Недостатки
Авторегрессионные	Простота в моделировании и анализа	Нельзя моделировать нелинейные процессы
Экспоненциальное сглаживание	Простота моделирования	Узкая применимость моделей
Нейросетевые	Большое разнообразие архитектур; нелинейность	Сложность в выборе архитектур; размер обучающей выборки; большие временные и ресурсоемкие затраты на обучение
Цепи Маркова	Единообразие проектирования	Узкая применимость моделей; нельзя моделировать долгосрочные процессы
Классификационно-регрессионные деревья	Простота обучения модели; возможность масштабирования	Сложность построения алгоритма дерева

1.2. Описание выбранной модели прогнозирования

Наиболее популярной моделью прогнозирования являются модели авторегрессии и проинтегрированного скользящего среднего ARIMA(Autoregressive Integrated Moving Average). Это - важный класс параметрических моделей позволяет описывать нестационарные ряды.

ARIMA использует три основных параметра, которые выражаются целыми числами: p , d , q . Поэтому модель принято записывать как ARIMA(p , d , q). Все эти параметры учитывают разные характеристик прогнозируемого временного ряда:

- p - порядок авторегрессии AR, который позволяет добавить предыдущие значения временного ряда;
- d - порядок интегрирования, он вычисляется как количество использования оператора взятия разности над временным рядом;
- q - порядок скользящего среднего MA. Позволяет установить погрешность модели как линейную комбинацию наблюдавшихся ранее значений ошибок.

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

$$(\Delta^d X_t) = \sum_{i=1}^p \phi_i(\Delta^d X_{t-i}) + \epsilon_t + \sum_{j=1}^q \theta_j(\Delta^d \epsilon_{t-j}), \epsilon \sim N(0, \sigma_t^2)$$

где Δ^d - оператор разности временного ряда порядка d , X_t - стационарный временной ряд, p - порядок компонента авторегрессия, MA, q - порядок компонента скользящего среднего, а d - порядок интегрированного временного ряда, ϕ_t, θ_t - коэффициенты, выражающие влияние значений на прогнозируемое.

Алгоритм построения модели приведен на рисунке 1:

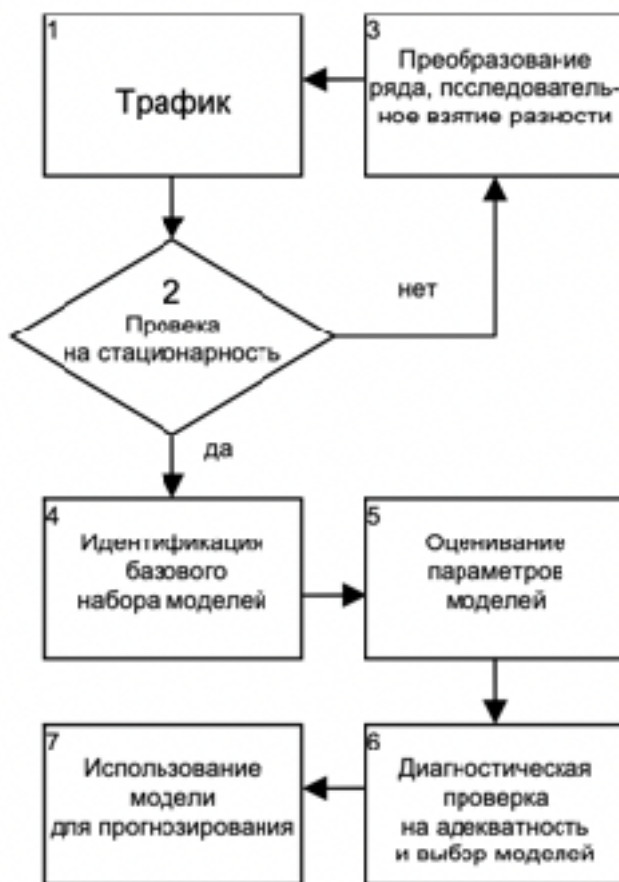


Рисунок 1 - Структурная схема подбора модели ARIMA

Первый систематический подход к построению и использованию статистической модели ARIAM был опубликован Боксом и Дженкинсом в 1976 году. Построение модели ARIMA для исследования и прогнозирования временного ряда включает в себя следующие основные этапы:

- идентификацию пробной модели;
- оценивание параметров модели и диагностическую проверку адекватности модели;
- использование модели для прогнозирования.

Далее мы подробно разберем каждый из этапов построения и использования модели ARIMA. Сначала в блоке 1-3 происходит обработка изучаемых данных. Из них нам необходимо построить временной ряд и привести его к стационарному виде. Для следует проводить анализ авто корреляционной функции(АКФ) и частной автокорреляционной функции(ЧАКФ). Быстрое затухание значений АКФ говорит нам о том, что ряде не является стационарным. Для более точной проверки ряда на стационарность мы применяем к нему тест статистической значимости Дики-Фуллера(ADF-тест). В его основе лежит нулевая гипотеза о том, что ряд не является стационарным.

Если после проведения теста Дики-Фуллера мы получили р-значение больше порога статистической значимости, равного 0.05, то мы делаем вывод, что нулевая гипотеза не была отвергнута. Следовательно исследуемый временной ряд не является стационарным. Для перехода от не стационарного вида к стационарному принято использовать оператор взятия последовательных разностей, тем самым мы определяем значение параметра d , который необходим для построения статистической модели прогнозирования $ARIMA(p, d, q)$.

В блоке 4 после получения стационарного временного ряда мы исследуем значения, полученные после проведения АКФ и ЧАКФ тестов, и выдвигаем гипотезы о значениях параметров p и q , где p - это порядок авторегрессии, а q - порядок скользящего среднего. На выходе 4 блока мы можем сформировать базовый набор, включающий в себя одну, две или более моделей для прогнозирования.

В блоке 5 после подбора параметров для модели происходит оценка этих параметров. Для этого этих целей используется метод максимального правдоподобия.

В блоке 6 для проверки модели на адекватность мы используем для обучения часть исследуемой выборки. После этого мы делаем наш прогноз и вычисляем квадратичную и среднюю абсолютную ошибки. Так же мы анализируем ее ряд остатков. У адекватной модели ряд остатков должен быть похож на белый шум. Для анализа ряда остатков используется Q-тест Льюнг-Бокса.

Если после проведения тестов в блоке 6 мы получили значения, которые нас устраивают, т. е. небольшое значение погрешности, а ряд остатков похож на белый шум, то мы переходим к блоку 7.

В блоке 7 мы получили модель прогнозирования, с помощью которой можем строить точный прогноз и интервальный прогноз на L шагов вперед.

Для сравнения эффективности подбора параметров описанным выше методом мы будем подбирать параметры для второй модели методом полного перебора. Параметры подбирались по информационному критерию Акаике (AIC — Akaike Information Criterion). AIC оценивает, насколько хорошо модель подходит под данные. Чем меньше AIC — тем точнее модель. Такой подход отнимает много времени для построения модели и ее анализа.

Для анализа модели, параметры которой были получены методом полного перебора мы так же вычисляем квадратичную и среднюю абсолютную процентную ошибки. Это необходимый шаг для сравнения разных способов подбора параметров для статистической модели прогнозирования.

Если обе полученные модели являются адекватными, т. е. их ряды остатков похожи на белый шум, а величины ошибок нас устраивают, то мы можем использовать обе эти модели для прогнозирования. В качестве итогового прогноза мы усредняем результаты обеих моделей. Таким образом мы уменьшаем величину погрешности, а значит увеличиваем точность прогнозирования трафика инфокоммуникационных сетей.

Целью данной выпускной квалификационной работы является сравнить два различных метода подбора параметров для модели, построение статистической модели прогнозирования временных рядов ARIMA, адекватно отражающей поведение сетевого трафика, на основе которой можно совершать достоверные краткосрочные прогнозы.

1.3. Оценка качества прогнозирования

Важным этапом работы с статистической моделью прогнозирования является оценка качества полученного прогноза. Такая оценка позволяет оценить приблизительную погрешность полученной модели, а так же проверить, что для прогноза используются все имеющиеся данные.

Для оценки погрешности вычисляется квадратичная и средняя абсолютная процентная ошибки. Значения этих ошибок показывают точность полученных моделей.

Под квадратичной ошибкой в данном случае понимается разность между фактическим значением выходной переменной и значением, оцененным моделью на данном наблюдении. Переход к квадратичной функции позволяет исключить отрицательные значения ошибки. Она имеет следующий вид:

$$E(w) = \frac{1}{n} \sum_{i=0}^n (d_i - y_i)^2,$$

где n - это количество элементов, d_i - фактическое i -ое значение, y_i - прогнозированный i -ый элемент.

Средняя абсолютная процентная ошибка является мерой оценки точности предсказания метода прогнозирования в статистике. Она отражает погрешность полученной модели прогнозирования и выражается следующей формулой:

$$MAPE = \frac{100}{n} \sum_{i=0}^n \left| \frac{(d_i - y_i)}{d_i} \right|,$$

где n - это количество элементов, d_i - фактическое i -ое значение, y_i - прогнозированный i -ый элемент.

Для оценки ряда остатков полученной модели прогнозирования используется Q-тест Льюнг-Бокса. Это тест представляет собой статистический критерий, предназначенный для нахождения автокорреляции временных рядов. Он проверяет на отличие от нуля сразу несколько коэффициентов автокорреляции. В качестве нулевой гипотезы используется утверждение, что данные являются случайными, т. е. представляют собой белый шум. Q-тест задается следующей формулой:

$$Q = n(n + 2) \sum_{i=1}^m \frac{\hat{p}_k}{n - k},$$

где n - число наблюдений, \hat{p}_k - корреляция k -го порядка, m - количество рассматриваемых лагов.

2. Этапы подбора данных для анализа

2.1. Оценка качества данных

Качество данных имеет важную роли при анализе и прогнозирование. Вбросы и искажения являются причиной появления ложных закономерностей и тенденций, что приводит к ошибке в прогнозирование.

По этой причине для исключения таких факторов проверка данных проводится на всех этапах работы с этими данными. Чтобы правильно подготовить данные к анализу выполняется оценка их качества.

Оценка качества данных - необходимый этап процесса подготовки данных к анализу, так как именно на этом этапе происходит выявление наиболее значимых проблем, которые оказывают влияние на результат нашего анализа и снижают его значимость.

Существует достаточно большой список проблем, в той или иной степени влияющих на качество анализируемых данных. Чтобы систематизировать

данные проблемы, а также определить степень их влияния и пригодности для анализа, выделяют три уровня качества данных, сравнительная характеристика которых приведена в таблице 2.

Таблица 2 - Уровни качества данных

Уровень	Факторы	Проявления
Технический	<ul style="list-style-type: none"> - нарушение в структуре данных; - некорректное наименование таблиц и полей; - некорректные форматы и кодировки; - нарушение полноты и целостности данных. 	Мешают выполнению консолидации и интегрированию данных, применению к ним алгоритмов обработки
Аналитический	<ul style="list-style-type: none"> - пропуски; - аномальные значения; - противоречия; - дубликаты 	Снижают достоверность данных и искажают результаты их анализа, мешают обнаружению закономерностей и тенденций
Концептуальный	Собранные и консолидированные данные в недостаточной мере отражают исследуемый процесс	Отсутствие или недостаток данных для анализа

Выбросы, пропуски и структурные нарушения можно обнаружить используя табличное представление данных. Но такой способ не эффективен,

так как при большом объеме информации мы можем легко упустить из виду пропуски, а также не заметить выбросы в нашей статистике.

Так же для выявления проблем в данных используются графики и диаграммы. Если данные содержат шумы и выбросы, то мы сможем наглядно увидеть их на графиках, так как они будут отображаться в виде резких скачков и отклонений.

Оценка качества данных является значимым этапом в подготовке данных к дальнейшему статистическому анализу, так как именно на них основаны результаты нашего анализа и значимость полученных результатов.

2.2. Предобработка данных

На этом этапе мы приводим наши данные к определенному виду, в соответствие с спецификацией нашей задачи. Если поступающая на вход модуля информация не будет обработана должным образом, то дальнейший анализ может быть бесполезен.

Предобработка данных может включать в себя два направления: очистку и оптимизацию.

Очистка данных - процесс, который исключает большинство факторов, которые снижают качество данных и мешающие дальнейшей корректной работе аналитической модели. Данный процесс включает в себя обработку дубликатов, фиктивных значений, выбросов, заполнение пропусков.

Оптимизация данных же обеспечивает снижение размерности, выявление и исключение незначущих признаков, что адаптирует данные для конкретной задачи и повышает эффективность анализа.

Главное отличие очистки и оптимизации данных заключается в том, что очистка исключает факторы, которые существенно снижают точность решения задачи, в отдельных случаях делают работу аналитического алгоритма невозможным.

Далее мы рассмотрим только восстановление пропущенных значений, так как именно это в полученной выборке является значительным фактором, влияющим на значимость нашего анализа.

Сами пропуски не влияют на информативность данных, но к этому может привести некорректное применение методов заполнения пропусков. Такие методы могут привести к появлению дубликатов, противоречия и аномалии.

Тем не менее, восстановления пропущенных значений необходимо, так как пропуски вызывают неопределенность в работе алгоритмов статистического анализа.

Разберем существующие методы восстановления пропущенных значений:

- Ручная обработка - заключается в заполнении пропусков наиболее вероятными значениями, основанными на личных знаниях и опыте;
- Подстановка констант - вместо пропущенных значений подставляется некоторое фиксированное значение;
- Подстановка среднего значения - замена пропусков средним по ряду данных или по нескольким его ближайшим соседям;
- Подстановка моды - вместо пропущенного значения вставляется наиболее популярное в ряду.

Довольно часто в больших наборах данных встречаются выбросы, т. е. значения которые сильно отличаются от окружающих данных. Такие выбросы можно разделить на искусственные, вызваны ошибками измерений данных, а так же на естественные - вызванные обстоятельствами, которые встречаются редко или в единичных случаях.

Эффективным способом обнаружения таких выбросов является визуализация данных с помощью графиков и таблиц.

Обнаруженные выбросы корректируются следующими способами:

- Удаление выброса;
- Ручная замена значения, если выбросов в выборке мало;
- Интерполяция, значения заменяются другими, на основании соседних;
- Замена на вероятное значение.

2.3. Представление данных

Для хранения данных были выбраны CSV таблицы. Такие таблицы представляют из себя текстовый файл, в котором хранится информация для каждой ячейки разделенная запятой. На рисунке 1 приведен пример таблицы, используемой для прогнозирования в данной модели.

```
date,data
2009-3-30-0:0,11682.93
2009-3-30-0:15,11697.34
2009-3-30-0:30,11709.92
2009-3-30-0:45,10388.29
2009-3-30-1:0,10445.75
2009-3-30-1:15,10540.36
2009-3-30-1:30,9940.78
2009-3-30-1:45,9919.54
2009-3-30-2:0,9848.27
2009-3-30-2:15,9808.58
2009-3-30-2:30,9664.25
2009-3-30-2:45,9517.67
2009-3-30-3:0,8820.23
2009-3-30-3:15,9487.88
2009-3-30-3:30,9241.88
2009-3-30-3:45,9195.72
2009-3-30-4:0,8610.56
2009-3-30-4:15,8165.18
2009-3-30-4:30,8703.04
2009-3-30-4:45,8608.17
2009-3-30-5:0,8368.44
2009-3-30-5:15,8567.68
2009-3-30-5:30,7948.10
2009-3-30-5:45,7735.26
2009-3-30-6:0,7708.49
2009-3-30-6:15,7741.83
2009-3-30-6:30,7650.01
2009-3-30-6:45,7818.37
2009-3-30-7:0,6844.44
2009-3-30-7:15,7528.90
2009-3-30-7:30,7012.25
2009-3-30-7:45,7472.01
2009-3-30-8:0,6353.24
2009-3-30-8:15,5932.76
2009-3-30-8:30,5946.49
2009-3-30-8:45,6284.71
2009-3-30-9:0,6978.49
2009-3-30-9:15,7067.06
```

Рисунок 2 - Пример CSV таблицы

3. Выбор инструментов для построения модели

Реализация модели прогнозирования является задачей из области Data Science. В ходе реализации необходимо считывать большие объемы однообразной информации, преобразовывать ее к нужному виду, анализировать пропуски. Так же необходимо применять различные статистические тесты к данным в ходе работы. Важной частью работы с данными является их визуальное представление.

В наше время для работы с данными самым популярным является язык программирования Python. На этом языке реализованы несколько удобных библиотек для работы с данными, статистикой и визуализации данных.

Для работы со статистическими тестами была использована библиотека statsmodels. Она предоставляет различные инструменты для анализа данных. Так же для реализации данной модели прогнозирования из данной библиотеки была взята ARIMA(p, d, q).

Следующей важной библиотекой является Pandas. Это библиотека с открытым исходным кодом, которая представляет эффективные и простые в использовании структуры данных и инструменты анализа данных. Она позволяет считывать данные из CSV файлов и создавать объект Python со строками и столбцами, который называется фреймом данных. Это делает данную библиотеку одной из базовых в области Data Science.

Так же была использована библиотека NumPy. Это универсальный пакет для работы с массивами. Он предоставляет эффективные объекты многомерных массивов и инструменты для работы с ними. Его основной задачей в этой работе является эффективная и быстрая обработка больших объемов данных.

Для визуализации данных используется библиотека Matplotlib. Она предоставляет API для построения различных графиков и отрисовки их в окне приложения.

4. Примеры работы модели прогнозирования трафика инфокоммуникационных сетей

4.1. Получение данных для экспериментов

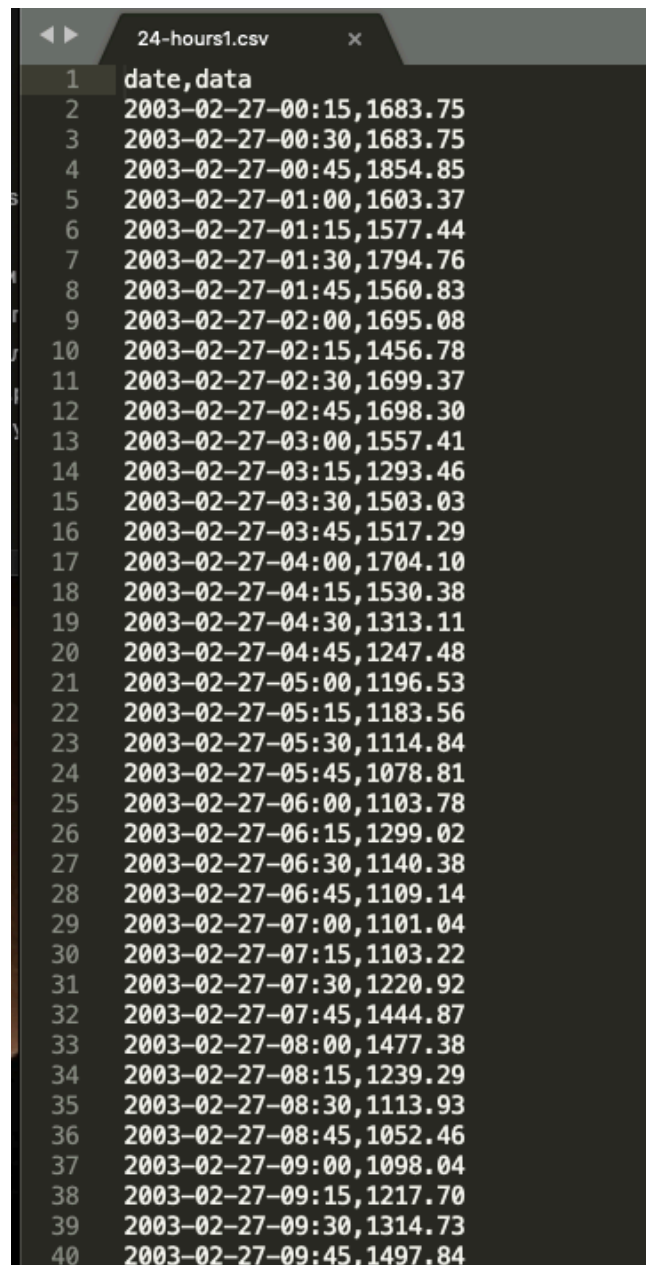
В качестве изучаемых выборок будут использоваться дампы трафика взятый с сайта <http://mawi.nezu.wide.ad.jp>. Эти дампы предоставлены группой the MAWI Working Group, которая занимается сбором и анализом трафика. Предоставленные данные выложены в открытый доступ для самостоятельного анализа и практики умений в области Data Science.

В предоставленных нет пропусков значений. Это позволяет сделать более точный прогноз, на основании данных выборок данных.

4.2. Прогнозирование на основе дампа трафика за 24 часа

В качестве исследуемой выборки будет использоваться непрерывный дампы трафика за 24 часа. Трафик разделен на временные промежутки по 15 минут. Из полученных записей трафика нас интересует общий объем информации, который был передан за наблюдаемое время.

Обработав полученные данные мы получили временной ряд, в котором каждой точке во времени соответствует одно значений - количество переданной информации за 15 минут в Мб. Временной ряд хранится в CSV таблице. Она приведена на рисунке 3.



1	date,data
2	2003-02-27-00:15,1683.75
3	2003-02-27-00:30,1683.75
4	2003-02-27-00:45,1854.85
5	2003-02-27-01:00,1603.37
6	2003-02-27-01:15,1577.44
7	2003-02-27-01:30,1794.76
8	2003-02-27-01:45,1560.83
9	2003-02-27-02:00,1695.08
10	2003-02-27-02:15,1456.78
11	2003-02-27-02:30,1699.37
12	2003-02-27-02:45,1698.30
13	2003-02-27-03:00,1557.41
14	2003-02-27-03:15,1293.46
15	2003-02-27-03:30,1503.03
16	2003-02-27-03:45,1517.29
17	2003-02-27-04:00,1704.10
18	2003-02-27-04:15,1530.38
19	2003-02-27-04:30,1313.11
20	2003-02-27-04:45,1247.48
21	2003-02-27-05:00,1196.53
22	2003-02-27-05:15,1183.56
23	2003-02-27-05:30,1114.84
24	2003-02-27-05:45,1078.81
25	2003-02-27-06:00,1103.78
26	2003-02-27-06:15,1299.02
27	2003-02-27-06:30,1140.38
28	2003-02-27-06:45,1109.14
29	2003-02-27-07:00,1101.04
30	2003-02-27-07:15,1103.22
31	2003-02-27-07:30,1220.92
32	2003-02-27-07:45,1444.87
33	2003-02-27-08:00,1477.38
34	2003-02-27-08:15,1239.29
35	2003-02-27-08:30,1113.93
36	2003-02-27-08:45,1052.46
37	2003-02-27-09:00,1098.04
38	2003-02-27-09:15,1217.70
39	2003-02-27-09:30,1314.73
40	2003-02-27-09:45,1497.84

Рисунок 3 - CSV таблица временного ряда за 24 часа

График данного временного ряда показан на рисунке 4.

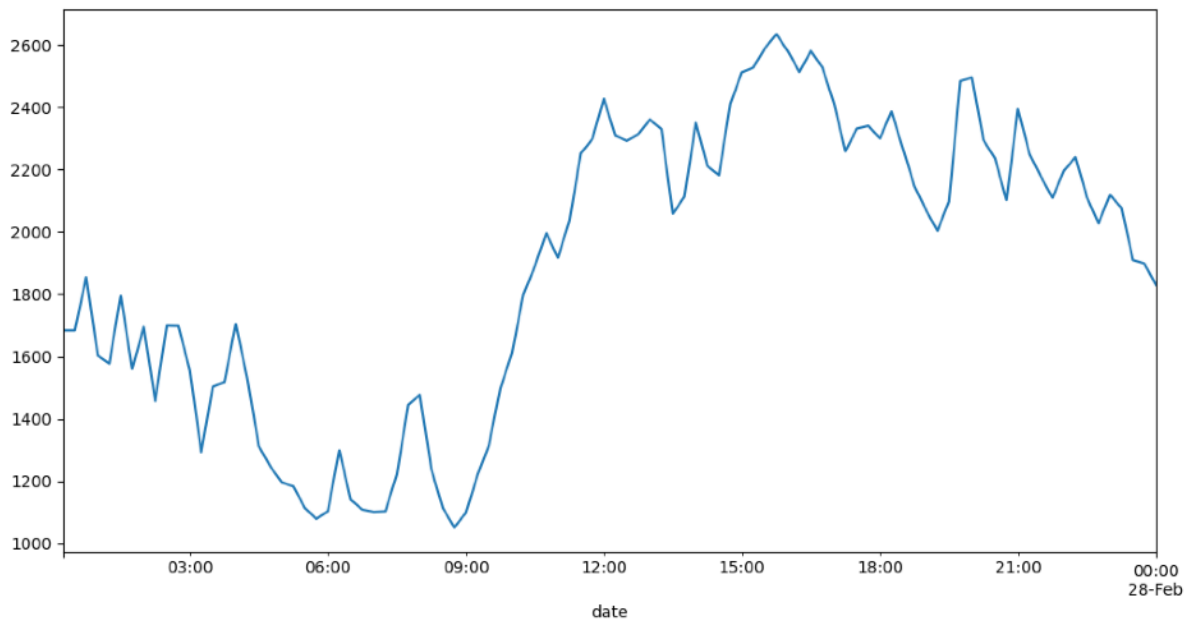


Рисунок 4 - Временной ряд на основе 24 часовой сессии

Следующим шагом прогнозирования будет приведение данного временного ряда к стационарному виду, так как это необходимо для корректных прогнозов. На полученном графике видно, что присутствуют резкие перепады значений. В особенности в промежутке с 9 до 12 часов. Это говорит нам, что рассматриваемый временной ряд скорее всего не стационарный. Для точной оценки мы проводим тест Дики-Фуллера, который определяет статистическую значимость нулевой гипотезы, что временной ряд не стационарен.

Для данного временного ряда в результате проверки нулевой гипотезы мы получили значение статистической значимости p равное 0.56940223. Это значение значительно превышает принятый порог значимости в 0.05, следовательно данный временной ряд не является стационарным.

Так как нам необходимо получить стационарный временной ряд, мы будем применять оператор взятия последовательных разностей, и повторно проводить ADF-тест. Данная последовательность действий будет повторяться, пока мы не получим стационарный ряд.

После проведения следующего теста мы получили p -значение равное $2.2072165750462306e-11$. Данное значение меньше принятой границы

значимости, это означает, что мы получили стационарный ряд. Мы так же нашли наше $d = 1$, необходимое для работы модели прогнозирования ARIMA. На рисунке 5 мы можем видеть полученный временной ряд после операции взятия последовательных разностей.

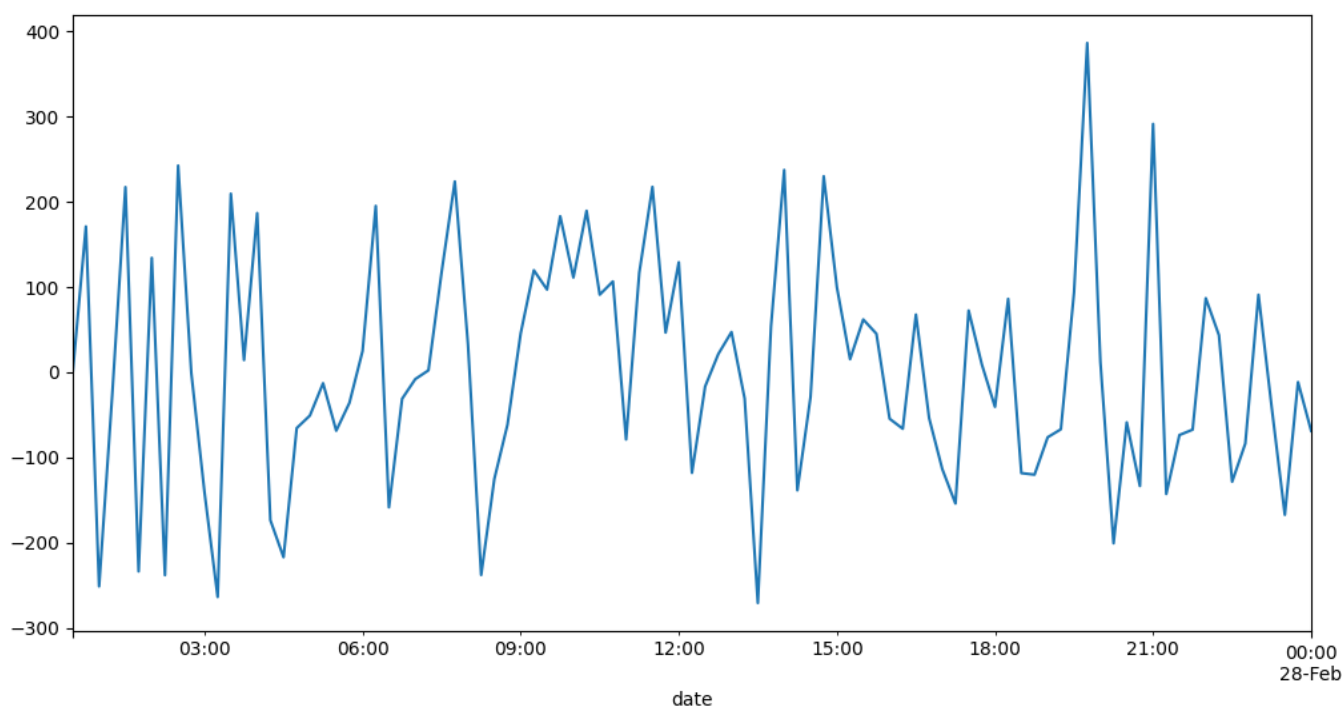


Рисунок 5 - Полученный стационарный временной ряд

Следующим шагом будет получение коэффициентов q и p . Для их получения мы проводим АКФ и ЧАКФ.

На рисунке 6 представлены графики АКФ и ЧАКФ для 45 лагов. На первом графике мы видим, что только один лаг сильно отличен от 0, значит что $p = 1$. На втором графике наблюдается аналогичная ситуация. Значит коэффициент $q = 1$.

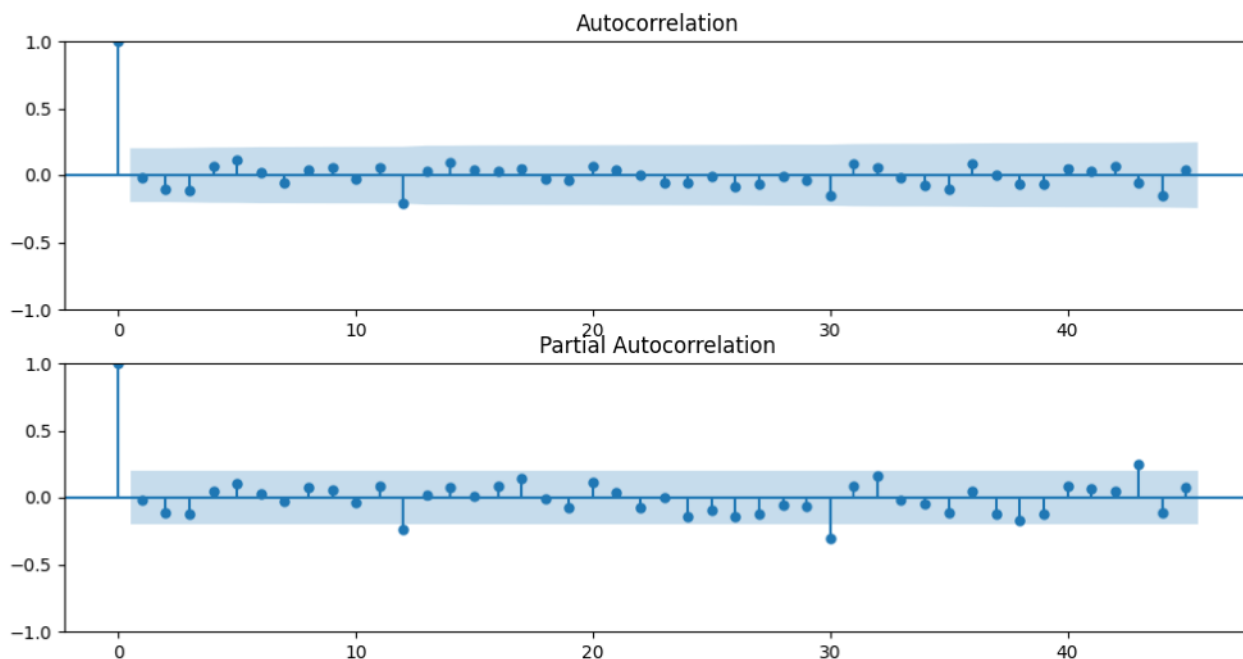


Рисунок 6 - Результаты АКФ и ЧАКФ тестов

После проведенным тестов мы нашли все необходимые для работы модели коэффициенты. Значит мы будем использовать модель $ARIMA(1, 1, 1)$.

Для сравнения эффективность такого подхода к подбору коэффициентов, мы построим вторую модель, в которой параметры были подобраны перебором. Параметры подбирались по информационному критерию Акаике (AIC — Akaike Information Criterion). AIC оценивает, насколько хорошо модель подходит под данные. Чем меньше AIC — тем точнее модель. Этим методом мы были получены параметры $p = 0$, $d = 2$, $q = 1$.

Для первоначальной оценки модели используется оценка ряда остатков Q-тестом Льюнг-Бокса. Данный тест проверяет, является ли ряд остатков случайным. Если он похож на “белый шум”, значит в модели используется максимум доступной информации. Тест проводится на обеих полученных моделях. Результаты теста приведены в таблице 3.

Таблица 3 - Результаты Q-тест Льюнг-Бокса на полученных моделях

	ARIMA(1, 1, 1)		ARIMA(0, 1, 2)	
	Q-stat	p-value	Q-stat	p-value
0	0.021647	0.883031	0.000596	0.980523
1	0.144928	0.930099	0.367486	0.832150
2	1.740700	0.627924	2.117409	0.548399
3	1.755060	0.780693	2.138518	0.710300
4	3.312875	0.651871	3.382465	0.641242
5	3.878129	0.693165	3.957369	0.682446
6	3.937656	0.786930	4.078753	0.770663
7	4.502906	0.809142	4.789650	0.779804
8	5.662342	0.773179	6.108815	0.728980
9	5.671201	0.842090	6.166934	0.801049
10	5.808005	0.885863	6.236483	0.857128
11	9.471981	0.662175	9.931758	0.621948
12	10.126358	0.683570	10.633591	0.641483
13	10.293914	0.740395	10.710752	0.708594
14	11.118487	0.744151	11.700751	0.701524
15	11.403277	0.783936	11.996858	0.744196
16	11.931198	0.804285	12.436869	0.772992
17	12.157516	0.838999	12.676440	0.810425
18	12.367459	0.869324	12.902463	0.843527

По таблице 3 мы видим, что ряд остатков обеих моделей похож на случайный. Так как для каждого элемента из таблицы мы получили коэффициент статистической значимости p больше, чем заданный порог в 0.5. Мы можем сделать вывод, что мы получили адекватные модели, пригодные для прогнозирования.

В результате работы полученных моделей были получены следующие временные ряды. Ряд для модели $ARIMA(1, 1, 1)$ представлен на рисунке 7, для модели $ARIMA(0, 2, 1)$ - на рисунке 8. Наложение прогнозов моделей произведено на рисунке 9.

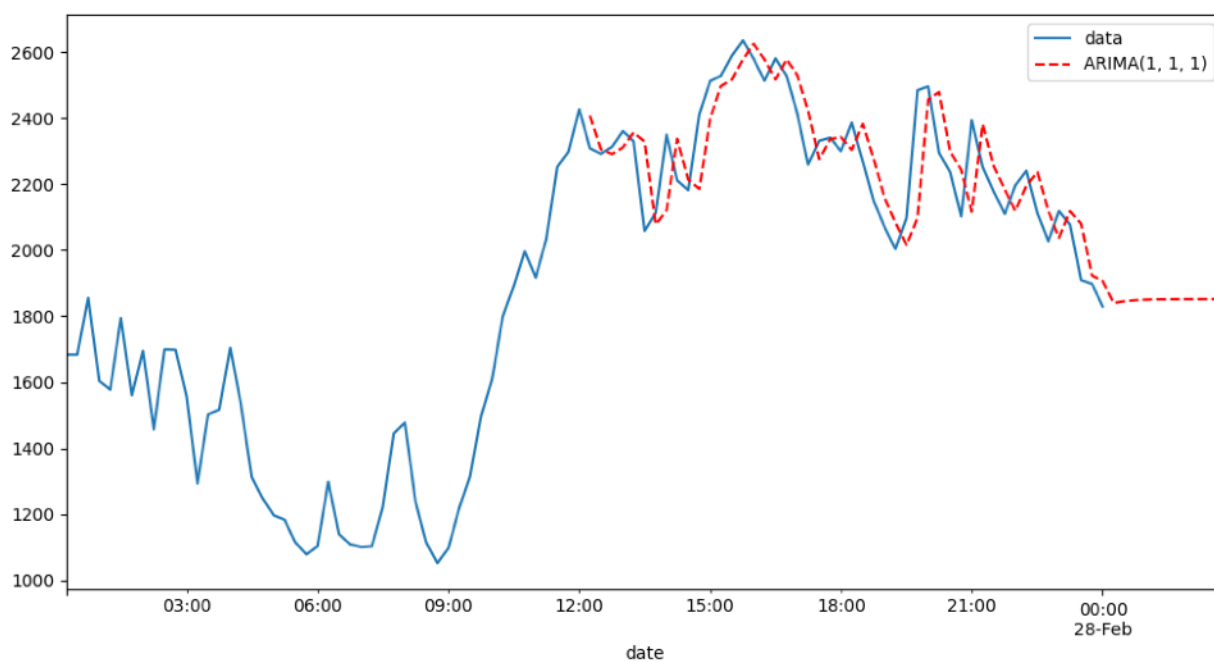


Рисунок 7 - График результатов прогнозирования трафика модели $ARIMA$ с параметрами $p = 1, d = 1, q = 1$

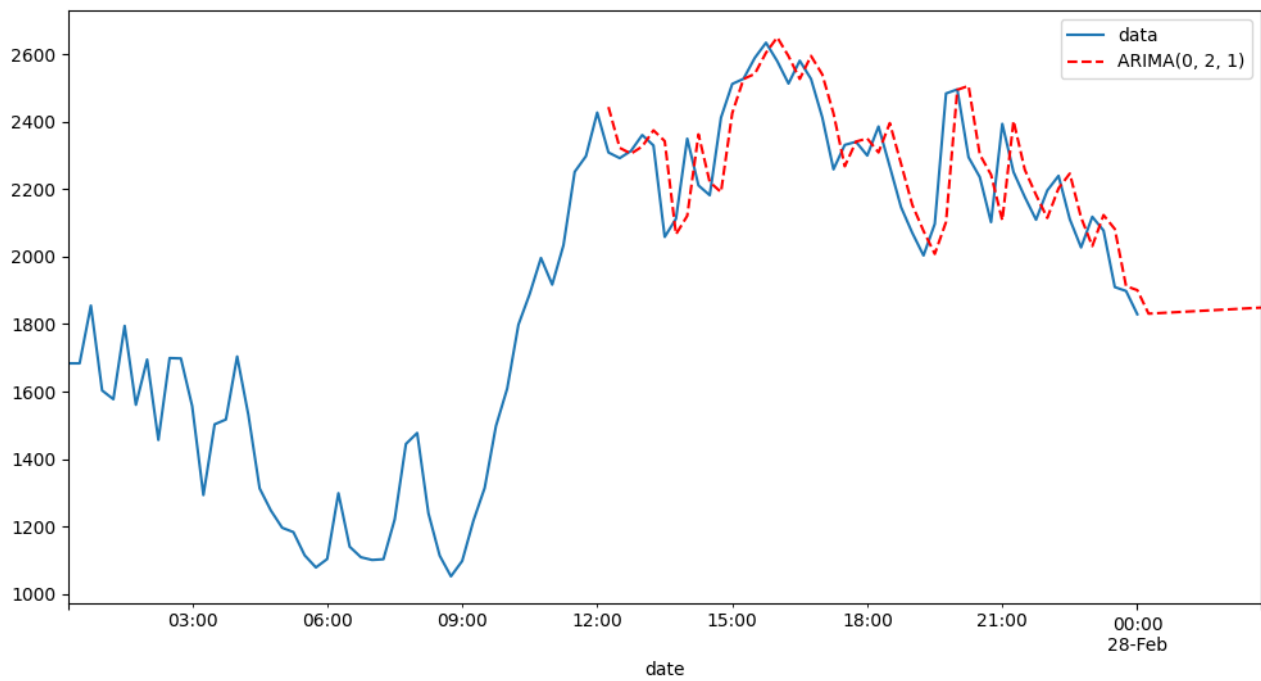


Рисунок 8 - Графики результатов прогнозирования трафика модели ARIMA с параметрами $p = 0$, $d = 2$, $q = 1$

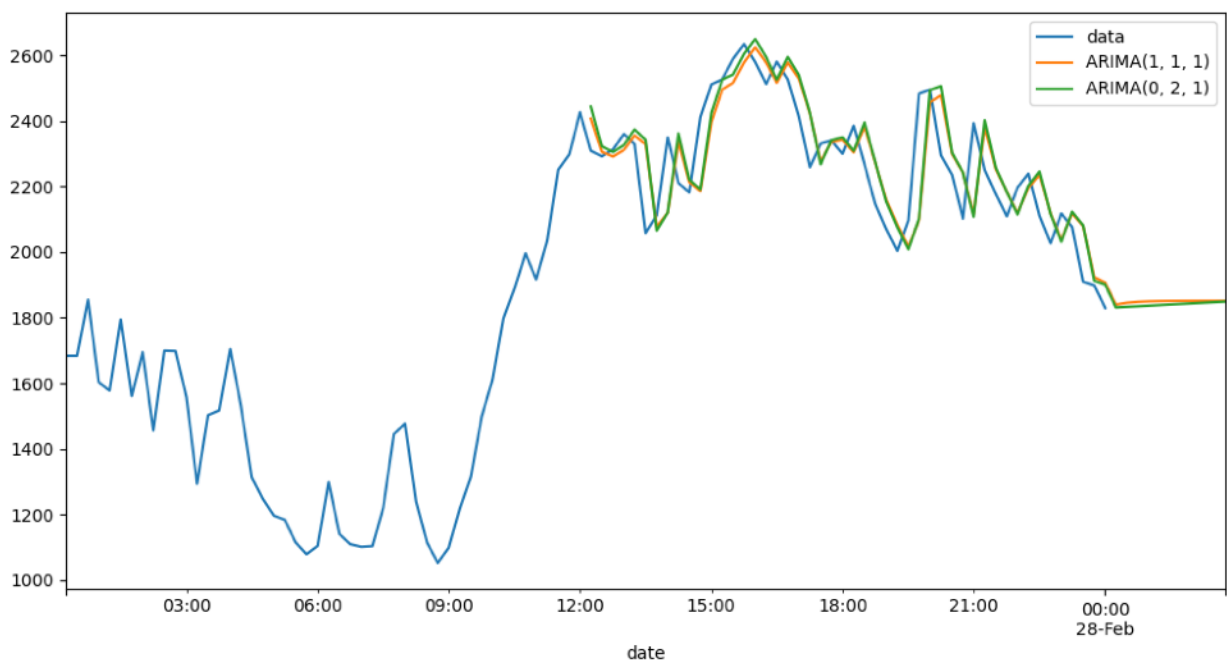


Рисунок 9 - График сравнения прогнозирования трафика моделей с разными параметрами

Для оценки точности прогнозирования полученных моделей были вычислены квадратичная ошибка и средняя абсолютная процентная ошибка. Результаты вычислений приведены в таблице 4.

Таблица 4 - Значения квадратичной и средней абсолютной процентной ошибок прогнозирования для моделей с разными параметрами

Модель ARIMA(p, d, q)	Квадратичная ошибка	Средняя абсолютная процентная ошибка
ARIMA(1, 1, 1)	123.75	4.331387
ARIMA(0, 2, 1)	127.44	4.422438
Усредненная модель	122.62	4.283928

По значениям из таблицы мы видим, что модель, параметры которой были подобраны с помощью статистических тестов является более точной, по сравнению с полным перебором.

Для итогового прогнозирования мы будем использовать обе полученные модели. Усредним полученные предсказания. Итоги такого комбинированного прогнозирования показаны на рисунке 10. Для данного прогноза мы также будем вычислять ошибки. Результат тестов занесены в таблицу 4.

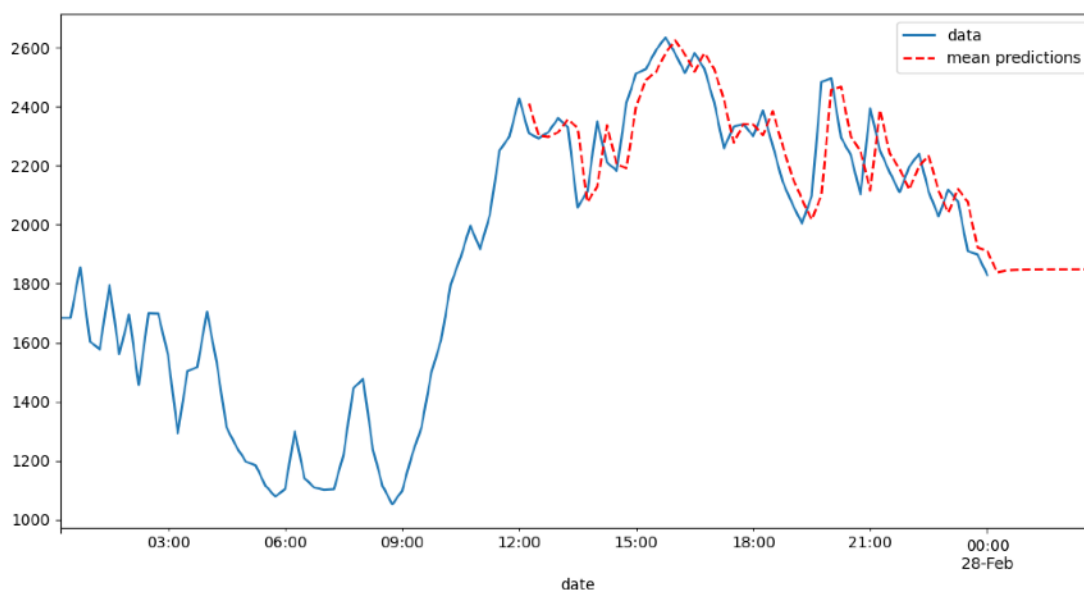


Рисунок 10 - График усредненного прогнозирования

Из полученных данным мы можем увидеть, что усреднив наш прогноз мы уменьшили квадратичную и среднюю абсолютную процентную ошибку. Мы можем сделать вывод, что такой подход дает реальный выигрыш перед использованием одной модели за раз.

4.3. Прогнозирование на основе дампа трафика за 96 часов

В данном примере мы будем рассматривать непрерывные дампы трафика для сессии в 96 часов. Она так же разбита на промежутки в 15 минут. В качестве измеряемого параметра используется суммарный объем данных, переданных за 15 минут.

После обработки дампов трафика был получен временной ряд длиной 384 элемента. Каждый элемент полученного ряда хранит объем переданных данных за 15 минут в Мб. CSV таблица с полученным временным рядом показана на рисунке 11.

	96-hours.csv
1	date,data
2	2009-3-30-0:0,11682.93
3	2009-3-30-0:15,11697.34
4	2009-3-30-0:30,11709.92
5	2009-3-30-0:45,10388.29
6	2009-3-30-1:0,10445.75
7	2009-3-30-1:15,10540.36
8	2009-3-30-1:30,9940.78
9	2009-3-30-1:45,9919.54
10	2009-3-30-2:0,9848.27
11	2009-3-30-2:15,9808.58
12	2009-3-30-2:30,9664.25
13	2009-3-30-2:45,9517.67
14	2009-3-30-3:0,8820.23
15	2009-3-30-3:15,9487.88
16	2009-3-30-3:30,9241.88
17	2009-3-30-3:45,9195.72
18	2009-3-30-4:0,8610.56
19	2009-3-30-4:15,8165.18
20	2009-3-30-4:30,8703.04
21	2009-3-30-4:45,8608.17
22	2009-3-30-5:0,8368.44
23	2009-3-30-5:15,8567.68
24	2009-3-30-5:30,7948.10
25	2009-3-30-5:45,7735.26
26	2009-3-30-6:0,7708.49
27	2009-3-30-6:15,7741.83
28	2009-3-30-6:30,7650.01
29	2009-3-30-6:45,7818.37
30	2009-3-30-7:0,6844.44

Рисунок 11 - CSV таблица временного ряда за 96 часо

График данного временного ряда показан на рисунке 12.

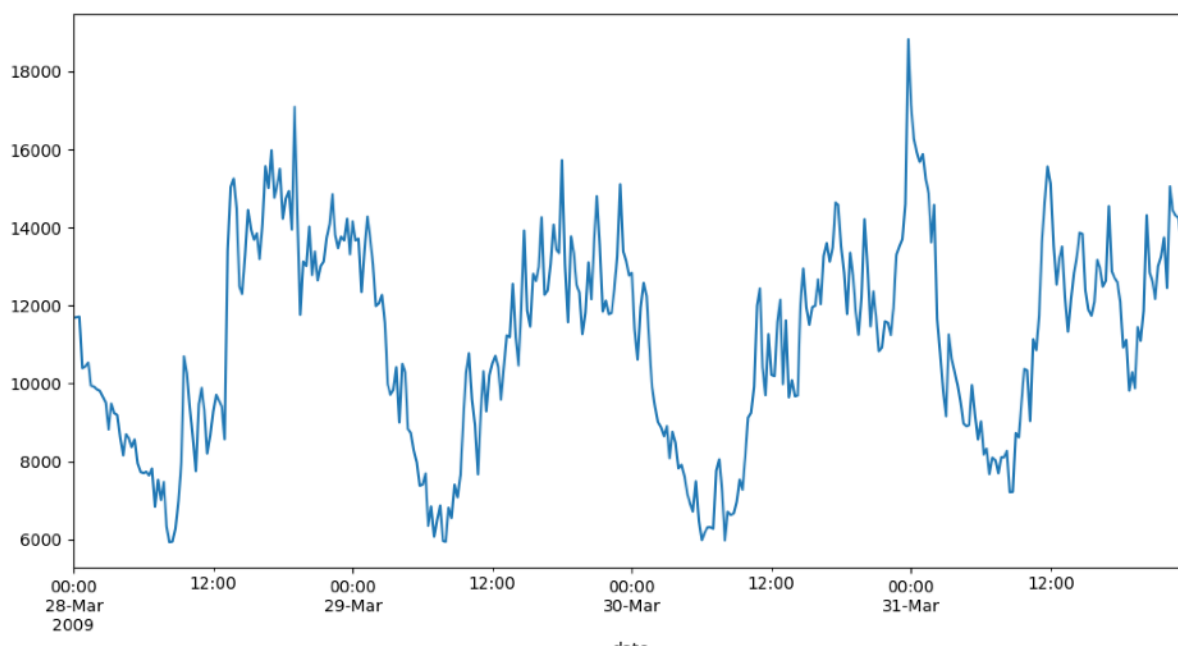


Рисунок 12- Временной ряд на основе 96 часовой сессии

В данном временном ряду мы можем наблюдать сезонность. Ее можно учитывать для более точного прогноза, но для этого необходимо реализовывать другую статистическую модель прогнозирования SARIMA. Так как в данной выпускной квалификационной работе рассматривается модель ARIMA, то мы никак не обрабатываем сезонность. Согласно тесту Дики-Фуллера мы получили значение коэффициента значимости p равное 0.0508322929678728. Так как полученное значение больше принятого порогового значения в 0.5. То полученный ряд не является стационарным.

Следующим шагом в работе с нашей моделью будет приведение временного ряда к стационарному виду. Для проверки на стационарность также используется тест Дики-Фуллера. Для приведения ряда к стационарному виду используется оператор взятие разности над временным рядом.

После проведения ряда тестов мы получили p значение, равное $6.073077676614103e-30$. Теперь мы можем сделать вывод, что ряд скорее всего является стационарным. При этом значение коэффициента p равняется 1. Полученный временной ряд показан на рисунке 13.

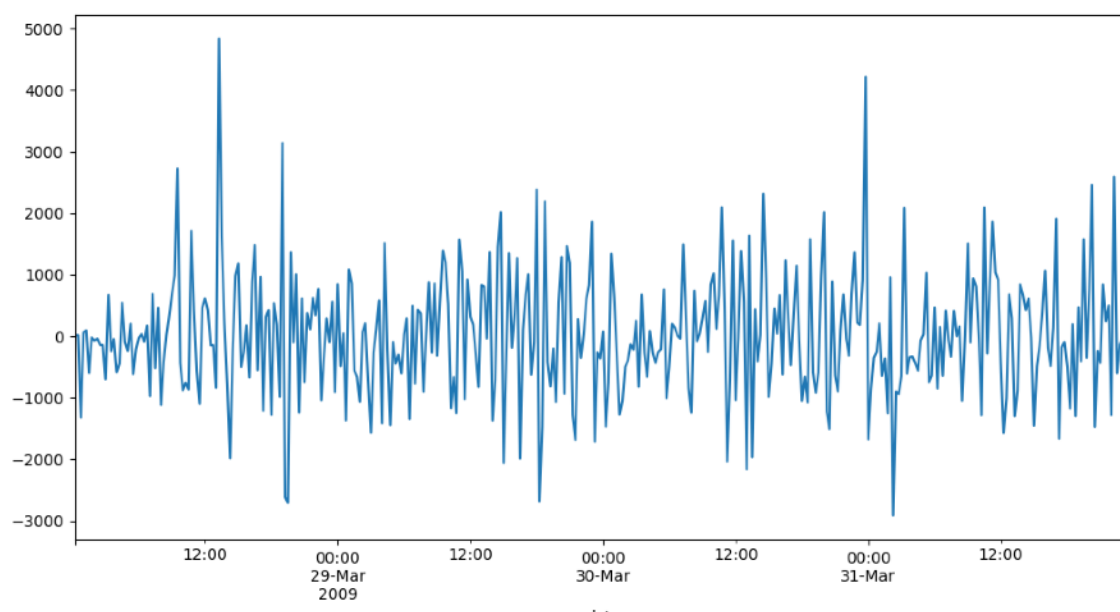


Рисунок 13 - Полученный стационарный временной ряд

Следующим шагом прогнозирования временного ряда будет проведения АФК и ЧАФК для 25 лагов над временным рядом. После проведения тестов были получены графики. Они показаны на рисунке 14.

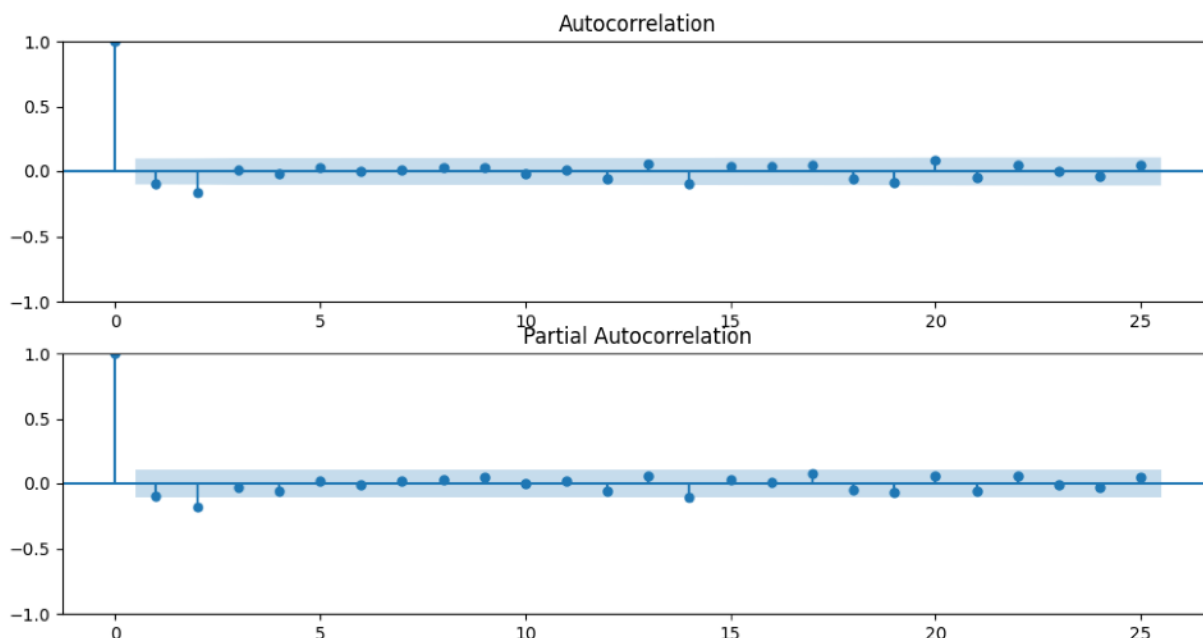


Рисунок 14 - Результаты АКФ и ЧАКФ тестов

На полученном графике АКФ мы можем видеть только 1 лаг, сильно отличающийся от 0. Это означает, что коэффициент p равен 1. На графике ЧАКФ наблюдается аналогичная ситуация. Лишь один лаг сильно отличается от 0. Значит коэффициент q также будет равен 1.

После проведенных тестов мы получили все необходимые параметры для работы нашей модели ARIMA: $p = 1$, $d = 1$, $q = 1$. Для сравнения эффективности подбора параметров данным методом мы подбираем другие параметры методом полного перебора. На каждом шаге мы оцениваем AIC полученной модели и выбираем параметры, для которых мы получили минимальное значение AIC. Данным методом были получены следующие параметры: $p = 0$, $d = 1$, $q = 2$.

Для проверки полученных моделей необходимо также провести Q-тест Льюнг-Бокса. С его помощью мы проверим ряд остатков временных рядов, на схожесть с белым шумом. Результаты данных тестов приведены в таблице 5.

Таблица 5 - Результаты Q-тест Льюнг-Бокса на полученных моделях

	ARIMA(1, 1, 1)		ARIMA(0, 1, 2)	
	Q-stat	p-value	Q-stat	p-value
0	0.463315	0.496079	0.022708	0.880218
1	1.742263	0.418478	0.093048	0.954542
2	1.743866	0.627226	0.231022	0.972432
3	1.753316	0.781011	0.272647	0.991511
4	2.177635	0.824060	0.466406	0.993302
5	2.178134	0.902608	0.490566	0.997951
6	2.277854	0.942877	0.591097	0.999040
7	2.464379	0.963371	0.699184	0.999529
8	2.672936	0.975865	0.931177	0.999580
9	2.755040	0.986616	1.098805	0.999735
10	2.755089	0.993583	1.119817	0.999911
11	3.877209	0.985552	2.771053	0.996956
12	4.642898	0.982237	3.774284	0.993364
13	6.419259	0.954793	5.901727	0.968908
14	6.574216	0.968407	6.259432	0.975053
15	6.684163	0.978879	6.280180	0.984777
16	6.838280	0.985548	6.413620	0.989947
17	7.356342	0.986761	6.742737	0.992122
18	9.156455	0.970790	8.642977	0.978951

Проанализировав результаты из таблицы 5, мы можем сделать вывод, что наши остатки похожи на “белый шум“, значит для полученных моделей были подобраны корректные параметры. Следовательно мы можем применять данные модели для прогнозирования.

После всех проведенных действий были получены 2 набора параметров для модели ARIMA. Мы также убедились в адекватности подобранных параметров. Значит мы можем переходить к оценке наших моделей на точность прогнозирования. Результаты прогнозирования модели ARIMA(1, 1, 1)

показаны на рисунке 15. Результаты прогнозирования модели ARIMA(0, 1, 2) приведены на рисунке 16.

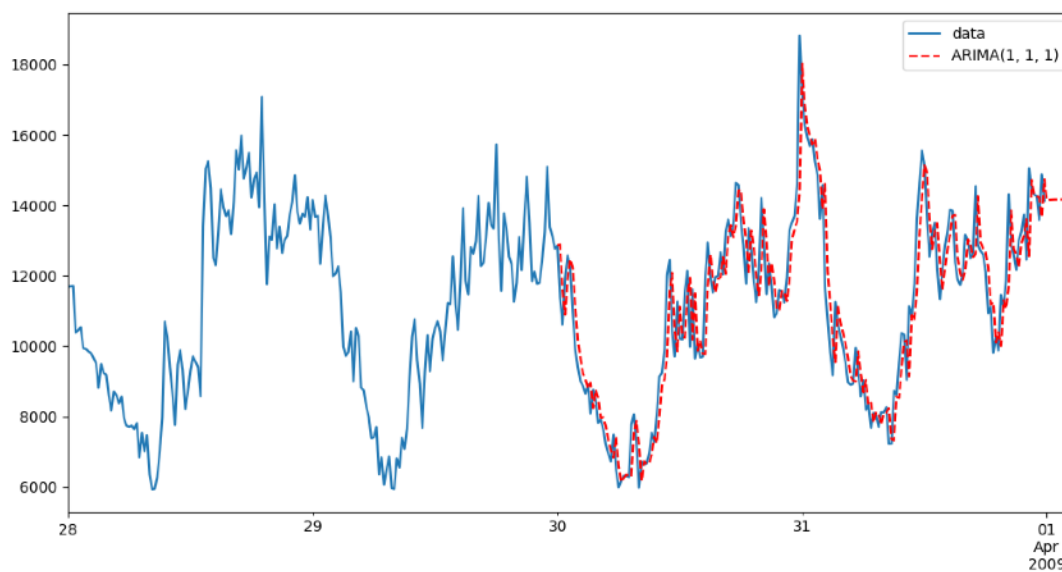


Рисунок 15 - График результатов прогнозирования трафика модели ARIMA с параметрами $p = 1$, $d = 1$, $q = 1$

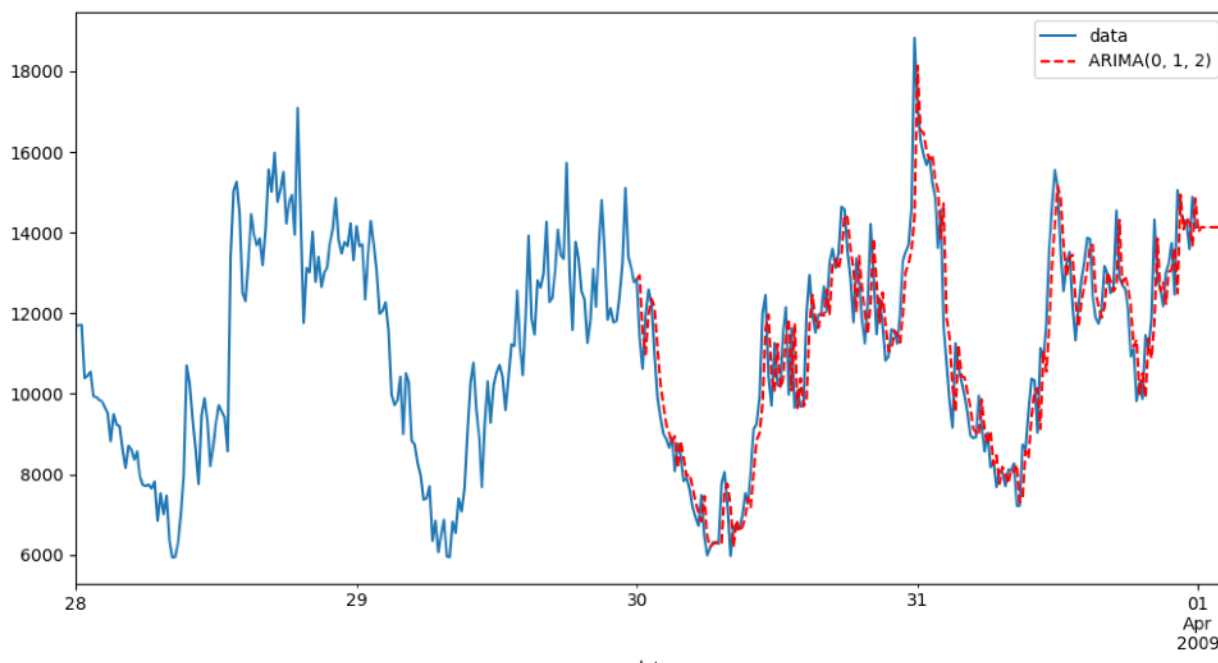


Рисунок 16 - График результатов прогнозирования трафика модели ARIMA с параметрами $p = 0$, $d = 1$, $q = 2$

На рисунке 17 произведено наложение прогнозов для их сравнения.

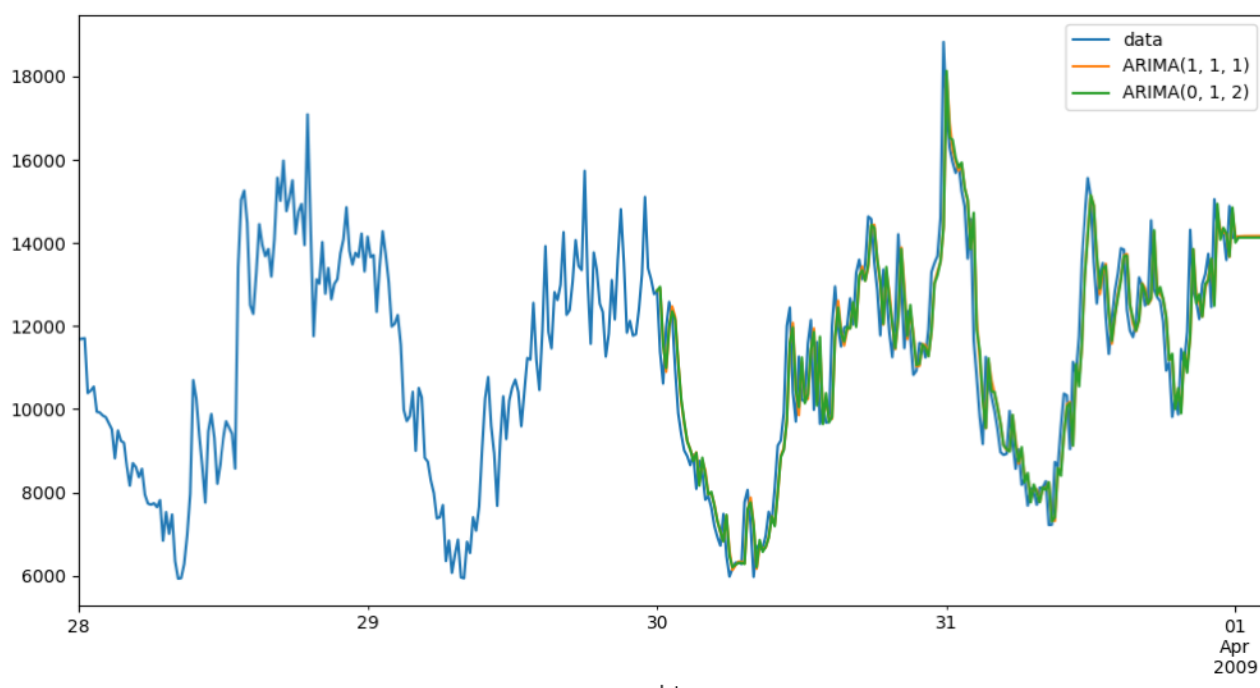


Рисунок 17 - График сравнения прогнозирования трафика моделей с разными параметрами

Для оценки точности прогнозирования полученных моделей были вычислены квадратичная ошибка и средняя абсолютная процентная ошибка. Результаты вычислений приведены в таблице 4.

Таблица 6 - Значения квадратичной и средней абсолютной процентной ошибок прогнозирования для моделей с разными параметрами

Модель ARIMA(p, d, q)	Квадратичная ошибка	Средняя абсолютная процентная ошибка
ARIMA(1, 1, 1)	985.05	6.799146
ARIMA(0, 1, 2)	983.70	6.798303
Усредненная модель	983.24	6.789006

По значениям из таблицы мы видим, полученные результаты близки по значениям. Значит оба способа пригодны для подбора параметров модели прогнозирования.

Для итогового прогнозирования мы будем использовать обе полученные модели. Усредним полученные предсказания. Итоги такого прогнозирования показаны на рисунке 18. Для данного прогноза мы также будем вычислять ошибки. Результат тестов занесены в таблицу 6.

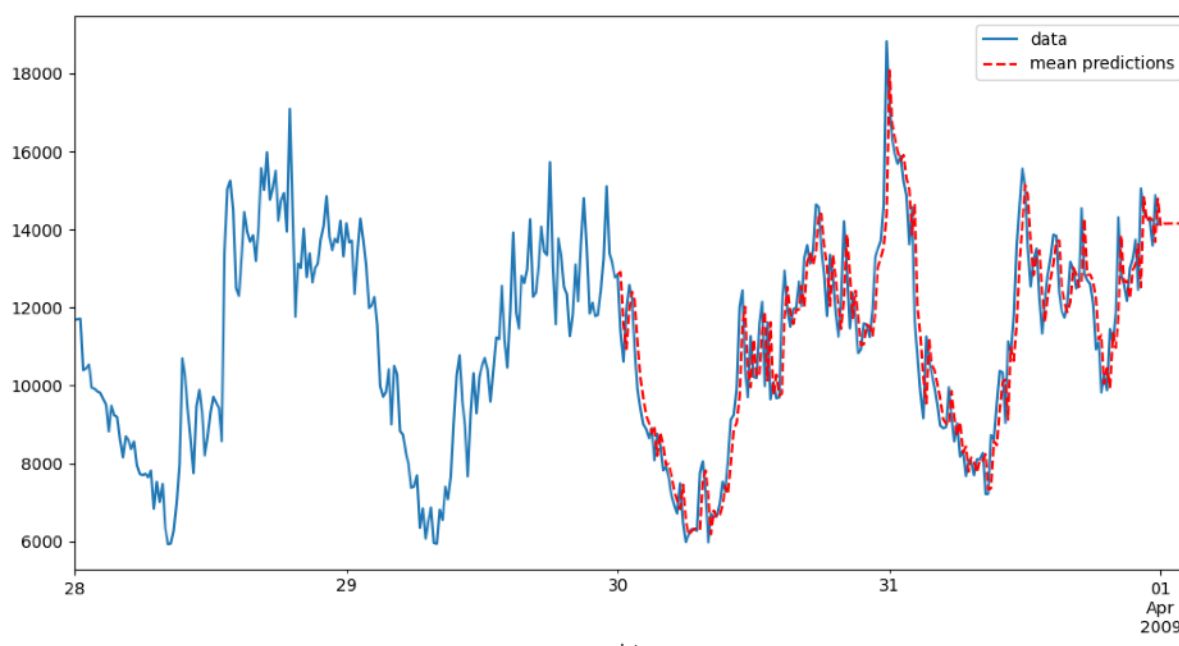


Рисунок 14 - График усредненного прогнозирования

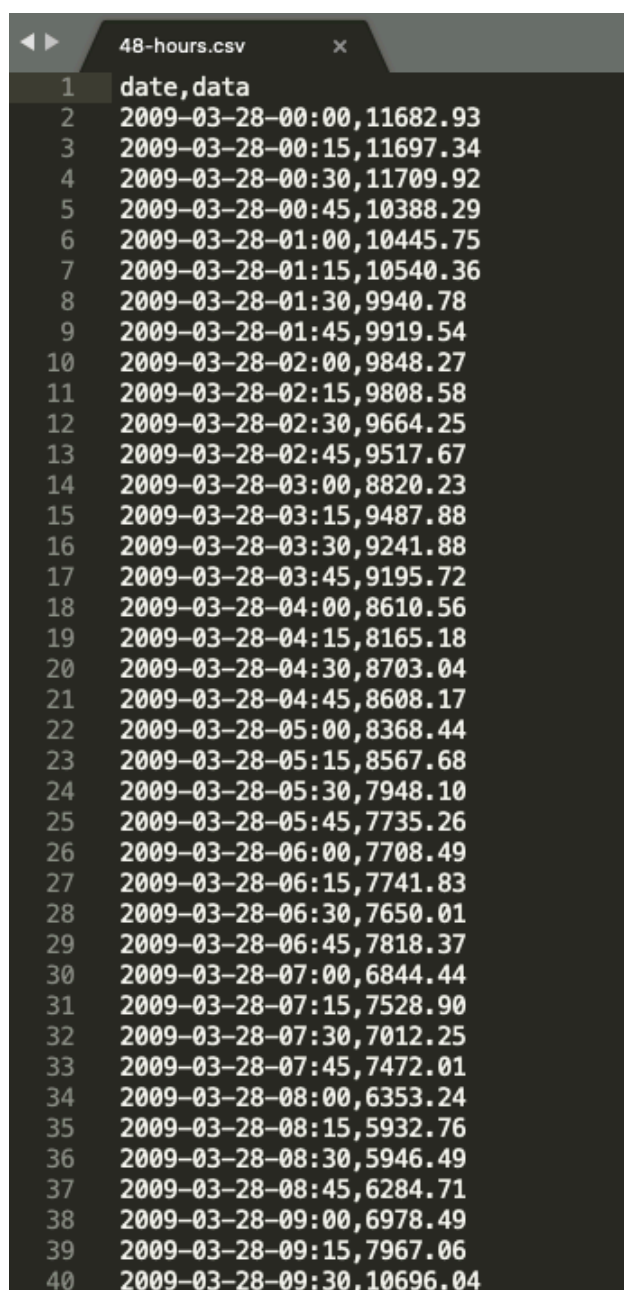
Проведя оценку точности для комбинированного использования прогнозов мы видим, что квадратичная и средняя абсолютная процентная ошибка уменьшились, значит такой подход дает выигрыш перед использованием моделей по отдельности.

4.4. Прогнозирование на основе дампа трафика за 48 часов

Для данного примера будет использован дам трафика за 48 часов непрерывной сессии. Трафик также разделен на временные промежутки по 15

минут. Из полученных записей трафика мы возьмем общий объем информации, который был передан в наблюдаемый промежуток времени.

После обработки данного дампа трафика мы получили временной ряд, который содержит 192 значения. Как и в предыдущих примерах объем трафика выражен в Мб. Данный временной ряд хранится в CSV таблице. Ее часть показана на рисунке 15.



1	date,data
2	2009-03-28-00:00,11682.93
3	2009-03-28-00:15,11697.34
4	2009-03-28-00:30,11709.92
5	2009-03-28-00:45,10388.29
6	2009-03-28-01:00,10445.75
7	2009-03-28-01:15,10540.36
8	2009-03-28-01:30,9940.78
9	2009-03-28-01:45,9919.54
10	2009-03-28-02:00,9848.27
11	2009-03-28-02:15,9808.58
12	2009-03-28-02:30,9664.25
13	2009-03-28-02:45,9517.67
14	2009-03-28-03:00,8820.23
15	2009-03-28-03:15,9487.88
16	2009-03-28-03:30,9241.88
17	2009-03-28-03:45,9195.72
18	2009-03-28-04:00,8610.56
19	2009-03-28-04:15,8165.18
20	2009-03-28-04:30,8703.04
21	2009-03-28-04:45,8608.17
22	2009-03-28-05:00,8368.44
23	2009-03-28-05:15,8567.68
24	2009-03-28-05:30,7948.10
25	2009-03-28-05:45,7735.26
26	2009-03-28-06:00,7708.49
27	2009-03-28-06:15,7741.83
28	2009-03-28-06:30,7650.01
29	2009-03-28-06:45,7818.37
30	2009-03-28-07:00,6844.44
31	2009-03-28-07:15,7528.90
32	2009-03-28-07:30,7012.25
33	2009-03-28-07:45,7472.01
34	2009-03-28-08:00,6353.24
35	2009-03-28-08:15,5932.76
36	2009-03-28-08:30,5946.49
37	2009-03-28-08:45,6284.71
38	2009-03-28-09:00,6978.49
39	2009-03-28-09:15,7967.06
40	2009-03-28-09:30,10696.04

Рисунок 15 - CSV таблица временного ряда за 48 часов

График полученного временного ряда приведен на рисунке 16.

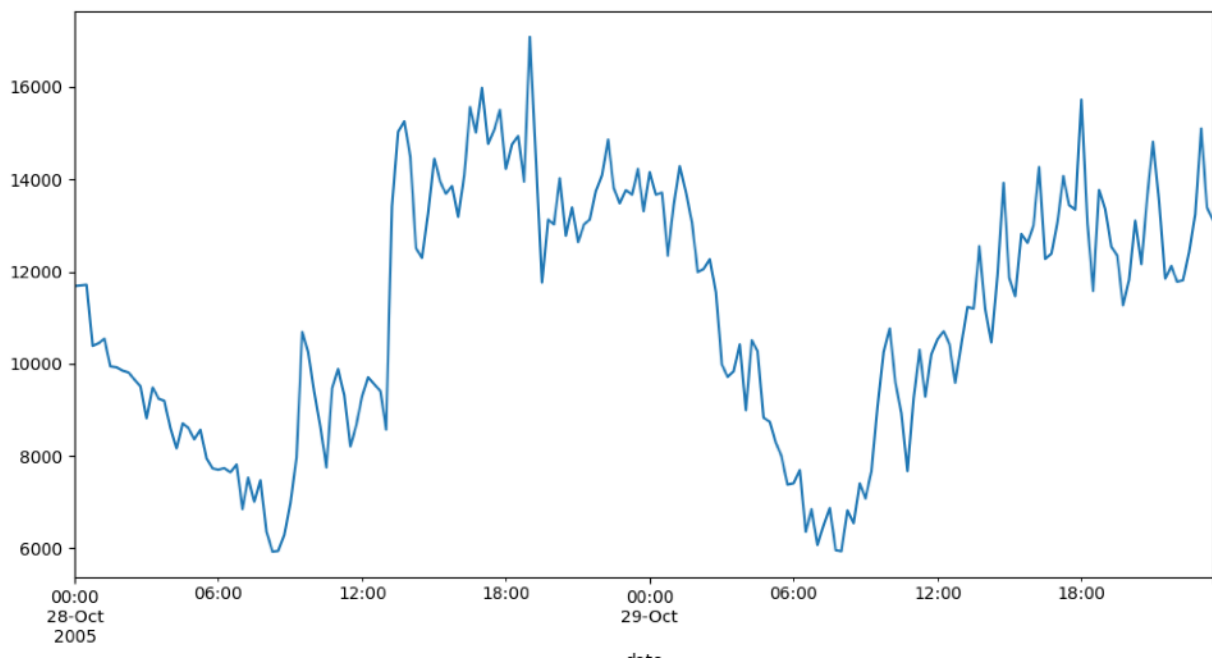


Рисунок 16 - Временной ряд на основе 48 часовой сессии

В данном временном ряду мы также видим некую сезонность, но мы не будем учитывать ее, так как это не используется в реализуемой в данной ВКР модели $ARIMA(p, d, q)$. Проверив это ряд на стационарность с помощью теста Дики-Фуллера мы получили коэффициент статистической значимости p равный 0.5941018228910662. Так как он почти в 12 раз больше порога значимости, мы делаем вывод, что ряд скорее всего не является стационарным.

Далее нам необходимо привести временной ряд к стационарному виду. Для этого мы будем использовать оператор взятия разности над временным рядом. Для оценки полученных рядов будет также использоваться тест Дики-Фуллера.

После одной операции взятия разности над временным рядом мы получили стационарный временной ряд. Об этом говорит значение p , полученное в результате ADF-теста. Оно равно $2.4819334765188016e-16$. Так как мы провели всего одну интеграцию, значения параметра d мы приравняем к 1. Полученный временной ряд показан на рисунке 17.

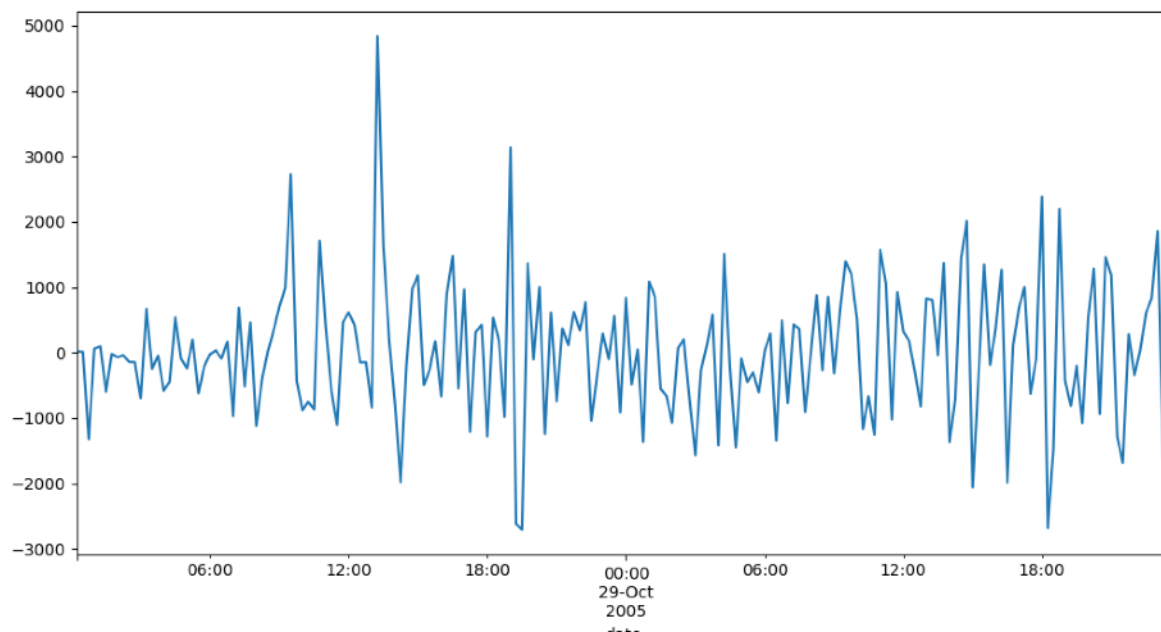


Рисунок 17 - Полученный стационарный временной ряд

Следующим шагом прогнозирования временного ряда будет проведения АФК и ЧАФК для 25 лагов над временным рядом. После проведения тестов были получены графики. Они показаны на рисунке 18.

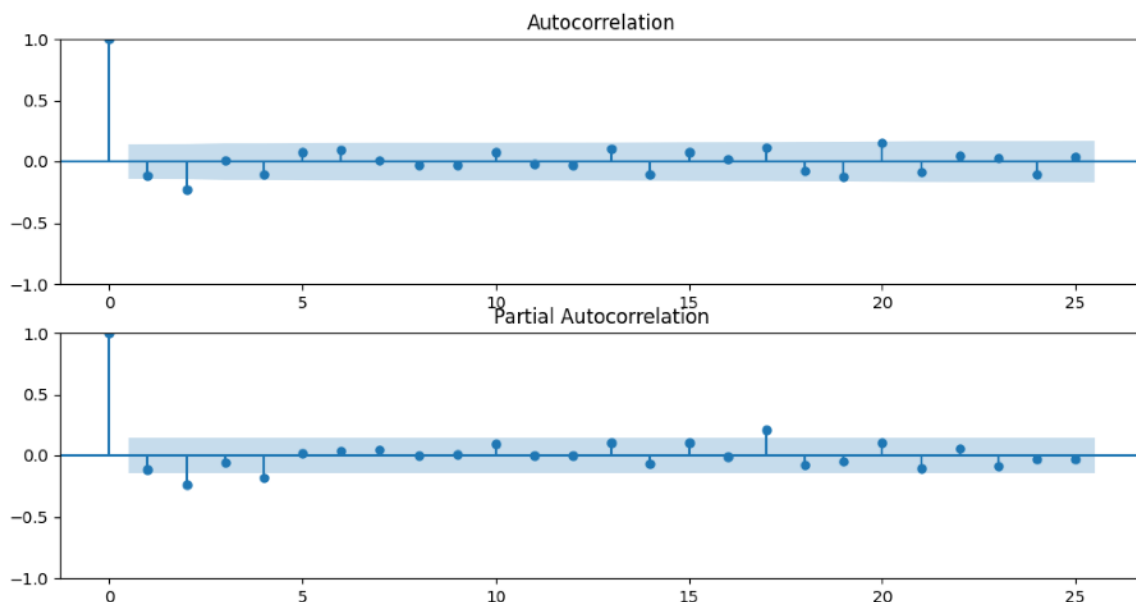


Рисунок 18 - Результаты АКФ и ЧАКФ тестов

На полученном графике АФК мы видим, что 2 лага сильно отличны от 0, значит порядок авторегрессии p мы приравниваем к 2. На графике ЧАКФ мы 3 лага, которые сильно отличаются от 0. Поэтому значение порядка скользящего среднего q мы приравниваем к 3.

После проведенным тестов мы нашли все необходимые для работы модели коэффициенты. Значит мы будем использовать модель $ARIMA(2, 1, 3)$. Для сравнения эффективности подбора параметров данным методом мы подбираем другие параметры методом полного перебора. На каждом шаге мы оцениваем AIC полученной модели и выбираем параметры, для которых мы получили минимальное значение AIC. Данным методом были получены следующие параметры: $p = 0$, $d = 1$, $q = 2$.

Для проверки полученных моделей необходимо также провести Q-тест Льюнг-Бокса. С его помощью мы проверим ряд остатков временных рядов, на схожесть с белым шумом. Результаты данных тестов приведены в таблице 7.

Таблица 7 - Результаты Q-тест Льюнг-Бокса на полученных моделях

	ARIMA(2, 1, 3)		ARIMA(0, 1, 2)	
	Q-stat	p-value	Q-stat	p-value
0	0.034486	0.852677	0.040874	0.839781
1	0.172980	0.917145	0.273252	0.872296
2	0.390149	0.942270	0.440007	0.931859
3	0.446439	0.978501	0.802898	0.938059
4	0.615138	0.987300	1.194861	0.945369
5	1.176183	0.978045	1.391522	0.966372
6	1.176694	0.991454	1.423206	0.984850
7	1.189337	0.996746	1.426584	0.993861
8	1.267248	0.998533	1.452160	0.997486
9	1.663312	0.998330	1.660929	0.998340
10	1.683231	0.999336	1.661170	0.999376
11	1.828938	0.999627	1.959542	0.999465

	ARIMA(2, 1, 3)		ARIMA(0, 1, 2)	
	Q-stat	p-value	Q-stat	p-value
12	3.094542	0.997571	3.624076	0.994573
13	3.592246	0.997461	4.405772	0.992488
14	4.212349	0.996955	5.414167	0.988001
15	4.218057	0.998474	5.465010	0.992914
16	4.521625	0.998833	5.843790	0.994175
17	4.616907	0.999342	5.939917	0.996434
18	5.863207	0.998233	7.233014	0.992864
19	6.706911	0.997541	8.136891	0.990923
20	7.389220	0.997239	8.871387	0.990190
21	7.405162	0.998419	8.887180	0.993899

Проанализирована результаты в таблице 7, можно сделать вывод, что на основе полученных параметров можно построить адекватные модели статистического прогнозирования.

Результаты прогнозирования модели ARIMA(2, 1, 3) показаны на рисунке 19, прогноз модели ARIMA(0, 1, 2) приведен на рисунке 20. Наложение прогнозов выполнено на рисунке 21.

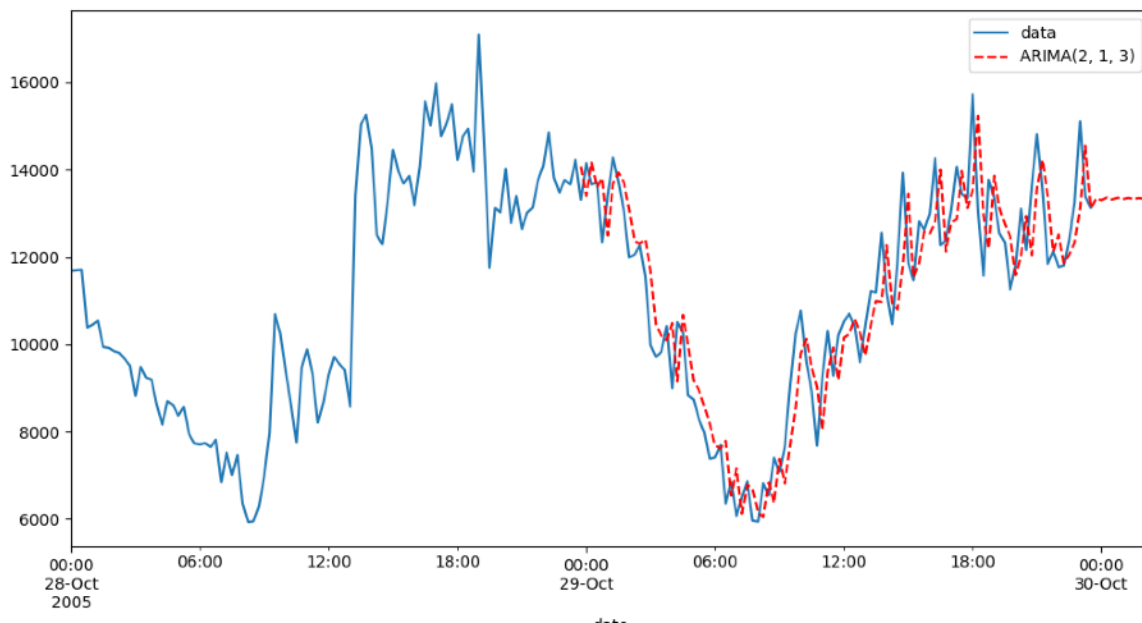


Рисунок 19 - График результатов прогнозирования трафика модели ARIMA с параметрами $p = 2$, $d = 1$, $q = 3$

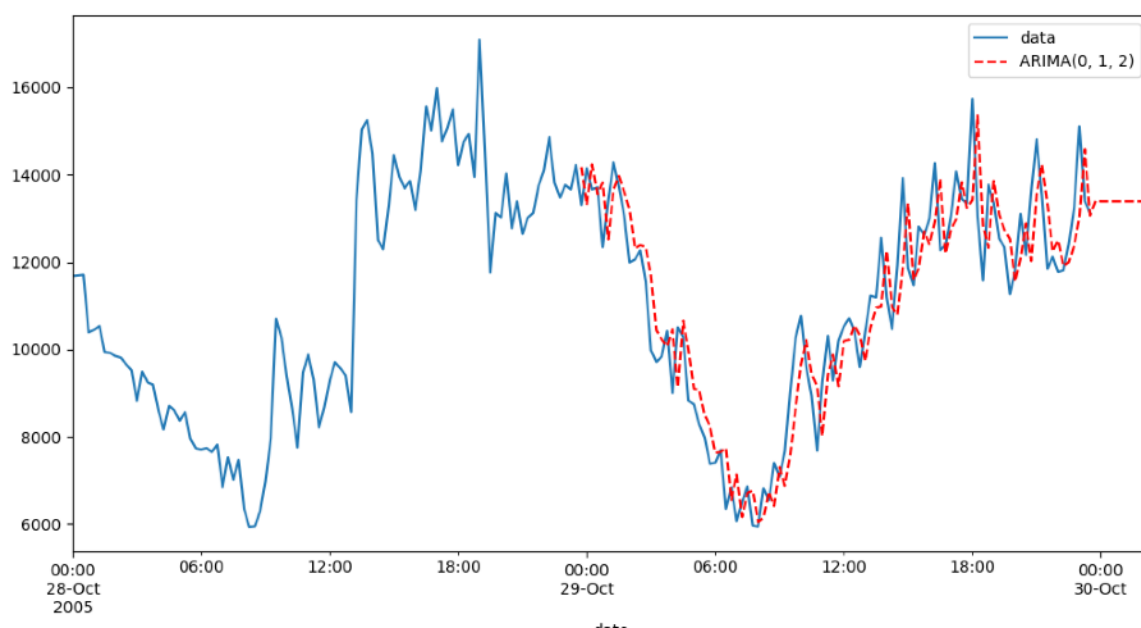


Рисунок 20 - График результатов прогнозирования трафика модели ARIMA с параметрами $p = 0$, $d = 1$, $q = 2$

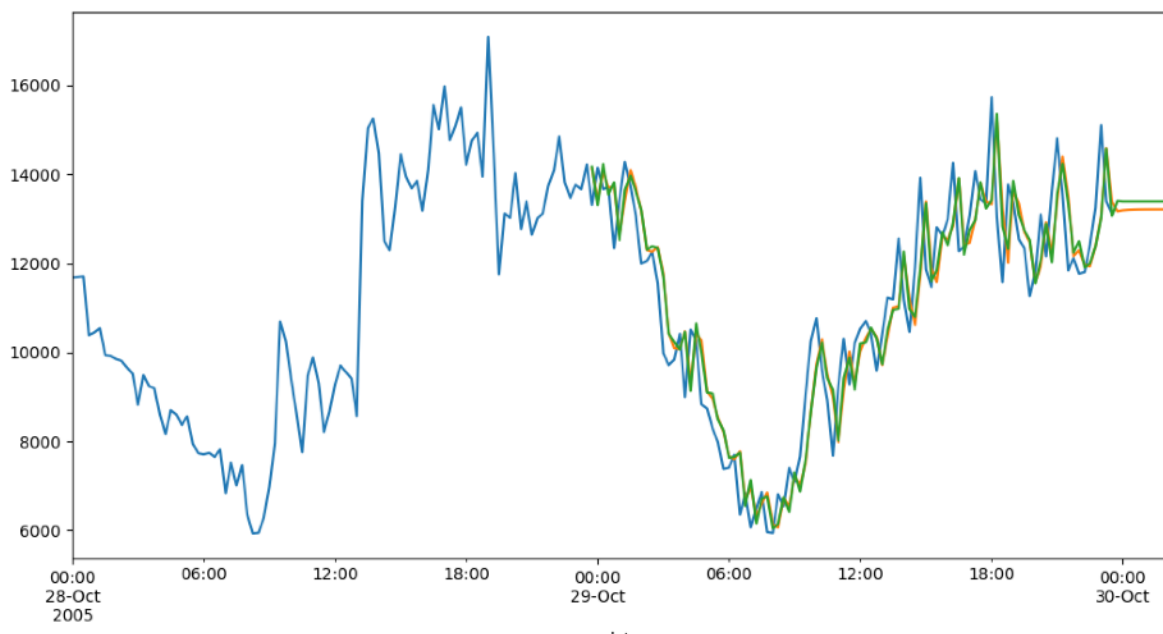


Рисунок 21 - График сравнения прогнозирования трафика моделей с разными параметрами

Для оценки точности прогнозирования полученных моделей были вычислены квадратичная ошибка и средняя абсолютная процентная ошибка. Результаты вычислений приведены в таблице 8.

Таблица 8 - Значения квадратичной и средней абсолютной процентной ошибок прогнозирования для моделей с разными параметрами-1

Модель ARIMA(p, d, q)	Квадратичная ошибка	Средняя абсолютная процентная ошибка
ARIMA(2, 1, 3)	973.25	7.515624
ARIMA(0, 1, 2)	971.80	7.469604
Усредненная модель	971.89	7.489096

По результатам проведенных тестов мы видим, что полученные значения близки к друг другу. Поэтому мы можем сделать вывод, что оба способа пригодны для подбора параметров для моделей прогнозирования.

Для сравнения, мы усредним оба прогноза, а для результата так же вычислим квадратичную и среднюю абсолютную процентную ошибки. Результат объединения двух прогнозов показан на рисунке 22. Результаты тестов занесены в таблицу 8.

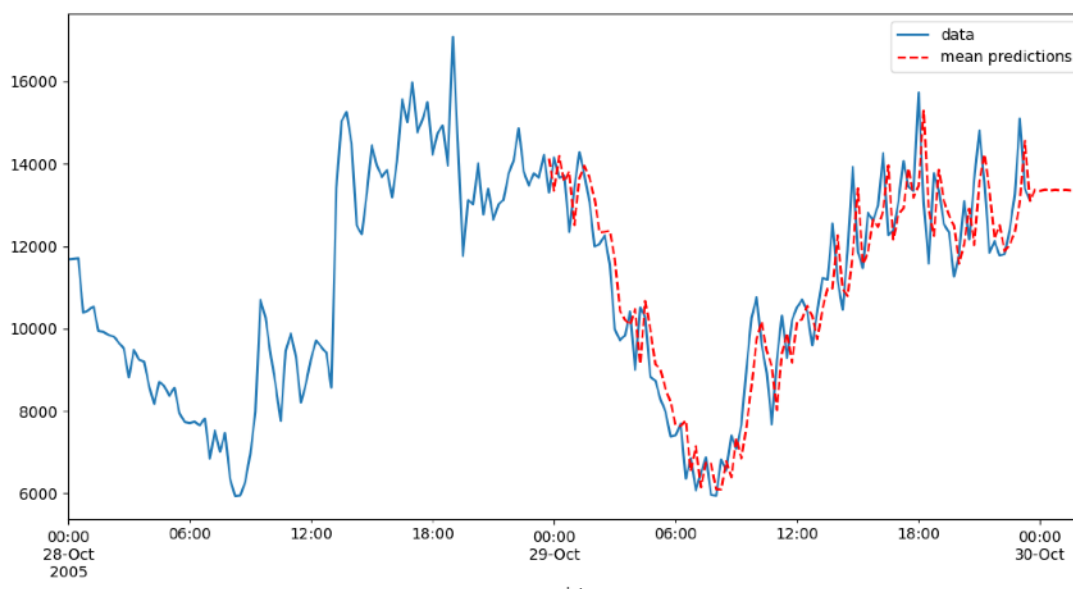


Рисунок 23 - График усредненного прогнозирования

По результатам наших тестов видно, что полученный прогноз получился более точным, по сравнению с прогнозом модели $ARIMA(2, 1, 3)$, но он уступает в точности прогнозу модели $ARMIA(0, 1, 2)$. Значит для дальнейшего прогнозирования следует использовать именно вторую модель.

Заключение

В ходе выполнения данной выпускной квалификационной работы была построена модель для прогнозирования трафика инфокоммуникационных сетей на основе статистической модели $ARIMA$. Данная модель позволяет строить точный и интервальный прогноз на L шагов вперед. Используемые для оценки модели тесты дают представления о точности модели.

Полученную модель мы проанализировали на трех разных дампах трафика, оценили точность их прогнозирования. В среднем погрешность

составил 6%. Я считаю, что данная погрешность не велика, с учетом того, что мы используем данное прогнозирование для предотвращения снижения пропускной способности инфокоммуникационной сети. Эта погрешность может уменьшиться, если для обучения модели использовать большие объемы данных одной сети.

Для подбора параметров модели $ARIMA(p, d, q)$ были использованы два различных метода. В первом варианте применяются статистические тесты. Для подбора коэффициента был использован оператор взятия разности временного ряда и тест статистической значимости Дики-Фуллера. А для нахождения коэффициентов p и q применялись тесты автокорреляционной функции и частной автокорреляционной функции. Для подбора параметров этим методом необходимо вручную проверять результаты статистических тестов, чтобы избежать лишних ошибок.

Второй способ не требует ручного анализа, так как программно перебираются разные параметры для модели в заданном диапазоне и оценивается по информационному критерию Акаике. В ходе анализа дампа трафика произошла ситуация, в которой более эффективной моделью оказалась та, для которой параметры были подобраны полным перебором. Тем не менее ее точность была выше лишь на $\sim 0.03\%$, а времени на подбор параметров было потрачено значительно больше.

После исследования прогнозов трех разных временных рядов можно сделать вывод, что первый метод подбора параметров требует ручного анализа, но на основании полученных коэффициентов мы строим более точную модель прогнозирования. Второй способ также позволяет найти параметры для модели, но с меньшей точностью. Также можно сделать вывод, что использование прогнозов моделей с разными параметрами и их усреднение может помочь уменьшить погрешность дальнейших прогнозов. Так произошло в 2-х из 3-х случаев. Но для точной оценки необходимо анализировать результаты тестов

квадратичной и средней абсолютной процентной ошибок, для выбора модели, с которой стоит работать.

Библиографический список

1. Дуброва, Т. А. Статистические методы прогнозирования.
2. Канторович Г.Г. Анализ временных рядов
3. Крюков, Ю. А. Исследование самоподобия трафика высокоскоростного канала передачи пакетных данных.
4. Rutka G. Network Traffic Prediction using ARIMA and Neural Networks Models
5. ARIMA – модель прогнозирования значений трафика Ю.А. Крюков, Д.В. Чернягин URL: http://www.isa.ru/jitcs/images/documents/2011-02/41_49.pdf
6. Дэвид Шпигельхалтер , Искусство статистики.
7. Q-тест Льюнг-Бокса URL: https://ru.wikipedia.org/wiki/Q-тест_Льюнг_Бокса
8. Квадратичная ошибка URL: <https://wiki.loginom.ru/articles/quadratic-error.html>
9. Библиотеки для работы с Data Science URL: <https://datastart.ru/blog/read/top-10-bibliotek-python-dlya-data-science>
10. Марк Л. Изучаем Python. Том 1.
11. ARIMA Forecasting in Python URL <https://towardsdatascience.com/arima-forecasting-in-python-90d36c2246d3>
12. Модель ARIMA в Python для прогнозирования временных рядов URL <https://pythonpip.ru/examples/model-arima-v-python>
13. Stastmodels info URL <https://www.statsmodels.org/stable/index.html>
14. Прогнозирование временных рядов с помощью ARIMA в Python 3 URL <https://www.8host.com/blog/prognozirovanie-vremennyx-ryadov-s-pomoshhyu-arima-v-python-3/>
15. Autoregressive integrated moving average URL https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average