

Министерство цифрового развития, связи и
массовых коммуникаций Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего
образования «Сибирский государственный университет телекоммуникаций и
информатики» (СибГУТИ)

Отчёт
по лабораторной работе №5
по дисциплине «**Прикладные задачи теории вероятностей**»

Выполнил:
студент гр. ИС-142
«__» декабря 2023 г.

/Григорьев Ю.В./

Проверил:
профессор кафедры В.С.,
«__» декабря 2023 г.

/Родионов А.С./

Оценка « _____ »

Новосибирск 2023

ВЫПОЛНЕНИЕ РАБОТЫ

Цель работы: провести анализ временного ряда для входного интернет-трафика и спрогнозировать различными методами следующие 5 точек в этом ряду.

Визуализация исходных данных



Для построения прогнозов я выбрал следующие методы:

1. Простое скользящее среднее (Simple Moving Average, SMA)

Статистический метод, используемый для сглаживания временного ряда путем создания средних значений за определенный период времени. Особенно эффективен для устранения краткосрочных колебаний и выявления долгосрочных трендов.

Используется для гладких временных рядов без сильных трендов или сезонности. Может отставать от текущих данных, особенно при больших размерах окна.

2. Экспоненциальное сглаживание

Метод прогнозирования, который также используется для сглаживания временных рядов. В отличие от SMA, он придает больший вес более недавним данным, что делает его более реактивным к изменениям.

Подходит для данных с некоторыми изменениями во времени, но без четко выраженной сезонности или тренда. Необходима настройка параметра α .

3. ARIMA (Авторегрессионная интегрированная скользящая средняя)

Более сложный статистический метод, который комбинирует авторегрессию (AR), интегрирование (I) и скользящие средние (MA) для прогнозирования будущих значений на основе прошлых данных.

Компоненты:

- AR (p): Авторегрессия, модель, которая использует зависимость между наблюдением и некоторым количеством лаговых наблюдений (количество лагов).
- I (d): Интегрирование, разность последовательных наблюдений для того, чтобы сделать временной ряд стационарным (степень дифференцирования для достижения стационарности).
- MA (q): Скользящее среднее, модель, которая использует зависимость между наблюдением и остаточными ошибками прогнозов (количество лагов ошибок прогноза MA-модели, window).

Эффективен для анализа временных рядов с трендами, сезонностью и другими сложностями. Требуется тщательного выбора параметров модели и предварительного анализа данных на стационарность.

Получение прогнозов

Используя Python-программу и библиотеки **pandas**, **statsmodels** и **matplotlib**, мной была написана программа для вычисления и визуализации прогнозов на основе предложенных моделей прогнозирования.

```
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.tsa.api import ARIMA, SimpleExpSmoothing

# Чтение данных
file_path = 'all-in.txt'
data = pd.read_csv(file_path, header=None)
data_series = data[0]

# Подготовка данных
train_data = data_series[:560]
test_data = data_series[560:565]
```

```

# Простое скользящее среднее (SMA)
window_size = 5
sma_560 = train_data.rolling(window_size).mean()
sma_predictions = sma_560.iloc[-1]

# Экспоненциальное сглаживание
exp_model_560 = SimpleExpSmoothing(train_data).fit(smoothing_level=1.1)
exp_predictions_560 = exp_model_560.fittedvalues
exp_forecast = exp_model_560.forecast(5)

# ARIMA
p = 5
d = 2
q = 5
arima_model_560 = ARIMA(train_data, order=(p,d,q)).fit()
arima_predictions_560 = arima_model_560.predict(start=1, end=len(train_data))
arima_forecast = arima_model_560.forecast(5)

# Расширение серий для включения прогнозов
sma_extended = pd.concat([sma_560, pd.Series([sma_predictions] * 5, index=range(560, 565))])
exp_extended = pd.concat([exp_predictions_560, exp_forecast])
arima_extended = pd.concat([arima_predictions_560, arima_forecast])

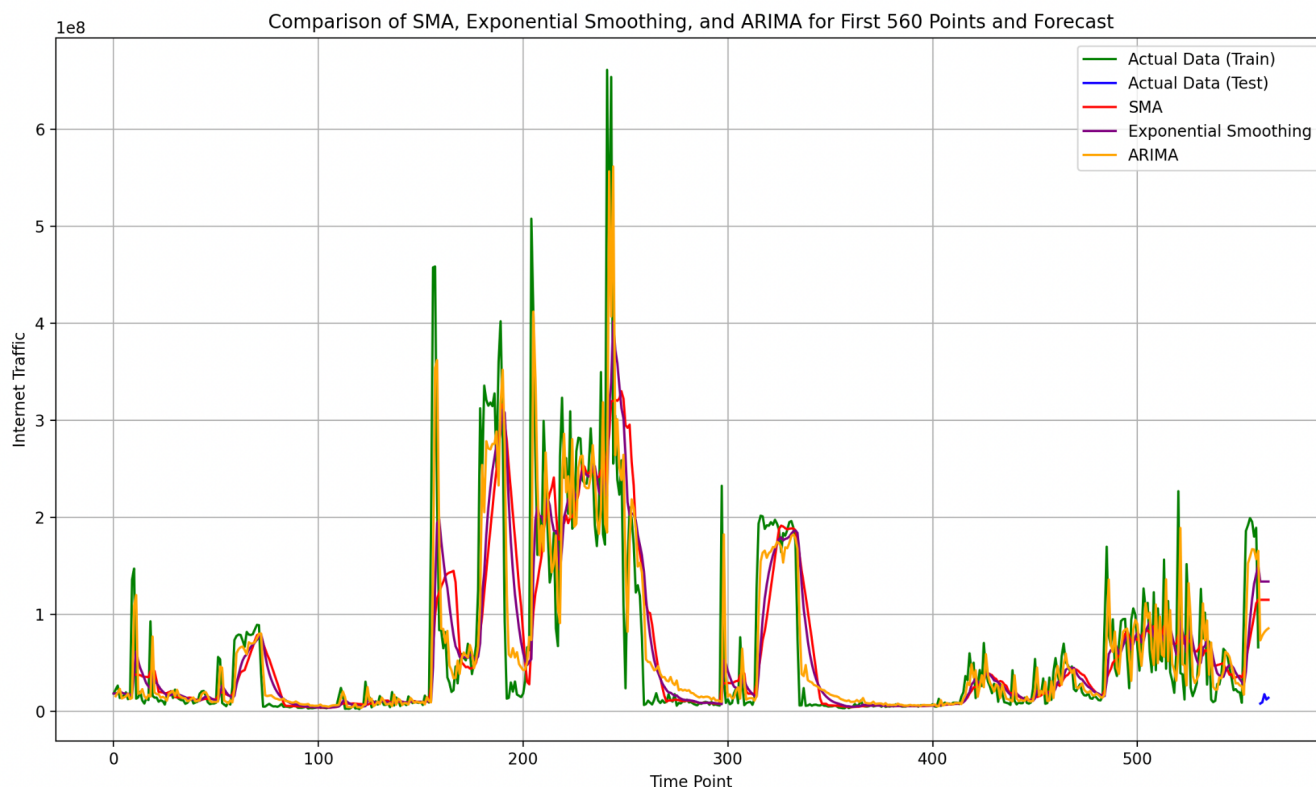
# Построение графика
plt.figure(figsize=(15, 8))
plt.plot(train_data, label='Actual Data (Train)', color='green')
plt.plot(test_data, label='Actual Data (Test)', color='blue')
plt.plot(sma_extended, label='SMA', color='red')
plt.plot(exp_extended, label='Exponential Smoothing', color='purple')
plt.plot(arima_extended, label='ARIMA', color='orange')
plt.title('Comparison of SMA, Exponential Smoothing, and ARIMA for First 560 Points and Forecast')
plt.xlabel('Time Point')
plt.ylabel('Internet Traffic')
plt.legend()
plt.grid(True)
plt.show()

```

Посмотрим, что будет происходить с прогнозами при разных значениях параметров в предложенных методах.

Итерация 1. (basic)

(параметры - SMA_window = 5, EXP_smooth(alpha) = 0.2, ARIMA = 1(p),1(d),1(q))



На представленном графике отображены результаты прогнозирования для первых 560 точек входящего интернет-трафика, используя методы Простого скользящего среднего (SMA), Экспоненциального сглаживания и ARIMA, а также прогнозы для следующих 5 точек.

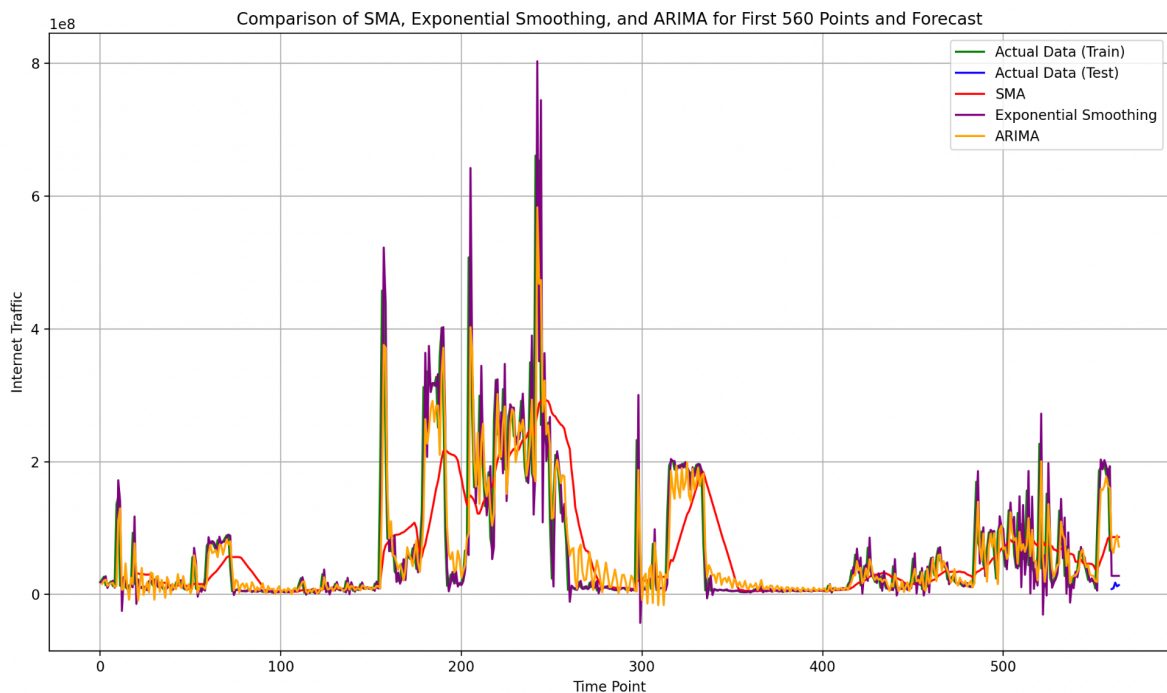
- Зеленая линия показывает реальные данные обучающей выборки (560 точек).
- Синяя линия - фактические значения тестовой выборки (5 точек).
- Красная линия - прогнозы SMA, которые являются сглаженными значениями для первых 560 точек, дополненными одинаковыми прогнозными значениями.
- Фиолетовая линия - прогнозы, сделанные с помощью метода экспоненциального сглаживания, как для обучающих, так и для прогнозируемых точек.
- Оранжевая линия - прогнозы, полученные с помощью ARIMA, также для обучающих и прогнозируемых точек.

В данный момент графики ведут себя достаточно предсказуемо в области реальных данных, но на этапе прогноза ведут себя как константные функции

(за исключением ARIMA, которая показывает линейный рост). Промежуточные выводы - экспоненциальное сглаживание ведет себя слишком «гладко», требуется изменение параметра EXP_smooth для наблюдений за поведением прогноза.

Итерация 2 (aggressive)

(SMA_window = 20, EXP_smooth = 1.3, ARIMA=5,1,5)



Экспоненциальное сглаживание начало вести себя более хаотично на пиках (и верхних, и нижних), но тем не менее дало более верный прогноз за счет поддержки нисходящего тренда графика.

ARIMA начало вести себя более хаотично на ровных участках, и это также заметно на прогнозе. То, что в прогнозируемое время реальные данные показывают похожий пик, связано с более явной автокорреляцией (авторегрессией) с пиками несколько лагов назад.

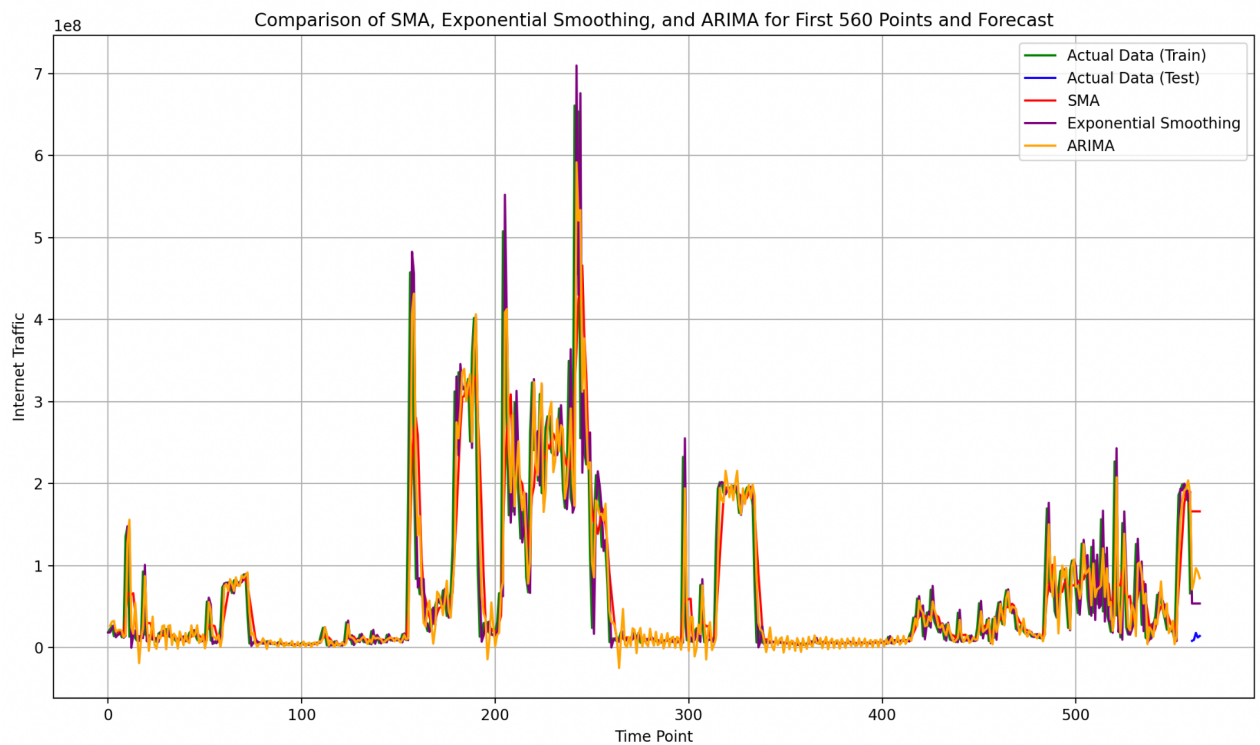
Простое скользящее среднее из-за повышения «окна» лага начало гораздо больше запаздывать и показывать похожий результат при прогнозе.

Промежуточные выводы - параметр window у SMA лучше продолжать держать ближе к 5, коэф. экспоненциального сглаживания - ближе к 1.0, параметры ARIMA - степень дифференцирования оставлять такую же (или пробовать повысить), менять окна (лаги) авторегрессии и скользящего среднего).

Примечание: после тестирования многих значений параметров **p** и **q** у модели ARIMA, не было выявлено сильной закономерности в построении прогноза от этих параметров.

Итерация 3. (final)

(SMA_window = 5, EXP_smooth = 1.0, ARIMA=5,2,5)



Выводы

Простое скользящее среднее (SMA) предсказывает одинаковое значение для всех пяти точек, что указывает на его ограниченность в адаптации к изменениям в данных. Чем больше окно скользящего среднего, тем, условно, больший отрезок оно усредняет, но из-за этого мы теряем в точности времени предсказания. На низких значениях окна оно показывает среднее значение «пика», с которым мы столкнулись в данных. **Плохой прогноз.**

Экспоненциальное сглаживание показало более гибкие прогнозы за счет чувствительности к резкому изменению данных и, вследствие чего, поддержки нисходящего тренда графика после заданного отрезка. **Самый точный, но немасштабируемый прогноз.**

ARIMA адаптируется к последним изменениям в данных и учитывает потенциальные тренды и сезонность. В данном случае этим методом был получен дополнительный пик после главного, который хоть и не приблизился к прогнозу экспоненциального сглаживания, но показал всю силу авторегрессии. **Не лучший, но масштабируемый прогноз.**

Исходя из проведенного анализа, можно сделать вывод, что для прогнозирования временного ряда входящего интернет-трафика модель экспоненциального сглаживания показала наибольшую точность, однако выбор метода зависит от конкретных характеристик временного ряда и целей анализа. В долгосрочной перспективе метод ARIMA будет показывать лучшие результаты из-за того, что тот опирается на периодичность данных.