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 №3
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 данной работы - исследовать и анализировать автокорреляционные функции для данных входного и выходного трафика. Основными задачами являются:

1. **Определение Внутренней Структуры Данных**: Используя автокорреляционные функции, мы стремимся выявить внутреннюю структуру данных, исследуя, как каждое значение временного ряда связано с его предыдущими значениями.

2. **Поиск Периодичности**: Особый интерес представляет определение наличия или отсутствия периодичности в данных. Это достигается путем анализа повторяющихся паттернов в автокорреляционных функциях на различных лагах.

3. **Практическое Применение**: Результаты этого анализа могут быть использованы для улучшения понимания поведения трафика, что в свою очередь может способствовать более эффективному планированию сетевой инфраструктуры и управлению трафиком.

4. **Разработка Методологии**: Разработка и демонстрация методологии анализа временных рядов, которая может быть применена для подобных данных в будущем.

Целью данной работы является не только идентификация текущих характеристик данных трафика, но и обеспечение понимания того, как эти данные могут быть интерпретированы и использованы для прогнозирования будущих трендов и планирования ресурсов.

Автокорреляционная Функция (ACF)

В анализе временных рядов автокорреляционная функция показывает степень линейной статистической связи между значениями временного ряда. Численно, автокорреляционная функция представляет собой последовательность коэффициентов корреляции между исходным рядом, и его копией, сдвинутой на заданное число интервалов ряда (это число называется лагом L):

$$f(L) = \sum_{L=0}^{n} r_{t,t-L},$$

где n — число членов (уровней) временного ряда, r — коэффициент корреляции.

При увеличении лага количестве элементов ряда, для которых вычисляется коэффициент корреляции, уменьшается. На практике, максимальный лаг не должен превышать четверти длины ряда, т.е. должно выполняться соотношение $L_{max} \leq n / 4$.

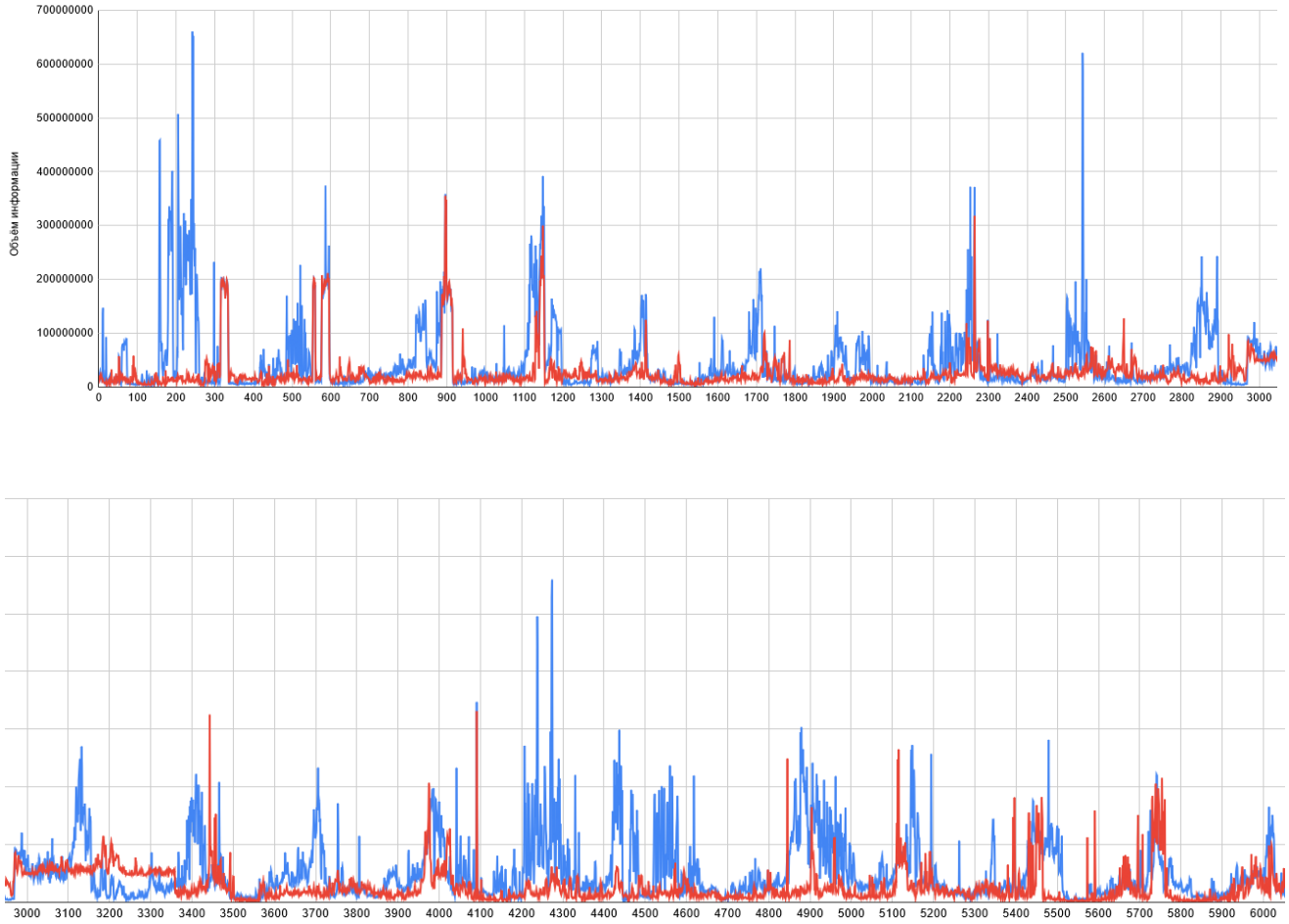
График зависимости коэффициента корреляции от лага называется коррелограммой. Она является эффективным инструментом исследования динамических свойств временных рядов.

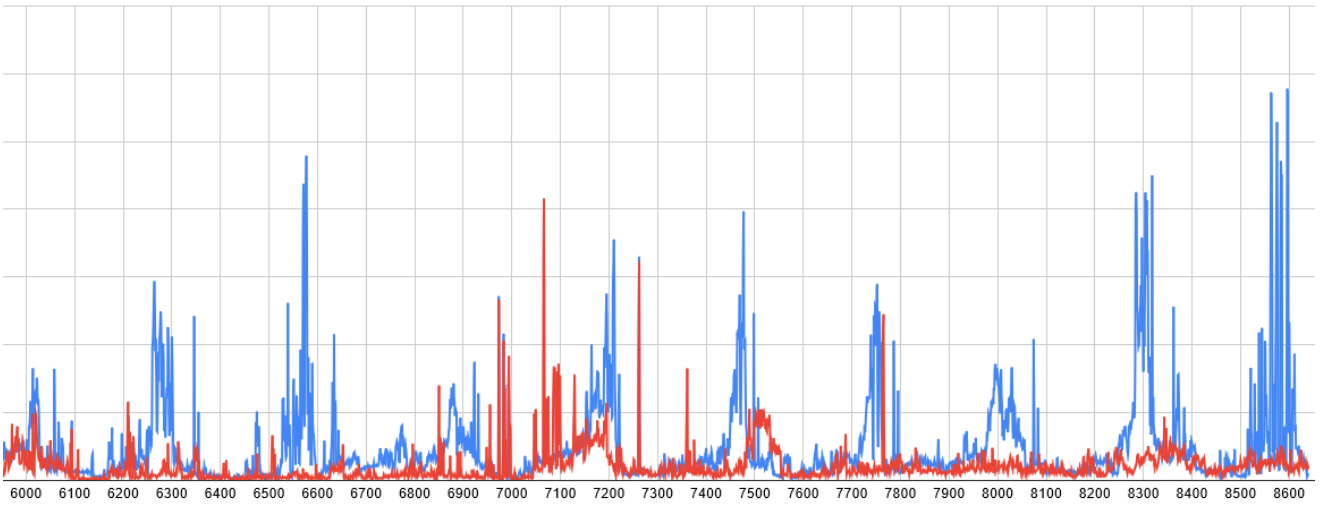
Корреляционная функция максимальна при L=0, когда ряд полностью коррелирован сам с собой. Поскольку значение коэффициента автокорреляция меняется от -1 до 1, то АКФ может принимать значения только из этого же диапазона.

Основное отличие временного ряда от случайного набора значений, зафиксированных в заданные, равноотстоящие моменты времени, является то, что значения членов ряда являются статистически зависимыми. Степень этой зависимости и определяется парным коэффициентом автокорреляции временного ряда.

Если на ряд действуют какие-либо долговременные внешние факторы, то это приводит к появлению в ряде трендов (тенденций) и циклической компоненты, которые и позволяют обнаруживать АКФ. Если выраженный максимум коррелограммы АКФ оказывается для лага L = k, то временной ряд содержит циклическую компоненту с периодом k.

Визуализация исходных данных





(входной трафик - синяя линия, выходной - красная)
(масштаб уменьшен для наглядности размера входных данных)

Шаги Расчета Автокорреляционных Функций

1. **Загрузка Данных**: Данные были загружены из текстовых файлов, содержащих временные ряды. Это было сделано с использованием функции `pd.read_csv` из библиотеки Pandas. Данные были загружены без заголовков, так как предполагается, что файлы содержат только числовые значения.

2. **Использование Statsmodels для ACF**: Для расчета и визуализации автокорреляционных функций была использована функция `plot_acf` из библиотеки Statsmodels. Эта функция автоматически вычисляет автокорреляционную функцию для временного ряда и отображает ее в виде графика.

3. **Настройка Параметров ACF**:
   - В качестве входных данных для `plot_acf` были переданы временные ряды из каждого файла.
   - Параметр `lags` был установлен на 40. Это означает, что функция рассчитывает автокорреляции для 40 задержек (lags). Выбор количества лагов зависит от длины временного ряда и интересующих временных интервалов.
   - Для каждого временного ряда был создан отдельный график.

4. **Визуализация**: Графики были нарисованы с использованием Matplotlib. Они показывают степень автокорреляции для различных лагов. Высокие значения в вертикальной оси указывают на сильную автокорреляцию на соответствующем лаге.

Приложение: исходный код программы на Python

```
import numpy as np
```

```
import matplotlib.pyplot as plt
from statsmodels.tsa.stattools import acf

# read the contents of the uploaded files to understand the data format
with open('all-in.txt', 'r') as file_in, open('all-out.txt', 'r') as
file_out:
    data_in = file_in.readlines()
    data_out = file_out.readlines()

# Displaying the first few lines of each file to understand their structure
data_in[:5], data_out[:5]

# Converting the data into numpy arrays
data_in_array = np.array([int(line.strip()) for line in data_in])
data_out_array = np.array([int(line.strip()) for line in data_out])

# Calculating the autocorrelation function for both series up to a lag of 2000
acf_in = acf(data_in_array, nlags=2000, fft=True)
acf_out = acf(data_out_array, nlags=2000, fft=True)

# Plotting the autocorrelation functions
plt.figure(figsize=(15, 6))

# Autocorrelation plot for input data
plt.subplot(1, 2, 1)
plt.plot(acf_in, marker='o', linestyle='-', color='blue')
plt.title('Autocorrelation Function - Input Data')
plt.xlabel('Lag')
plt.ylabel('Autocorrelation')
plt.xlim(0, 2000)

# Autocorrelation plot for output data
plt.subplot(1, 2, 2)
plt.plot(acf_out, marker='o', linestyle='-', color='red')
plt.title('Autocorrelation Function - Output Data')
plt.xlabel('Lag')
plt.ylabel('Autocorrelation')
plt.xlim(0, 2000)

plt.tight_layout()
plt.show()
```

Ключевые Замечания по Расчету ACF

- **ACF и Значения Лагов**: ACF измеряет, насколько хорошо значение временного ряда соответствует своему собственному предыдущему значению на каждом лаге. Например, ACF с лагом 1 измеряет корреляцию между каждым значением и его непосредственно предшествующим значением.

- **Интерпретация Графиков**: На графиках ACF, близкие к 1 значения автокорреляции указывают на сильную положительную корреляцию, значения около -1 указывают на сильную отрицательную корреляцию, а значения около 0 означают отсутствие корреляции.

- **Периодичность**: Периодичность в данных может проявляться через регулярные пики на графике ACF. Например, если вы замечаете пики на каждом 5-м лаге, это может указывать на 5-периодичный цикл в данных.

Определение Периодичности

Для более точного определения наличия периодичности в данных, необходимо рассмотреть следующие аспекты:

- **Расположение Пиков**: Если пики на графике ACF расположены через равные интервалы, это является хорошим индикатором периодичности. Например, если пики наблюдаются каждые 10 лагов, это может указывать на 10-периодический цикл в данных.

- **Высота и Затухание Пиков**: Периодические данные обычно показывают высокие пики на начальных лагах, которые постепенно уменьшаются с увеличением лага. Если пики остаются высокими и стабильными на больших лагах, это может указывать на более сложную структуру временного ряда.

Результаты

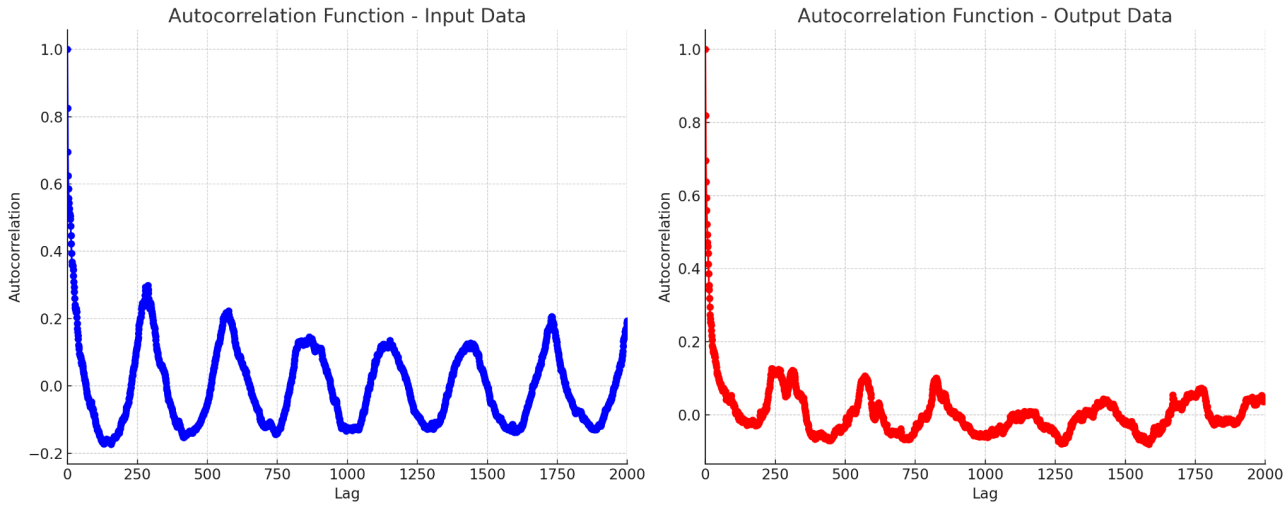


Рисунок 1 - Полученные автокорреляционные функции до лага в 2000 ед. (n / 4)

На основании вышеуказанных графиков автокорреляционных функций (ACF) для данных входного и выходного трафика, можно сделать следующие выводы:

1. **Автокорреляционная Функция для Входного Трафика**: На графике ACF для входного трафика видны пики на некоторых лагах. Это может указывать на наличие периодичности или зависимости данных от их предыдущих значений около 300/600/900 единиц измерения назад.

2. **Автокорреляционная Функция для Выходного Трафика**: Похожая картина наблюдается и для выходного трафика с пиками на лагах в 300/600/800 ед.

Выводы

1. **Возможная периодичность в 24 часа (1 день)**: Анализ автокорреляционных функций (ACF) показывает наличие ярко выраженной периодичности в данных интернет-трафика с циклом в 300 единиц (максимальный пик достигается именно на этом значении лага), что скорее всего соответствует периоду в 24 часа (1 день). Такие временные зависимости могут быть связаны с дневной и ночной активностью пользователей.

2. **Убывающая корреляция**: С увеличением лага корреляция немного уменьшается. Это означает, что данные на более отдаленных временных интервалах немного меньше зависят друг от друга, что типично для многих временных рядов, однако, если мы предполагаем, что эти пики значат интервалы в 2,3,4… дней, то вполне закономерно, что они не будут сильно убывать.

3. **Отсутствие сильной корреляции на других лагах**: Хотя корреляция на лаге 300 наибольшая, на других лагах (600 и 900) корреляция значительно меньше. Это может свидетельствовать о том, что влияние предыдущих данных присутствует в течение примерно половины дня (12 часов), но становится менее существенным на более длинных временных интервалах.

4. **Полезность анализа ACF**: Анализ автокорреляционных функций (ACF) позволяет выявлять периодичность и зависимости во временных рядах данных. В данном случае, он подтверждает наличие дневной периодичности в данных интернет-трафика, что может быть важной информацией при принятии решений и прогнозировании поведения пользователей или проведения профилактических работ на серверах интернет-провайдера.

В целом, проделанная работа позволяет лучше понять структуру данных интернет-трафика и выявить периодичность, что может быть полезно для оптимизации и принятия управленческих решений.