Министерство науки и высшего образования

Российской Федерации

Федеральное государственное бюджетное образовательное

учреждение высшего образования

«НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Кафедра Теоретической и Прикладной Информатики

Лабораторная работа №1 по предмету

«**КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ»**

Факультет: прикладной математики и информатики

Группа: ПМИ-12

Бригада: 10

Студенты: Швадченко А. В.

Субботин Д. А.

Преподаватели: Черникова О.С.

Карманов В.С.

Новосибирск, 2024

**Исходные данные к лабораторной работе**

Временной ряд  – полугодовой курс Шведской кроны 26.03.2024-26.09.2024.

**Цель лабораторной работы**:

Предварительная обработка и анализ данных.

**Задание лабораторной работы**:

1. Удалить выбросы методом LOF

2. Фильтрация шума:

a. Вычислить статистику Бокса-Пирса

b. Очистить данные от шума алгоритмом Эллиптического фильтра

3. Сглаживание значений ряда алгоритмом экспоненциального скользящего среднего

4. Проверить ряды на стационарность до и после выполнения выше перечисленных пунктов с помощью тестов ADF и KPSS

5. Проверить ряды на наличие тренда методом Валлиса и Мура

**Ход работы**

**1.** **Удаление выбросов.** Выявить выбросы и восстановить значения. Используемый метод: LOF.

Выбросы – это наблюдения, которые значительно отличаются от остальных данных. Локальный коэффициент выброса основан на концепции локальной плотности, где местоположение задается через k ближайших соседей, расстояние до которых используется для оценки плотности. Сравнивая локальную плотность объекта с локальными плотностями его соседей, можно идентифицировать области с аналогичной плотностью и точки, которые имеют существенно меньшую плотность, чем их соседи. Они считаются выбросами.

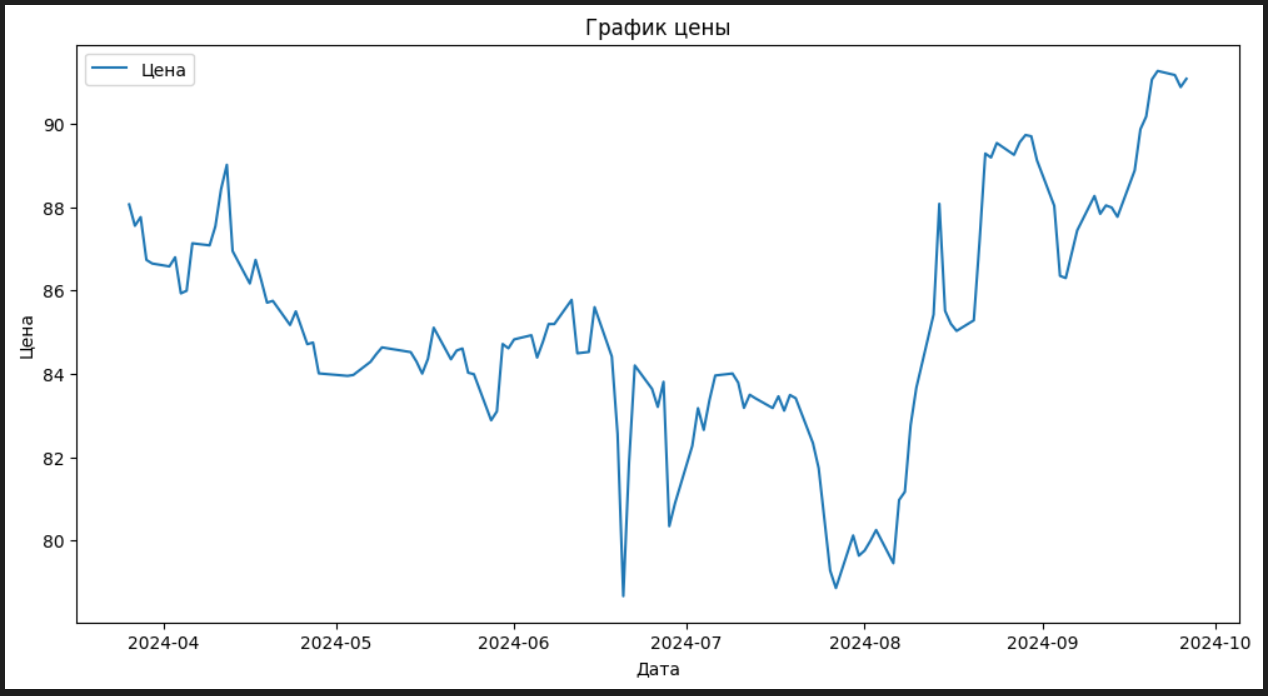
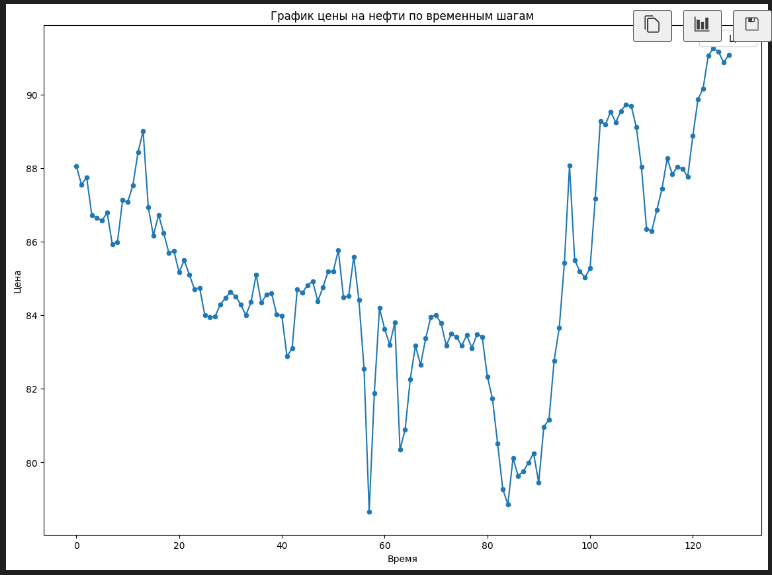


Рис. 2. График выбросов временного ряда

После обнаружения выбросов их значения нужно восстановить. Мы воспользовались функцией KNNimputer. KNNImputer работает путем поиска k ближайших соседей (на основе заданной метрики расстояния) для точек данных с отсутствующими значениями. Затем он вычисляет недостающие значения, используя среднее значение соседних точек данных. Ключевое преимущество этого подхода заключается в том, что он сохраняет взаимосвязи между функциями, что может привести к повышению производительности модели.



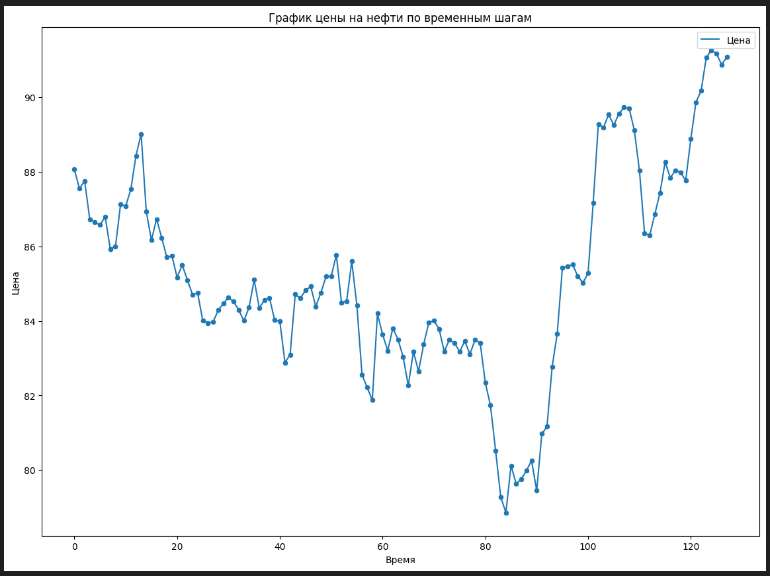


Рис. 3. График сравнения исходного и обработанного временных рядов

**2.1.** **Фильтрация шума.** Вычислить статистику Бокса-Пирса для тестирования на наличие белого шума.

Статистика Бокса-Пирса (Box-Pierce test) – это критерий для проверки гипотезы о том, что временной ряд является белым шумом. Белый шум – это временной ряд, в котором значения не коррелированы между собой, и их среднее значение равно нулю. Нулевая гипотеза: временной ряд является белым шумом. Альтернативная гипотеза: временной ряд не является белым шумом.

Данная статистика используется для проверки белого шума в данных и рассчитывается по формуле:

,

где *n* – количество наблюдений; *m* – количество лагов (временных отступов), для которых вычисляются автокорреляции; – выборочная автокорреляция для лага *k*.

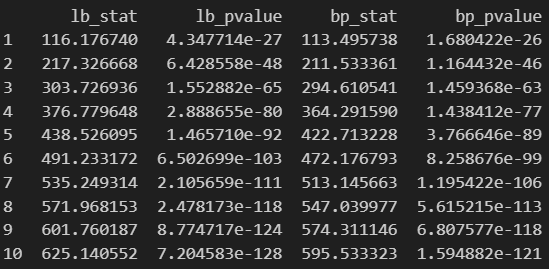
Выборочная автокорреляция – это мера корреляции между временным рядом и его сдвигом на определенный временной лаг. Формула для выборочной автокорреляции на лаге выглядит следующим образом:

,

где ** – значение временного ряда в момент времени *t*;  – среднее значение временного ряда; *n* – количество наблюдений во временном ряде.

Статистика *Q* имеет распределение  с *m* степенями свободы. Если *Q* превышает критическое значение  для выбранного уровня значимости, то нулевая гипотеза отвергается, и ряд не является белым шумом.

Критический уровень для данного теста: 0.05.



bp\_stat – значения статистики Бокса-Пирса для каждого лага и bp\_ pvalue – p-values для статистики Бокса-Пирса.

По результатам тестирования мы видим, что для всех лагов bp\_value < 0.05, следовательно мы можем утверждать, что на каждом лаге есть автокорреляция и наш временной ряд не является белым шумом.

**2.2. Фильтрация шума.** Очистить данные от шума следующим алгоритмом Эллиптического фильтра.

Эллиптические фильтры — это эффективные цифровые фильтры, которые используют для фильтрации шума и разделения частотных диапазонов в сигнале.

Алгоритм:

• Настройка параметров:

Порядок фильтра (order): Определяет крутизну перехода между полосами.

Затухание в полосе пропускания (rp): Допустимые колебания амплитуды сигнала в полосе пропускания.

Затухание в полосе подавления (rs): Минимальное затухание, которое должно быть достигнуто в полосе подавления.

Частота среза (cutoff): Определяет границу между полосой пропускания и полосой подавления.

• Применение фильтра к сигналу:

Когда фильтр применяется, он пропускает сигналы в пределах полосы пропускания с минимальными искажениями, одновременно подавляя частоты, выходящие за пределы заданной частоты среза. Мы не будем рассматривать первые несколько значений ряда, так как, из-за запаздывания фазово-частотной характеристики, они принимают неправильные значения.

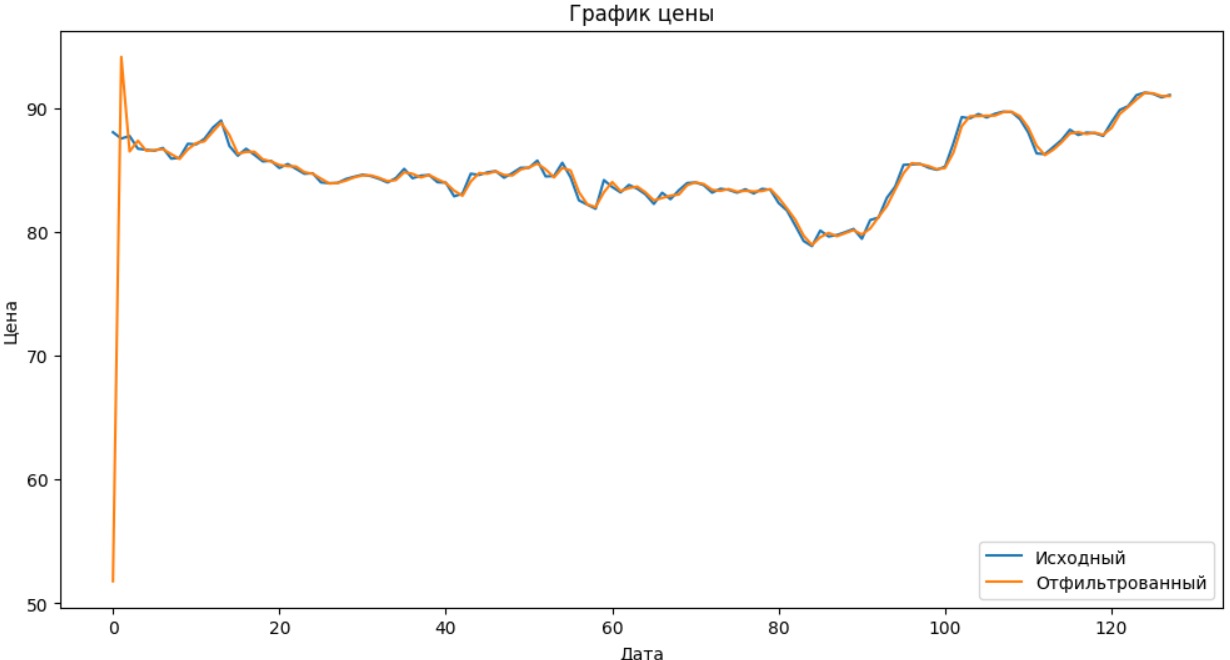
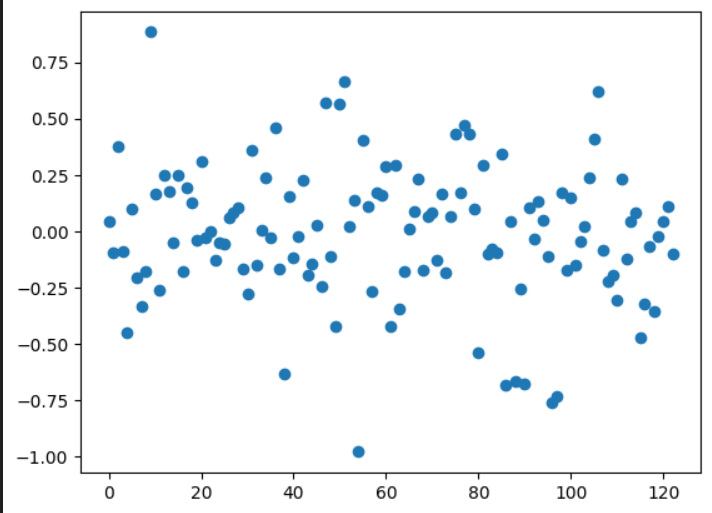
****

График остатков:

****

**3. Сглаживание значений ряда** алгоритмом экспоненциального скользящего среднего.

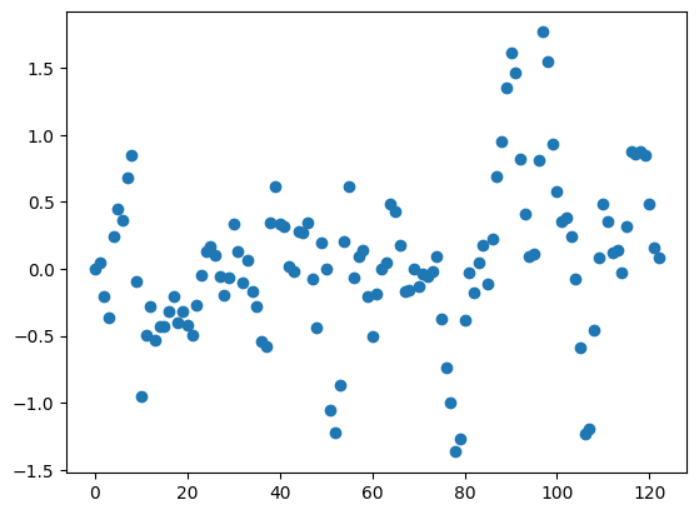
Экспоненциальное скользящее среднее - это статистический расчет, который сглаживает данные о ценах за определенный период. За счет большего внимания к последним ценам и постепенного снижения значимости старых данных EMA быстро адаптируется к изменениям на рынке.

Этот расчет включает отслеживание значений EMA для каждого торгового периода, начиная с начального значения, и применение формулы для непрерывного обновления значений EMA. Коэффициент сглаживания определяет скорость уменьшения весов для каждой последующей точки данных.

Формула EMA учитывает цену закрытия текущего периода, значение EMA с предыдущего дня и коэффициент сглаживания. Включая предыдущее значение EMA, формула гарантирует, что расчет EMA является рекурсивным процессом, где каждое новое значение EMA зависит от предыдущего.



График остатков:



**4.** **Проверить ряды на стационарность** до предварительной обработки (исходные данные) и после выполнения п.1-3, используя тест Дикки-Фуллера (ADF-тест), тест Квятковского–Филлипса–Шмидта–Шина (KPSS- тест).

Временные ряды могут быть стационарными и нестационарными. Временной ряд называется стационарным, если его вероятностные характеристики (математическое ожидание, дисперсия) постоянны. Временной ряд называется нестационарным, если хотя бы одна из вероятностных характеристик непостоянна. Многие часто встречающиеся на практике временные ряды являются нестационарными.

Два разных типа нестационарных по отношению к среднему временных рядов могут быть приведены к стационарному виду с помощью взятия последовательных разностей. Это ряды с детерминированным трендом (TS – trend stationary) и ряды, имеющие стохастический тренд (DS – difference stationary).

Принципиальное различие между этими двумя типами рядов выражается в том, что TS ряд также можно привести к стационарному виду с помощью выделения линейного тренда, тогда как вычитание детерминированной составляющей из DS ряда оставляет его нестационарным

ADF-тест используется для проверки гипотезы о наличии единичного корня во временном ряде. Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд. Нулевая гипотеза состоит в том, что временной ряд имеет единичный корень (то есть нестационарен). Альтернативная гипотеза состоит в том, что ряд стационарен или стационарен вокруг детерминированного тренда.

Вид используемого авторегрессионного уравнения первого порядка:

,

где  - оператор разности первого порядка ,  - константа, ,  - коэффициент, значение которого и проверяет тест (если  = 1, то временной ряд имеет единичный корень),  - значение временного ряда в момент ,  - ошибка в момент .

Для такого вида уравнения существует свой вид критических значений DF-статистики (статистика теста), которые берутся из специальной таблицы Дики-Фуллера. Если значение статистики лежит левее критического значения при данном уровне значимости, то нулевая гипотеза отклоняется и процесс признается стационарным, в противном случае гипотеза не отвергается и процесс может содержать единичные корни.

Как результат, функция возвращает следующие параметры:

ADF стастистика: это значение статистики теста. Чем меньше это значение, тем сильнее аргумент против нулевой гипотезы.

p-value (уровень значимости): это вероятность получения наблюдаемой статистики при условии, что нулевая гипотеза верна. Если p-value меньше выбранного уровня значимости (например, 0.05), то нулевая гипотеза отвергается, и ряд считается стационарным.

Критические значения: это критические значения для различных уровней значимости (1%, 5%, 10%). Если ADF статистика меньше критического значения, то нулевая гипотеза отвергается.

KPSS-тест (Kwiatkowski-Phillips-Schmidt-Shin test) – это статистический тест для проверки стационарности временного ряда. В отличие от ADF-теста, который проверяет наличие единичного корня и нестационарность, KPSS-тест проверяет стационарность ряда вокруг детерминированного тренда. Нулевая гипотеза звучит так: временной ряд стационарен вокруг детерминированного тренда (тренд-стационарен). Альтернативная гипотеза же: временной ряд нестационарен (имеет стохастический тренд).



,

где – детерминированный тренд,  – стохастический тренд, – стационарная ошибка (белый шум).

Нулевая гипотеза:  является TS-рядом.

Альтернативная гипотеза:  не является TS-рядом.

Как результат, функция возвращает следующие параметры:

KPSS статистика: Это значение статистики теста. Чем больше это значение, тем сильнее аргумент против нулевой гипотезы.

p-value: Это вероятность получения наблюдаемой статистики при условии, что нулевая гипотеза верна. Если p-value меньше выбранного уровня значимости (например, 0.05), то нулевая гипотеза отвергается, и ряд считается нестационарным.

Критические значения: Это критические значения для различных уровней значимости (10%, 5%, 2.5%, 1%). Если KPSS Statistic больше критического значения, то нулевая гипотеза отвергается.

(уровень значимости 0.05)

***Проверка на стационарность исходных данных***:

ADF Test:

ADF Statistic: -1.5332934963505378

p-value: 0.516981976206809

Ряд не стационарный.

Critical Values:

1%: -3.482920063655088

5%: -2.884580323367261

10%: -2.5790575441750883

KPSS Test:

KPSS Statistic: 0.3959270278462816

p-value: 0.07891076385936138

Ряд стационарный.

Critical Values:

10%: 0.347

5%: 0.463

2.5%: 0.574

1%: 0.739

***Проверка на стационарность данных после 1-3п.:***

ADF Test:

ADF Statistic: -1.5332934963505378

p-value: 0.516981976206809

Ряд не стационарный.

Critical Values:

1%: -3.482920063655088

5%: -2.884580323367261

10%: -2.5790575441750883

KPSS Test:

KPSS Statistic: 0.3959270278462816

p-value: 0.07891076385936138

Ряд стационарный.

Critical Values:

10%: 0.347

5%: 0.463

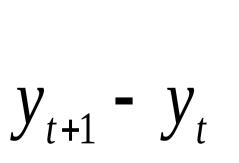
2.5%: 0.574

1%: 0.739

Исходя из полученных результатов мы делаем вывод, что ***обработанный временной ряд не является стационарным***.

**5. Проверить ряды на наличие тренда**. Метод Валлиса и Мура.

Этот критерий позволяет отличить отклонения последовательности уровней ряда от чисто случайной последовательности. Если тренд отсутствует, то знаки разностей значений уровней

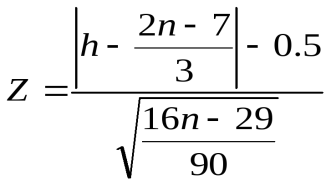
образуют случайную последовательность.

С помощью критерия Валлиса-Мура проверяется гипотеза:

последовательность знаков разностей имеет случайный характер.

Альтернативной к ней является гипотеза:

последовательность знаков разностей значимо отличается от случайной.

 где h – число плюсовых и минусовых разностей

Исходя из полученных результатов мы делаем вывод, что ***обработанный временной имеет тренд***.

**Вывод**

Мы обработали наш исходного временной ряд при помощи LO, применили эллиптический фильтр шума. Для уже обработанного временного ряда методом экспоненциального скользящего среднего сгладили значения. С помощью тестов ADF и KPSS выявили, что и исходный, и обработанный временные ряды не являются стационарными. Методом Валлиса и Мура убедились в наличии тренда.