«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
ИМЕНИ М.В. ЛОМОНОСОВА

Факультет Вычислительной Математики и Кибернетики

Кафедра исследования операций

Отчет по заданию №2

Студенты:	Липатова Александра Смирнов Михаил Разумова Вера
Преподаватель:	Гусева Юлия
Группа:	312

Постановка задачи.

1. Считать данные из **training.xlsx.** Ответы на тестовой выборке **testing.xlsx** не следует использовать ни в каких экспериментах, кроме финального. Проверить является ли ряд стационарным в широком смысле.

2. Разложить временной ряд на **тренд, сезональность, остаток** в соответствии с аддитивной, мультипликативной моделями. Визуализировать их, оценить стационарность получившихся рядов, сделать выводы.

3. Проверить является ли временной ряд интегрированным порядка k. Если является, применить к нему модель **ARIMA**, подобрав необходимые параметры с помощью функции автокорреляции и функции частичной автокорреляции. Выбор параметров обосновать. Отобрать несколько моделей. Предсказать значения для тестовой выборки. Визуализировать их, посчитать r2 score для каждой из моделей. Произвести отбор наилучшей модели с помощью информационного критерия Акаике. Провести анализ получившихся результатов.

Основные понятия.

Под **временным рядом** понимаются последовательно измеренные через некоторые (зачастую равные) промежутки времени данные.

$$y_1, ..., y_T, ..., y_t \in \mathbb{R}$$

Прогнозирование временных рядов заключается в построении модели для предсказания будущих событий, основываясь на известных событиях прошлого.

Анализ временных рядов — совокупность математико-статистических мето- дов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования.

Дисперсия выборки - это среднее арифметическое квадратов отклонений.

Отклонение - это разность числа и некоторой точки отчёта, чаще всего это среднее арифметическое или медиана. Например, если у нас есть следующий ряд чисел: 1; 2; 3; 4; 5; 6; 7, то его среднее арифметическое это сумма чисел ряда, деленная на их количество, то есть $(1 + 2 + 3 + 4 + 5 + 6 + 7) : 4 = 28 : 7 = 4$ (здесь среднее арифметическое набора чисел совпадает с медианой). Тогда найдём отклонения. Они будут соответственно -3; -2; -1; 0; 1; 2; 3. Тогда квадраты отклонений будут 9; 4; 1; 0; 1; 4; 9. Найдём их среднее арифметическое: $(9 + 4 + 1 + 0 + 1 + 4 + 9) : 7 = 28 : 7 = 4$. Получаем, что дисперсия данного набора равна 4.

Количественной характеристикой сходства между значениями ряда в соседних точках является **автокорреляционная функция** (или просто автокорреляция), которая задаётся следующим соотношением:

$$r_\tau = \frac{\mathbb{E}((y_t - \mathbb{E}y)(y_{t+\tau} - \mathbb{E}y))}{\mathbb{D}}$$

Ряд называется **слабо стационарным** или стационарным в широком смысле, если его среднее значение и дисперсия не зависят от времени, а ковариационная функция зависит только от сдвига. Если нарушается хотя бы одно из этих условий, то ряд является нестационарным.

Тест Дики — Фуллера — это методика, которая используется в прикладной статистике и эконометрике для анализа временных рядов для проверки на стационарность. Является одним из тестов на единичные корни (Unit root test).

Временной ряд **имеет единичный корень**, или порядок интеграции один, если его первые разности образуют стационарный ряд.

Как и большинство других видов анализа, анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент. Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей.

Таким образом, каждый уровень временного ряда может формироваться из трендовой T, циклической или сезонной компоненты (S), а также случайной E компоненты.

Модели, где временной ряд представлен в виде суммы перечисленных компонентов называются аддитивными, если в виде произведения – мультипликативными моделями.

Аддитивная модель имеет вид: Y = T + S + E

Мультипликативная модель имеет вид: Y = T * S * E

Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени.

Из этого определения следует, что ряды, в которых присутствует тренд, являются нестационарными: в зависимости от расположения окна изменяется средний уровень ряда. Кроме того, нестационарны ряды с сезонностью: если ширина окна меньше сезонного периода, то распределение ряда будет разным, в зависимости от положения окна.

Сезонность – строго периодические и связанные с календарным периодом отклонения от тренда:

- **Аддитивная сезонность** – амплитуда сезонных колебаний не имеет ярко

выраженной тенденции к изменению во времени.

- **Мультипликативная сезонность** – амплитуда сезонных колебаний имеет

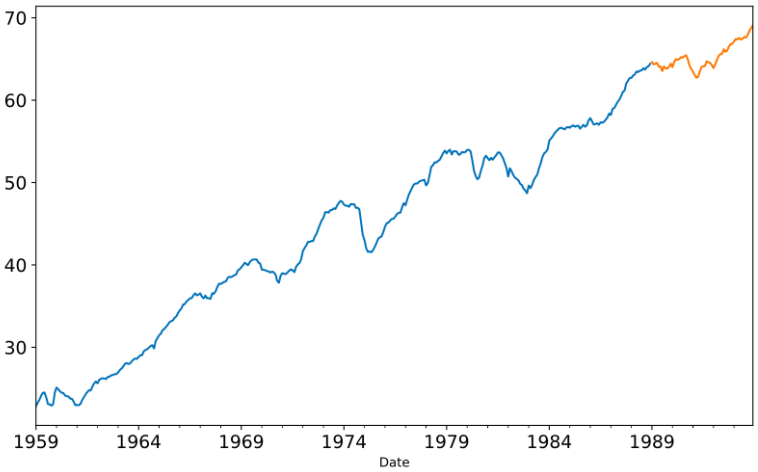
выраженную тенденцию к изменению во времени.

Скользящая статистика — общее название для семейства функций, значения которых в каждой точке определения равны среднему значению исходной функции за предыдущий период. Скользящая статистика обычно используются с данными временных рядов для сглаживания краткосрочных колебаний и выделения основных тенденций или циклов.

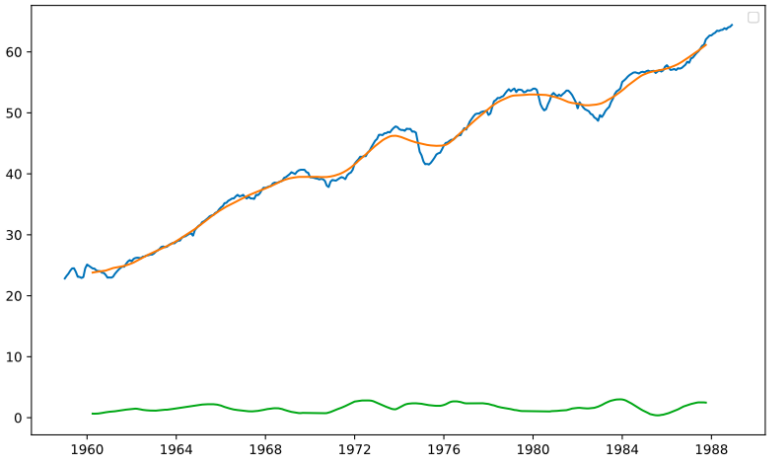
Описание программы.

Цель: провести анализ временного ряда и попробовать предсказать значения для последующих месяцев.

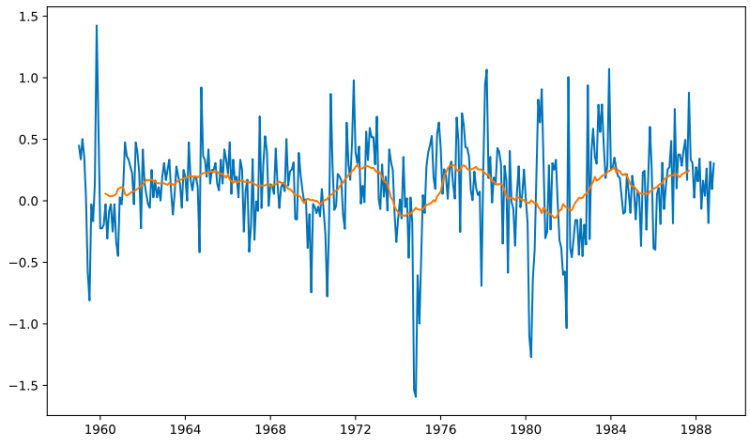
1. С помощью **matplotlib** строится график данного временного ряда.



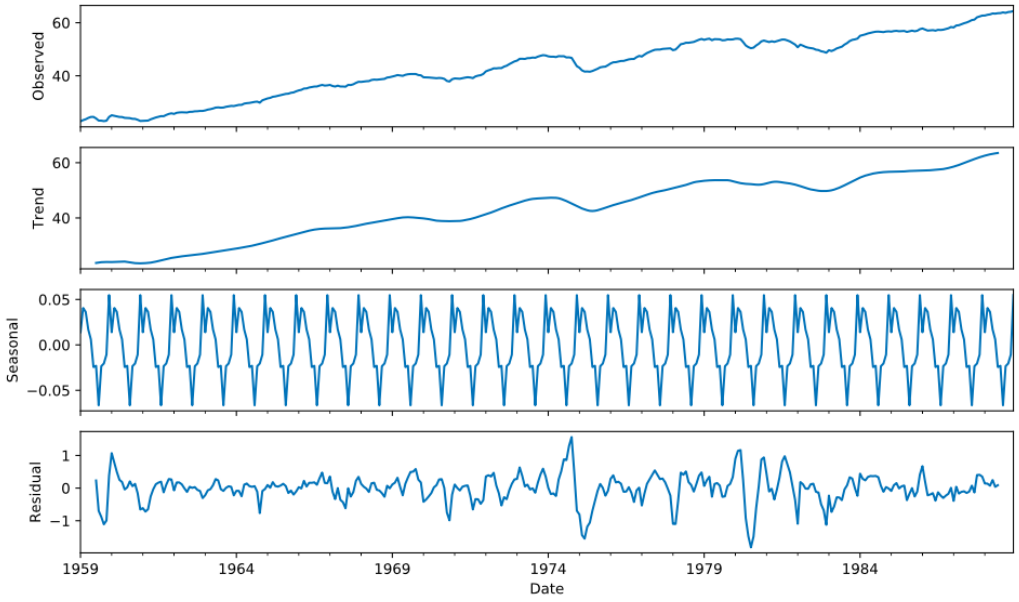
2. Далее осуществляется проверка на стационарность временного ряда. С помощью функций **Series.rolling.mean()** и **Series.rolling.std()** находятся среднее и стандартное отклонения соответственно, строятся их графики вместе с оригинальным.



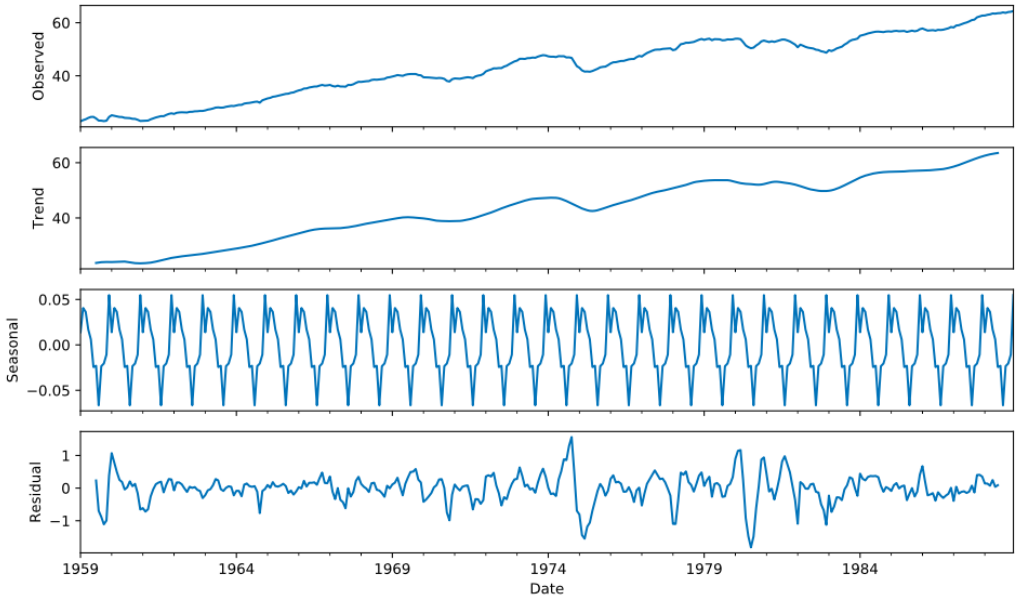
3. Проведем дифференцирование первого порядка.



4. Разлагаем ряд на тренд, сезонность и остаток с помощью функции из модуля **statmodels seasonaldecompose** с параметрами **model = 'additive' и model = 'multiplicate'** для аддитивной и мультипликативной моделей соответственно.Далее идет проверка на стационарность рядов. Наблюдается тренд, что означает, что ряд не является стационарным

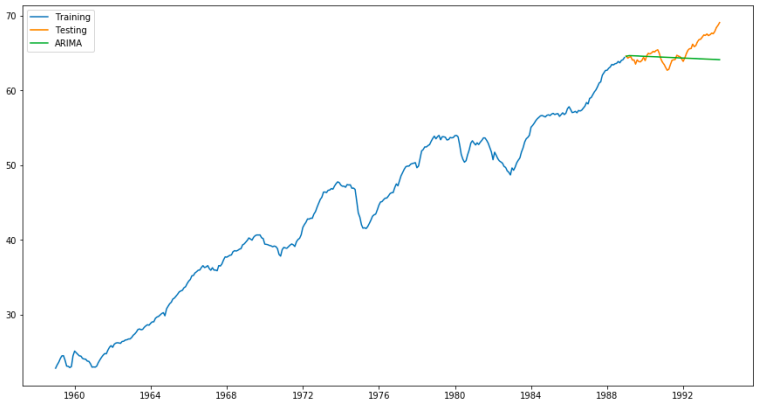


5. Ряд проверяется на интегрируемость порядка k. Поэтому для проверки стационарности проведем **обобщенный тест Дики-Фуллера** на наличие единичных корней. Для этого в **модуле statsmodels** есть функция **adfuller()**. Если проведенный тест подтвердил предположения о не стационарности ряда, для нахождения k берется разность рядов. Если первые разности ряда стационарны, то он называется интегрированным рядом первого порядка. В нашем случае k = 1 и ряд интегрируем, значит с помощью функций автокорреляции и частичной автокорреляции подбираются параметры для модели ARIMA. Для построения модели нужно знать ее порядок, состоящий из 3-х параметров: p — порядок компоненты AR, d — порядок интегрированного ряда, q — порядок компонетны MA.



Параметр d равен 1, осталось определить p и q. Для их определения надо изучить автокорреляционную(ACF) и частично автокорреляционную(PACF) функции для ряда первых разностей. ACF поможет определить q, т. к. по ее коррелограмме можно определить количество автокорреляционных коэффициентов сильно отличных от 0 в модели MA. PACF поможет определить p, т. к. по ее коррелограмме можно определить максимальный номер коэффициента сильно отличного от 0 в модели AR. Чтобы построить соответствующие коррелограммы, в **пакете statsmodels** имеются функции: **plotacf() и plotpacf()**. Они выводят графики **ACF** и **PACF**, у которых по оси X откладываются номера лагов, а по оси Y значения соответствующих функций. В первой модели ACF экспоненциально затухает, начиная с первого лага, причем затухание может носить монотонный или колебательный характер. PACF затухает экспоненциально, монотонно или колебательно. Это означает, что $p = 1$, а $q = 3$. Во второй модели мы ссылаемся на стандартный подход по выбору параметров: q - номер последнего лага, при котором автокорреляция значима, p - номер последнего лага при котором частичная автокорреляция значима. Получаем $p = 12$, $q = 3$. $p = 1$, $q = 4$ выбираем по тому же принципу.

Далее строятся модели ARIMA и осуществляется прогноз. Строится график, на котором изображены данные из файла **testing.xlsx** и построенный про- гноз для каждой из моделей. Для каждой модели считается R2 - коэффициент детерминации, чтобы понять какой процент наблюдений описывает данная модель, и критерий Акаике (AIC), выбирающий наилучшую модель.



Необходимые компоненты.

• **Библиотеки**

- matplotlib - пакет, используемаый для отрисовки графиков

- statsmodels - пакет Python, который позволяет пользователям иссле- довать данные, оценивать статистические модели и выполнять стати- стические тесты. Он дополняет модуль статистики SciPy. Мы исполь- зуем ее для проведения теста Дики-Фуллера, а так же для построения модели ARIMA.

- sklearn - пакет для машиного обучения. Мы используем ее для оценки r2value.

- pandas - библиотека предназначенная для хранения таблиц. Так же содержит огромное количество универсальных функций для их ком- фортной обрабоки.

- pylab - большой универсальный пакет питон. Мы используем для за- дания параметров отрисовки.

• **Программы**

- Jupyter Notebook

Вклад участников.

• Липатова Александра - разложение временного ряда на тренд, сезональность и шум, полная сборка программы,Diki-fuller, составление ReadMe

• Разумова Вера - проверка ряда на стационарность посредством визуализации статистик, полная сборка программы,Diki-fuller, составление ReadMe

• Смирнов Михаил - построение прогнозирующей модели, Diki-fuller, составление ReadMe