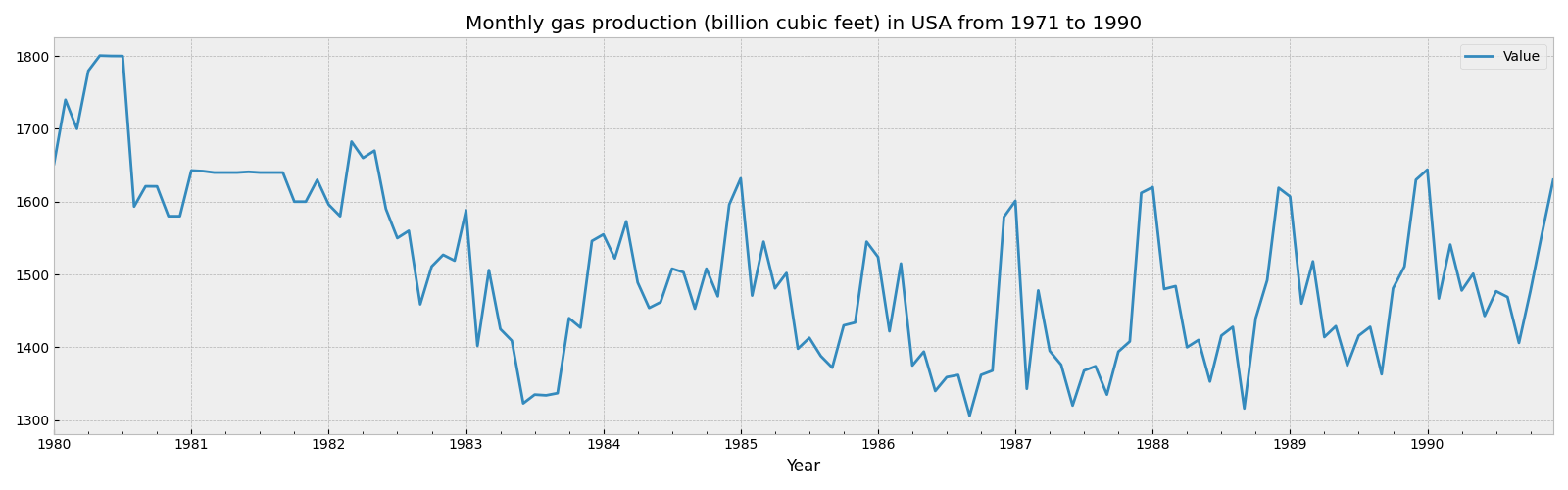
**Анализ временных рядов на примере добычи газа в США с 1980 по 1990 годы**

1. **График соответствующего временного ряда. Графики ACF и PACF (автокорреляционная и частная автокорреляционная функции)**



Предположения:

- График не имеет ярко выраженного убывающего или возрастающего тренда

- Имеет место аддитивный сезонный эффект

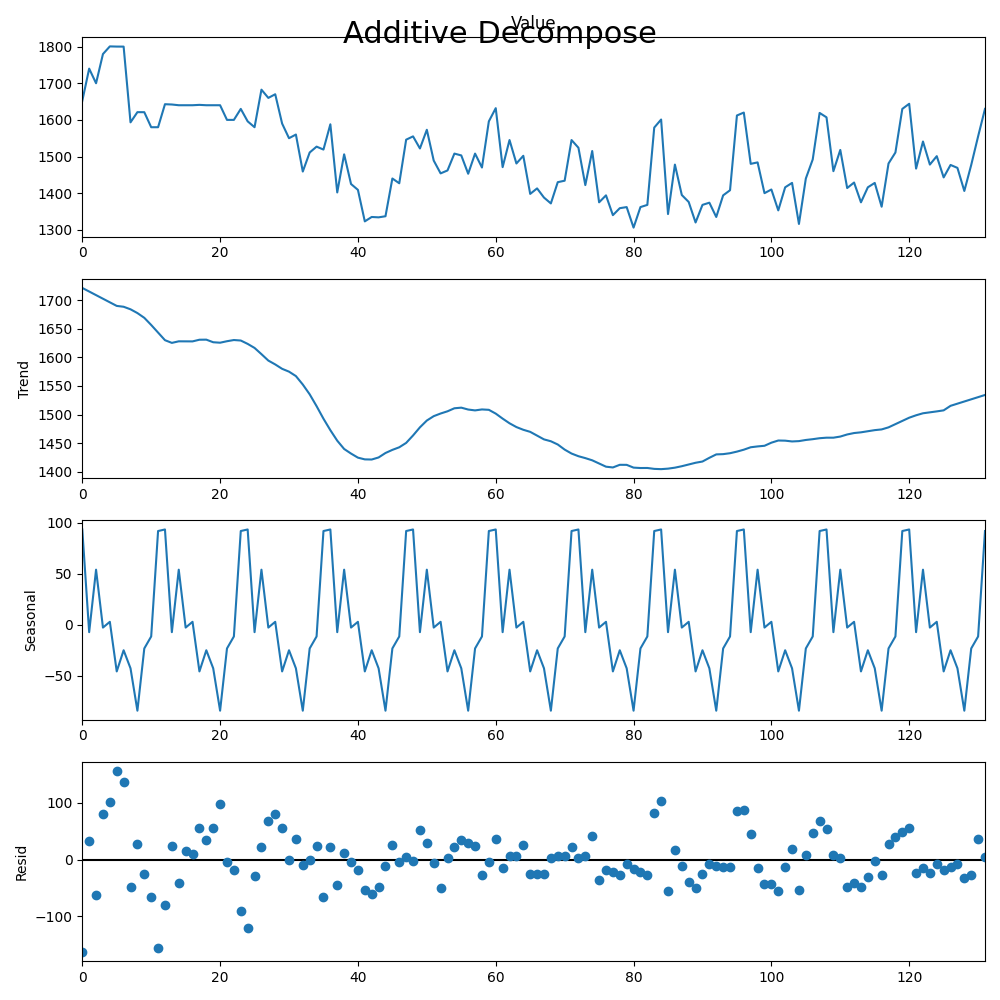


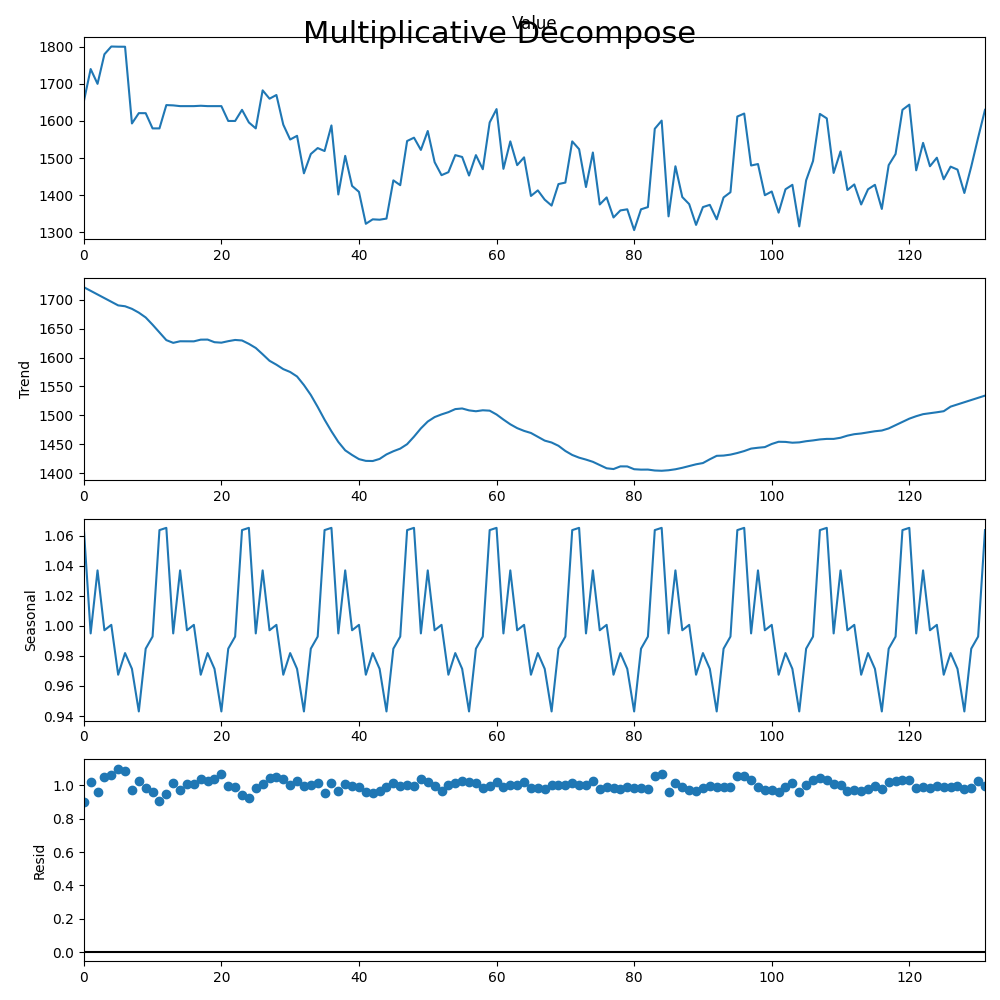
Глядя на АКФ данного процесса, можем заметить, что она постепенно затухает, что свидетельствует о нестационарности данного процесса.

PACF имеет выбросы на 7 и 13 лаге, что говорит о том, что текущее значение ряда коррелирует с аналогами на 7 и 13 лаге, что также может быть связано с сезонными выбросами в сентябре и январе.

1. **Выделение тренда, сезонности и остатков с помощью STL**

Рассмотрим аддитивную и мультипликативную модели. По графикам видно, что в случае с аддитивной моделью, остатки выглядят более случайными, поэтому данная модель предпочтительнее.







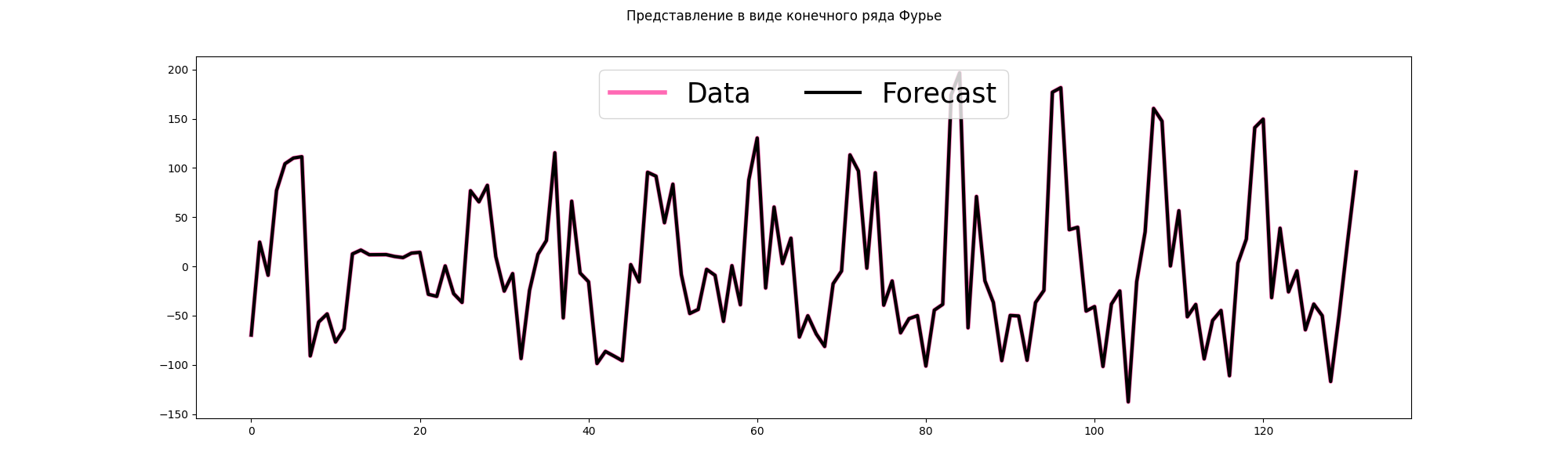
Мультипликативная

Аддитивная

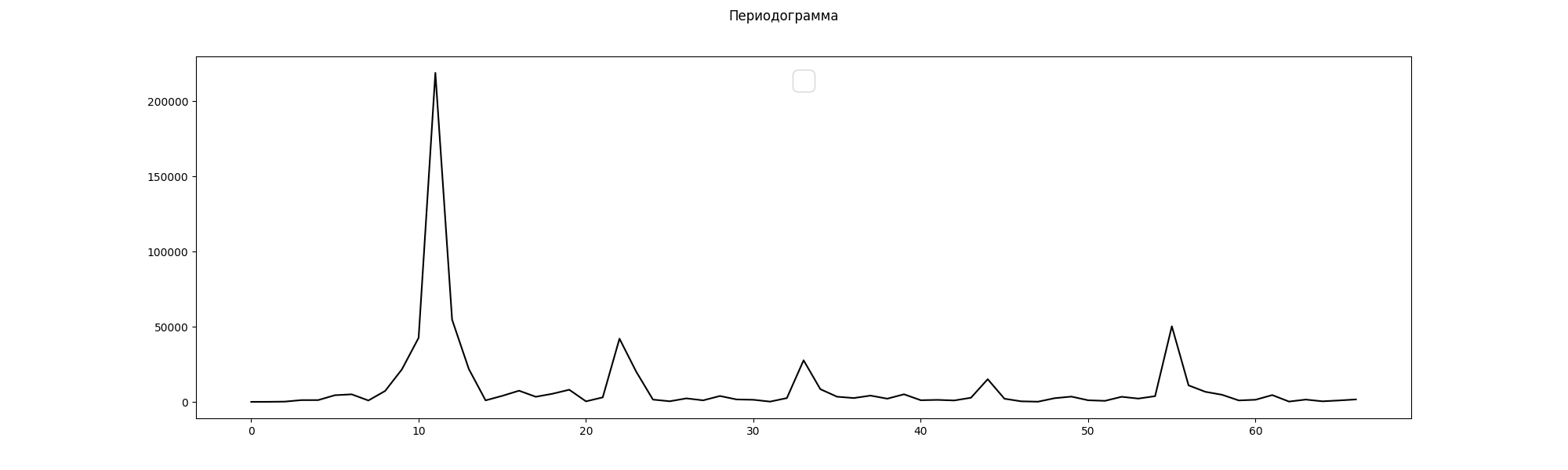


1. **Периодограмма на сезонной компоненте и остатках STL. Разложение ряда сезонности с остатками STL в ряд Фурье.**

Выделим сезонную компоненту и остатки в мультипликативной модели и разложим их в ряд Фурье.



Как можно заметить, графики исходных данных и с разложением Фурье совпали.



На 12 значении лага наблюдается пик, что свидетельствует о вероятной годовой сезонности.

1. **Критерий Спирмена проверяет ряд на наличие временного тренда, стационарность. Рассчитаем с помощью библиотеки Scipy коэффициент ранговой корреляции Спирмена и статистику на его основе:**

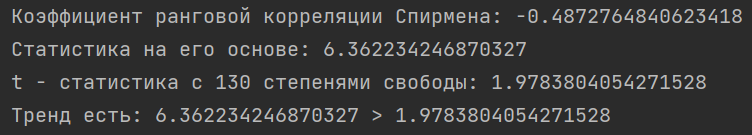
Одним из способов проверки исходного ряда на стационарность является коэффициент ранговой корреляции Спирмена. Чтобы его рассчитать, нам необходимо проранжировать исходный ряд по возрастанию и присвоить каждому элементу ранг.

Сам ранговый коэффициент корреляции рассчитывается по формуле:

Чтобы определить наличие тренда в исходном ряду, необходимо рассчитать значение статистики, основанной на коэффициенте :

Выдвинем две гипотезы:

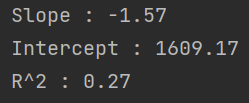
В качестве нулевой гипотезы мы говорим, что коэффициент ранговой корреляции Спирмена равен нулю и тренда нет, следовательно, ряд является стационарным. При подтверждении альтернативной гипотезы говорим о присутствии тренда и нестационарности ряда.



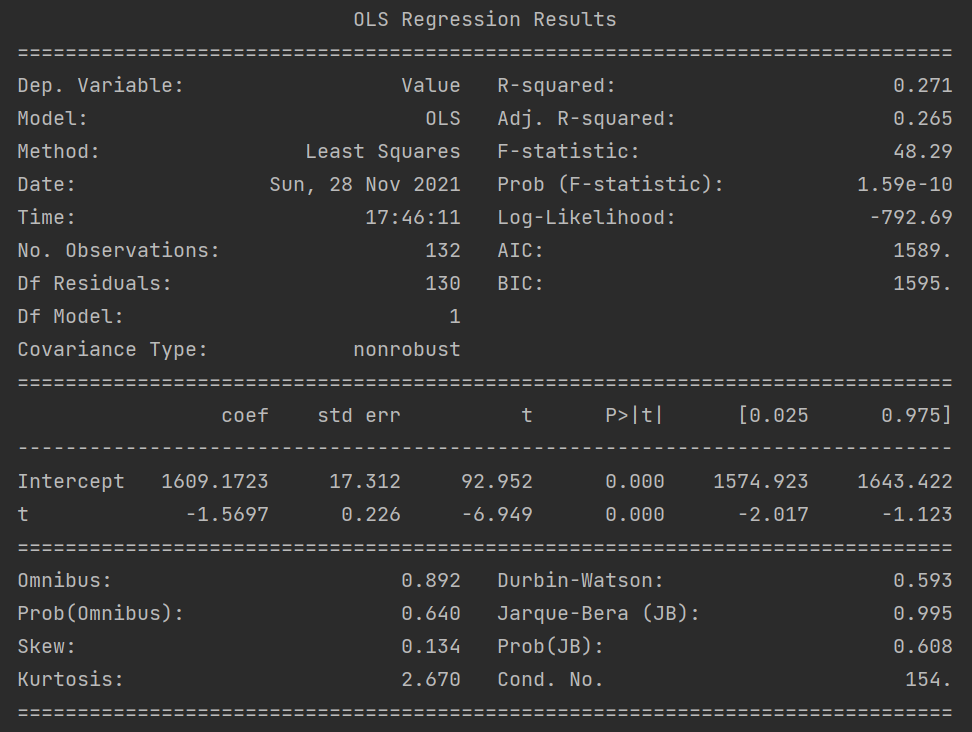
Так как статистика, полученная на основе коэффициента Спирмена, больше, чем t-статистика с T-2 (то есть 130), то нулевая гипотеза об отсутствии тренда и стационарности временного ряда отклоняется. Следвательно, исходный ряд нестационарен.

1. **Построение (оценка) модели линейного тренда**

Построим модель линейного тренда для исходного временного ряда, где Slope – коэффициент a1 перед t, показывающий угол наклона линии тренда (в каждом следующем месяце добыча газа снижается на 1,57); Intercept – коэффициент a0, константа (объем добычи газа в США, независящий от года); – коэффициент детерминации, один из показателей, определяющих точность модели(составляет 0,27, что говорит о том, что линейный тренд плохо описывает исходный ряд. Однако коэффициент детерминации не является самым главным критерием при выборе модели).



Проверим значимость коэффициента при переменной тренда и значимость модели в целом:

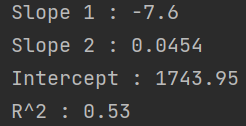


Исходя из таблицы выше, коэффициент при переменной тренда является значимым.

Проверим значимость регрессии через F – статистику. Для этого сравним расчетное значение F – статистики с табличным. , следовательно, регрессия в целом значима.

1. **Построение (оценка) модели полиномиального тренда**

Построим модель квадратичного тренда:, где Slope 1 – коэффициент a1 перед t; Slope 2 – коэффициент a2 перед ; Intercept – константа, коэффициент a0. Также можно заметить, что коэффициент детерминации увеличился, что говорит о том, что точность модели повысилась.



Посмотрим на значимость коэффициентов модели и значимость регрессии в целом:

В силу особенностей Python значения коэффициентов и коэффициента детерминации, посчитанные через LinearRegression() (скрин выше) и через библиотеку statmodels.formula.api отличаются, однако регрессия все равно является значимой в целом (расчетная F статистика больше табличной - ). Однако коэффициенты регрессии не являются значимыми, так как допускают ошибки 13,8% и 26%.

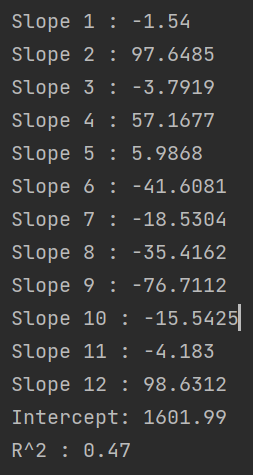


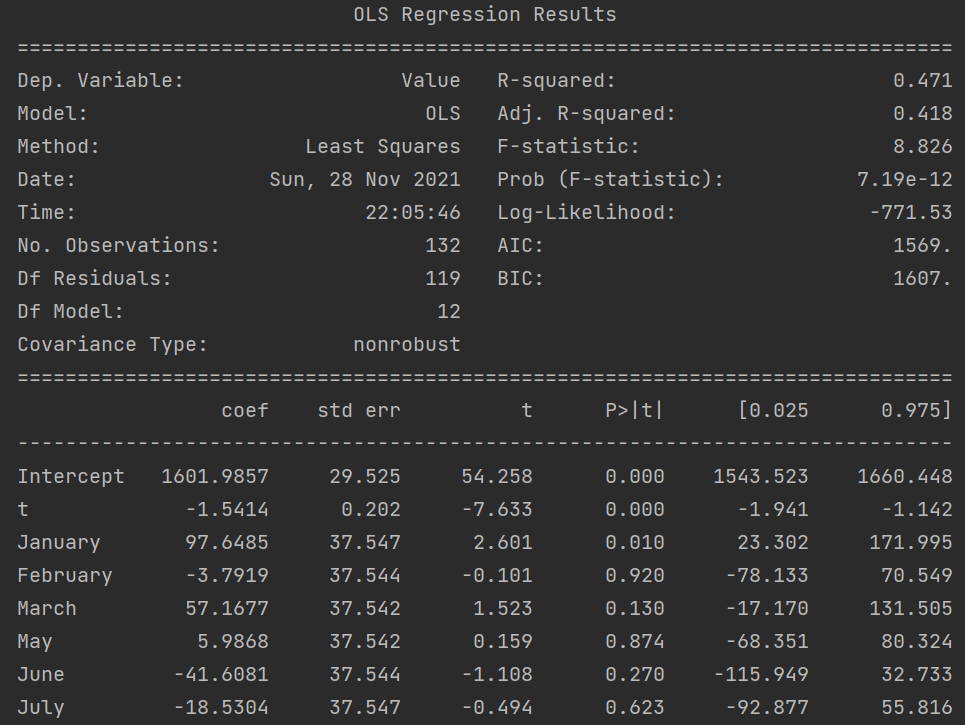
1. **Добавление в модели линейного и полиномиального трендов (p≥2) сезонных компонент (фиктивных переменных).**

|  |  |  |
| --- | --- | --- |
| Год | Минимум | Максимум |
| 1980 | ноябрь, дек | май |
| 1981 | ноябрь, окт | янв |
| 1982 | окт | март |
| 1983 | июнь | январь |
| 1984 | сент | дек |
| 1985 | сент | янв |
| 1986 | сент | дек |
| 1987 | сент | дек |
| 1988 | сент | янв |
| 1989 | сент | дек |
| 1990 | сент | янв |
|  |  |  |

Сделав анализ графика из пункта один, составим таблицу, в которой в столбец минимум – это месяц, в котором объем добычи газа был минимальным, а максимум – месяц, в котором он был максимальным. Таким образом, можно заметить, что чаще всего минимальная добыча наблюдалась в сентябре, а максимальная – в январе и декабре.

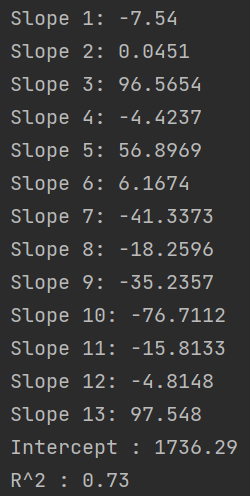
Введем 11 фиктивных переменных (все месяца, кроме апреля) и построим с ними модель линейного тренда:

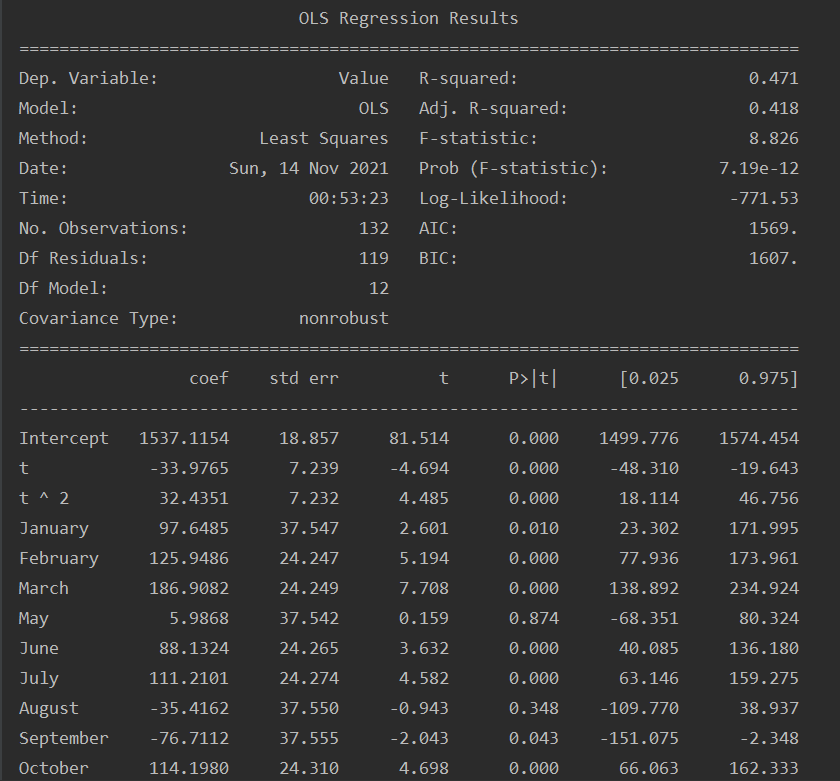


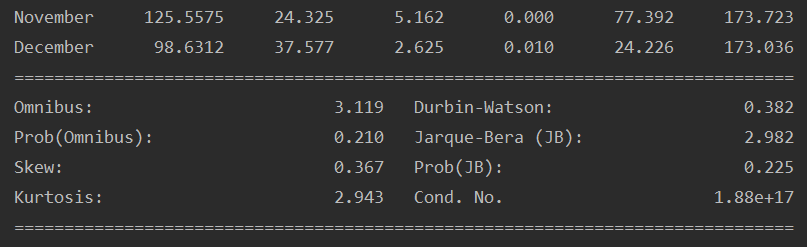




Исходя из таблицы выше, коэффициенты при переменных Февраль, Март, Май, Июнь, Июль, Август, Ноябрь, Октябрь являются незначимыми. В целом регрессия значима: . Проделаем то же самое для модели квадратичного тренда:



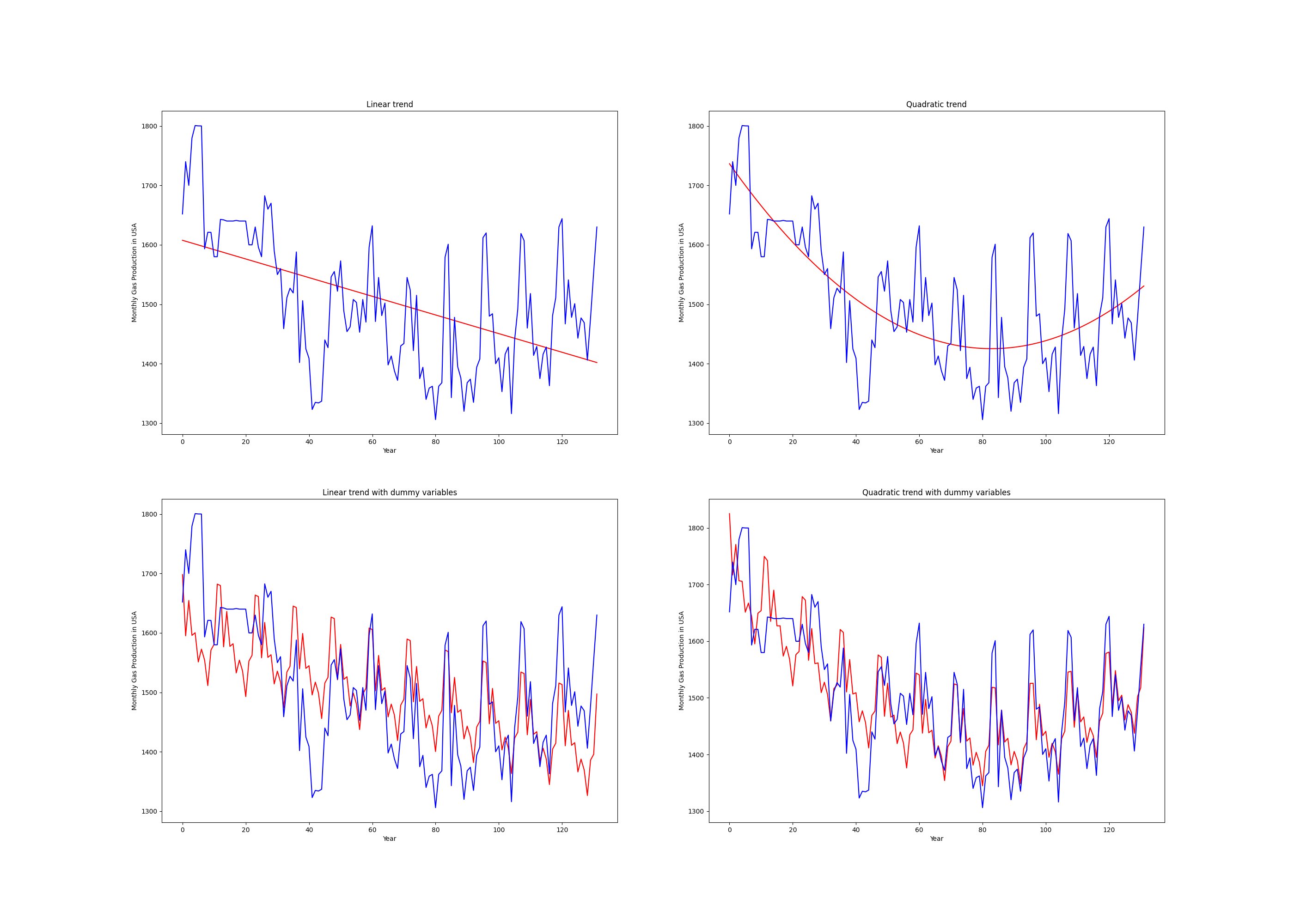




Как можно заметить из таблицы, все коэффициенты при переменных являются значимыми. Регрессия в целом также значима: .

1. **Расчет для каждой модели (линейный тренд, полиномиальный тренд (p>2), линейный тренд с сезонными фиктивными переменными, полиномиальный тренд с сезонными фиктивными переменными) R2, AIC, BIC, построение графиков для моделей. Определение, какая модель лучше аппроксимирует исходный временной ряд**

Построим графики всех четырех моделей: линейного, квадратичного, линейного с сезонным переменными и квадратичного трендов



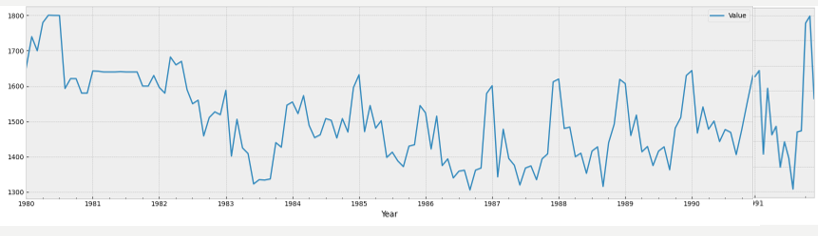
Как можно заметить, квадратичный тренд наиболее хорошо описывает исходный временной ряд. Также это подтверждает показатель коэффициента детерминации, который у данной модели самый высокий.

Сравним 4 модели по информационному критерию Акаике (чем он меньше, тем лучше модель). Самое низкое значение данный критерий принимает в моделях с линейным трендом и сезонными переменными и квадратичным трендом и сезонными фиктивными переменными.

Сравним 4 модели по информационному критерию BIC (Байесовский информационный критерий). Минимальное значение он принимает в модели линейного тренда без сезонных фиктивных переменных. Однако в данной модели малое количество переменных, что вероятно объясняет его малое значение. Поэтому предпочтительнее все-таки взять модель с квадратичным трендом и сезонными фиктивными переменными.

1. **Построение точечного прогноза на тестовой выборке, график.**

**Построение прогноза на 15 шагов вперед (то есть на 15 месяцев: январь 1991 – март 1992) на тестовой выборке с помощью модели квадратичного тренда с сезонными фиктивными переменными.**



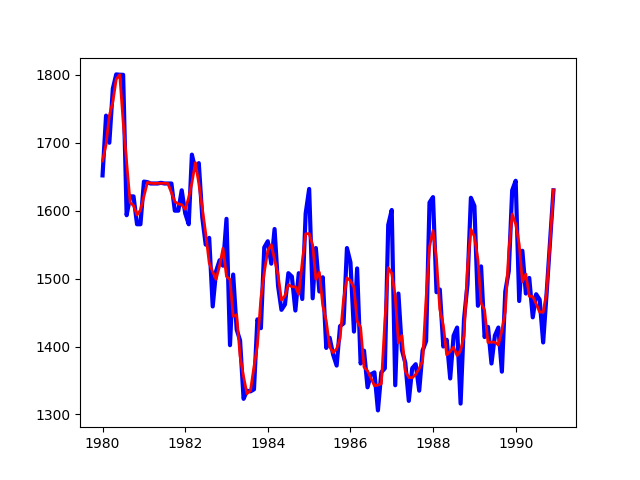
**Часть 2: Сглаживание временного ряда**

1. **Полиномиальное сглаживание (Savitzky-Golay Filter) c параметрами (m, p) = {(1,1); (1,2); (1,3); (1,1); (2,1); (3,1)}.**

При полиномиальном сглаживании окно скольжения рассчитывается как 2m + 1, тогда рассмотрим следующие случаи:

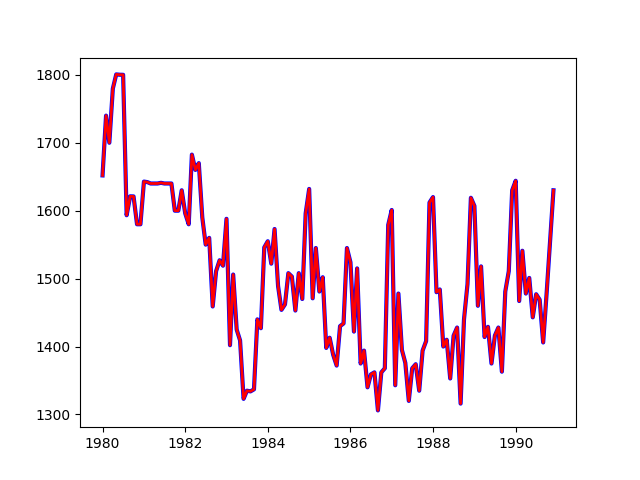
* 1. (m,p) = (1,1).

Окно скольжения = 3, степень полинома = 1:



* 1. (m,p) = (1,2).

Окно скольжения = 3, степень полинома = 2:

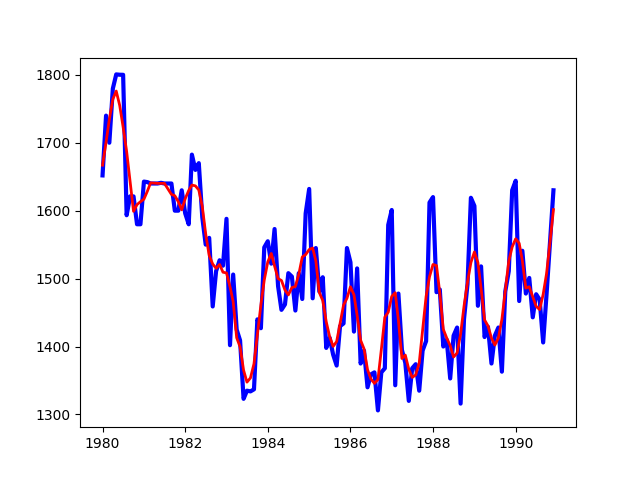


12.3 (m,p) = (1,3).

Окно скольжения = 3, степень полинома = 3, следовательно, полиномиальное сглаживание с данными параметрами невозможно, так как степень полинома должна быть строго меньше длины окна скольжения.

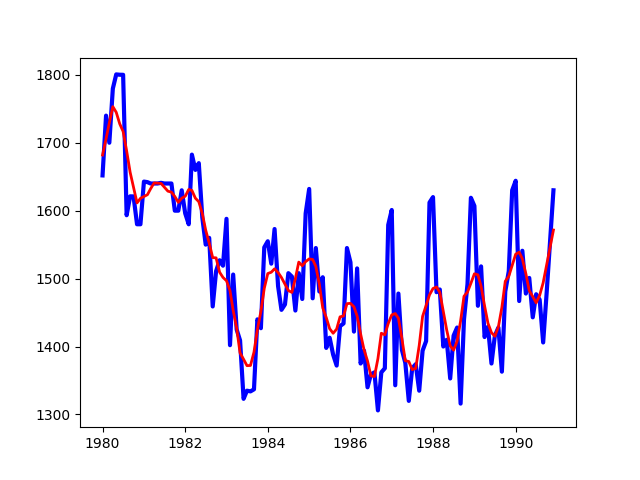
* 1. (m,p) = (2,1)

Окно скольжения = 5, степень полинома = 1:



* 1. (m, p) = (3, 1)

Окно скольжения = 7, степень полинома = 1:

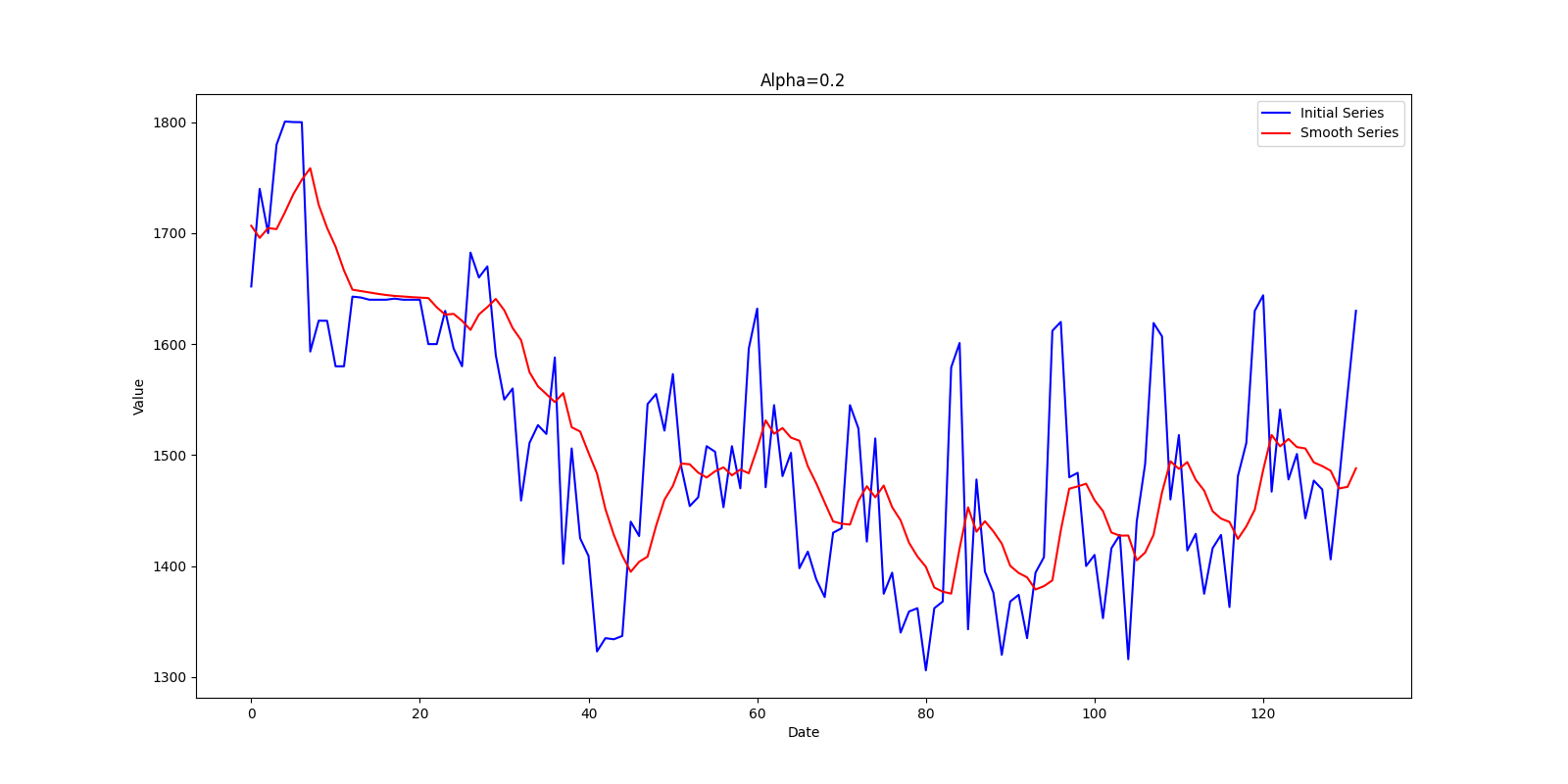


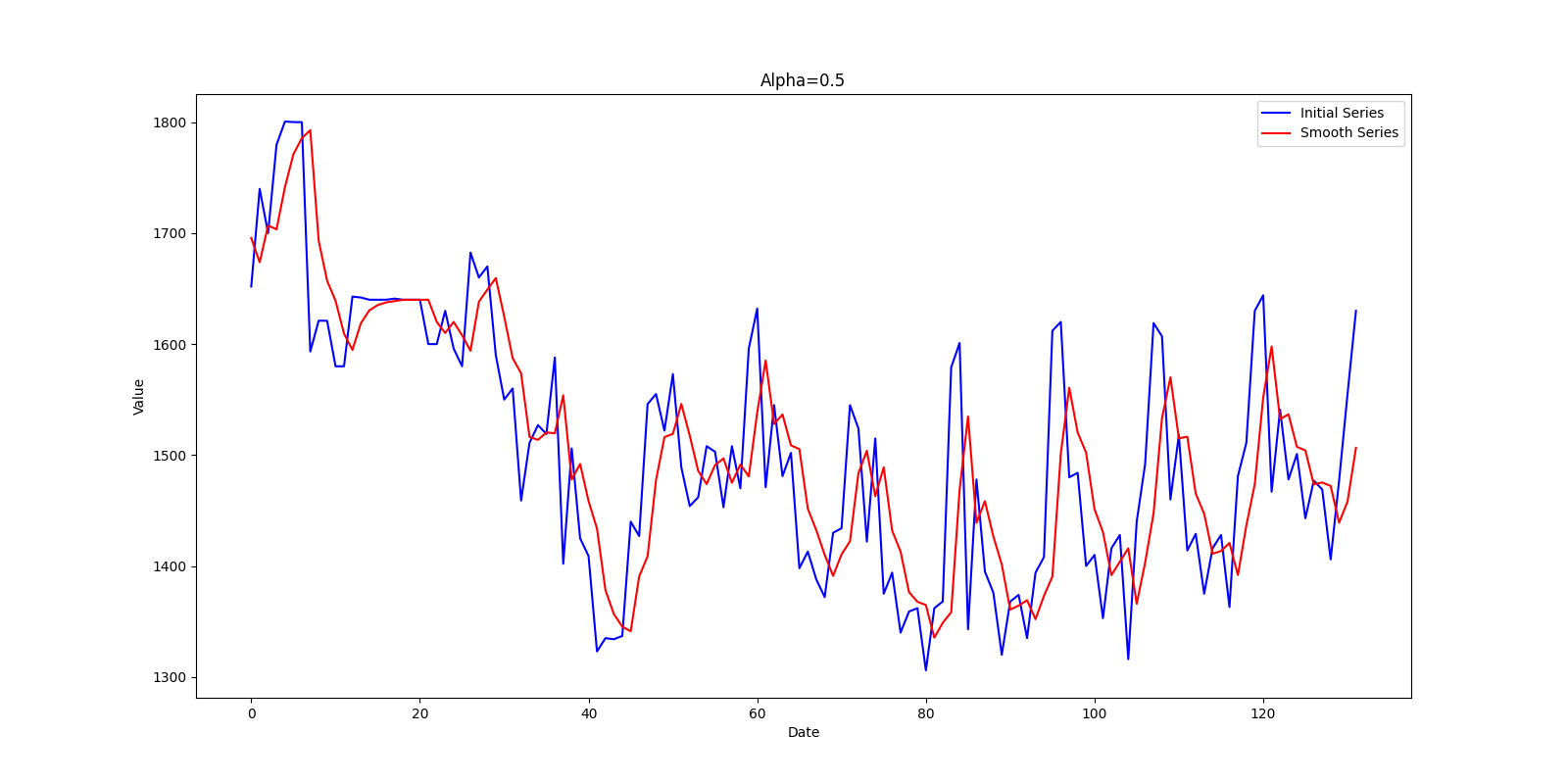
Как можно заметить из полученных выше графиков, при увеличении окна скольжения и степени полинома, равной 1, ухудшается аппроксимация ряда.

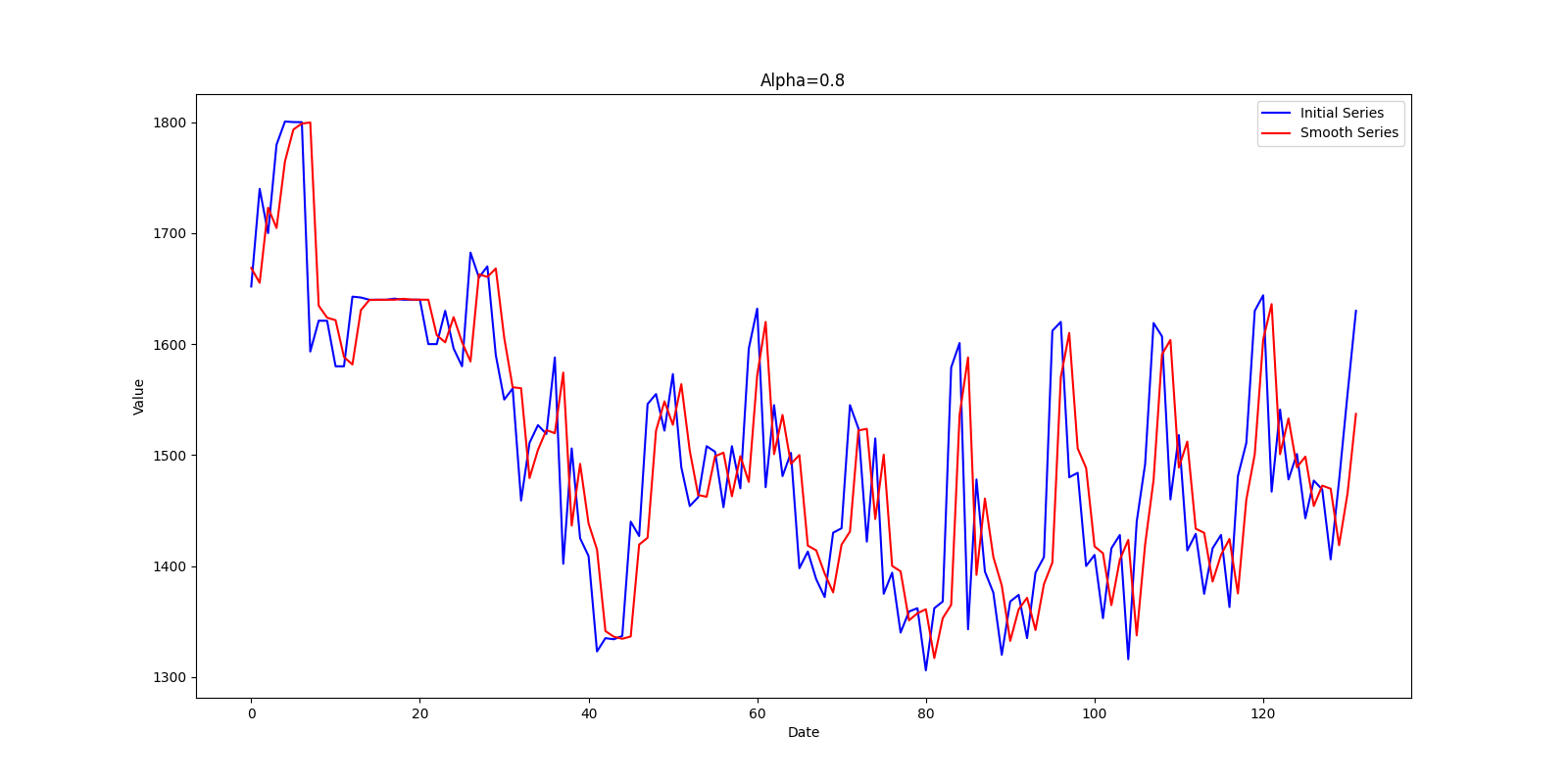
1. **Экспоненциальное сглаживание временного ряда (Simple Exponential Smoothing) для α = {0.2, 0.5, 0.8}. График. Вывод о влиянии соответствующего параметра на динамику ряда.**

Суть метода – фильтрация ряда с помощью взвешенных экспоненциальных средних. Взвешенная средняя рассчитывается по формуле: , где

Использование метода оправдано при наличии небольшой сезонности (в моем случае: сезонность ярко-выраженная).







Чем ближе значение параметра сглаживания альфа к единице, тем меньше расхождение между исходным рядом и сглаженным.

**Часть 3: Модели ARIMA**

1. **Проверка исходного ряда на стационарность с использованием:**

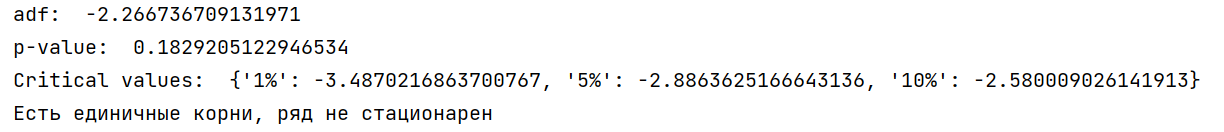
- Автокорреляционной функции (ACF):



Автокорреляционная функция, как видно из графика, является постепенно затухающей, что свидетельствует о том, что исходный ряд нестационарен.

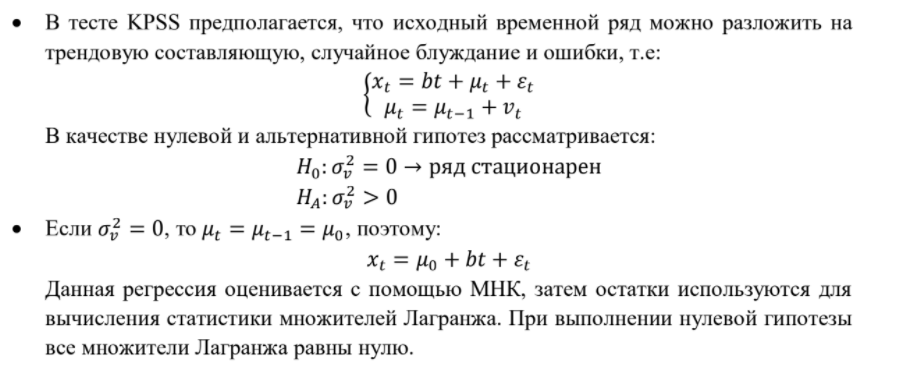
- Тест Дики-Фуллера (ADF):

Одним вариантов проверки ряда на стационарность является тест Дики-Фуллера. Его суть заключается в проверке процесса на стационарность с поиском единичных корней. Однако он является слишком ограничительным, так как говорит о том, что переменная следует авторегрессионному процессу первого порядка, и ошибки не коррелированы между собой. Поэтому воспользуемся дополненным тестом Дики-Фуллера (ADF):

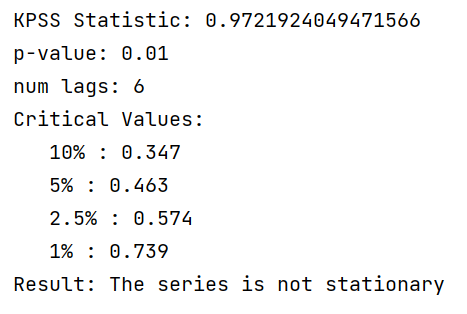


Посчитаем статистику для ADF-теста, она рассчитывается как обычная t-статистика для коэффициента δ: . Далее сравниваем расчетную t-статистику с табличным значением из таблицы, рассчитанной эмпирически методом Монте-Карло. (Сравниваем не с обычной t-статистикой, так как статистика Дики-Фуллера имеет нестандартное распределение). Как видно из теста: , следовательно, есть единичные корни и ряд не стационарен.

- Тест KPSS (тест Квятковского – Филлипса – Шмидта – Шина)



Проверим исходный ряд на стационарность через тест KPSS:



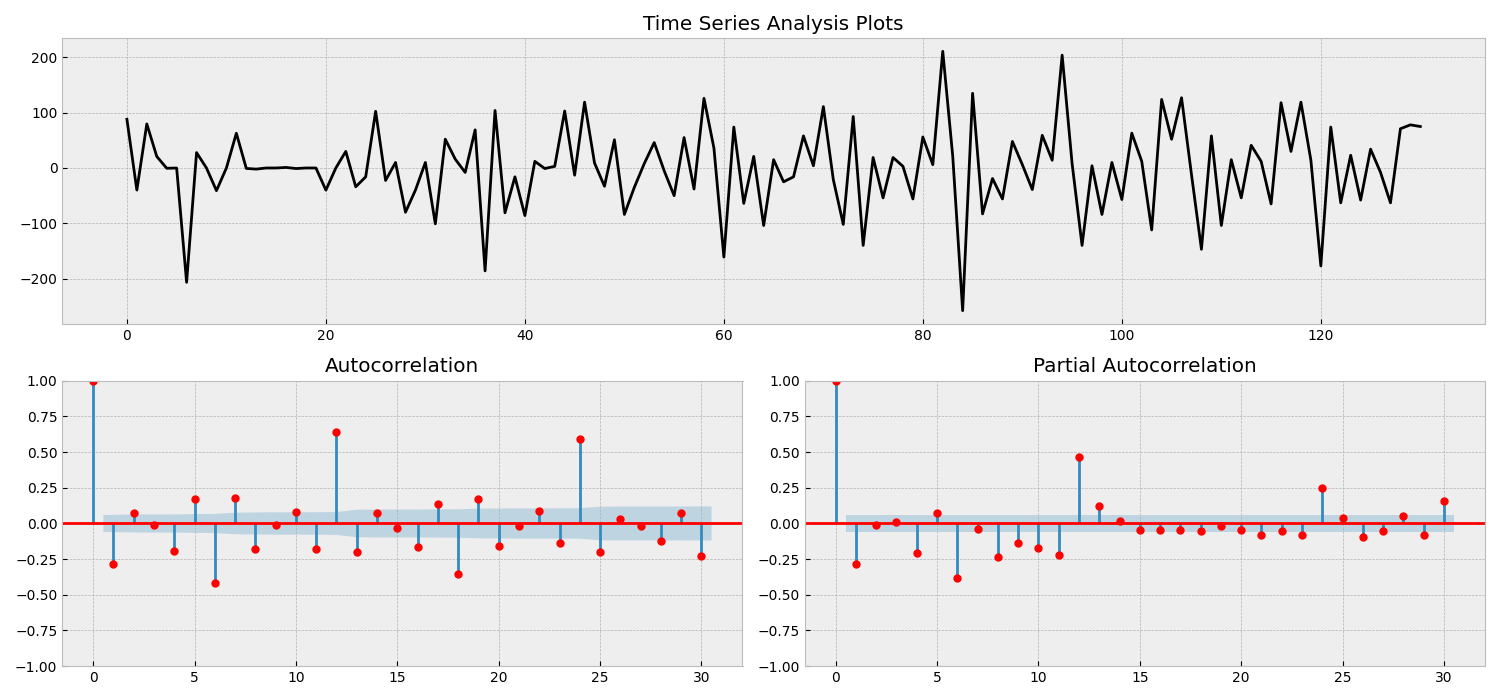
Тестовая статистика больше, чем значение табличной статистики на уровне значимости 5%, следовательно, нулевая гипотеза о стационарности ряда отклоняется, ряд не стационарен.

Как можно заметить, все три способа говорят о нестационарности исходного ряда.

**2. Проверка ряда первых и вторых разностей на стационарность с использованием**

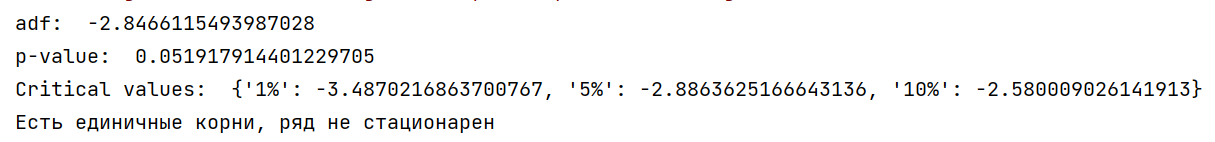
Проверим ряд первых разностей на стационарность:

- Автокорреляционной функции (ACF)



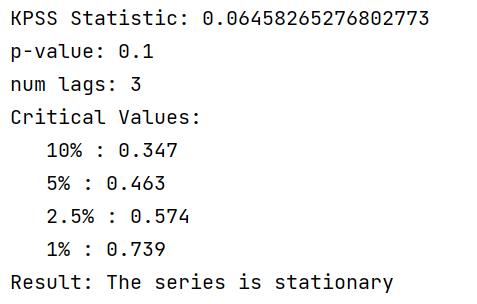
Исходя из графика автокорреляционной функции, автокорреляции нет (так как она обрывается после нулевого лага). Наблюдается годовая сезонность (по пиковым значениям на 12 и 24 лагах).

- Тест Дики-Фуллера (ADF)



t-статистика > τ => ряд не стационарен

- Тест KPSS:

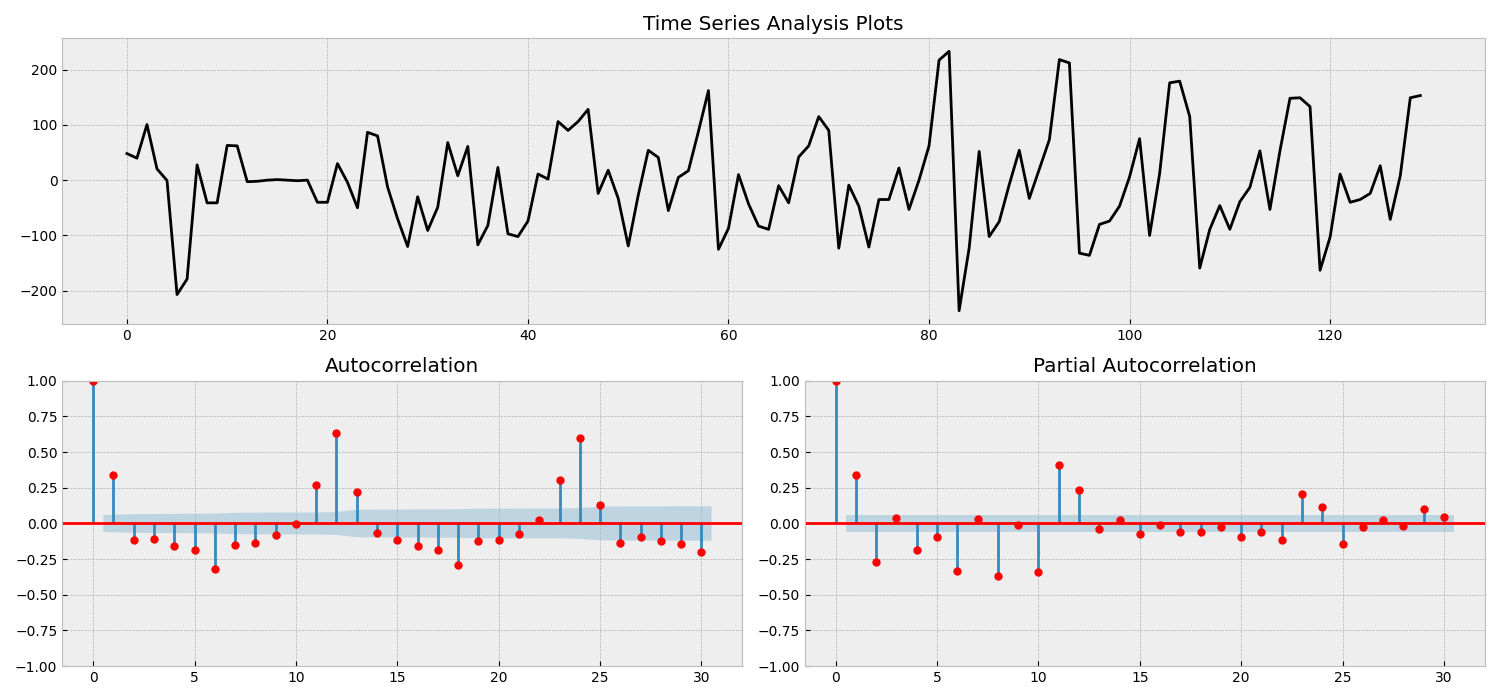


Статистика, рассчитанная в тесте KPSS меньше, чем табличная, следовательно, принимается нулевая гипотеза о стационарности ряда.

Результаты тестов не совпадают, ADF – ряд нестационарный, KPSS – ряд стационарный. В таком случае можно сделать, что ряд стационарен относительно тренда. В статистическом анализе из временной ряд , тренд-стационарный процесс - это стохастический процесс , из которого можно удалить базовый тренд (функция только времени), оставив стационарный процесс.

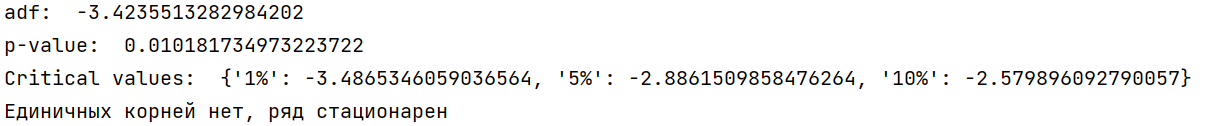
Проверим ряд вторых разностей:

- Автокорреляционная функция (ACF)



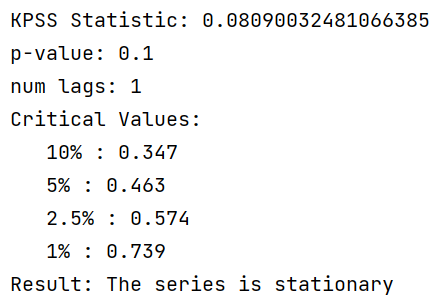
Годовая сезонность, МА(1) (так как обрывается после первого лага)

-Тест Дики-Фуллера (ADF)



=> ряд стационарен

-Тест KPSS

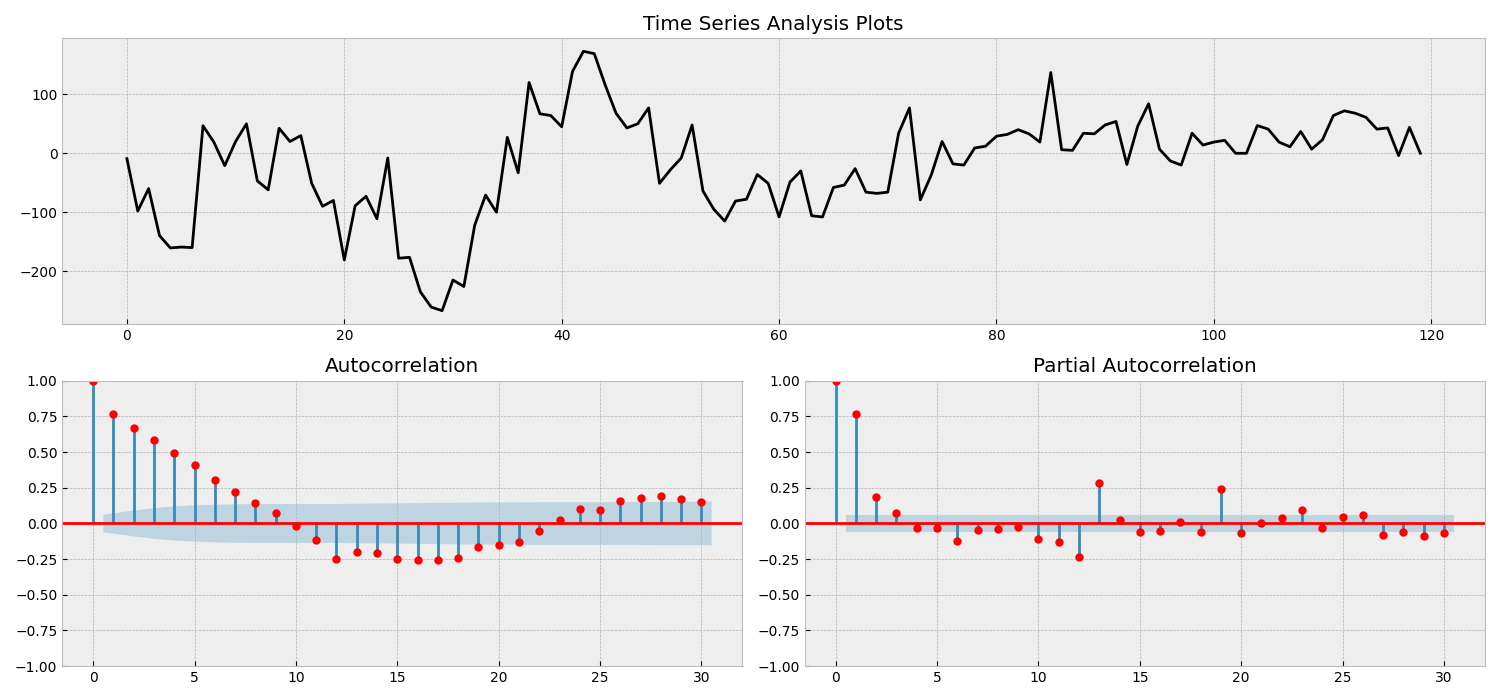


Тестовая статистика < табличное значение статистики => ряд стационарен, принимается нулевая гипотеза о стационарности ряда.

**3. Проверка ряда сезонных разностей на стационарность**

Предположим, что у нас годовая сезонность, тогда:

- Автокорреляционная функция (ACF)



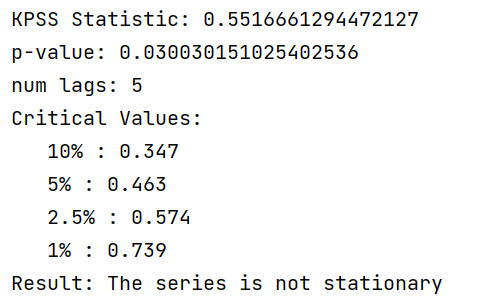
Постепенно затухает, следовательно, ряд сезонных разностей не стационарен.

- Тест Дики-Фуллера (ADF)



=> ряд не стационарен

- Тест KPSS

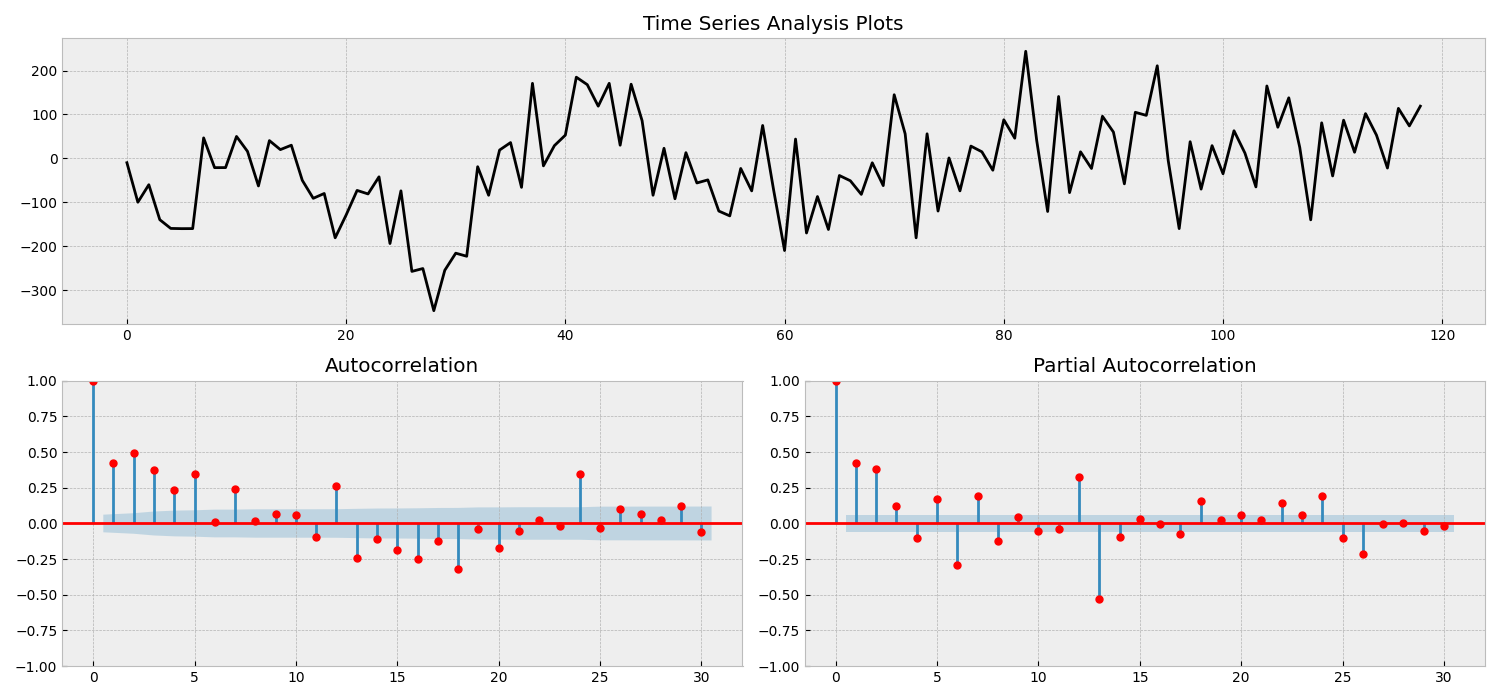


Тестовая статистика > табличного значения статистики => ряд не стационарен, отклоняется нулевая гипотеза о стационарности ряда.

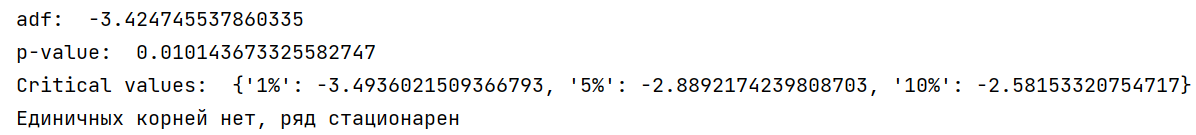
Результаты трех тестов совпадают, ряд не стационарен.

Рассмотрим ряд первых разностей:

-Автокорреляционная функция (АКФ)



-Тест Дики-Фуллера



-Тест KPSS не рассматривается, так как дальше при проверки различных разностей ряда на стационарность по этому тесту ряд не является стационарным.

**4. По результатам предыдущих пунктов выберите порядок интегрирования d\*:**

Исходя из результатов тестов выше, ряд является стационарным при разностях второго порядка. Следовательно, d\* = 2.

**5. По результатам предыдущих пунктов выберите порядок сезонных разностей D\*:**

Исходя из результатов тестов с сезонными разностями выше, можем сделать вывод, что ряд сезонных разностей является стационарным при первых разностях. Следовательно, D\* = 1.

**6. Выбор модели ARIMA (p, d∗, q).**

Чтобы определить параметры p и q, посмотрим на графики ACF и PACF, полученные выше, и составим сетку параметров, которые могут быть у нашей ARIMA модели:

Проанализируем графики автокорреляционной и частичной автокорреляционной функции для вторых разностей. По графику PACF, соответствующему процессу авторегрессии, мы видим, что функция обрывается после первого лага, что говорит о том, что p = 1. Следовательно, AR (1). По графику ACF, соответствующему процессу скользящего среднего, видно, что функция так же обрывается после первого лага, что говорит о том, что q = 1. Следовательно, MA (1).

- Скользящая кросс-валидация (1-я модель):

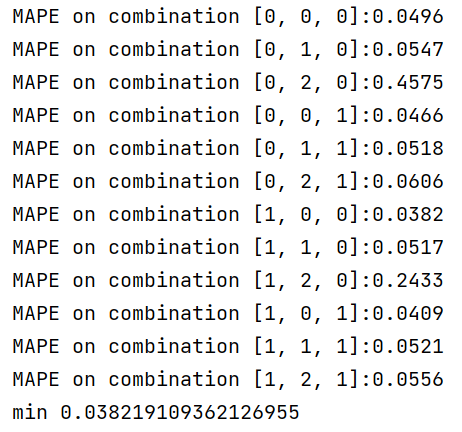
Скользящая кросс-валидация - это способ выделения независимых данных, при котором вышеописанные проблемы практически не возникают.

Выделим из исходных 132 наблюдений 10 наблюдений для тестовой выборки, остальные 122 – для обучающей. На обучающей строим модель, а на тестовой проверяем ее качество.

Выделим из исходных 132 наблюдений первые 112 наблюдений для обучающей выборки, а следующие 10 (т.е с 113 по 122) – для тестовой выборки.

Затем из исходных 132 наблюдений выделяем первые 102 наблюдения для обучающей выборки, а следующие 10 (т.е со 103 по 112) – для тестовой выборки.

То есть мы разбиваем исходную выборку на train и test 3 раза – на 3 фолда. После этого берем среднее из оцененных метрик и получаем усредненное качество модели на независимых данных. Далее посмотрим, какие значения MAPE получились для каждой ARIMA модели:



Минимальное значение MAPE достигается в модели ARIMA (1, 0, 0).

- Информационные критерии:

Сделаем сетку параметров и посмотрим значения информационного критерия AIC для различных p, d, q:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | (p, d, q) | (0,1,0) | (0,2,0) | (0,0,1) | (0,1,1) | (0,2,1) | (1,0,0) | (1,1,0) | (1,2,0) | | AIC | 1507,83 | 1618,99 | 1564,22 | 1499,13 | 1504,48 | 1509,36 | 1498,81 | 1553,15 | |

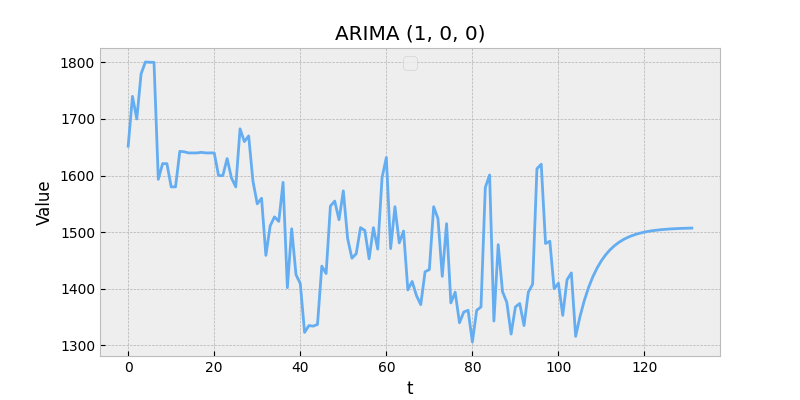
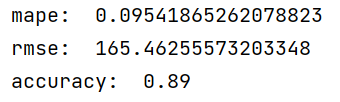
|  |  |  |  |
| --- | --- | --- | --- |
| (p, d, q) | (1,0,1) | (1,1,1) | (1,2,1) |
| AIC | 1507,20 | 1500,80 | 1496,04 |

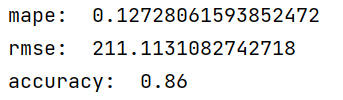
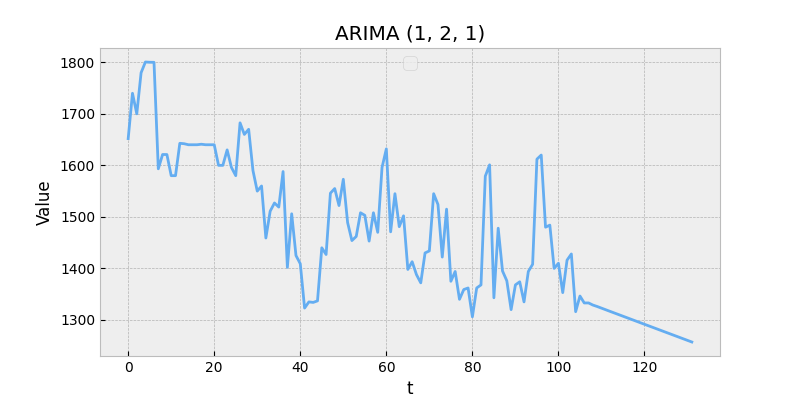
Минимальное значение критерия AIC наблюдается у модели ARIMA (1, 2, 1).

Результаты не совпадают, так как лучшей моделью в плане минимальной процентной ошибки является ARIMA (1,0,0), а с точки зрения минимального значения информационного критерия – ARIMA (1, 2, 1)

**7. Построение точечного прогноза для ARIMA на H шагов вперёд по выбранным моделям. Оценка качества прогнозов.**

ARIMA(1,0,0):



Как видно из графиков и значений метрик качества прогноза, прогнозы по моделям ARIMA являются не правдоподобными

**8. Выбор модели SARIMA(p, d∗, q, P, D,∗, Q).**

SARIMA – модель ARIMA с учетом сезонной составляющей. Определим ее параметры p, q, P, Q.

Проанализируем графики автокорреляционной и частичной автокорреляционной функции для сезонных разностей. По графику PACF, соответствующему процессу авторегрессии, мы видим, что функция обрывается после третьего лага, что говорит о том, что p = 3. По графику ACF, соответствующему процессу скользящего среднего, видно, что функция обрывается после пятого лага, что говорит о том, что q = 5. Теперь определим значения P и Q. Определим Q из графика автокорреляционной функции: отметим «пики» сезонности на 12 и 24 лагах, следовательно Q=2. Проделаем аналогичную процедуру с графиком PACF для определения P: так же отметим «пики» сезонности на 12 и 24 лагах и сделаем из этого вывод, что P=2.

Составим сетку параметров, используя полученные выше значения и определим с помощью скользящей кросс-валидации и информационного критерия Акаике наиболее предпочтительные модели SARIMA.

- Скользящая кросс-валидация (3-я модель)

Для q взят range (2, 5), так как PyCharm не прогружает, если брать от нуля.







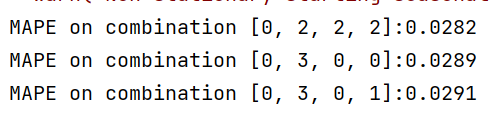














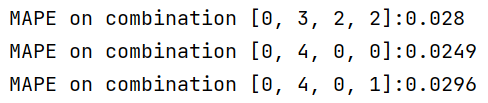


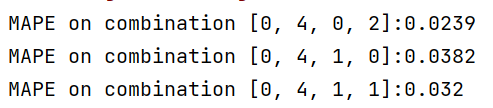
























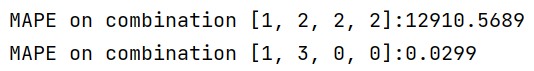






















































































































































































































Минимальное значение MAPE: 0,231 – SARIMA (1, 3, 2, 0, 2, 1)

- Информационные критерии:

SARIMA AIC

[0, 5, 0, 0] 2054.9219026962646

[0, 5, 0, 1] 2054.9219026962646

[0, 5, 0, 2] 2054.9219026962646

[0, 5, 1, 0] 1897.7744153599992

[0, 5, 1, 1] 1897.7744153599992

[0, 5, 1, 2] 1897.7744153599992

[0, 5, 2, 0] 1751.8408739377896

.

.

.

[2, 5, 2, 0], 1594.4159757643606

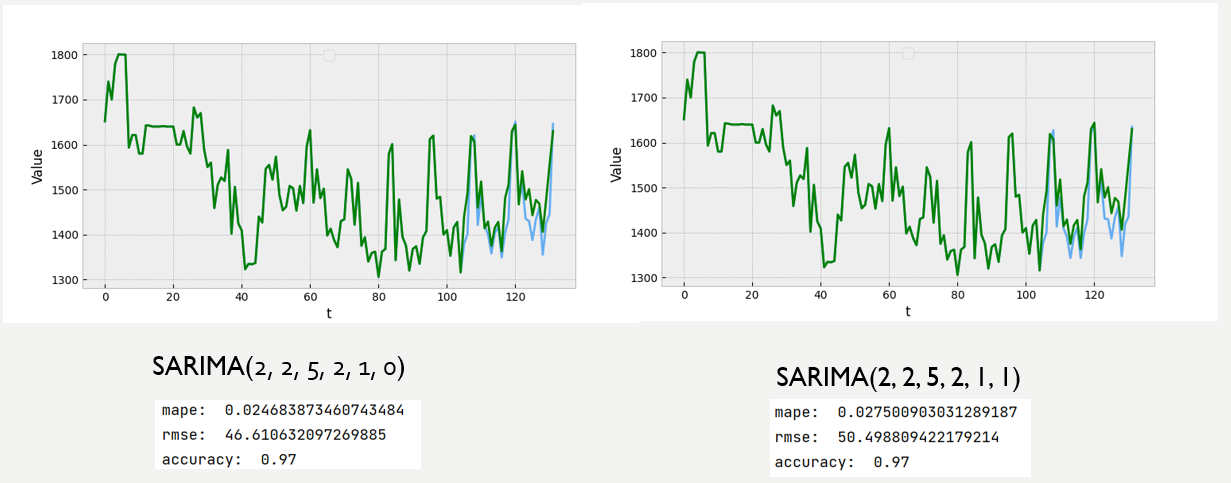
[2, 5, 2, 1], 1594.4159757643606

[2, 5, 2, 2], 1594.4159757643606

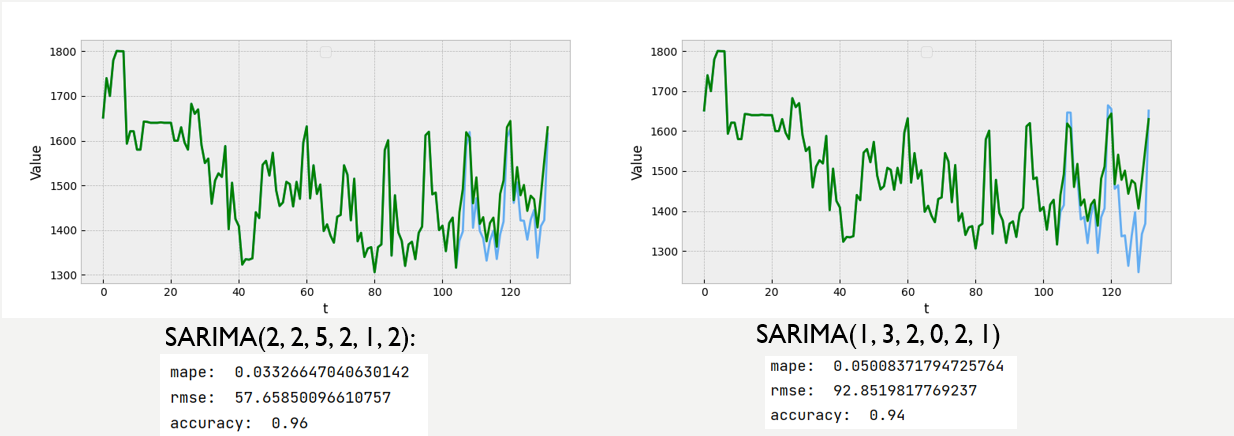
Минимальными являются значения AIC при SARIMA (2, 2, 5, 2, 1, 0), (2, 2, 5, 2, 1, 1), (2, 2, 5, 2, 1, 2)

Результаты по информационному критерию и по кросс-валидации не совпадают

**9. Построение точечного прогноза для SARIMA на H шагов вперёд по выбранным моделям. Оценка качества прогнозов.**



Как видно из графиков и значений метрик качества прогноза, модели минимально отличаются друг от друга. Но у первой модели стандартное отклонение и ошибка в процентах меньше.



Исходя из визуального анализа и анализа метрик качества прогноза (MAPE и RMSE у третьей и четвертой модели SARIMA выше, чем у первой модели), можем сделать вывод, что наилучшей моделью для построения прогноза среди моделей ARIMA и SARIMA является модель SARIMA (2, 2, 5, 2, 1, 0)