**Федеральное государственное бюджетное образовательное учреждение высшего образования**
**"Уфимский государственный авиационный технический университет"**

**Кафедра** Высокопроизводительных вычислительных технологий и систем

**Дисциплина:** Теория случайных процессов и математическая статистика

# Отчет по лабораторной работе № 2

**Тема:** «Прогнозирование временных рядов с помощью ARIMA»

| Группа ПМ-453 | Фамилия И.О. | Подпись | Дата | Оценка |
|---|---|---|---|---|
| Студент | Шамаев И.Р. | | | |
| Принял | Маякова С.А. | | | |

**Уфа 2022**

**Цель:** научиться восстанавливать трендовую составляющую методом простой скользящей средней временных рядов и создать прогнозы временных рядов с помощью ARIMA.

## Теоретический материал

Временной ряд – это последовательность наблюдений (измерений, отсчетов) $x_1, x_2, \dots x_n$, упорядоченная во времени, т.е. $x_k = x(t_k), k = 1, 2, \dots n$. Будем рассматривать временные ряды, в которых наблюдения ведутся через равные промежутки времени.

Временной ряд $x_1, x_2, \dots x_n$ имеет два главных отличия от случайной выборки, образованной из наблюдений переменной X.

1. Наблюдения $x_1, x_2, \dots x_n$, рассматриваемые как случайные величины, в большинстве случаем статистически зависимы, т.е. коррелированы. Поэтому значение наблюдения в момент времени $t_k, k = 1, 2 \dots n,$ может зависеть от того, какие значения были зарегистрированы до этого момента времени. Следовательно, имеется принципиальная возможность изучения и прогнозирования статистических свойств ряда.
2. Наблюдения временного ряда в общем случае не образуют стационарной последовательности, т.е. при изменении момента времени $t_k$ меняются числовые характеристики случайной величины $x(t_k)$, в частности ее математическое ожидание $\mu_x = \mu_x(t_k)$ и дисперсия $\sigma_x = \sigma_x(t_k)$.

Функция $\mu_x = \mu_x(t_k)$, описывающая зависимость математического ожидания от времени, называется трендом.

***Разложение временного ряда***

Разложение ряда на составляющие осуществляется для отдельного изучения составляющих ряда. Ряд может быть представлен как смесь четырех компонент:

$$X = T + C + S + \varepsilon,$$

где $T$ – тренд, или долгосрочное движение; $C$ – циклическая составляющая, более или менее регулярные колебания относительно тренда; $S$ – сезонная компонента (регулярные колебания не слишком большого периода; сезонность отражает внутригодичные колебания, связанные с погодой, праздниками, обычаями); $\varepsilon$ – остаток, или несистематический случайный эффект (Рис. 1).

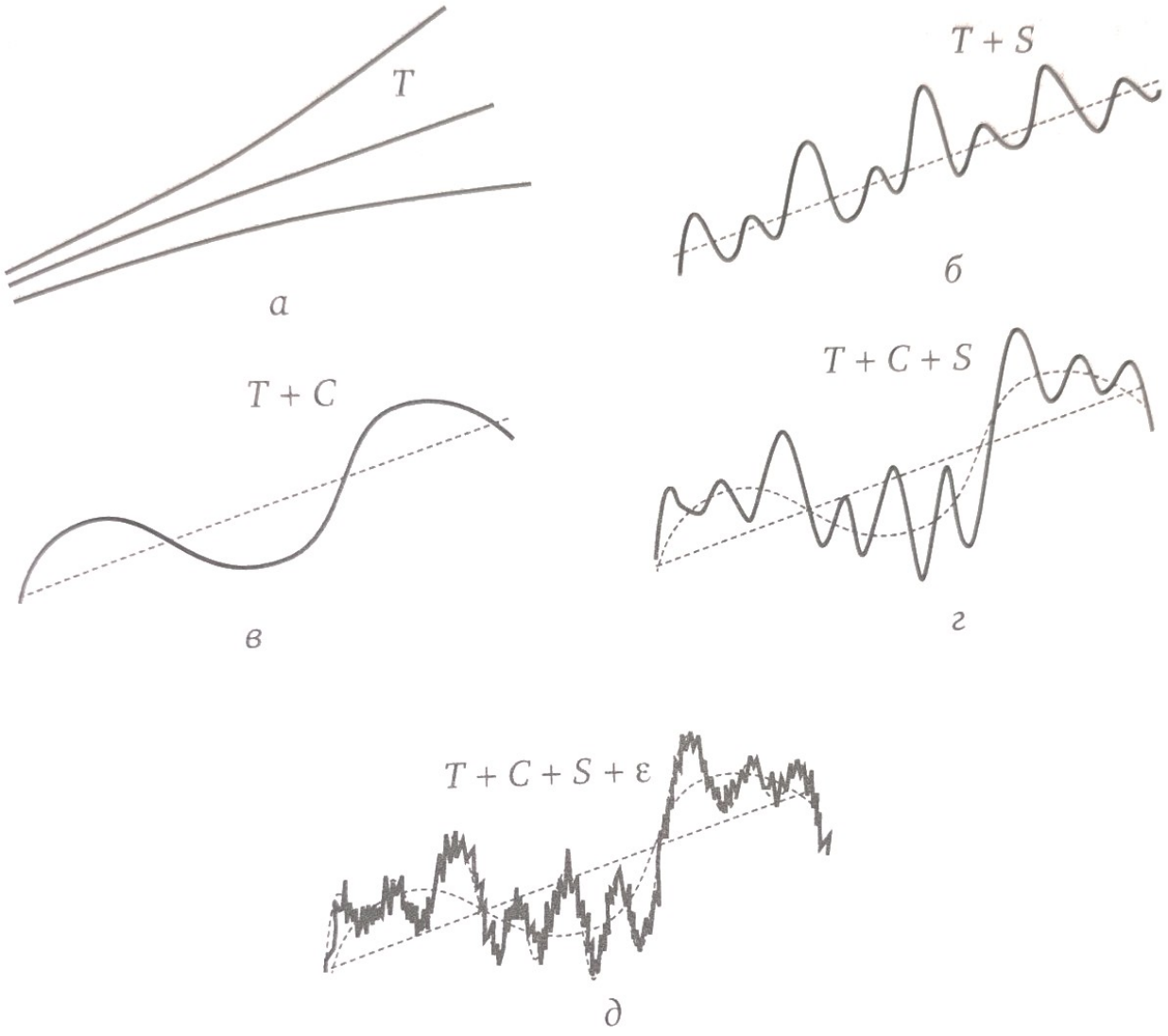


Рисунок 1. Базовые модели (компоненты) временных рядов

Для выделения циклической компоненты необходимо исключить из ряда тренд. Для получения цикла в чистом виде нужно использовать тренд с точкой

перегиба, т.е. линию скользящей средней. Тогда сглаженный ряд представляется моделью вида $Y=T+C$, в которой случайная компонента сглаживается вместе с сезонной. Отсюда циклическую компоненту получают исключением тренда.

Выделение сезонных колебаний возможно, когда во временных рядах отсутствуют (или устранены с помощью фильтрации) явно выраженные циклические колебания.

Случайная компонента выражается через разность между исходным рядом X и моделью сглаженного и выровненного ряда.

Например, разложение ряда, не имеющего циклической компоненты, может иметь вид, представленный на рисунке 2.

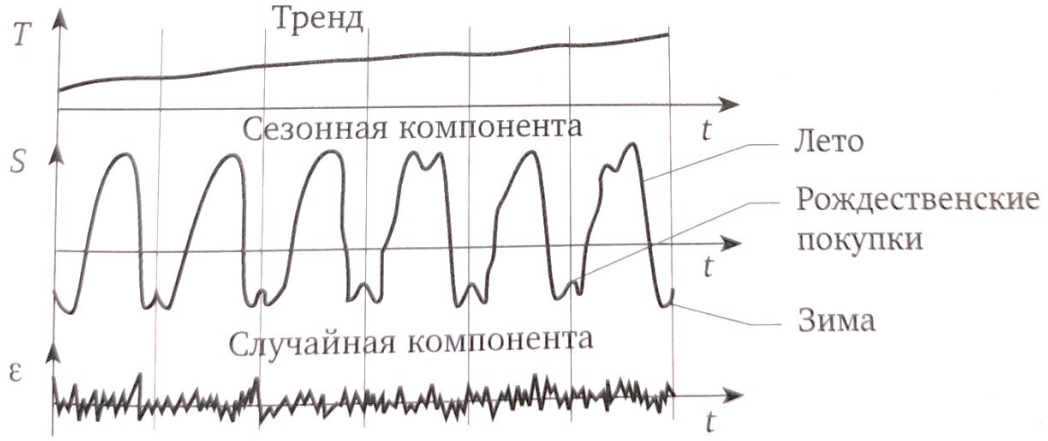


Рисунок 2. Пример: разложение временного ряда для объема продаж футболок

***Модель скользящей средней (скользящего окна)***

Модель скользящей средней основана на том, что за сглаженное значение ряда в любой дискретной точке t принимают среднее значение в некоторой окрестности. При изменении момента времени окрестность скользит вдоль оси t, чем и объясняется название модели. Модель позволяет получить для всех точек исходного временного ряда $\{x_t\}$ последовательность $\{y_t\}$, которая является сглаженным рядом исходной последовательности.

Скользящие средние могут быть взвешенными и простыми. Модель взвешенной средней имеет вид

$$y_t = \sum_{k=-m}^{m} \alpha_k x_{t+k}, t = m+1..n-m,$$

где число $2m+1$ – размер окрестности (окна), также называемое порядком скользящей средней. Веса $\alpha_k$ предполагаются нормированными, так что

$$\sum_{k=-m}^{m} \alpha_k = 1.$$

При $\alpha_k = \frac{1}{2m+1}$ получаем простую скользящую среднюю $2m+1$ порядка.

### *Модель ARIMA*

ARIMA использует три основных параметра (p, d, q), которые выражаются целыми числами. Потому модель также записывается как ARIMA(p, d, q). Вместе эти три параметра учитывают сезонность, тенденцию и шум в наборах данных:

- p – порядок авторегрессии (AR), который позволяет добавить предыдущие значения временного ряда. Этот параметр можно проиллюстрировать утверждением «завтра, вероятно, будет тепло, если в последние три дня было тепло».
- d – порядок интегрирования (I; т. е. порядок разностей исходного временного ряда). Он добавляет в модель понятия разности временных рядов (определяет количество прошлых временных точек, которые нужно вычесть из текущего значения). Этот параметр иллюстрирует такое утверждение: «завтра, вероятно, будет такая же температура, если разница в температуре за последние три дня была очень мала».
- q – порядок скользящего среднего (MA), который позволяет установить погрешность модели как линейную комбинацию наблюдавшихся ранее значений ошибок.

Параметр s определяет периодичность временного ряда (4 – квартальные периоды, 12 – годовые периоды и т.д.).

**Практическая часть**

**Задание № 1**

Восстановить трендовую составляющую методом простой скользящей средней для любого из представленных временных рядов.

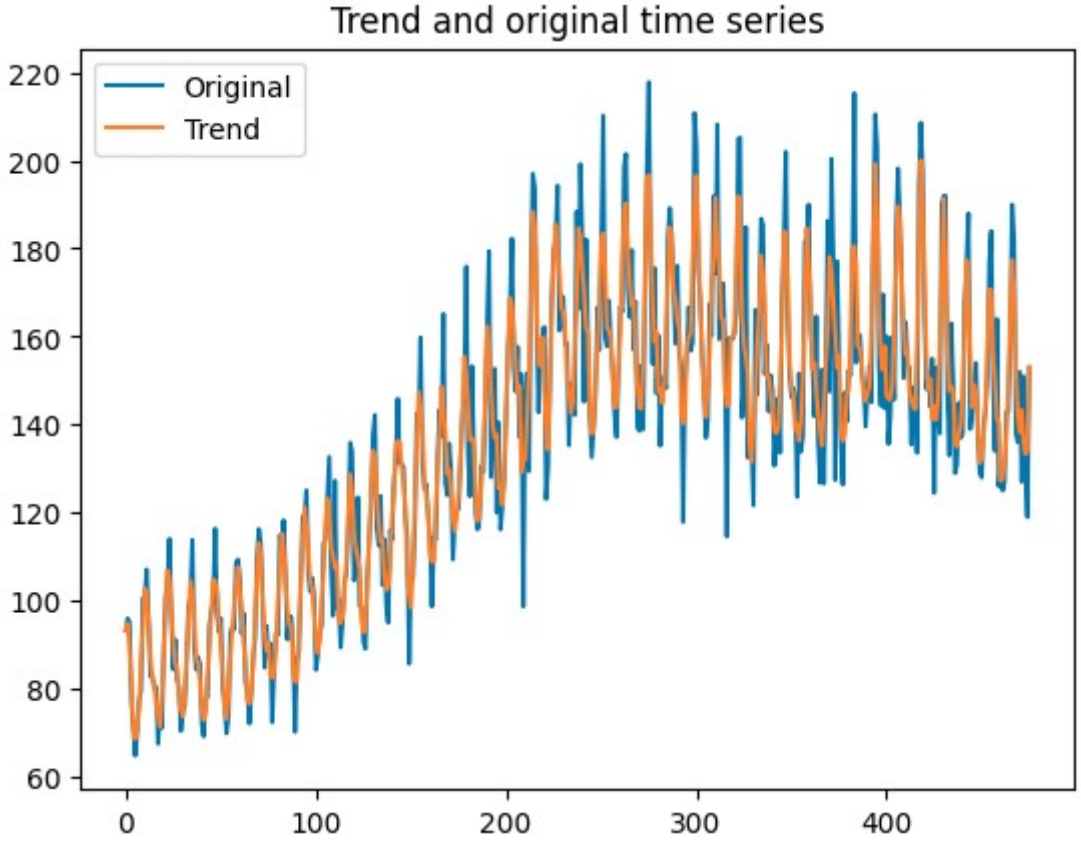


Рисунок 1. Результат работы программы с размерами

Трендовая составляющая была восстановлена методом простой скользящей средней с размером окрестности равным 3. Результат представлен на рисунке ниже.

**Задание № 2**

Создать прогнозы временных рядов. Аналогично построить прогноз для любого другого временного ряда (можно выбрать любой из представленных). Проанализировать полученные результаты.

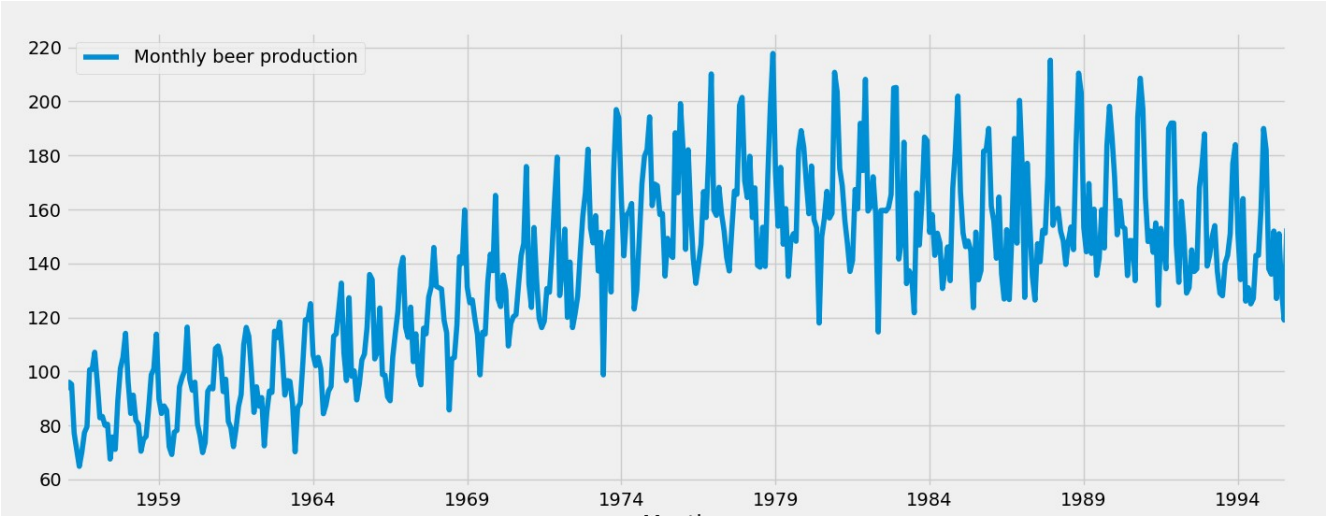


Рисунок 2. Временной ряд с количеством произведённого количества

Далее подберём оптимальные параметры для модели подбором так, чтобы коэффициент AIC был минимален.

Получим значения $p=0, q=1, d=1, s=12$ с AIC = 3370.13.

Теперь проведём диагностику модели.

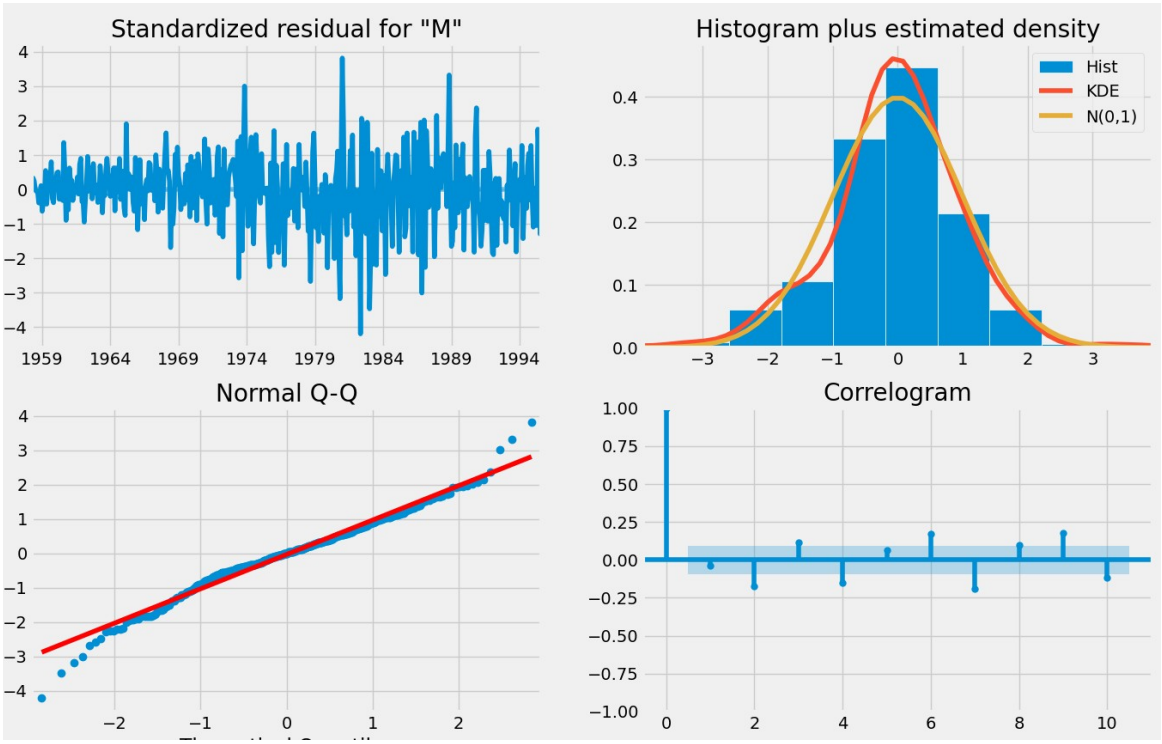


Рисунок 3. Диагностика модели.

Из гистограммы следует, что остатки распределены нормально, а упорядоченное распределение остатков (график слева снизу) следует линейному тренду выборок. График автокорреляции (внизу справа) показывает, что остатки временных рядов имеют низкую корреляцию с запаздывающими данными.

В этом случае диагностика показала, что остатки модели правильно распределяются:

- На верхнем правом графике красная линия KDE находится близко к линии N (0,1) (где N (0,1) является стандартным обозначением нормального распределения со средним 0 и стандартным отклонением 1) . Это хороший признак того, что остатки нормально распределены.
- График в левом нижнем углу показывает, что упорядоченное распределение остатков (синие точки) следует линейному тренду выборок, взятых из стандартного распределения N (0, 1). Опять же, это признак того, что остатки нормально распределены.
- Остатки с течением времени (верхний левый график) не показывают явной сезонности и кажутся белыми шумами. Это подтверждается графиком автокорреляции (внизу справа), который показывает, что остатки временных рядов имеют низкую корреляцию с запаздывающими данными.

Эти графики позволяют сделать вывод о том, что выбранная модель (удовлетворительно) подходит для анализа и прогнозирования данных временных рядов.

Теперь выполним прогнозирование с 1992 года.

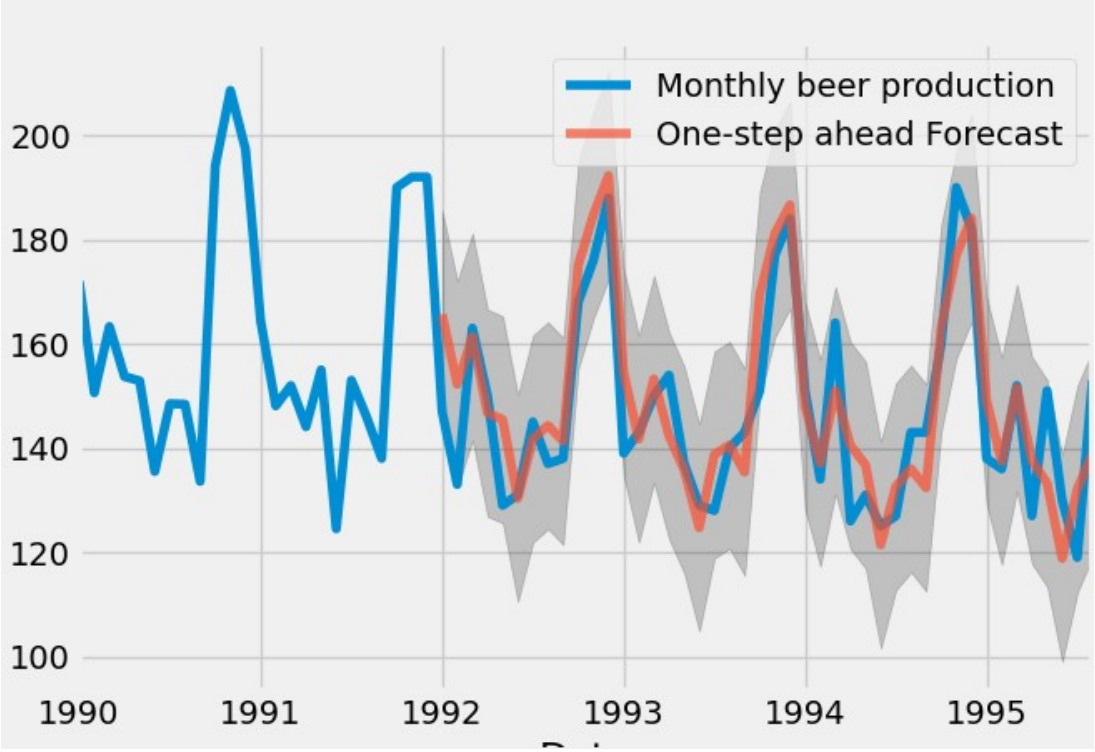


Рисунок 4. Сравнение прогнозируемых значений с реальными значениями временного ряда.

В целом, прогнозы соответствуют истинным значениям, демонстрируя общий тренд. Среднеквадратическая ошибка прогноза равна 89.14.

Более точное представление точности прогнозирования может быть получено с помощью динамических прогнозов. В этом случае нужно использовать только информацию из временных рядов до определенной точки; затем прогнозы сгенерируются с помощью значений из предыдущих прогнозируемых временных точек.

Используем модель ARIMA для прогнозирования будущих значений, выбрав значение равным 500 шагов вперёд.

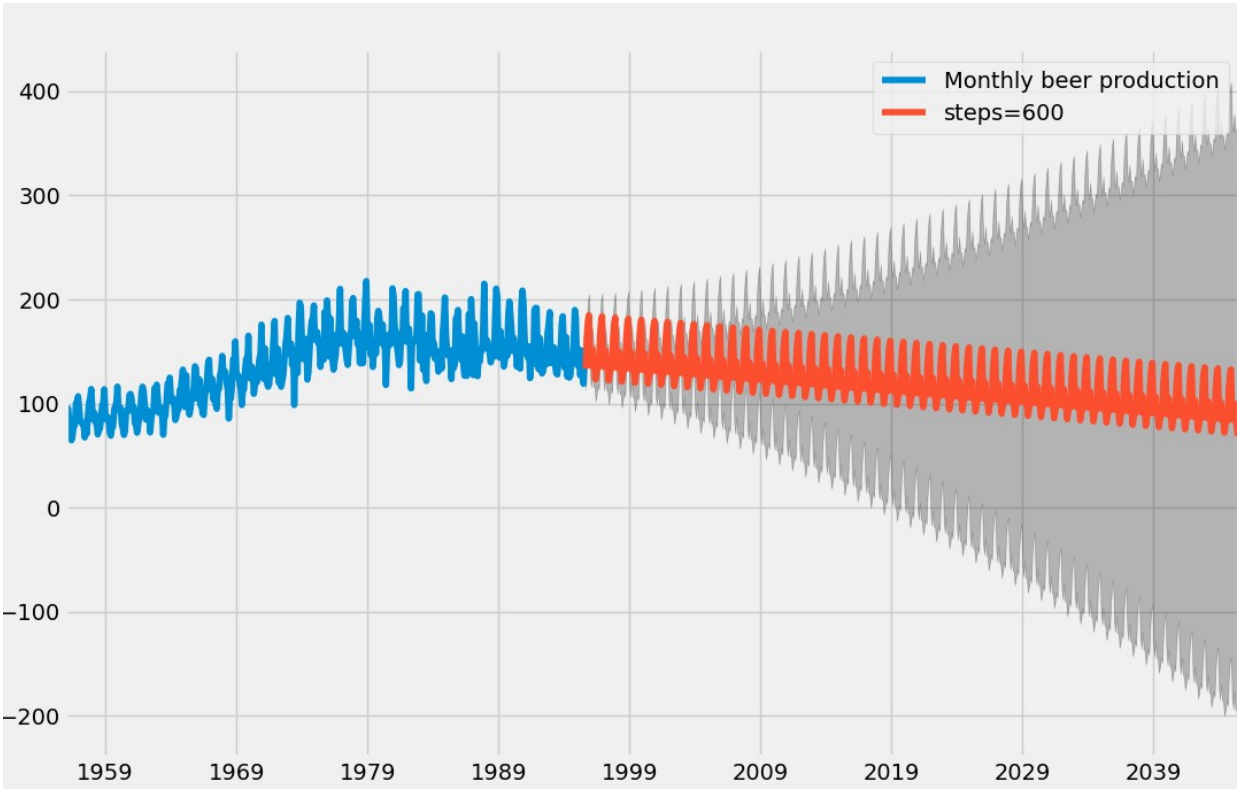


Рисунок 5. Результаты прогнозирования.

Из рисунка 5 видно, что нет смысла делать такой большой прогноз для этих данных, так как он будет неточным. Но первые точки прогноза показывают тренд на спад производства

**Список литературы**

1. Гмурман В. Е. Теория вероятностей и математическая статистика: Учебное пособие. – М: Высшая школа, 2003. – 479 с.
2. Степанов С. С. Стохастический мир, 2009. – 376 с.

**Вывод**

В ходе выполнения лабораторной работы было изучено применение метода простой скользящей средней для прогнозирования временных рядов и построены прогнозы временных рядов с помощью инструмента ARIMA.

## Приложение

### Задание 1

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


# define moving average function
beer = pd.read_csv("monthly-beer-production-in-austr.csv")
monthly_beer_production_mean = beer['Monthly beer production'].mean()

x = beer['Monthly beer production'].to_numpy()

plt.plot(x, label = 'old')
w1 = 25
w2 = 50
MA25 = beer['Monthly beer production'].rolling(window = w1).mean()
MA50 = beer['Monthly beer production'].rolling(window = w2).mean()
plt.plot(MA25, label = f'w= {w1}')
plt.plot(MA50, label = f'w= {w2}')
plt.legend()
plt.show()
```

### Задание 2

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


import warnings
import itertools
import statsmodels.api as sm
plt.style.use('fivethirtyeight')
beer = pd.read_csv("monthly-beer-production-in-austr.csv", index_col = 0, parse_dates
= ["Month"])
print(beer.info())
data = sm.datasets.co2.load_pandas()
y = beer

# bfill значит, что нужно использовать значение до заполнения пропущенных значений
y = y.fillna(y.bfill())
print(y)
y.plot(figsize = (15, 6))
plt.show()

# Определите p, d и q в диапазоне 0 - 2
p = d = q = range(0, 2)
# Сгенерируйте различные комбинации p, q и q
pdq = list(itertools.product(p, d, q))
# Сгенерируйте комбинации сезонных параметров p, q и q
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]
print('Examples of parameter combinations for Seasonal ARIMA...')
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[4]))
warnings.filterwarnings("ignore") # отключает предупреждения
for param in pdq :
for param_seasonal in seasonal_pdq :
try :
    mod = sm.tsa.statespace.SARIMAX(y,
```

```
        order = param,
        seasonal_order = param_seasonal,
        enforce_stationarity = False,
        enforce_invertibility = False)
    results = mod.fit(disp = 0)
    print('ARIMA{}x{}12 - AIC:{}'.format(param, param_seasonal, results.aic))
    except:
continue

mod = sm.tsa.statespace.SARIMAX(y,
    order = (1, 1, 1),
    seasonal_order = (1, 1, 1, 12),
    enforce_stationarity = False,
    enforce_invertibility = False)
    results = mod.fit(disp = 0)
    print(results.summary().tables[1])

    results.plot_diagnostics(figsize = (15, 12))
    plt.show()


    pred = results.get_prediction(start = pd.to_datetime('1992-01-01'), dynamic =
False)
    pred_ci = pred.conf_int()


    ax = y['1990':].plot(label = 'observed')
    pred.predicted_mean.plot(ax = ax, label = 'One-step ahead Forecast', alpha = .7)
    ax.fill_between(pred_ci.index,
        pred_ci.iloc[:, 0],
        pred_ci.iloc[:, 1], color = 'k', alpha = .2)
    ax.set_xlabel('Date')
    ax.set_ylabel('Beer production')
    plt.legend()
    plt.show()

    pred_dynamic = results.get_prediction(start = pd.to_datetime('1992-01-01'),
dynamic = True, full_results = True)
    pred_dynamic_ci = pred_dynamic.conf_int()
    ax = y['1990':].plot(label = 'observed', figsize = (20, 15))
    pred_dynamic.predicted_mean.plot(label = 'Dynamic Forecast', ax = ax)
    ax.fill_between(pred_dynamic_ci.index,
        pred_dynamic_ci.iloc[:, 0],
        pred_dynamic_ci.iloc[:, 1], color = 'k', alpha = .25)
    ax.fill_betweenx(ax.get_ylim(), pd.to_datetime('1992-01-01'), y.index[-1],
        alpha = .1, zorder = -1)
    ax.set_xlabel('Date')
    ax.set_ylabel('Beer production')
    plt.legend()
    plt.show()


    # Получить прогноз на 500 шагов вперёд
    pred_uc = results.get_forecast(steps = 600)
    # Получить интервал прогноза
    pred_ci = pred_uc.conf_int()


    ax = y.plot(label = 'observed', figsize = (20, 15))
    pred_uc.predicted_mean.plot(ax = ax, label = 'steps=600')
    ax.fill_between(pred_ci.index,
        pred_ci.iloc[:, 0],
        pred_ci.iloc[:, 1], color = 'k', alpha = .25)
    ax.set_xlabel('Date')
    ax.set_ylabel('Beer production')
```

```
plt.legend()
plt.show()
```