МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ

УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ

ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

**институт информационных технологий и технологического образования**

**кафедра информационных технологий и электронного обучения**

Основная профессиональная образовательная программа

Направление подготовки 09.03.01 Информатика и вычислительная техника

Направленность (профиль) «Технологии разработки программного обеспечения»

форма обучения – очная

**Курсовая работа**

по дисциплине «Пакеты прикладных программ для статистической обработки и анализа данных»

Анализ и прогнозирование численности населения на основе временных рядов

Работу выполнил:

Шардт Максим Александрович

очная форма обучения

курс: 3; группа: 3об\_ИВТ-1/21

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Научный руководитель:

профессор

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Власова Елена Зотиковна

«\_\_\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2024г.

Санкт-Петербург

2024

# Оглавление

[Оглавление 2](#_gjdgxs)

[Введение 3](#_30j0zll)

[1. Теоретическая часть 4](#_1fob9te)

[1.1 Временные ряды 4](#_r5xerrqszfzs)

[1.2 Анализ временных рядов 5](#_2et92p0)

[1.2.1. Визуальный анализ 5](#_6zmtme48ss3f)

[1.2.2. Стационарность 6](#_p0qaphx8eszq)

[1.2.3. Моделирование и прогнозирование 6](#_bz0csnf1z3t)

[1.3 Подход Бокса-Дженкинса 7](#_4t09u5wsh417)

[1.3.1 Выбор порядка модели 7](#_u2qgbfmt4v6e)

[1.3.2 Оценка модели выбранного порядка 8](#_egv51ap0j0xl)

[1.3.3 Проверка “Адекватности” 8](#_xezcjb9myfw4)

[2. Практическая часть 9](#_tyjcwt)

[2.1 Первичная обработка данных 9](#_3dy6vkm)

[2.2 Анализ данных 10](#_1t3h5sf)

[Заключение 18](#_4d34og8)

[Литература 19](#_2s8eyo1)

[Приложение А 20](#_17dp8vu)

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Введение

Демографические данные играют важную роль в понимании и анализе состояния населения в различных регионах мира. Одним из таких регионов является Санкт-Петербург, второй по численности город Российской Федерации. Изучение демографических характеристик Санкт-Петербурга позволяет не только сформировать представление о текущем состоянии населения, но и провести прогнозы для планирования социально-экономического развития города.

Целью данного исследования является анализ демографических данных Санкт-Петербурга с использованием методов анализа временных рядов и построение модели для прогнозирования изменений численности населения города.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Провести первичную обработку данных о демографии.

2. Построить график динамики численности населения Санкт-Петербурга для визуальной оценки изменений в течение рассматриваемого периода.

3. Проверить стационарность временного ряда численности населения с использованием соответствующих статистических тестов.

4. Подобрать параметры модели ARIMA для прогнозирования численности населения на основе имеющихся данных.

5. Обучить модель ARIMA на тренировочных данных и получить прогнозы для сравнения с тестовыми данными.

Осуществление данных задач позволит получить более глубокое понимание динамики численности населения Санкт-Петербурга и разработать модель, которая может быть использована для прогнозирования будущих изменений в демографической ситуации города.

# 1. Теоретическая часть

## 1.1 Временные ряды

Временные ряды представляют собой последовательность наблюдений, измеренных в разные моменты времени и упорядоченных в хронологическом порядке. Они используются для анализа и моделирования изменений величин, которые меняются со временем.

Ключевая особенность временных рядов заключается в том, что каждое наблюдение в ряду связано с определенным моментом времени. Это позволяет выявить различные временные зависимости, такие как тренды, циклы, сезонность и случайные колебания.

1. *Тренд* представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени.
2. *Циклы* представляют собой повторяющиеся колебания величины в течение определенного периода времени. Они могут быть связаны с экономическими циклами, бизнес-циклами или другими регулярными паттернами. Циклы обычно имеют более длительную продолжительность, чем сезонность, и могут быть менее регулярными.
3. *Сезонность* связана с повторяющимися паттернами величины в течение года или другого фиксированного периода времени. Например, продажи мороженого могут иметь сезонность, когда они возрастают летом и снижаются зимой. Сезонность может быть связана с климатическими условиями, праздниками, школьными каникулами и другими факторами.
4. *Случайные колебания* представляют собой непредсказуемые флуктуации величины, которые не могут быть объяснены трендами, циклами или сезонностью. Они могут быть вызваны случайными событиями или факторами, которые не учтены в модели.

Анализ временных рядов позволяет выделить и оценить эти различные компоненты временного ряда. Это может быть полезно для понимания причинно-следственных связей, прогнозирования будущих значений, а также для принятия решений и разработки стратегий на основе исторических данных.[1]

## 1.2 Анализ временных рядов

Для анализа и моделирования временных рядов используются различные статистические методы, включая методы сглаживания, авторегрессионные модели (AR), скользящие средние модели (MA), авторегрессионные интегрированные скользящие средние модели (ARIMA) и другие более сложные модели.

Одной из основных задач анализа временных рядов является прогнозирование. На основе исторических данных можно построить модель, которая будет предсказывать будущие значения временного ряда. Прогнозирование может быть полезным для принятия решений, планирования, оптимизации ресурсов и других задач.

### 1.2.1. Визуальный анализ

Визуальный анализ временных рядов является первым этапом в изучении данных. Он позволяет наглядно представить основные характеристики ряда и выявить возможные временные зависимости и паттерны. Используются следующие методы визуального анализа временных рядов:

1. *Линейные графики*. На оси абсцисс отображается время, а на оси ординат - значения ряда. Это позволяет увидеть общую динамику ряда, его тренд, сезонность и случайные колебания. Линейные графики могут быть полезны для общего представления и оценки основных компонентов временного ряда.
2. *Диаграммы рассеяния* позволяют исследовать связь между двумя переменными во времени. Например, можно построить диаграмму рассеяния, где на одной оси отображаются значения временного ряда в момент t, а на другой оси - значения временного ряда в момент t+1. Это позволяет увидеть взаимосвязь между текущим и следующим наблюдениями и выявить возможные корреляции или зависимости

### 1.2.2. Стационарность

Понятие стационарного временного ряда означает, что его среднее значение не изменяется во времени, т.е. временной ряд не имеет тренда. Кроме того, ковариация между разными элементами временного ряда (как между случайными величинами) зависит только от того, насколько сильно они отдалены друг от друга во времени. Величина h, характеризующая разницу во времени между элементами временного ряда, называется лаговой переменной или запаздыванием. [3]

Временной ряд называется стационарным (в широком смысле), если

1. Ext ≡ const (среднее постоянно во времени);

2. cov(xt, xt+h) = γ(h) (ковариация зависит только от лага h). (1)

### 1.2.3. Моделирование и прогнозирование

Для анализа и прогнозирования временных рядов используются следующие модели: авторегрессионные модели (AR), скользящие средние модели (MA), авторегрессионные интегрированные скользящие средние модели (ARIMA) и другие. Модели позволяют описать и предсказать поведение временного ряда на основе его предыдущих значений и структуры данных. [3]

Общий вид модели ARIMA(p, q):

Используя лаговый оператор L () и два многочлена степени p и q модель можно записать в следующем виде:

(2)

φ(z) называется авторегрессионным многочленом,

φ(L)xt называется авторегрессионной частью модели,

θ(L)ut – частью скользящего среднего.

Модель MA(q) или модель скользящего среднего, записывается следующим образом:

xt = μ + ut + θ1ut−1 + · · · + θq ut−q , (3)

где ut ∼ WN(0, σ2u), θq ≠ 0.

Модель AR(p) или модель авторегрессии имеет вид:

xt = μ + φ1xt−1 + · · · + φpxt−p + ut, (4)

где ut ∼ WN(0, σ2u), φp ≠ 0.

## 1.3 Подход Бокса-Дженкинса

Метод Бокса-Дженкинса является одним из наиболее распространенных подходов к моделированию и прогнозированию временных рядов.

Работа будет проходить в следующем виде:

1. Выбор порядка модели
2. Оценка модели выбранного порядка
3. Проверка “адекватности” [2]

### 1.3.1 Выбор порядка модели

Для выбора порядка модели будет вычислена выборочная и частная автокорреляция.

* Выборочные коэффициенты автокорреляции

(5)

* Выборочные коэффициенты автокорреляции (ACF)

(6)

### 1.3.2 Оценка модели выбранного порядка

Оценивание модели ARIMA(p, q) происходит с помощью метода максимального правдоподобия.

* Проверка значимости коэффициента: тестовая статистика

или (7)

### 1.3.3 Проверка “Адекватности”

Для проверки адекватности модели используется метод исследования, предложенный Боксом и Льюнгом на основе Q-статистик. Тесты подразумевают проверку гипотезы [1]:

* Тестовая Q-статистика Льюинга-Бокса

. (8)

где – автокорреляционная функция ACF для ряда остатков.

# 2. Практическая часть

## 2.1 Первичная обработка данных

Первичная обработка данных включает в себя несколько шагов, которые позволяют подготовить данные для дальнейшего анализа. В данном случае рассматривается обработка данных о демографии в Санкт-Петербурге.

Для обработки данных будут использоваться библиотеки pandas, statsmodels и matplotlib на языке Python. Данные были получены с официального сайта Федеральной службы государственной статистики в формате xlsx.

Используя функцию read\_excel() из библиотеки pandas для чтения данных из файла. Указываем, что нам нужно использовать только столбцы "A" и "B" Мы также указываем формат даты "%Y" для правильного разбора года, и устанавливаем столбец "Год" в качестве индекса данных.

| demography = pd.read\_excel(  "Демография СПб.xlsx",  names=header,  usecols="A:B",  date\_format="%Y",  parse\_dates=[0],  index\_col=[0], ) |
| --- |

Листнг 1. Чтение и первичная обработка данных

Далее создается новый индекс даты с помощью функции date\_range(). Мы задаем начальную и конечную даты с частотой "AS" (годовая начальная дата), чтобы создать годовые интервалы.

| new\_index = pd.date\_range(start="1926-01-01", end="2023-01-01", freq="AS") |
| --- |

Листинг 2. Создание нового индекса

Затем используется метод reindex() для переиндексации данных демографии с использованием нового индекса. Это позволит добавить отсутствующие годы в данные.

| demography = demography.reindex(new\_index) |
| --- |

Листинг 3. Создание дового индекса

Далее мы выполняем интерполяцию значений столбца "Численность" с помощью метода interpolate(). Интерполяция позволяет заполнить пропущенные значения между существующими значениями, чтобы получить более плавную и непрерывную последовательность данных, а также для корректной работы моделей далее.

Наконец, мы преобразуем индекс данных в объект DatetimeIndex с помощью функции DatetimeIndex() и устанавливаем ему частоту "AS" (годовая начальная дата). Это позволяет нам работать с временными данными и выполнять различные операции, связанные с временными рядами.

| demography.index = pd.DatetimeIndex(demography.index, freq="AS") |
| --- |

Листинг 4. Присваивание нового индекса

Таким образом, после применения всех этих шагов, данные о демографии в Санкт-Петербурге будут готовы для дальнейшего анализа и моделирования временных рядов.

## 2.2 Анализ данных

Для построения графика численности населения Санкт-Петербурга используется библиотека matplotlib.

График можно построить используя метод plot библиотеки matplotlib. Передадим ему следующие значения и настроим график:

| import matplotlib.pyplot as plt plt.plot(demography.index, demography['Численность']) plt.xlabel('Год') plt.ylabel('Численность населения') plt.title('Динамика численности населения Санкт-Петербурга') plt.grid(True) plt.show() |
| --- |

Листинг 5. Создания графика численности населения Санкт-Петербурга

Этот код построит следующий график:



График 1. Визуализация численности населения Санкт-Петербурга

С помощью функции индексирования из библиотеки pandas возможно приблизить график на участок с 2000 года:



График 2. Визуализация численности населения Санкт-Петербурга c 2000 по 2023 года

Для определения стационарности временного ряда можно воспользоваться воспользоваться условиями (1) или функцией adfuller из библиотеки statsmodels следующим образом:

| adf\_test = adfuller(demography1["Численность"]) print("p-value = " + str(adf\_test[1])) |
| --- |

Листинг 6. Определение стационарности временного ряда



Рис. 1. Результат вычисления уровня значимости

Для анализа стационарности временного ряда были сформулированы следующие гипотезы:

* Нулевая гипотеза (H0): Временной ряд является стационарным.
* Альтернативная гипотеза (H1): Временной ряд не является стационарным.

Для проверки данных гипотез был использован тест на стационарность с уровнем значимости 0.05. Полученное значение равно 0.62917, что превышает уровень значимости. Следовательно, на уровне значимости 0.05 первая гипотеза (H0) отвергается, а нулевая гипотеза принимается. Это означает, что временной ряд не является стационарным.

Отсутствие стационарности временного ряда может свидетельствовать о наличии тренда. Тренд может указывать на долговременное изменение в динамике временного ряда, такое как постепенный рост или снижение значений.

Для получения значений значимости стационарности ряда для разных разностей рядов применяется функция diff, которая работает по формулам (3) и (4), а затем функция dropna для удаления пустых значений. Значения на первых и вторых уровнях значимости равны соответственно. Далее эти значения будут использоваться для подбора параметров модели ARIMA.



Рис. 2. Уровни значимости первой и второй разности рядов

Также можно применить экспоненциальное сглаживание для снижения влияния случайных колебаний на графике следующим образом:

| alpha = 0.2  exp\_smoothing = [demography1["Численность"][0]]  for i in range(1, len(demography1["Численность"])):  exp\_smoothing.append(  alpha \* demography1["Численность"][i] + (1 - alpha) \* exp\_smoothing[i - 1]  )  demography1.insert(1, "Экспоненциальное Сглаживание", exp\_smoothing, True) |
| --- |

Листинг 7. Экспоненциальное сглаживание

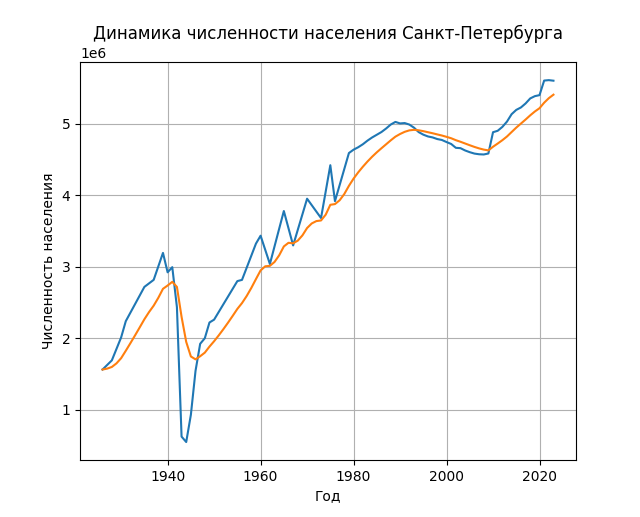


График 3. Визуализация экспоненциального сглаживания

Для корректного определения аргументов для модели ARIMA необходимо составить автокорреляцию ряда. Это можно сделать с помощью функций lag\_plot и autocorrelation\_plot библиотеки pandas, а также plot\_acf библиотеки statsmodels:

| lag\_plot(demography)  plot\_acf(demography)  plt.show() |
| --- |

Листинг 8. Создание автокорреляционных графиков

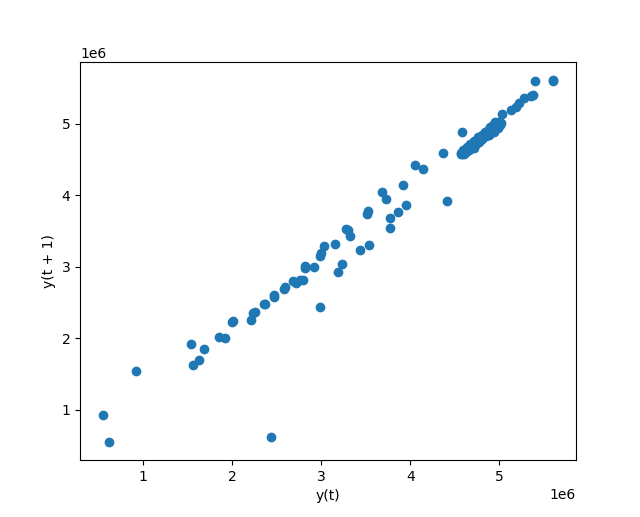


График 4. График автокорреляции библиотеки pandas

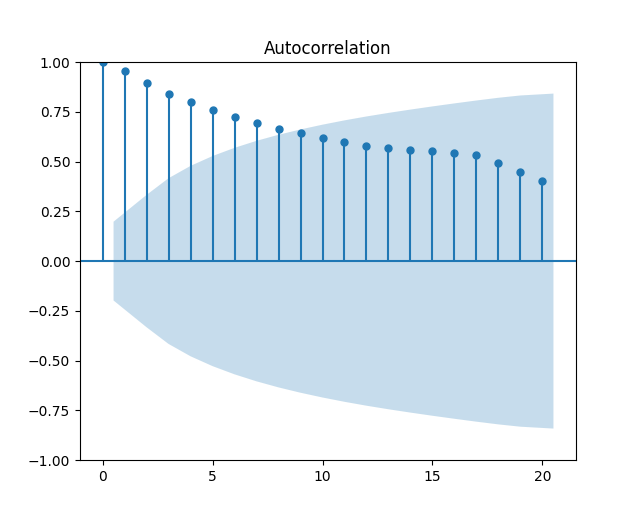


График 5. График автокорреляции библиотеки statsmodels

Создание модели ARIMA проходит в несколько этапов:

1. Выборка тренировочных данных

Модель будет обучаться на данных с 1950 по 2018 год, сравниваться прогноз модели будет с данными после 2018 года.

| train = demography["1950":"2018"]['Численность'] test = demography["2018":]['Численность'] |
| --- |

Листинг 9. Создание тестовых данных

1. Подбор параметров модели

Подбор параметров модели осуществляется с помощью уже полученных данных авторегрессии, а также максимального сходства прогноза модели с существующими данными.

В данном случае параметр p = 5, d = 2, q = 20.

1. Создание модели по формуле (2)

Модель создается и обучается с помощью класса ARIMA из библиотеки statsmodel. Затем вызывается метод predict для создания прогноза с начала данных для сравнения до их конца.

| model = ARIMA(train, order=(5, 2, 10)).fit()  start = len(train) - 1  end = len(train) + len(test) - 1  predictions = model.predict(start, end) |
| --- |

Листинг 10. Создание модели ARIMA

1. Визуализация модели

Визуализация модели происходит с помощью библиотеке matplotlib следующим образом:

| plt.plot(train, color="black")  plt.plot(test, color="red")  plt.plot(predictions, color="green")  plt.xlabel("Год")  plt.ylabel("Численность населения")  plt.title("Анализ численности населения")  plt.grid(True)  plt.show() |
| --- |

Листинг 11. Визуализация модели ARIMA



График 6. График с прогнозом модели

Значение Q-статистики Льюинга-Бокса по формуле (8) составляет 0.02, что может говорит о том, что модель хорошо соответствует представленным данным.

Полный код программы приведен в Приложении А.

# Заключение

В данной работе была проведен анализ демографических данных Санкт-Петербурга с использованием методов анализа временных рядов и моделирования. Целью исследования было изучение динамики численности населения города и создание модели ARIMA для прогнозирования изменений в демографической ситуации.

В результате проведенного анализа было выявлено, что численность населения Санкт-Петербурга имеет тенденцию к росту, однако наблюдаются колебания в течение рассматриваемого периода.

С использованием модели ARIMA были получены прогнозы для численности населения на основе имеющихся данных. Однако необходимо учитывать, что модель может иметь ограничения и не всегда точно предсказывать будущие значения. Значение Q-статистики Льюинга-Бокса указывает на хорошее соответствие модели представленным данным.

Данные результаты позволяют получить более глубокое понимание динамики численности населения Санкт-Петербурга и использовать модель ARIMA для прогнозирования будущих изменений в демографической ситуации города. Это может быть полезным для планирования социально-экономического развития Санкт-Петербурга и принятия соответствующих решений в области инфраструктуры, образования, здравоохранения и других сфер, связанных с населением.

# Литература

1. Мишулина О. А. Статистический анализ и обработка временных рядов. — М.: МИФИ, 2004.
2. Айвазян С. А. Прикладная статистика. Основы эконометрики. Том 2. — М.: Юнити-Дана, 2001.
3. Н. В. Артамонов. Введение в анализ временных рядов. Учебное пособие для вузов — М.: МГУ им. М. В. Ломоносова, 2021
4. StatSoft.ru. Анализ временных рядов. [Онлайн] Доступно по: http://statsoft.ru/home/textbook/modules/sttimser.html [Дата обращения: 3.12.2023].

# Приложение А

Код программы

header = ["Год", "Численность"]

demography = pd.read\_excel(

"Демография СПб.xlsx",

names=header,

usecols="A:B",

date\_format="%Y",

parse\_dates=[0],

index\_col=[0],

)

new\_index = pd.date\_range(start="1926-01-01", end="2023-01-01", freq="AS")

demography = demography.reindex(new\_index)

demography["Численность"] = demography["Численность"].interpolate()

demography.index = pd.DatetimeIndex(demography.index, freq="AS")

# Стационарность временного ряда

adf\_test = adfuller(demography["Численность"])

print("p-value = " + str(adf\_test[1]))

adf\_test = adfuller(demography["Численность"].diff().dropna())

print("p-value = " + str(adf\_test[1]))

adf\_test = adfuller(demography["Численность"].diff().diff().dropna())

print("p-value = " + str(adf\_test[1]))

lag\_plot(demography)

autocorrelation\_plot(demography)

plot\_acf(demography)

plt.show()

# Экспоненциальное сглаживание

alpha = 0.2

exp\_smoothing = [demography["Численность"][0]]

for i in range(1, len(demography["Численность"])):

exp\_smoothing.append(

alpha \* demography["Численность"][i] + (1 - alpha) \* exp\_smoothing[i - 1]

)

demography.insert(1, "Экспоненциальное Сглаживание", exp\_smoothing, True)

plt.plot(demography)

plt.xlabel('Год')

plt.ylabel('Численность населения')

plt.title('Динамика численности населения Санкт-Петербурга')

plt.grid(True)

plt.show()

train = demography["1950":"2018"]["Численность"]

test = demography["2018":]["Численность"]

model = ARIMA(train, order=(5, 2, 30)).fit()

start = len(train) - 1

end = len(train) + len(test) - 1

predictions = model.predict(start, end)

print(model.summary())

plt.plot(train, color="black")

plt.plot(test, color="red")

plt.plot(predictions[:-1], color="green")

plt.legend(["Тренировочные данные", "Данные для сравнения", "Прогноз"])

plt.xlabel("Год")

plt.ylabel("Численность населения")

plt.title("Анализ численности населения")

plt.grid(True)

plt.show()