Некоторые статистические методы прогнозирования заболеваемости COVID-19 на региональном уровне

# Введение

Прогнозирование заболеваемости различными вирусными инфекциями, наряду с моделированием указанных процессов, является важной задачей, особенно в условиях пандемии и эпидемий. Обоснованный прогноз позволяет оптимально использовать наличные ресурсы и спланировать ограничительные меры таким образом, чтобы минимизировать негативные последствия для экономики и здоровья населения региона.

Объектом исследования является временной ряд, отражающий заболеваемость COVID-19 в регионе. Так как методы анализа и прогнозирования должны соответствовать характеру данных, то анализу подвергается либо количество заболевших накопленным итогом, либо посуточная заболеваемость.

В настоящей работе не рассматриваются особенности сбора статистических данных и их влияние на результаты моделирования.

# Инструментальные средства

Настоящая работа рассматривает некоторые методы анализа и прогнозирования заболеваемости на примере Псковской области, как типичного региона Север-Запада Российской Федерации и их реализацию средствами таких распространенных и доступных свободно средств как универсальный язык программирования Python ([www.python.org](http://www.python.org)) и ориентированный на статистические исследования язык R ([www.r-project.org](http://www.r-project.org)).

Авторы использовали дистрибутив Anaconda ([www.anaconda.com](http://www.anaconda.com)) и редактор Visual Studio Code ([www.code.visualstudio.com/download](http://www.code.visualstudio.com/download)) а также RStudio ([www.posit.co/products/open-source/rstudio/](http://www.posit.co/products/open-source/rstudio/)).

Тексты программ приводятся в приложениях и доступны в репозитории Github ().

# Обзор и классификация методов прогнозирования

## Методы, основанные на регрессионном анализе

Временные ряды представляют собой данные, расположенные через равные промежутки времени. Таким образом, они содержат, как минимум, две переменных – время и анализируемую величину. В традиционном статистическом анализе регрессором является количество заболевших, а предиктором – время, то есть дата наблюдения. Примером функций времени, которые могут быть использованы в регрессионном анализе временных рядов являются линейный тренд, полином или логистическая кривая, в частности линейная функция времени может быть использована для моделирования общего направления изменения временного ряда, а логистическая – для процессов, имеющих горизонтальную асимптоту (например, предельное количество заболевших).

Необъясненные трендом остатки могут быть проанализированы другими методами.

## Специальная модель регрессии – скрытая марковская модель

Представляет особый интерес, поскольку позволяет оценить по наблюдаемым данным скрытые от наблюдателя состояния системы. Предполагается, что наблюдаемая величина (количество заболевших) является следствием некоторого дискретного состояния (например, состояние волны/плато в период пандемии или действующего штамма вируса).

Существуют и более сложные версии марковской регрессии, в которых в зависимости от дискретных состояний меняются коэффициенты уравнений регрессии.

## Методы, предназначенные для классического анализа временных рядов

Для применения методов этой группы желательно, чтобы временной ряд носил стационарны характер (имел постоянное среднее и стандартное отклонение), либо мог быть приведен к такой форме.

Классический анализ временных серий предполагает, что независимая переменная в момент наблюдения является следствием своих предыдущих значений, например, средним за предыдущую неделю или линейной функцией от предыдущих значений (скользящее среднее и регрессия соответственно).

### ARIMA

Метод ARIMA комбинирует три основных компонента: авторегрессию (AR - Autoregressive), интеграцию (I - Integrated) и скользящее среднее (MA - Moving Average). Каждая из этих компонент имеет свою роль в модели ARIMA: авторегрессия (AR) используется для моделирования зависимости текущего значения ряда от предыдущих значений ряда через линейную комбинацию предыдущих значений; интеграция (I) используется для обеспечения стационарности ряда путем вычитания разности между последовательными значениями; скользящее среднее (MA) моделирует зависимость текущего значения ряда от прошлых ошибок прогноза (разницы между прогнозом и фактическим значением). MA-компонента позволяет учесть корреляцию в ошибках прогноза.

В наиболее развитой форме указанный метод был предложен Боксом и Дженкинсом. Временной ряд рассматривается как совокупность тренда (линейной функции), на который наслаивается цикличная во времени сезонная компонента и необъясненный первыми двумя составляющими случайный шум.

### Экспоненциальное сглаживание

Метод используется для прогнозирования будущих значений на основе исторических наблюдений. Он основан на идее, что более новые данные имеют большую важность при прогнозировании, чем более старые данные. Таким образом, каждое последующее значение ряда вычисляется как взвешенная сумма предыдущего значения ряда и актуального наблюдения. Веса определяются коэффициентом сглаживания, который задает уровень влияния новых данных на прогноз. Формула для расчета значений экспоненциально сглаженного ряда выглядит следующим образом: , где: - значение экспоненциально сглаженного ряда в момент времени , - актуальное наблюдение (текущее значение ряда) в момент времени , - значение экспоненциально сглаженного ряда в предыдущий момент времени , - коэффициент сглаживания, принимающий значения от 0 до 1.

Часто используется модификация, известная как метод Винтерса (Holt-Winters' method), которая расширяет базовый метод экспоненциального сглаживания, вводя компоненты тренда и сезонности. Уровень представляет собой фиксированную базовую составляющую ряда, тренд отражает долгосрочное изменение в данных, а сезонность объясняет повторяющиеся паттерны или колебания в ряде.

### Facebook Prophet

Открытая библиотека прогнозирования временных рядов. Она предоставляет простой в использовании и эффективный инструмент для анализа и прогнозирования временных рядов. Она предлагает автоматическую обработку таких задач, как обнаружение трендов, сезонности и праздников, а также учет неопределенности и выбросов данных. Основные особенности и Facebook Prophet: простота использования, гибкость, автоматизация. Однако стоит отметить, что Prophet не является универсальным инструментом и может иметь свои ограничения, особенно для сложных временных рядов с нерегулярной структурой или высокой степенью шума.

## Методы, основанные на применении искусственных нейронных сетей

# Использование методов прогнозирования для анализа заболеваемости COVID-19 в Псковской области

Для большинства последующих частей статьи необходимо загрузить данные о заболеваемости COVID-19, что делает скрипт load\_csv.py (Приложение 1. Загрузка данных)

## Анализ методом скрытой марковской регрессии

Методы, основанные на регрессии

Анализ временного ряда

Анализ результатов

Выводы

# Приложения

## Приложение 1. Загрузка данных

# %%

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

file = "COVID\_PSK.csv"

df = pd.read\_csv(file,sep=';',encoding='utf-8',index\_col='Date')

df["inf\_day"] = df["Infections"].diff(periods=1)

print(df.head())

print(df.columns.tolist())

#%%

# Assuming your DataFrame is called 'df'

df['Date'] = df.index

#%%

# Convert Date column to numerical format

df['Date'] = pd.to\_datetime(df['Date'],format='%d.%m.%Y',dayfirst=True)

print(df['Date'])

#%%

# Convert Date column to numerical format and convert from nanoseconds to days

df['Date'] = pd.to\_numeric(df['Date']-df['Date'].min()) / (24\*60\*60\*1000000000)

print(df['Date'])

#%%

if \_\_name\_\_ == '\_\_main\_\_':

    #plot the data using seaborn versus Date

    fig, axes = plt.subplots(2,1, figsize=(10, 8))

    sns.lineplot(data= df, ax =axes[0])

    sns.lineplot(data= df, x="Date", y="inf\_day", ax =axes[1])

    plt.show()

    print("Save to csv for using next time1")

    df.to\_csv("df.csv")