МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**КУРСОВАЯ РАБОТА**

**ПОДХОДЫ К МОДЕЛИРОВАНИЮ ЭПИДЕМИЙ**

Работу выполнил\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_И.В.Толстиков

(подпись)

Направление подготовки 02.03.02 — «Фундаментальная информатика и информационные технологии» курс 3

Направленность (профиль) «Вычислительные технологии»

Научный руководитель

доц., канд.т.н.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Т.А.Приходько

(подпись, дата)

Нормоконтролер

преподаватель\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.А.Климец

(подпись, дата)

Краснодар

2020

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 3](#_Toc58666358)

[1 Подходы к моделированию эпидемий 5](#_Toc58666359)

[1.1 Модели на графах 5](#_Toc58666360)

[1.1.1 Формальная постановка задачи 5](#_Toc58666361)

[1.1.2 Модель SI 5](#_Toc58666362)

[1.1.3 Модель SIR 6](#_Toc58666363)

[1.1.4 Модель SEIR 7](#_Toc58666364)

[1.1.5 Модель SEIRD 9](#_Toc58666365)

[1.2 Стратегии многошагового прогнозирования временных рядов 11](#_Toc58666366)

[1.2.1 Формальная постановка задачи 11](#_Toc58666367)

[1.2.2 DirRec 11](#_Toc58666368)

[2 Программная реализация моделей 13](#_Toc58666369)

[2.1 Инструменты анализа данных 13](#_Toc58666370)

[2.2 Реализация моделей на графах 13](#_Toc58666371)

[2.3 Реализация прогнозирования временного ряда 14](#_Toc58666372)

[2.4 Данные для реализации и проверки моделей 14](#_Toc58666373)

[3 Выделение лучшей модели на графах 15](#_Toc58666374)

[4 Сравнение алгоритма взвешенной скользящей средней с моделью SEIRD 20](#_Toc58666375)

[5 Тестирование на реальных данных 24](#_Toc58666376)

[ЗАКЛЮЧЕНИЕ 26](#_Toc58666377)

[**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27**](#_Toc58666378)

[**ПРИЛОЖЕНИЕ А Код программы 28**](#_Toc58666379)

# **ВВЕДЕНИЕ**

Эпидемия — прогрессирующее во времени и пространстве распространение инфекционного заболевания среди людей, значительно превышающее обычно регистрируемый на данной территории уровень заболеваемости и способное стать источником чрезвычайной ситуации [1]. Это одно из самых губительных для человека опасных природных явлений. Статистика свидетельствует о том, что инфекционные заболевания унесли больше человеческих жизней, чем войны. Поэтому при возникновении очага инфекционного заражения на пораженной территории вводится карантин или обсервация. Однако с учётом того, как карантин влияет на общественную жизнь и экономику государства, важно определить, когда и на какой период нужно вводить карантин, чтобы минимизировать последствия от возникшей эпидемии. С этой целью строится модель прогнозирования, функциональное представление, адекватно описывающее исследуемый процесс и являющееся основой для получения его будущих значений. Построение и изучение моделей реально существующих объектов, процессов или явлений с целью получения объяснений этих явлений, а также для предсказания явлений называется моделированием.

Целью курсовой работы является изучение подходов моделирования эпидемий. В частности, рассматривается задача моделирования COVID-19.

Курсовая работа состоит из пяти глав:

Первая глава содержит общие теоретические сведения о задаче моделирования эпидемии.

Вторая глава описывает программную реализацию моделей прогнозирования эпидемий.

В третьей главе происходит выделение лучшей модели на графах из рассматриваемых.

В четвёртой главе происходит сравнение стратегии многошагового прогнозирования временных рядов с лучшей моделью на графах.

В пятой главе проводится тестирование лучшей модели на графах на реальных данных.

# **1 Подходы к моделированию эпидемий**

В качестве подходов рассматриваются прогнозирование с помощью моделей на графах и прогнозирование временных рядов.

## 1.1 Модели на графах

### 1.1.1 Формальная постановка задачи

Рассмотрим граф, в котором вершинам сопоставлены люди, а ребрам – наличие социальной связи. Такого рода графы часто называют социальными сетями. В момент времени ( – дискретная величина) для графа определены:

* – множество восприимчивых людей (susceptible)
* – множество больных людей (infected)

Цель – предсказание появления новых заболевших к моментам времени на основе истории заболеваний за предшествующих моментов времени.

### 1.1.2 Модель SI

Модель подробно описана в [1,2]. Данный подход основывается на утверждении: для любого промежутка времени верно, что количество человек, присоединившихся к больным, равно количеству человек, переставших быть восприимчивыми. Формальный вид модели SI представлен на рисунке 1.

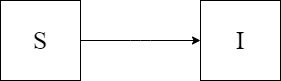


Рисунок 1 – Модель SI

Пусть

* – вероятность заражения восприимчивого при контакте с больным
* – количество восприимчивых людей в момент времени
* – количество больных людей в момент времени
* – общее количество людей

Тогда процесс распространения болезни описывается системой дифференциальных уравнений (1).

### 1.1.3 Модель SIR

Модель SIR подробно описана в [1,2]. Данный подход учитывает возможность перехода человека из множества больных () в множество здоровых (). Человек, попавший в множество здоровых, не может вернуться в множество больных. Формальный вид модели SIR представлен на рисунке 2.



Рисунок 2 – Модель SIR

Пусть

* – вероятность заражения восприимчивого при контакте с больным
* – вероятность выздоровления больного
* – количество восприимчивых людей в момент времени
* – количество больных людей в момент времени
* – количество здоровых в момент времени t
* – общее количество людей

Тогда процесс распространения болезни описывается системой дифференциальных уравнений (2).

### 1.1.4 Модель SEIR

Модель SEIR подробно описана в [3]. Данный подход дополняет модель SIR тем, что переходом между множеством восприимчивых (S) и множеством больных (I) служит множество (E) людей, у которых болезнь находится в инкубационном периоде. Формальный вид модели SEIR представлен на рисунке 3.

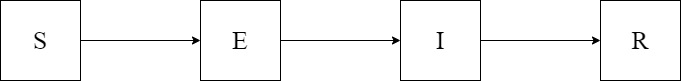


Рисунок 3 – Модель SEIR

Пусть

* – вероятность заражения восприимчивого при контакте с больным
* – скорость перехода болезни из инкубационной стадии в открытую
* – вероятность выздоровления больного
* – количество восприимчивых людей в момент времени
* – количество людей, у которых болезнь находится в инкубационном периоде
* – количество больных людей в момент времени
* – количество здоровых в момент времени t
* – общее количество людей

Тогда процесс распространения болезни описывается системой дифференциальных уравнений (3).

Можно заметить, что данная модель не учитывает возможность заражения восприимчивых людей людьми из множества , что для задачи моделирования COVID-19 является большим минусом. Поэтому в дальнейшем будет использоваться усовершенствованная модель SEIR.

Пусть – вероятность заражения человека из множества человеком из множества . Тогда процесс распространения болезни описывается системой дифференциальных уравнений (4).

### 1.1.5 Модель SEIRD

Данный подход дополняет модель SEIR наличием множества людей, умерших от болезни. Формальный вид модели SEIRD представлен на рисунке 4.

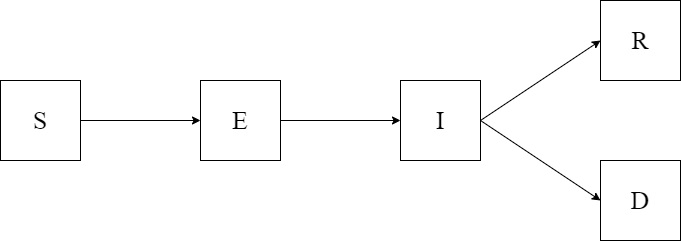


Рисунок 4 – Модель SEIRD

Пусть

* – вероятность заражения восприимчивого при контакте с больным
* – скорость перехода болезни из инкубационной стадии в открытую
* – вероятность выздоровления больного
* – вероятность летального исхода больного
* – количество восприимчивых людей в момент времени
* – количество людей, у которых болезнь находится в инкубационном периоде
* – количество больных людей в момент времени
* – количество здоровых в момент времени t
* – количество умерших в момент времени t
* – общее количество людей

Тогда процесс распространения болезни описывается системой дифференциальных уравнений (5).

Снова модифицируем модель, добавив возможность заражения восприимчивых людей людьми, у которых болезнь находится в инкубационном периоде.

Пусть – вероятность заражения человека из множества человеком из множества . Тогда процесс распространения болезни описывается системой дифференциальных уравнений (6).

## 1.2 Стратегии многошагового прогнозирования временных рядов

### 1.2.1 Формальная постановка задачи

Задача многошагового прогнозирования временного ряда (долговременного прогнозирования) состоит в получении прогнозов на шагов вперед временного ряда , состоящего из наблюдений, где определяет горизонт прогнозирования (количество шагов вперед для предсказания).

### 1.2.2 DirRec

DirRec стратегия, подробно описанная в [4,5], комбинирует архитектуры и принципы прямого и итеративного подходов ("direct" (Dir) – "прямой" в переводе с английского, "recursive" (Rec) – "рекурсивный"). DirRec предполагает обучение моделей, каждая для прогнозирования своего горизонта, и на каждом шаге модель дополняет множество входных наблюдений прогнозом с предыдущего шага. Таким образом длина учитываемой истории значений модели различается для разных горизонтов. Другими словами, DirRec стратегия обучает моделей на основе наблюдений ряда. Стратегия DirRec описывается уравнением (7).

где и .

В качестве алгоритма этой стратегии будет использоваться взвешенная скользящая среда. Этот алгоритм позволяет узнать прогноз на следующий день, суммируя последних наблюдения, каждому из которых назначается вес , зависящий от того, какие дни больше влияют на прогноз. Взвешенная скользящая средняя описывается уравнением (8).

# **2 Программная реализация моделей**

## 2.1 Инструменты анализа данных

Одним из наиболее популярных языков программирования для анализа данных является Python. Для него создано большое количество библиотек и фреймворков. В работе используется среда разработки PyCharm. Для визуализации в работе используются библиотеки matplotlib. Хранение данных осуществляется в оперативной памяти с помощью библиотеки pandas. Pandas – программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. Для решения дифференциальных уравнений используется метод odeint класса integrate из модуля scipy. Руководство по грамотной работе с перечисленными библиотеками описано в [6].

## 2.2 Реализация моделей на графах

Модели на графах реализованы в модуле graph\_model.py в виде методов с названием моделей, которые они реализуют. Пример реализации моделей на графах изображен на рисунке 5.

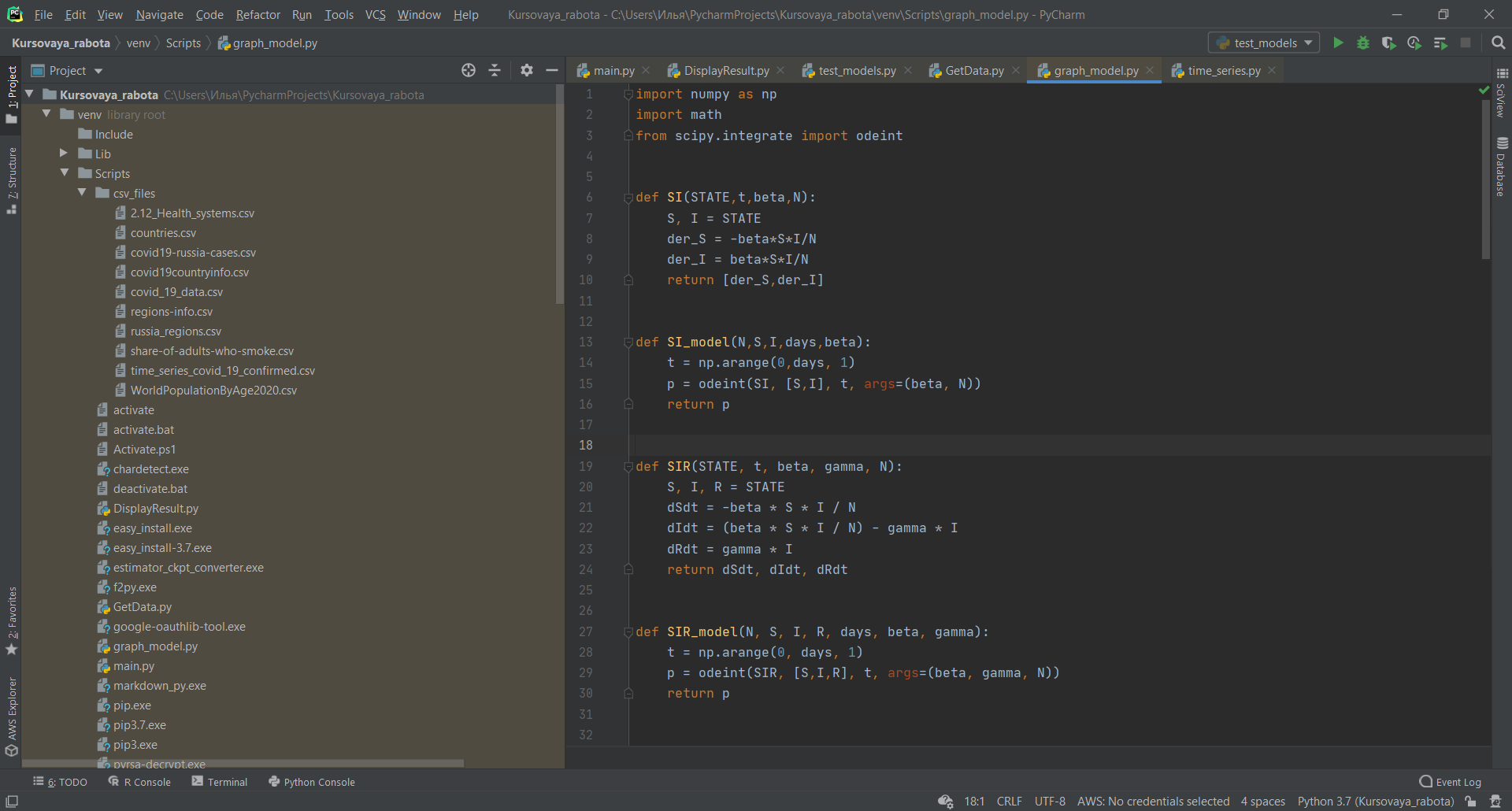


Рисунок 5 – Пример кода реализации моделей на графах

## 2.3 Реализация прогнозирования временного ряда

Алгоритмы прогнозирования временных рядов реализован в модуле time\_series.py. Реализации алгоритмов изображена на рисунке 6. Использовались скользящая средняя и взвешенная скользящая средняя.

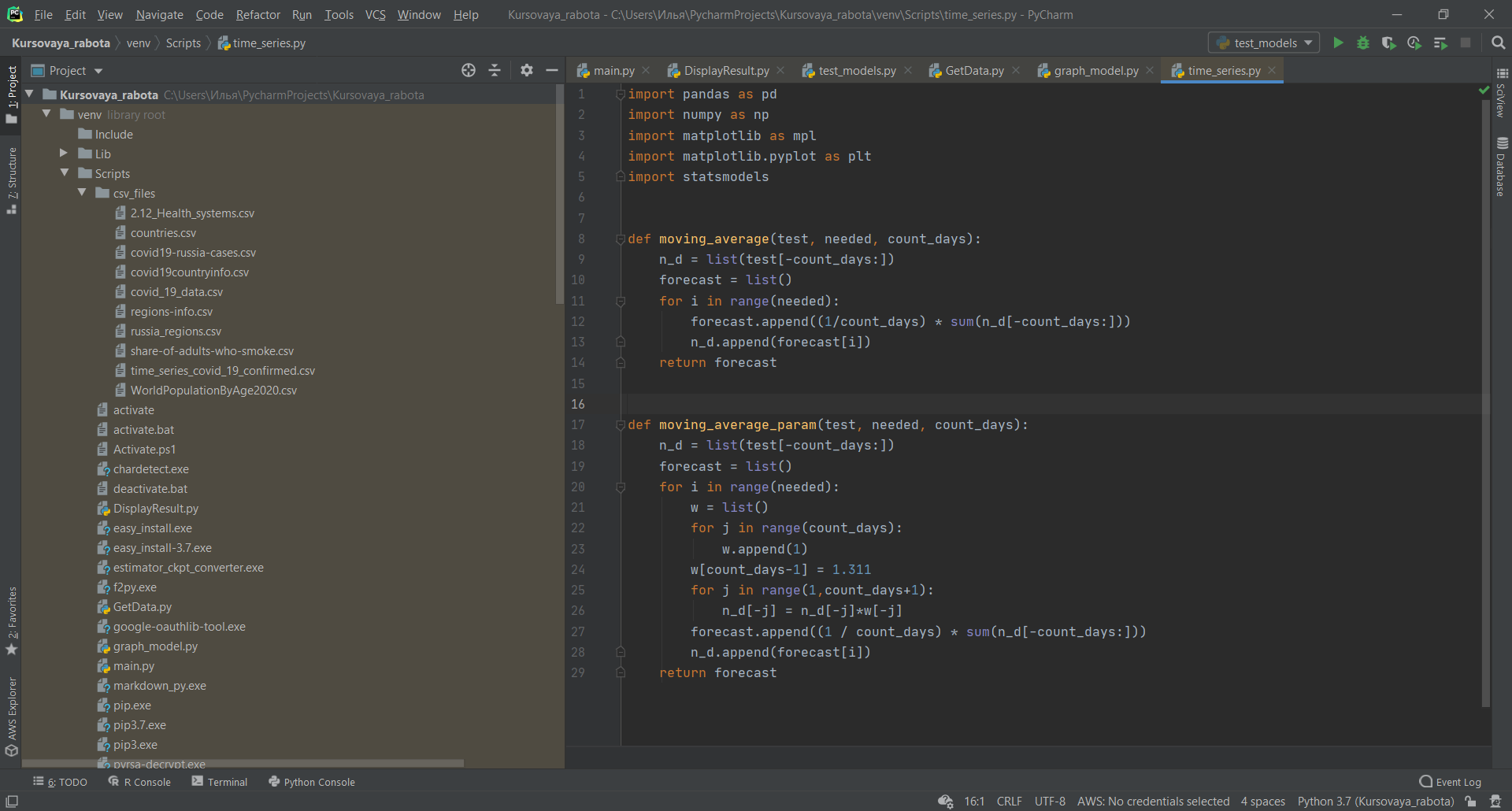


Рисунок 6 – Пример кода реализации алгоритма прогнозирования временных рядов

## 2.4 Данные для реализации и проверки моделей

В качестве начальных и проверочных данных используется информация о ситуации с COVID-19 по разным странам, которая была взята из [7,8]. Вся информация хранится в файле covid\_19\_data.csv. Для грамотного анализа данных использовались советы, описанные в [9].

# **3 Выделение лучшей модели на графах**

Тестирование моделей выполнено на основе данных по России, затрагивающих период с 11/03/2020 до 20/04/2020.

Параметры моделей принимают следующие начальные значения:

* За общее количество людей () принято население России за 11/03/2020.
* Количество больных людей () принято равным количеству больных COVID-19 в России за 11/03/2020.
* Количество выздоровевших людей () принято равным количеству выздоровевших от COVID-19 в России за 11/03/2020.
* Количество людей с инкубационной стадией болезни () принято равным произведению количества больных () на среднее количество людей, которое заражает человек, больной COVID-19, равное 2.5. Это количество называется индексом репродукции.
* Вероятность заражения восприимчивого при контакте с больным ( принята равной произведению вероятности выздоровления больного () на 4. Значение, большее, чем текущий индекс репродукции, берётся с целью показать, как быстро развивалась болезнь на начальной стадии.
* Скорость перехода болезни из инкубационной стадии в открытую () по умолчанию принята равной величине, обратной среднему инкубационному периоду. Для COVID-19 инкубационный период в среднем составляет 5 дней.
* Вероятность выздоровления больного () по умолчанию положена равной величине, обратной средней длительности болезни. Для COVID-19 средняя длительности болезни составляет 14 дней.
* Вероятность летального исхода () положена равной произведению вероятности выздоровления больного () на 0.1. Такое значение берётся, поскольку процент смертности от COVID-19 меньше, чем процент выздоровления.

Начнём анализ с самой простой модели. Модель SI с указанными параметрами даёт прогноз числа заболевших, изображенный на рисунке 7.

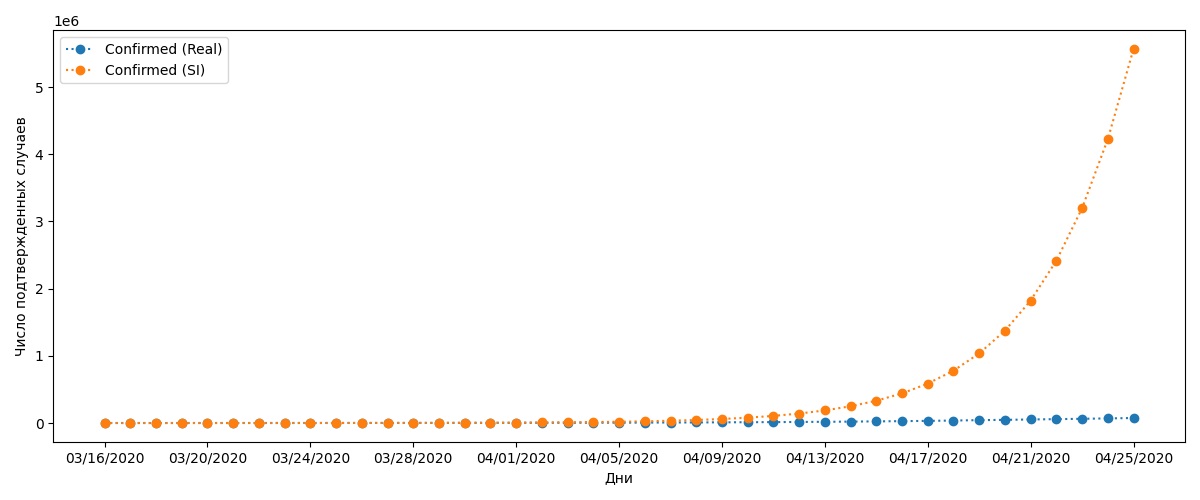


Рисунок 7 – Прогноз модели SI (на 41 день)

Модель даёт крайне неоптимистичный прогноз, сильно отличающийся от действительной ситуации. Это связано с тем, что существует всего два множества и и количество людей из множества убывает с такой же скоростью, с которой увеличивается количество людей из множества , что можно увидеть на рисунке 8.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 8 – Прогноз модели SI (на 60 дней)

Исходя из эксперимента, проведенного выше, можно сделать вывод о том, что модель SI не подходит для моделирования COVID-19.

В модели SIR, в отличие от модели SI, присутствует вероятность выздоровления, за счёт чего количество заболевших в прогнозе станет значительно меньше, что можно наблюдать на рисунке 9.

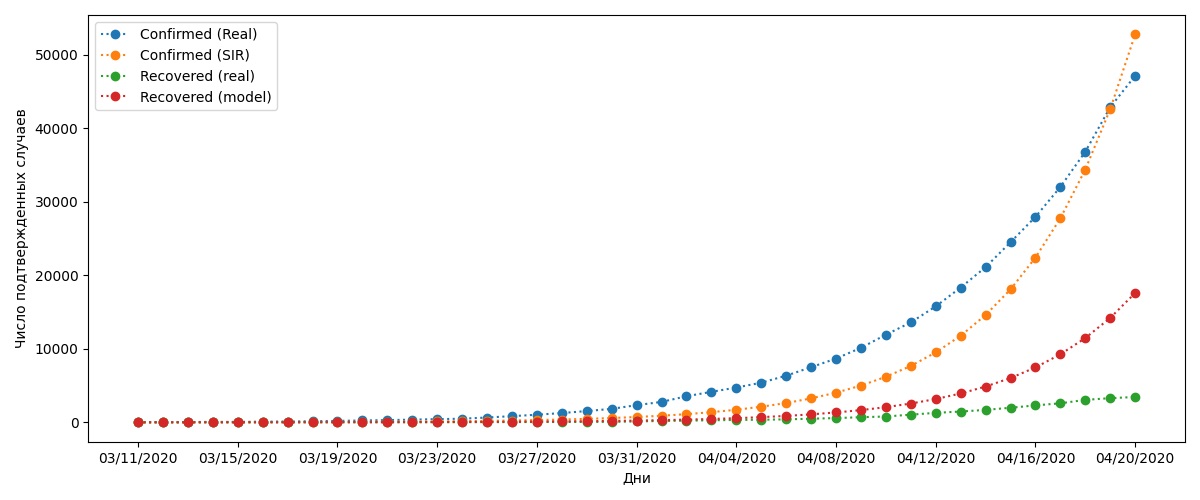


Рисунок 9 – Прогноз модели SIR (на 41 день)

Прогноз модели SIR уже больше приближен к реальным данным, чем прогноз модели SI, однако можно заметить, что рост заболевших происходит гораздо быстрее, из-за чего в дальнейшем, как показано на рисунке 10, ошибка будет только увеличиваться:

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 10 – Прогноз модели SIR (на 60 дней)

Это связано с тем, что заразившийся человек сразу переходит в множество , исключая инкубационный период, за время которого он может выздороветь.

Модель SEIR учитывает инкубационный период, поэтому прогноз, составленный этой моделью, изображенный на рисунке 11, является более реалистичным, чем прогнозы рассмотренных выше моделей:

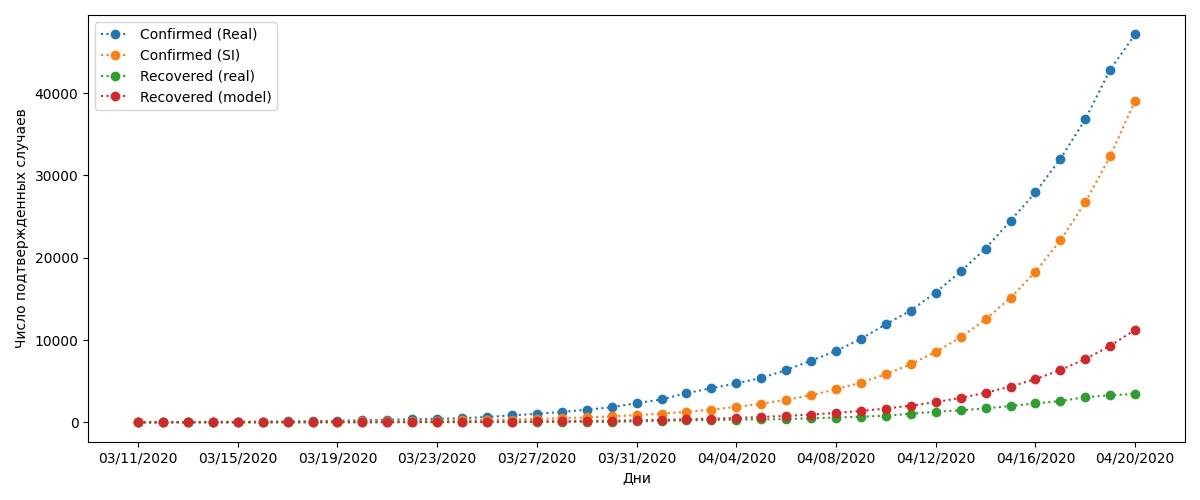


Рисунок 11 – Прогноз модели SEIR (на 41 день)

Конечно, этот прогноз нельзя назвать точным, но стоит учитывать, что все параметры подбирались из общепринятых значений для всех стран. Чтобы построить модель, дающую точный прогноз, необходимо учитывать особенности рассматриваемой страны.

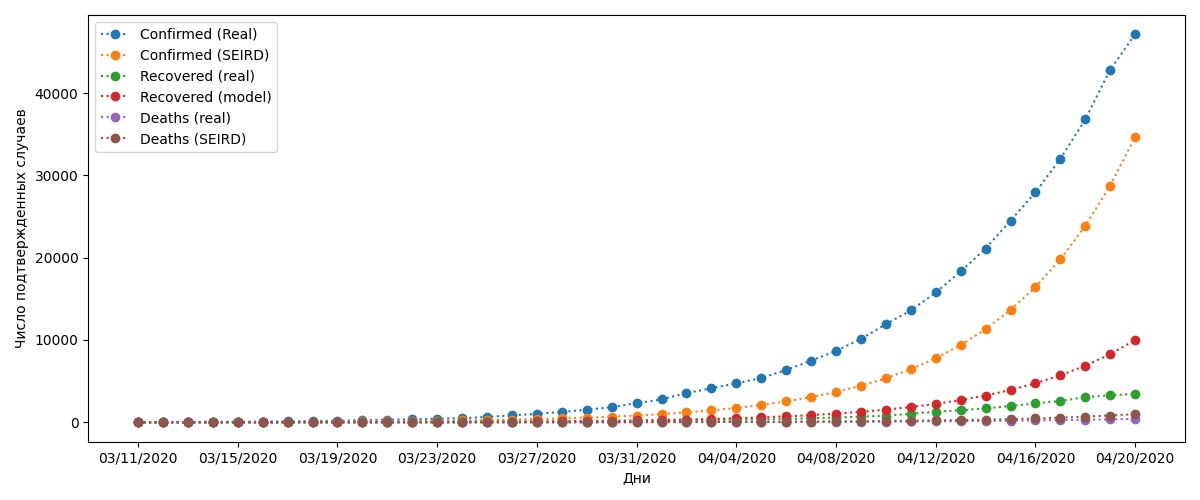


Рисунок 12 – Прогноз модели SEIRD (на 41 день)

В случае с COVID-19 важно так же учитывать множество людей, умерших от болезни, поэтому для моделирования COVID-19 будет использоваться модель SEIRD, дающая прогноз, показанный на рисунке 12.

# **4 Сравнение алгоритма взвешенной скользящей средней с моделью SEIRD**

Для начала модели тестировались на данных по России, затрагивающих период с 11/03/2020 до 20/04/2020.

Сначала были выполнены прогнозы с помощью обеих моделей, позже проведен их анализ. Прогноз, полученный с помощью алгоритма взвешенной скользящей средней, суммирующего последние 2 значения, изображен на рисунке 13.

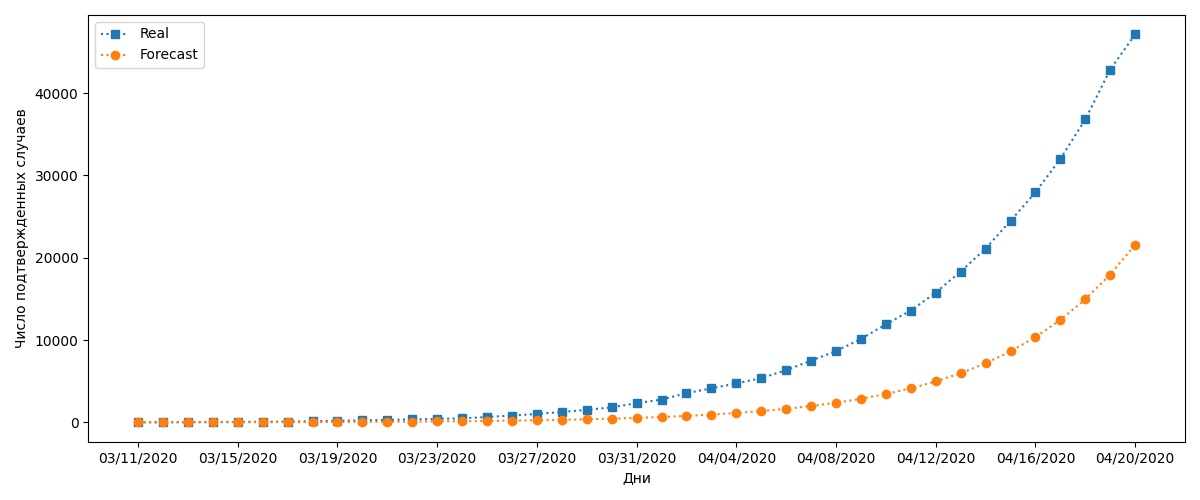


Рисунок 13 – Прогноз алгоритма скользящей средней (на 41 день)

Прогноз, полученный с помощью модели SEIRD изображен на рисунке 14.

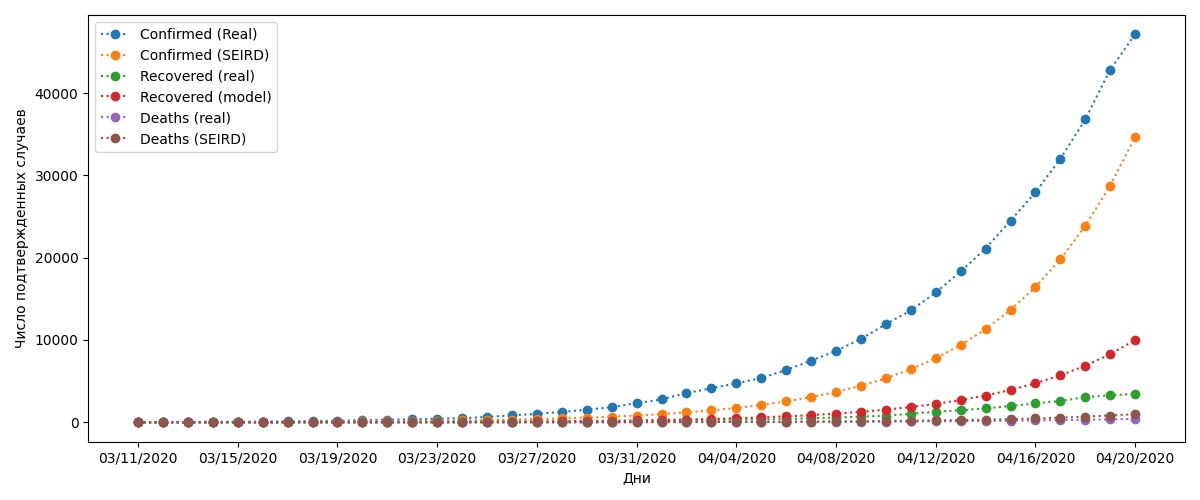


Рисунок 14 – Прогноз модели SEIRD (на 41 день)

Кажется, что эти два прогноза мало чем отличаются друг от друга, однако стратегия DirRec в прогнозировании временных рядов подразумевает то, что в качестве обучающих данных выступают предыдущие прогнозы, а значит ошибка с каждым разом накапливается, тогда как в моделях на графах можно задать различные модификаторы, симулирующие реальную ситуацию.

К примеру, увеличим горизонт прогнозирования с 41 дня до 60 дней и изменим параметры модели SEIRD.

Прогноз, полученный с помощью алгоритма взвешенной скользящей средней, суммирующего последние 2 значения, изображен на рисунке 15.

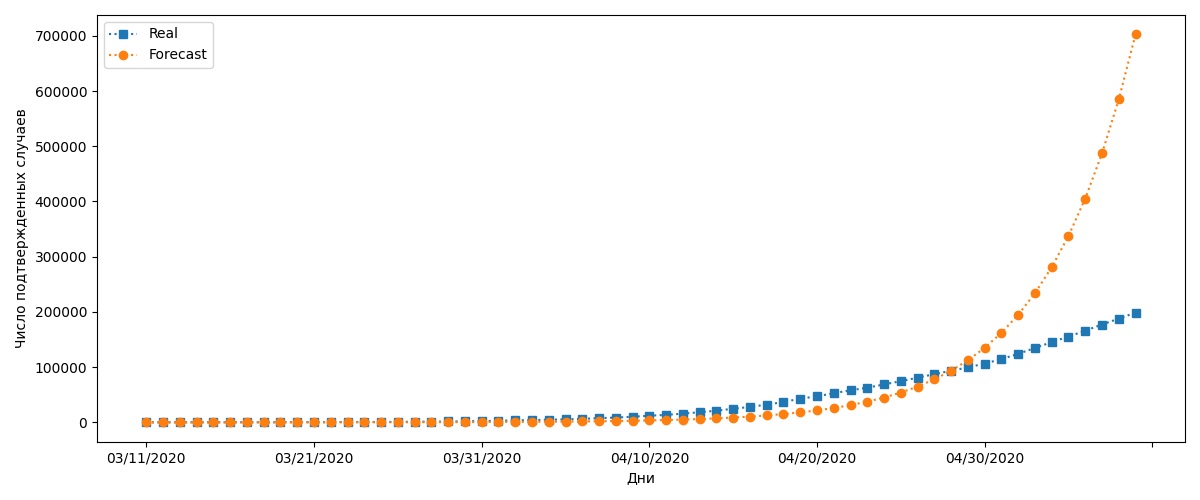


Рисунок 15 – Прогноз алгоритма скользящей средней (на 60 дней)

Модифицируем модель SEIRD, введя в дифференциальные уравнения функцию, зависящую от количества больных людей на момент вычисления. Таким образом будет моделироваться ситуация с карантином: на графиках можно заметить, что рост числа больных снижается с экспоненциального до линейного, за счёт того, что люди начинают серьёзней относиться к мерам предосторожности. Исходя из этого в качестве функции нужно взять ту, которая при малом числе заболевших должна быть близка к 1, при большом — должна асимптотически стремится к нулю. В качестве такой функции будет использоваться (9).

подбором параметров и можно компенсировать экспоненциальный рост.

Тогда модель SEIRD будет описываться системой дифференциальных уравнений (10).

Прогноз, сделанный модифицированной моделью SEIRD, изображен на рисунке 16.

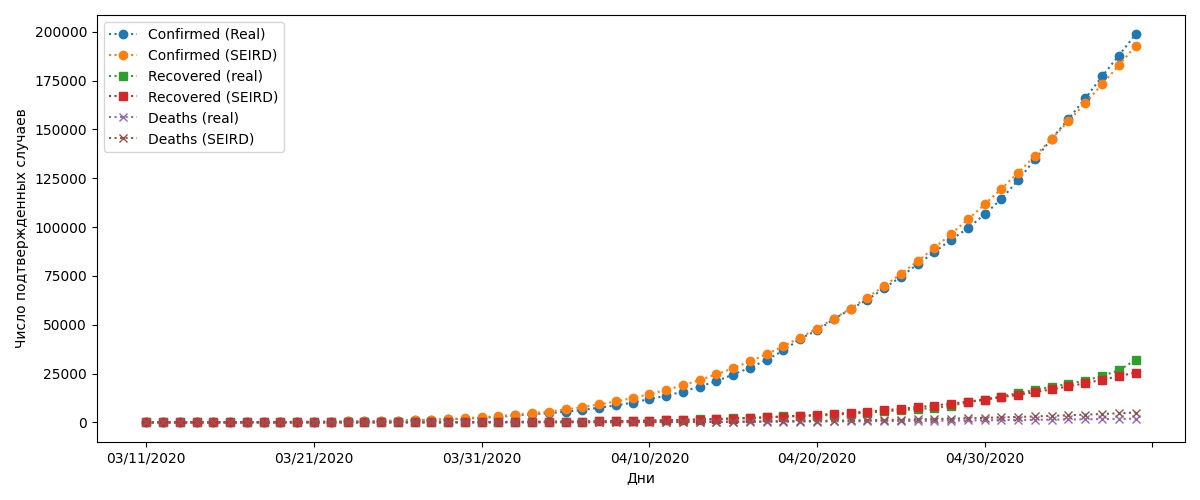


Рисунок 16 – Прогноз модели SEIRD (на 60 дней)

Прогноз модифицированной модели SEIRD очень точно предсказал развитие болезни в течение 60-ти дней, тогда как количество больных в прогнозе алгоритма взвешенной скользящей средней далеко от реальных данных.

Исходя из эксперимента, проведенного выше, в дальнейшем для моделирования COVID-19 будет использоваться модифицированная модель SEIRD.

# **5 Тестирование на реальных данных**

С помощью изменения значений параметров можно получить верный прогноз для любой страны. Для примера получим прогноз для следующих стран: Россия, Италия, Испания.

Модель тестировалась на данных, затрагивающих период с 11/03/2020 до 20/04/2020.

Прогноз для Россия изображен на рисунке 17.

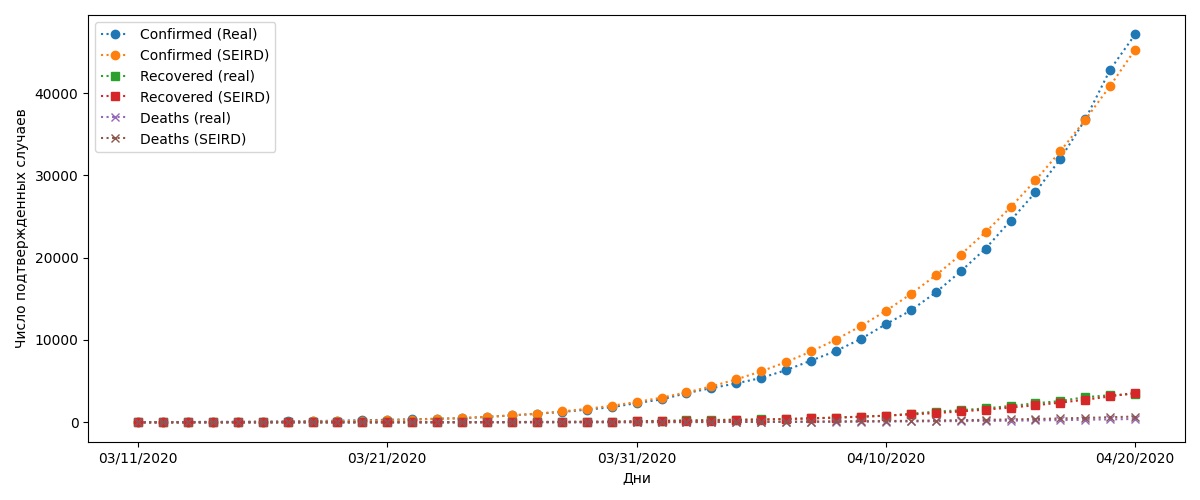


Рисунок 17 – Прогноз для России (на 41 день)

Прогноз для Италии изображен на рисунке 18.

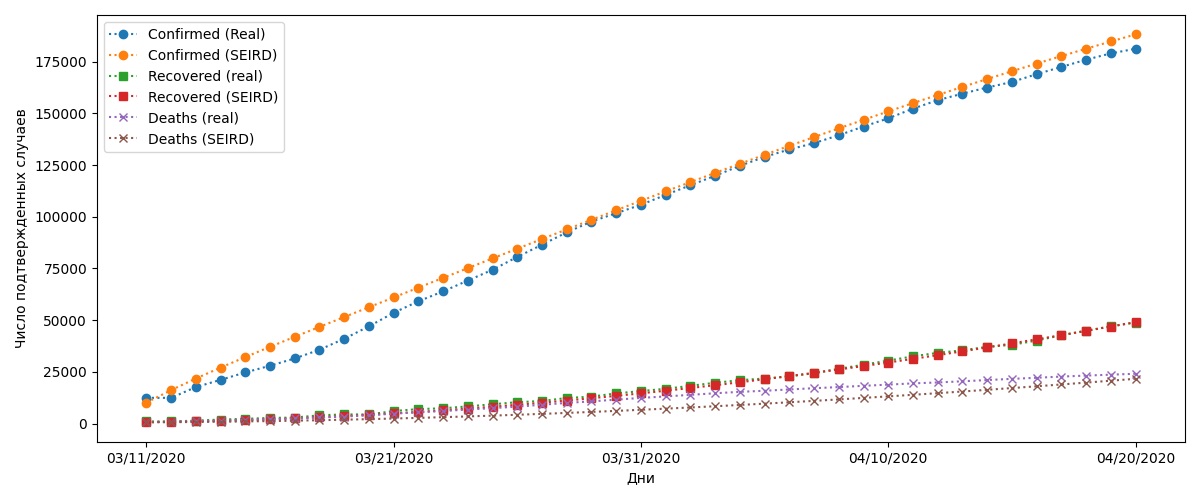


Рисунок 18 – Прогноз для Италии (на 41 день)

Прогноз для Испании изображен на рисунке 19.

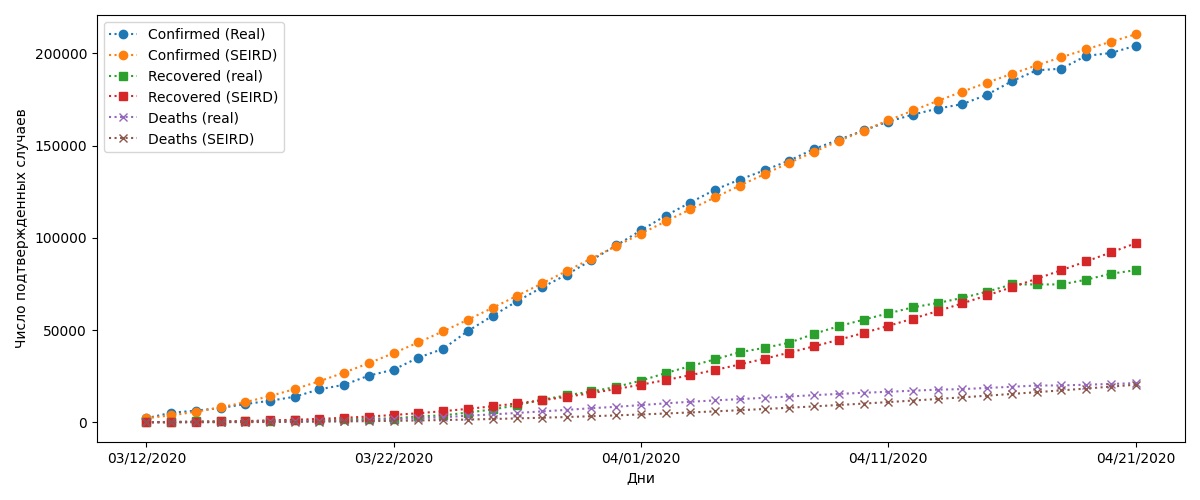


Рисунок 19 – Прогноз для Испании (на 41 день)

# **ЗАКЛЮЧЕНИЕ**

В результате проведенного анализа моделей развития эпидемий были выделены модели, дающие максимальное приближение к реальной обстановке. Модели были протестированы для нескольких стран. Для того, чтобы проверить результаты предсказаний были найдены и использованы большие массивы данных о течении эпидемии из общедоступных источников.

Исследование позволило сделать следующие вводы.

Чтобы автоматизировать процесс подбора параметров для модели SEIRD необходимо учитывать множество различных факторов, начиная от системы здравоохранения и заканчивая средним возрастом населения. Так же необходимо учитывать отдельные особенности каждой страны, например менталитет, возможность путешествовать в другие страны, уровень социальной активности и уровень контактов между людьми. Поэтому можно наблюдать существенную разницу в характере и степени развития COVID-19 между Россией и Италией.

Однако за сложной реализацией такой модели стоит осведомленность о дальнейшем ходе болезни. С помощью полученных прогнозов можно выяснить, когда и насколько вводить карантин, какие возрастные группы и слои населения больше подвержены риску, следует ли закрывать границы. Все эти знания дают неоспоримые преимущества перед развивающейся эпидемией, если правильно ими воспользоваться.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1 Anderson R. M., May R. M. Infectious diseases of humans, dynamics and control. – Oxford University Press, 1991. – 757 p.

2 Allen L. J. S. Some Discrete-Time SI, SIR, and SIS Epidemic Models. – Mathematical Biosciences, 1994. – P. 83-105.

3 Driessche P. Reproduction numbers of infectious disease models [Electronic resource]. – URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6002118/pdf/main.pdf (дата обращения: 22.09.2020).

4 Taieb S. B., Bontempi G., Atiya A., Sorjamaa A. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. – 2011. – 49 p.

5 Bontempi G. Machine learning strategies for time series prediction. [Electronic resource]. – URL: http://www.ulb.ac.be/di/map/gbonte/ftp/time\_ser.pdf (дата обращения 22.09.2020).

6 Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. 2-е издание / пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 302 с.

# **ПРИЛОЖЕНИЕ А Код программы**