ПРОГНОЗ РАЗВИТИЯ ЭПИДЕМИИ COVID-19 С ПОМОЩЬЮ МЕТОДА НЕЛИНЕЙНОЙ РЕГРЕССИИ

FORECAST FOR THE DEVELOPMENT OF THE EPIDEMIC COVID-19 USING THE METHOD OF NONLINEAR REGRESSION

УДК 004.5

Макаров Д.А., студент, makdenis.1997@gmail.com
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана, кафедра «Системы обработки информации и управления»
Шибанова А.Д., студент, alin.shibanova@yandex.ru
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана, кафедра «Системы автоматизированного проектирования»

Makarov D.A., student, makdenis. 1997@gmail.com
Russia, 105005, Moscow, MSTU N.E. Bauman, Department of Information Processing and Management Systems
Shibanova A.D., student, alin.shibanova@yandex.ru
Russia, 105005, Moscow, MSTU N.E. Bauman, Department of Computer-Aided Design Systems

Аннотация

Основная ценность математики заключается в возможности применения ее методов для анализа реальных объектов и явлений. Для оценки развития эпидемий обычно применяются эпидемиологические математические модели, такие как: SIR, SIRS, SEIR, SIS, MSEIR. В данной статье будет произведена попытка произвести аппроксимацию реальных данных о количестве инфицированных людей с помощью модели экспоненциального роста (на начальных этапах) и SIS-модели, используя метод нелинейной регрессии. В

качестве страны-примера будет рассматриваться Швейцария, однако приведенные ниже математические формулы позволят произвести подобную оценку для любой другой страны. В результате удастся оценить соответствие реальной эпидемиологической ситуации идеальной математической модели, а также сделать приблизительный прогноз о количестве дней до окончания эпидемии.

Summary

The main value of mathematics lies in the possibility of applying its methods to analyze real objects and phenomena. Epidemiological mathematical models such as SIR, SIRS, SEIR, SIS, MSEIR are commonly used to assess the development of epidemics. This article will attempt to approximate real data on the number of infected people using the exponential growth model (at the initial stages) and the SIS model using the nonlinear regression method. Switzerland will be considered as an example country, but the mathematical formulas below will allow a similar estimate to be made for any other country. As a result, it will be possible to assess the compliance of the real epidemiological situation with the ideal mathematical model, as well as to make an approximate forecast of the number of days before the end of the epidemic.

Ключевые слова: нелинейная регрессия, модель экспоненциального роста, SIS-модель, аппроксимация, метод наименьших квадратов, коронавирусная инфекция.

Key words: nonlinear regression, exponential growth model, SIS model, approximation, least squares method, coronavirus infection.

С помощью метода нелинейной регрессии попробуем предсказать динамику развития коронавирусной инфекции в Швейцарии.

Рассмотрим модель экспоненциального роста:

$$\frac{dI}{dt} = (\beta - \gamma)I, \qquad (1)$$

где I – количество инфицированных людей, β – среднее число контактов, приходящихся на человека в единицу времени, γ – среднее число выздоровевших, приходящихся на человека в единицу времени.

Рассмотрим также SIS-модель:

$$\frac{dI}{dt} = (\beta - \gamma)I - \frac{\beta}{N}I^2, \quad (2)$$

где N – число людей в популяции (например, число жителей страны).

Решение экспоненциальной модели будет иметь следующий вид (примем $I(0) = 0, \chi = (\beta - \gamma)$ и решим задачу Коши):

$$I(t) = I_0 e^{\chi t}, \quad (3)$$

Решение SIS-модели будет иметь следующий вид (I_∞ обозначает предельное значение $I(t)$ при $t \longrightarrow \infty$, т. е. $I_\infty = \lim_{t\to\infty} I(t)$):

$$I(t) = \frac{I_\infty}{1 + (\frac{I_\infty}{I_0} - 1)e^{-\chi t}}, \quad (4)$$

Реальные данные о количестве инфицированных в зависимости от времени были получены с официального сайта, предоставляющего статистические сведения [1]. Файл с данными был загружен 21.04.2020, этот день и является последней учтенной в выборке датой. В итоге получилась выборка, содержащая данные за 54 дня. График зависимости количества случаев заражения от даты представлен на рисунке 1.

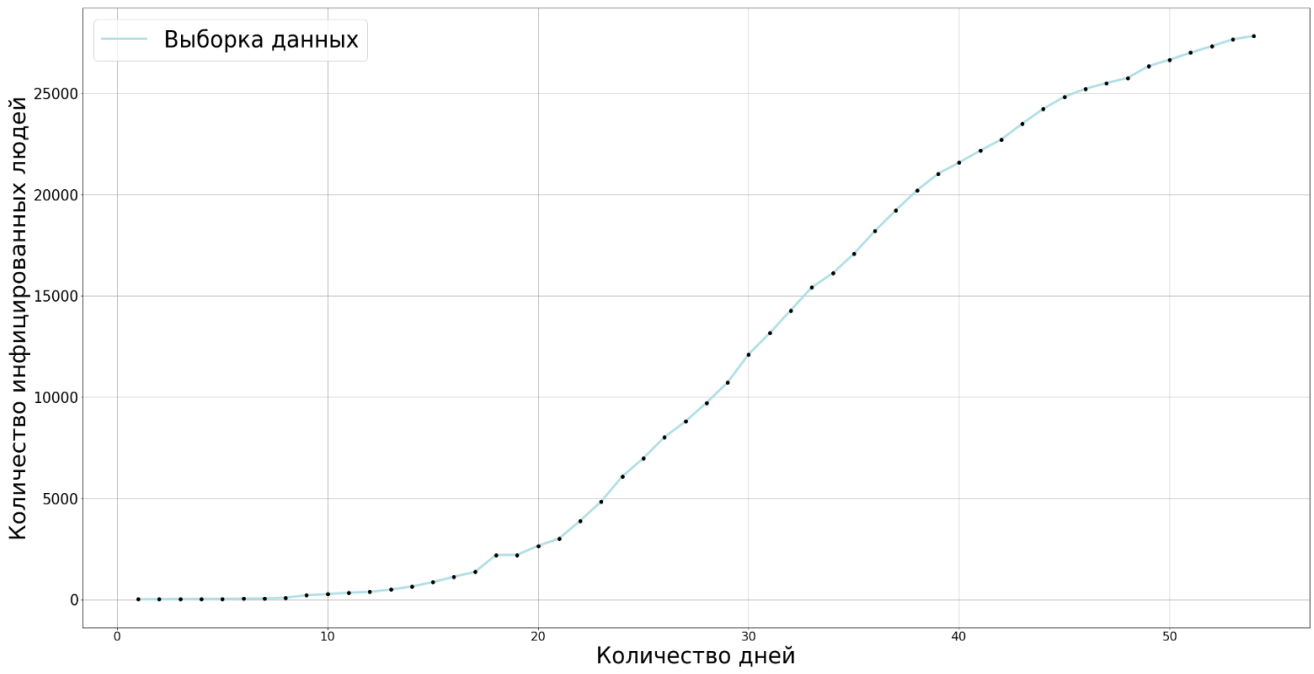

Рис. 1. График зависимости количества инфицированных людей от даты (54 дня)

Проанализируем получившийся график: можно предположить, что наилучшая аппроксимация с помощью модели экспоненциального роста получится, если первым днем будем считать 9-й день из выборки, когда количество инфицированных равнялось 209 ($I_0 = 209$), данные за предыдущие дни учитываться в дальнейшем рассмотрении не будут (их можно считать начальными флуктуациями).

Экспоненциальная модель обычно используется для анализа распространения эпидемии в первые ее недели (когда скорость роста графика зависимости количества инфицированных от времени наиболее высока). Чтобы оценить количество дней, в течение которых оправдано использование модели экспоненциального роста, рассмотрим график зависимости количества инфицированных от времени для 30 первых дней (этот график представлен на рисунке 2). Визуальная оценка позволяет предположить, что для аппроксимации экспоненциальной моделью стоит использовать интервал с 1-го по 16-й день, т. к. именно в этот период наблюдается постоянное увеличение скорости роста графика.

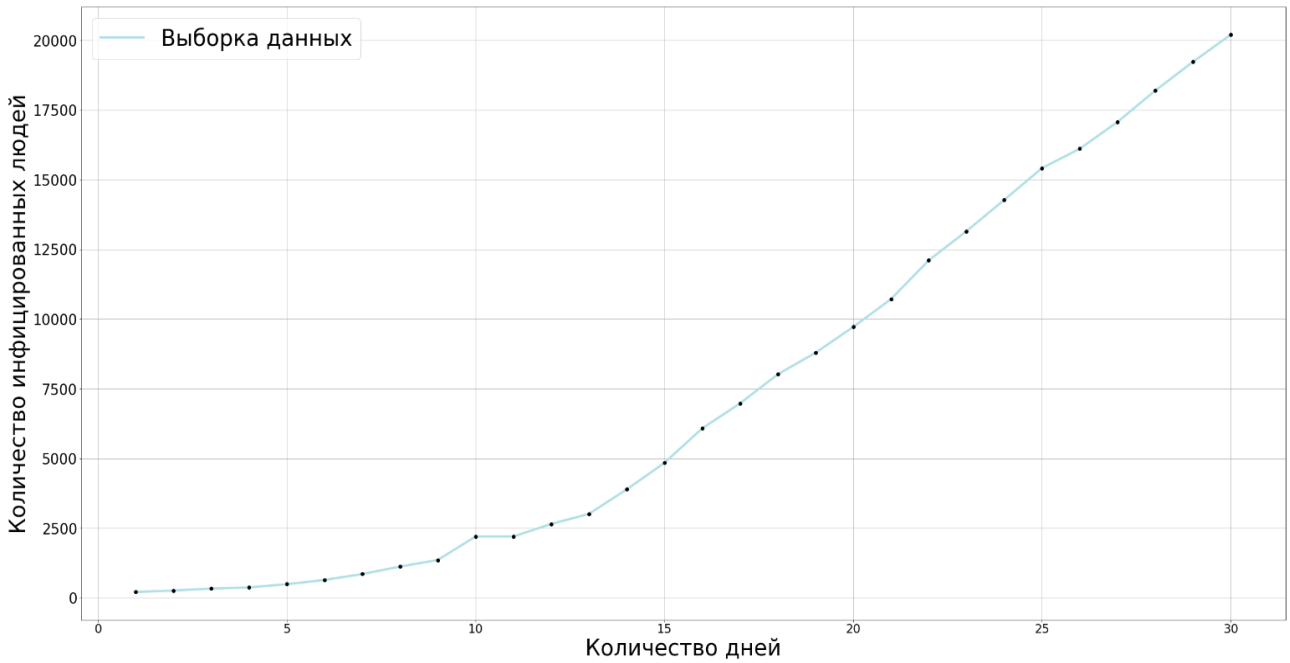

Рис. 2. График зависимости количества инфицированных людей от даты (30 дней)

Вернемся к рассмотрению графика, представленного на рисунке 1: можно заметить, что скорость роста графика замедляется ближе к последним числам

апреля. Вероятно, далее прирост графика по ординате будет становиться все меньше, а потом, скорее всего, график превратится в «плато» (не будет прироста по ординате вообще). Но если карантин будет прекращен преждевременно, то возможна вторая волна экспоненциального роста количества инфицированных.

Оценим значение $\chi = (\beta - \gamma)$ с помощью нелинейной регрессии и нормального уравнения, взяв в качестве аппроксимирующей функции решение экспоненциальной модели (3). Коэффициент χ входит в решение экспоненциальной модели нелинейно, чтобы исправить это, возьмем логарифм от $I(t)$:

$$\ln \frac{I}{I_0} = \chi t$$

$$\tilde{I} = \ln I - \ln I_0 = \chi t \qquad (5)$$

В сформированной выборке содержатся данные о количестве инфицированных в зависимости от времени:

$$Y = \{y_i\}, i = 1 \dots n, \qquad (6)$$

где n – количество дней, а y_i – количество инфицированных людей в i-й день.

Выполним следующие преобразования:

$$\tilde{y_i} = \ln y_i - \ln I_0, i = 1 \dots n \qquad (7)$$

Соответственно:

$$\tilde{Y} = \{\tilde{y_i}\}, i = 1 \dots n, \qquad (8)$$

В данном случае задачу минимизации (метод наименьших квадратов) можно сформулировать так:

$$\min_{\chi} \sum_{i=1}^{n} \left(\tilde{y_i} - \tilde{I}_i(\chi, t)\right)^2 \qquad (9)$$

Чтобы определить оптимальное значение χ, требуется найти решение нормального уравнения:

$$\chi = (X^T X)^{-1} X^T \tilde{Y}, \qquad (10)$$

где матрица $X = \begin{bmatrix} t_1 \\ t_2 \\ \dots \\ t_n \end{bmatrix}$, транспонированная матрица $X^T = [t_1 \quad t_2 \quad \dots \quad t_n]$, $\tilde{Y} = \begin{bmatrix} \widetilde{y_1} \\ \widetilde{y_2} \\ \dots \\ \widetilde{y_n} \end{bmatrix}$. Запишем решение нормального уравнения:

$$\chi = \frac{\sum_{i=1}^{n} t_i \widetilde{y_i}}{\sum_{i=1}^{n} {t_i}^2} \tag{11}$$

Теперь значение χ, определенное по формуле (11), можно подставить в (3). В результате вычислений χ получилось равным ≈ 0.22.

Графический результат представлен на рис. 3.

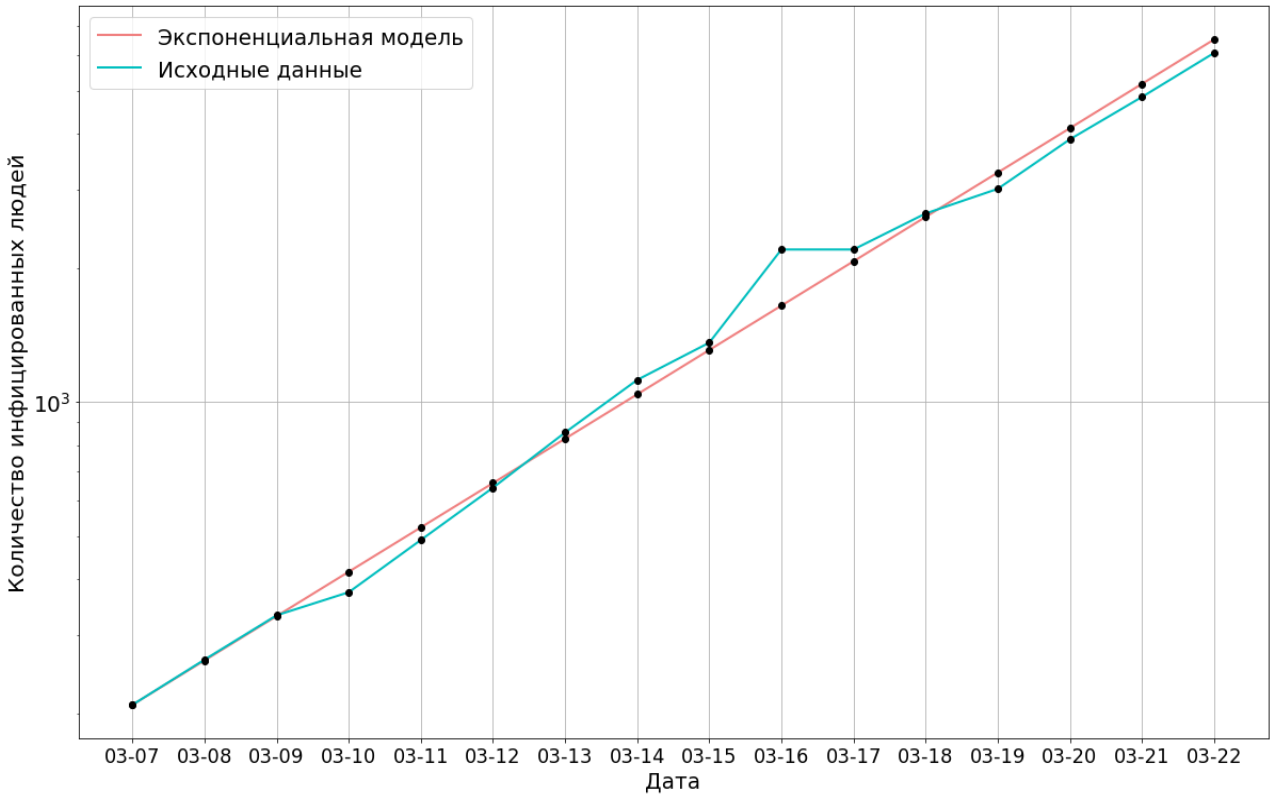

Рис. 3. Аппроксимация экспоненциальной моделью и исходные дискретные данные

Разумеется, абсолютного совпадения графика, построенного по исходным данным о числе инфицированных, и графика аппроксимирующей функции не наблюдается. Но нельзя забывать, что в данном случае мы имеем дело с реальными данными, которые никогда не будут полностью соответствовать идеальной математической модели. К тому же, известные данные о числе заболевших (из которых состояла выборка) могут (иногда даже значительно)

отличаться от действительных данных, т. к. не все заболевшие пойдут к врачу, не всегда тест на наличие коронавирусной инфекции покажет правильный результат и т. д. (факторов множество). Поэтому результат, представленный на рисунке 3, можно считать вполне удовлетворительным, а также можно утверждать, что данные за первые 16 дней рассматриваемого промежутка действительно описываются моделью экспоненциального роста.

Теперь определим значение I_∞ с помощью нелинейной регрессии, взяв в качестве аппроксимирующей функции решение SIS-модели. Значение χ было определено в предыдущем пункте. В данном случае будем рассматривать весь промежуток значений из выборки.

Коэффициент I_∞ входит в уравнение (4) нелинейно, чтобы исправить это, выполним переход к обратной зависимости:

$$\tilde{I} = \frac{1}{I(t)} = \frac{1}{I_\infty} + (\frac{I_\infty}{I_\infty I_0} - \frac{1}{I_\infty})e^{-\chi t}$$

$$\tilde{I} = \frac{1}{I(t)} = \frac{1}{I_\infty} + \frac{I_\infty - I_0}{I_\infty I_0}e^{-\chi t} \quad (12)$$

Обозначим $a_1 = \frac{1}{I_\infty}, a_2 = \frac{I_\infty - I_0}{I_\infty I_0}$. Тогда можем записать (12) в следующем виде:

$$\tilde{I} = a_1 + a_2 e^{-\chi t} \quad (13)$$

Определим $\tilde{y} = \frac{1}{y}$. В данном случае задачу минимизации (метод наименьших квадратов) можно сформулировать так:

$$\min_a \sum_{i=1}^{n} (\tilde{y}_i - \tilde{I}(t_i))^2 \quad (14)$$

Запишем нормальное уравнение:

$$a = (X^T X)^{-1} X^T \tilde{y}, \quad (15)$$

где $X = \begin{bmatrix} 1 & e^{-\chi t_1} \\ \dots & \dots \\ 1 & e^{-\chi t_n} \end{bmatrix}$, $X^T = \begin{bmatrix} 1 & \dots & 1 \\ e^{-\chi t_1} & \dots & e^{-\chi t_n} \end{bmatrix}$, $\tilde{y} = \begin{bmatrix} \frac{1}{y_0} \\ \dots \\ \frac{1}{y_n} \end{bmatrix}$, $a = \begin{bmatrix} \frac{1}{I_\infty} \\ \frac{I_\infty - I_0}{I_\infty I_0} \end{bmatrix}$.

В результате вычислений I_∞ получилось равным 33650 (при этом $I_0 = 209$). Графический результат представлен на рис. 4.

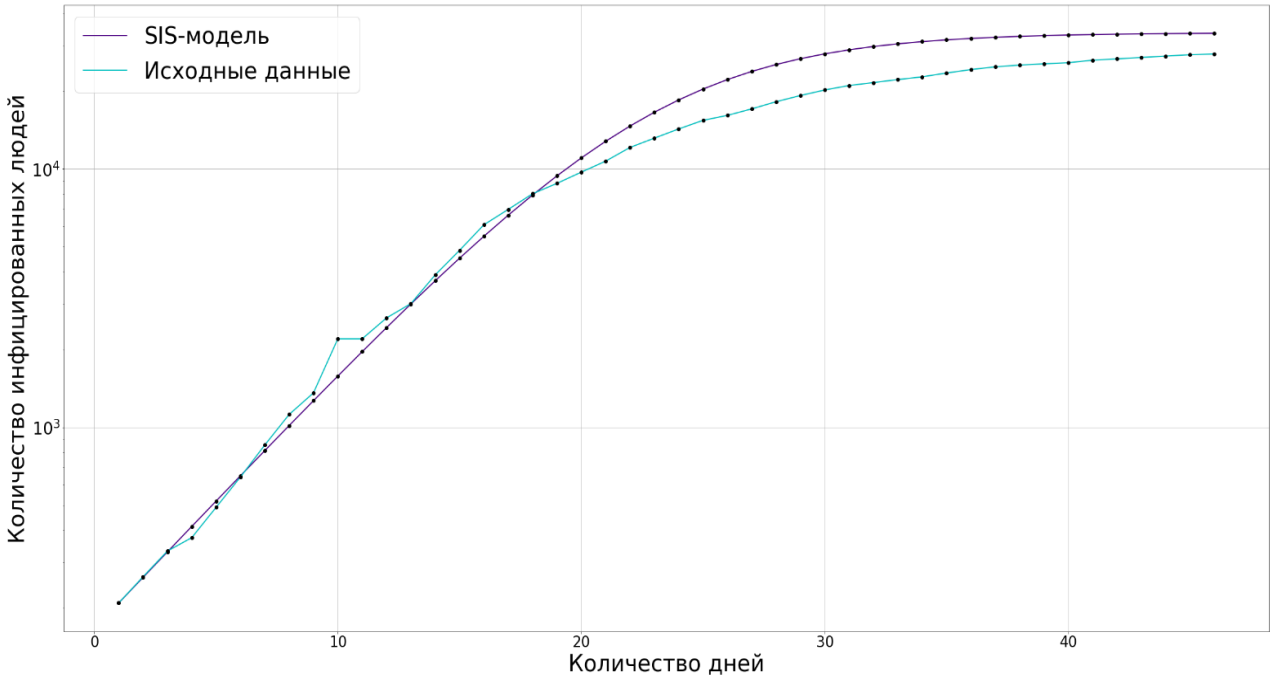

Рисунок 4. Аппроксимация SIS-моделью и исходные дискретные данные

Легко заметить, что аппроксимация получилась не очень точной. Но, тем не менее, результат предсказуемый: SIS-модель не учитывает приобретение иммунитета [3], а в случае с коронавирусной инфекцией наличие/отсутствие иммунитета к заболеванию у выздоровевших пока не доказано. Также SIS-модель, как и многие другие эпидемические модели, не учитывает дополнительные (замедляющие стремительный рост числа инфицированных) факторы: карантин, изоляция инфицированных людей, применение дополнительных способов защиты (ношение респираторных масок, например).

Стоит заметить, что в начале графики реальной зависимости и аппроксимации SIS-моделью довольно близки (первые 18 дней). Возможно, это связано с тем, что как раз в этот период коронавирусная инфекция распространялась в почти «естественных» условиях (не было карантина). Но далее, когда был введен режим самоизоляции, рост графика реальной зависимости замедлился. А график математической модели показывает, какой была бы ситуация, если бы государством не были приняты дополнительные меры борьбы с вирусом.

Попробуем улучшить аппроксимацию, определив значение I_∞ с помощью метода оптимизации scipy.optimize.minimize. Графический результат представлен на рис. 5.

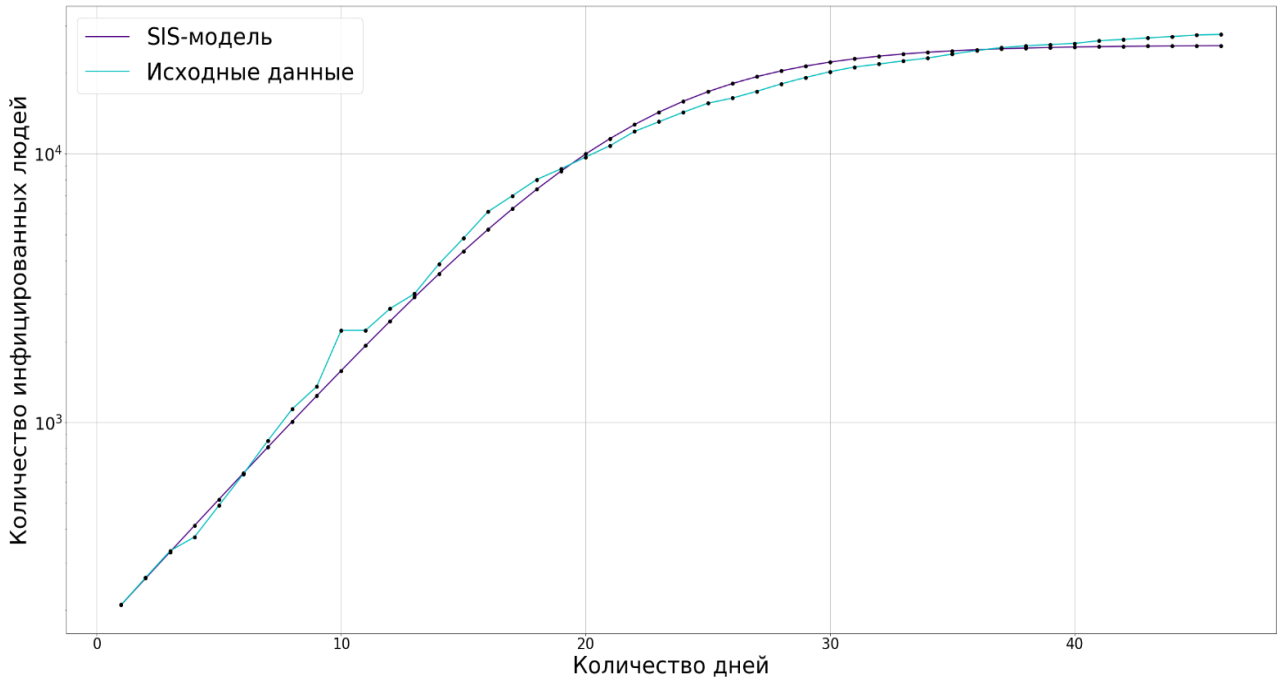

Рис. 5. Аппроксимация SIS-моделью (после оптимизации I_∞) и исходные данные

Теперь аппроксимация получилась более точной. Значение I_∞ после оптимизации стало равным 25355. Конечно, это значение меньше, чем должно быть: 21 апреля (последний день из выборки) было зафиксировано 27826 случаев заражения коронавирусной инфекцией. Но, как уже было сказано, график исходной зависимости близок к тому, чтобы превратиться в «плато». И оптимизированная аппроксимация SIS-моделью это подтверждает. Так что, вероятно, итоговое количество инфицированных людей в Швейцарии будет равным примерно 29000 – 30000 (очень грубая оценка).

Определим погрешность аппроксимации в пространстве L_∞ по формуле [2]:

$$\left|\left|f(t) - L(t)\right|\right|_\infty = \max_{t\in[t_1;t_2]} |f(t) - L(t)|, \qquad (16)$$

где $f(t)$ – количество инфицированных людей в момент времени t (данные из исходной выборки), $L(t)$ – значение аппроксимирующей функции (решение SIS-модели) в момент времени t, $[t_1; t_2]$ – временной промежуток (46 дней).

Определим погрешность для двух случаев: для аппроксимации без оптимизации значения I_∞ и для аппроксимации с оптимизацией значения I_∞.

В первом случае погрешность $error \approx 8367$, во втором случае погрешность $error_{opt} \approx 2571$. Полученные значения соответствуют графическому результату: погрешность действительно должна была получиться значительной в обоих случаях, но оптимизированная аппроксимация должна была дать меньшую ошибку. Учитывая, что SIS-модель в принципе не может дать очень точного результата, если условия распространения эпидемии не являются «естественными» (т. е. присутствуют сдерживающее распространение инфекции факторы), полученные результаты можно считать удовлетворительными.

Для определения погрешности в пространстве L_2 воспользуемся формулой [4]:

$$\left|\left|f(t) - L(t)\right|\right|_2 = \left(\frac{1}{n}\sum_{t=1}^{n}\left(f(t) - L(t)\right)^2\right)^{\frac{1}{2}} \qquad (17)$$

Определим погрешность для двух случаев: для аппроксимации без оптимизации значения I_∞ и для аппроксимации с оптимизацией значения I_∞.

В первом случае погрешность $error \approx 5010$, во втором случае погрешность $error_{opt} \approx 1188$. После оптимизации значение погрешности стало меньше. Поэтому для оценки количества дней до конца эпидемии будем использовать результаты, полученные с помощью оптимизированной аппроксимации.

Число инфицированных людей при $t \longrightarrow \infty$ уже было оценено – от 29000 до 30000. Конечно, это очень грубая оценка, к тому же, есть очень много факторов, которые могут повлиять на ее изменение (например, если карантин закончится раньше, чем следовало бы, то итоговое количество инфицированных в стране будет значительно больше).

Предположим, что эпидемия заканчивается тогда, когда прирост общего числа заболевших за один день становится меньше 1% от общего числа заболевших за предыдущий день. Тогда:

$$\frac{1+\left(\frac{I_\infty}{I_0}-1\right)e^{-\chi t}}{1+\left(\frac{I_\infty}{I_0}-1\right)e^{-\chi(t+1)}}<1.01$$

$$e^{-\chi t}\left(\frac{I_\infty}{I_0}-1\right)(1-1.01e^{-\chi})<0.01$$

Значит, время можно оценить так:

$$t>-\frac{ln\left(\frac{0.01}{\left(\frac{I_\infty}{I_0}-1\right)(1-1.01e^{-\chi})}\right)}{\chi} \qquad (18)$$

В результате расчета получаем, что окончание эпидемии в Швейцарии наступит через 44 дня (для определения t использовалось оптимизированное значение I_∞).

Итак, прогнозируемая дата окончания эпидемии коронавирусной инфекции в Швейцарии – 4 июня 2020 года.

И модель экспоненциального роста, и SIS-модель показывают довольно точные результаты при анализе эпидемии, распространяющейся в естественных природных условиях. В современном мире, разумеется, появляется значительное количество факторов, сдерживающих развитие инфекции. Но, как удалось выяснить, применение математических эпидемических моделей все же оправдано: они позволяют сделать приблизительный прогноз относительно динамики распространения вируса и предсказать возможную дату окончания эпидемии.

Литература

1. Coronavirus Source Data [Электронный ресурс] URL: https://ourworldindata.org/coronavirus-source-data (Дата обращения: 26.04.2020)

2. Белоцерковский О.М., Холодов А.С. Компьютерные модели и прогресс медицины. М.: Наука, 2001. 300 с.
3. Леоненко В.Н. Математическая эпидемиология: учебно-методическое пособие по выполнению лабораторных работ. СПб.: Университет ИТМО, 2018. 38 с.
4. Численные методы [Электронный ресурс] URL: http://orloff.am.tpu.ru/chisl_metod_labs/Lab3/norma.htm (Дата обращения: 26.04.2020)

Literature

1. Coronavirus Source Data [Electronic resource] URL: https://ourworldindata.org/coronavirus-source-data (Date of treatment: 04/26/2020)
2. Belotserkovsky O.M., Kholodov A.S. Computer models and medical progress. M .: Nauka, 2001.300 s.
3. Leonenko V.N. Mathematical epidemiology: a teaching aid for laboratory work. SPb .: ITMO University, 2018.38 s.
4. Numerical methods [Electronic resource] URL: http://orloff.am.tpu.ru/chisl_metod_labs/Lab3/norma.htm (Date of appeal: 04/26/2020)