Ссылка на статью:
// Ученые записки УлГУ. Сер. Математика и информационные технологии. УлГУ. Электрон. журн. 2022, № 1, с. 1-7.

Поступила: 06.04.2022
Окончательный вариант: 13.05.2022

УДК 004.942

Прогноз и исследование зависимостей временных рядов на примере роста заболеваемости COVID-19

Бутов А.А., Шабалин А.С.*

*alexshabalin73@gmail.com
УлГУ, Ульяновск, Россия

В работе проводится анализ зависимости возникновения числа новых случаев заболевания вирусом SARSCoV-2 от значений в предыдущие дни. Выявлены значимые лаги для заболеваемости в России, представлена попытка объяснить их. Проводится сравнение коррелограм для разных стран. Представлена модель линейной регрессии для предсказания числа новых случае заболевания, а также критерии качества, обуславливающие данную модель.

***Ключевые слова**: временной ряд, автокорреляционная функция, машинное обучение, линейная регрессия, covid-19.*

Введение

Вспышка заболеваемости COVID-19, вызванная штаммом вируса SARSCoV-2, привела к большому числу новых исследований, связанных с ним, в том числе и работ по математическому моделированию [4, 6]. В данной работе проводится исследование зависимости числа новых случаев заболевания от значений в предыдущие дни. Цель работы – построить модель машинного обучения, которая сможет предсказывать число новых случаев заболеваемости COVID-19, оценить её качество, а также выявить зависимости, которые обуславливают предсказания. На основе полученных зависимостей предложена модель линейной регрессии, способной предсказывать число новых выявленных случаев заболевания.

Постановка задачи

Пусть задано множество $X = \{x_1, x_2, \dots, x_n\}$ – последовательность дат ($x_1 = 31.01.2020$, $x_n = 14.03.2022$), также задано множество $Y = \{y_1, y_2, \dots, y_n\}$ – количество новых случаев заражения covid-19 в соответствующие даты. Последовательность вида $Y =$

$\{Y_t,\ t \in X\}$ представляет собой временной ряд. В качестве примера приведем график роста заболеваемости в России (рис. 1).

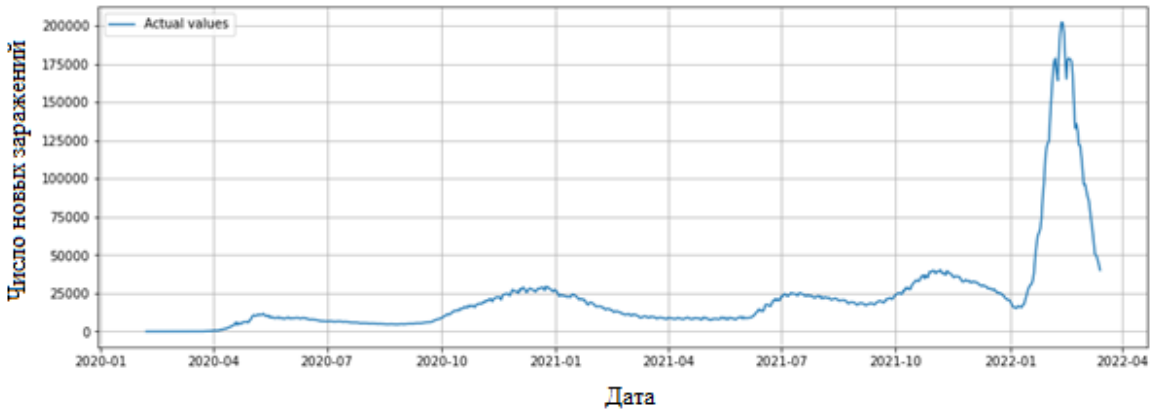

Рис. 1. График роста заболеваемости covid-19 в России

В работе представлена попытка выявить некоторые зависимости рассматриваемого временного ряда. При изучении процесса во времени довольно часто приходится строить оценки взаимосвязи в изменениях уровней рядов, связанных между собой. Одним из подходов, позволяющих решить такую задачу, является метод последовательных разностей [1].

Перейдем от исходного временного ряда к последовательности первых разностей, то есть исходный ряд фактически заменяется относительными величинами изменения уровня заболеваемости в единицу времени:

$$\Delta = \{\Delta_i\} = \{y_i - y_{i-1}\}, \quad (1)$$

где $i=2...n$, примем $\Delta_1 = 0$. Последовательность значений первых разностей $\Delta = \{\Delta_i\}$ представлены на рис. 2.

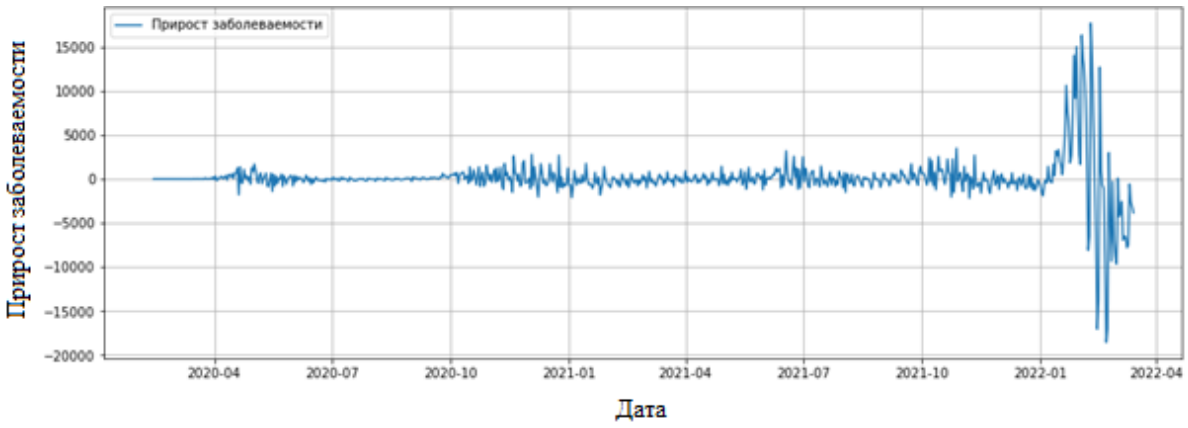

Рис. 2. График прироста заболеваемости covid-19 в России

Анализ автокорреляционной функции

Рассмотрим автокорреляционную функцию $R(t,s)$ зависимости взаимосвязи между функцией и ее сдвинутой копией от величины временного сдвига [2]:

$$R(t,s) = \frac{E[(y_t-\mu_t)(y_s-\mu_s)]}{\sigma_i\sigma_s}, \quad (2)$$

где $EY_t = \mu_t, DY_t = \sigma^2$.

Коэффициент автокорреляции первого порядка – это линейный коэффициент корреляции между уровнями исходного временного ряда и уровнями того же ряда, сдвинутыми на один момент времени. Лаг – степень запаздывания для автокорреляционной функции. Построим график автокорреляции для (1), с количеством лагов 35, то есть посмотрим, как обусловлена зависимость текущего уровня изменения заболеваемости от значений в предыдущие 35 дней.

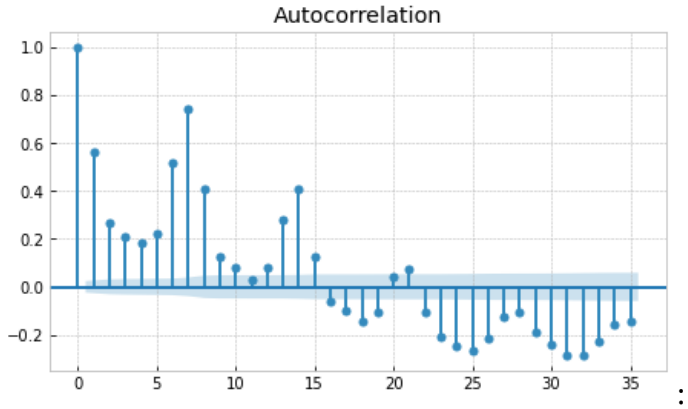

Рис. 3. График автокорреляции для 35 лагов

Согласно графику, представленному на рис. 3, наиболее высокий коэффициент автокорреляции, наблюдается на седьмом лаге, что может свидетельствовать о циклическом колебании с периодичностью 7 дней. Значимые значения коэффициента автокорреляции сохраняется в плоть до 14 лага. Наблюдаемая выраженная корреляционная зависимость с временным лагом равным латентному и инкубационному периоду (равному в сумме семи дням в среднем для штаммов SARSCoV-2 дельта и омикрон), а также периоду инфекционному (равному 14-15 дней в среднем для этих двух штаммов [5]).

На рис. 4-5 представлены графики автокорреляции для США и Соединенного королевства.

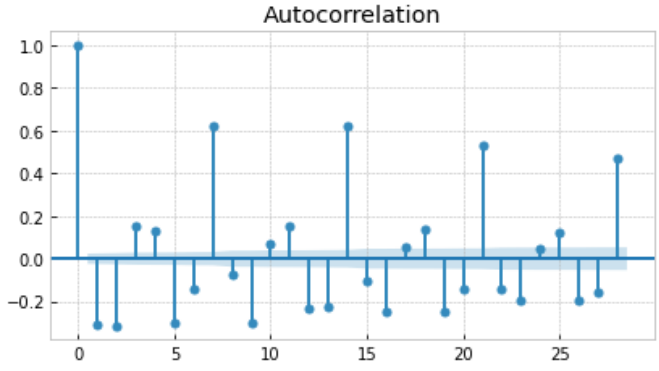

Рис. 4. График автокорреляции в США для 35 лагов

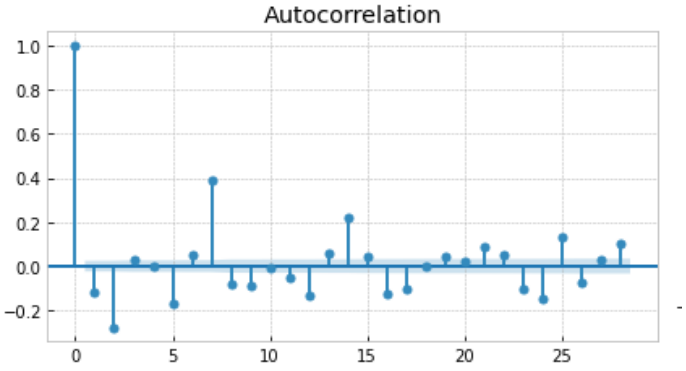

Рис. 5. График автокорреляции в Соединённого королевства для 35 лагов

Седьмой и четырнадцатый лаги остаются значимыми и для представленных стран. Стоит заметить, что значения лагов от 1 до 6 и от 8 до 13 не принимают значимых значений, некоторые из них даже отрицательны, что обуславливает обратную зависимость. Отсутствие схожести в статистических данных между Россией и представленными странами, вероятно, объясняется существенной генетической и иммунологической гетерогенностью населения этих стран.

Модель линейной регрессии

Выявление значимых лагов позволяет перейти к дальнейшему предсказанию числа новых выявленных случаев заболевания основанной на модели авторегрессии порядка *p:*

$$y_i = \sum_{i=1}^{p} \omega_i \cdot y_{t-i} + \xi_i, \quad (3)$$

где $\xi_i \sim N(0, \sigma^2)$, не зависит от *y*. Для таких моделей текущее значение временного ряда выражается через линейную комбинацию *p* предыдущих значений, с учетом некоторой случайной ошибки. Применимость такой модели может быть обусловлена стационарностью временного ряда (2), которую проверим применением теста Дики-Фуллера о наличии единичных корней [3]. Соответствующий тест показал вероятность *p= 0.002173,* согласно которому можно отвергнуть нулевую гипотезу о нестационарности рассматриваемого ряда. Для проведения теста Дики-Фуллера применялся модель adfuller(), библиотеки statsmodels языка Python.

Таким образом, задачу предсказания (3) можно рассматривать как задачу машинного обучения линейной регрессии. Исходя из данных представленных на рис. 3 можно выдвинуть гипотезу, что наилучшее предсказание регрессии (задача минимизации ошибки предсказания) могут быть достигнуты по 14 входным параметрам. Задача сводится к следующей:

$$y_i = f(\omega, y_j) + \xi_i = \omega_0 + \omega_1 y_{i-1} + \omega_2 y_{i-2} + \cdots + \omega_p y_{i-p}, \quad (4)$$

где y_j – j-ый лаг, $\xi_i \sim N(0, \sigma^2)$, не зависит от *y*. В таком предположении параметры ω_i вычисляются методом наименьших квадратов, оптимальные значения которых достигаются методом градиентного спуска.

Для обучения модели будем использовать библиотеку SkLearn, языка программирования Python. Данные выявления новых случаев разделим на обучающую и тестовую выборку. Обучающая выборка составит 30% от исходных данных, на этой выборке будут подбираться оптимальные параметры $\boldsymbol{\omega_i}$, минимизирующие среднеквадратичную ошибку. На тестовой выборке проверим качество полученной модели. В качестве критерия качества модели выберем среднеквадратичное отклонение. Кросс-валидация при решении данной задачи не проводилась, так как случайное перемешивание отдельных частей временного ряда без сохранения самой его структуры невозможно, иначе в процессе потеряются все взаимосвязи наблюдений друг с другом (заметим, что можно было использовать так называемую кросс-валидацию на скользящем окне).

Для проверки гипотезы об оптимальном количестве лагов проведем сравнительный анализ ошибок модели, данные представлены в таблице 1.

Таблица 1. Сравнение качества моделей регрессии.

Число лагов	MAE (mean absolute error)	RMSE (Root Mean Square Error)
5	2448	4669
8	1710	3313
13	1715	3288
14	1701	3229
15	1738	3335

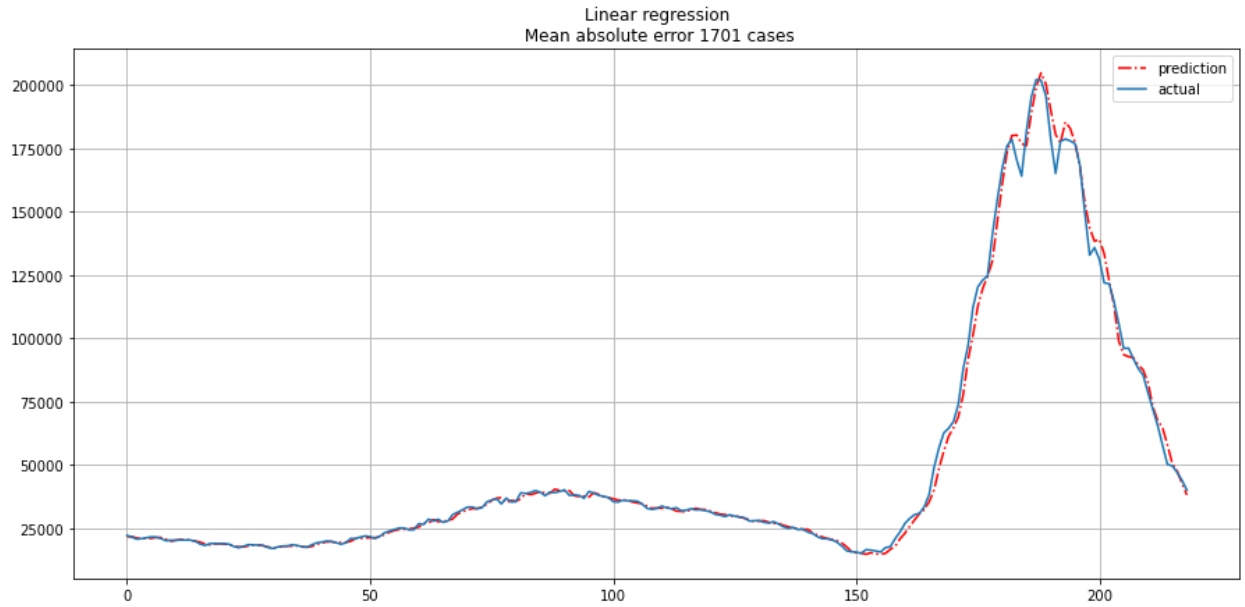

Рис. 6. Предсказания модели регрессии (пунктирная линия) и актуальные значения заболеваемости на тестовой выборке

Заключение

Исходя из данных таблицы 1, можно сделать вывод о том, что, как и предполагалось выше, оптимальным количеством лагов для нашей модели оказалось равным четырнадцати, при таком значении средняя абсолютная и квадратичная ошибки оказываются минимальными на тестовой выборке. Модель линейной регрессии показывает неплохой результат предсказания, что продемонстрировано на рис. 6. Кривая регрессии описывает реальные данные довольно четко, без видимых выбросов в значениях. Проведение прогноза “в будущее” представляется всего на несколько дней вперед (2-3 дня), так как в этом случае модель регрессии уже будет основываться на данных, спрогнозированных самой моделью, а самым значимым коэффициентом ω_i в (4) оказался ω_1, то есть наиболее весомый вклад в прогнозируемое значение вносит первый лаг.

Таблица 2. Спрогнозированные и реальные значения числа новых случаев заболеваемости.

Дата	Спрогнозированное значение	Реальное значение
15.03.2022	34322	36080
16.03.2022	28962	35969
17.03.2022	25103	34172

Анализ коррелограммы, представленной на рис. 3 для числа новых случаев заболевания вирусом SARSCoV-2 в России, показывает, что значимыми остаются значения за предыдущие 14 дней. Это может быть связано с тем, что средний инфекционный период различных штаммов вируса составляет 13-15 дней.

Список литературы

1. Афанасьев В.Н., Юзбашев М.М. *Анализ временных рядов и прогнозирование*: учебник. 2-е изд., перераб. и доп. М.: Финансы и статистика, 2012. 320 с.
2. Canale A., Ruggiero M. Bayesian nonparametric forecasting of monotonic functional time series // *Electronic Journal of Statistics*. 2016, v. 10(2), p. 3265-3286.
3. Dickey D.G. Dickey-Fuller Tests // In: Lovric M. (eds) *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-04898-2_2104.
4. Mazurek J., Neničková Z. Predicting the number of total COVID-19 cases in the USA by a Gompertz curve [Электронный ресурс]. 2020, Apr 18. Режим доступа: http://rgdoi.net/10.13140/RG.2.2.19841.81761 (дата обращения 02.04.2022).
5. Nichita E., Pietrusiak M.-A., Xie F., Schwanke P., Pandya A. Modeling COVID-19 Transmission using IDSIM, an Epidemiological-Modelling Desktop App with Multi-Level Immunization Capabilities // arXiv preprint, arXiv:2112.15252.

6. Pérez Abreu C.R., Estrada S., de-la-Torre-Gutiérrez H. A. Two-Step Polynomial and Nonlinear Growth Approach for Modeling COVID-19 Cases in Mexico //Mathematics. 2021, v. 9, № 18, p. 2180. https://doi.org/10.3390/math9182180.

Forecast and study of time series dependencies on the example of an increase in the incidence of COVID-19

Butov, A.A., Shabalin, A.S.*

*alexshabalin73@gmail.com
Ulyanovsk State University, Ulyanovsk, Russia

The paper analyzes the dependence of the number of new cases of the SARSCoV-2 virus on the values in previous days. Significant lags for incidence in Russia were identified and an attempt has been made to explain them. Correlograms for different countries are compared. For predicting the number of new cases of the disease, a linear regression model is presented as well as the quality criteria that determine this model.

***Keywords**: time series, autocorrelation function, machine learning, linear regression, covid-19.*