

АНАЛИЗ И ОЦЕНКА РАСПРОСТРАНЕНИЯ COVID-19 НА ОСНОВЕ МОДЕЛИ ARIMA

Ларькин А.Д.

Белорусский государственный университет информатики и радиоэлектроники,
г. Минск, Республика Беларусь

Научный руководитель: Тонкович И.Н. – канд. хим. наук, доцент, доцент кафедры ПИКС

Аннотация. Необходимо проанализировать и спрогнозировать эпидемическую ситуацию, чтобы в дальнейшем сформировать эффективную политику борьбы с пандемией. Модель ARIMA была разработана для прогнозирования кумулятивных подтвержденных, мертвых и выздоровевших случаев соответственно, поэтому идеально подходит для данного анализа.

Ключевые слова: COVID-19, модель ARIMA, оценка, временные ряды

Введение. Скорость заражения и распространения инфекции COVID-19 довольно высока по сравнению с другими вирусными инфекциями, встречавшимися до сегодняшнего дня. В силу быстрого прогрессирования COVID-19 и высокой контагиозной силе за короткий промежуток времени для достоверного анализа необходимо регулярно проводить интенсивные исследования. Успех борьбы с инфекцией COVID-19 зависит от выявления значимой информации, особенно в раннем периоде при очень ограниченных данных. Для этого необходимо правильно отслеживать случаи и повышать достоверность будущих прогнозов с каждым новыми данными.

В настоящее время проведено большое количество научных изысканий и исследований по прогнозированию эпидемических заболеваний [1]. Несмотря на существование известных эпидемиологических моделей, ни одна не может с абсолютной точностью спрогнозировать развитие эпидемии COVID-19 из-за множества факторов.

Цель данного исследования – провести оценку модели ARIMA для построения краткосрочных прогнозов в условиях непредсказуемости поведения анализируемой инфекции.

Основная часть. Авторегрессивная интегрированная скользящая средняя (ARIMA) – интегрированная модель авторегрессии – скользящего среднего – модель и методология анализа временных рядов. Используется для прогнозирования временных рядов. В машинном обучении модель ARIMA обычно представляет собой класс статистических моделей, которые формируют выходные данные и линейно зависят от их предыдущих значений в комбинации стохастических факторов. Временными рядами называют дискретные последовательности чисел, характеризующих состояние объекта наблюдения в отдельные моменты времени.

ARIMA использует структурные методы анализа временного ряда. Эти методы, в отличие от неструктурных, ориентированы на оценку по небольшому количеству параметров при дополнительных предположениях о свойствах наблюдаемых величин.

ARIMA работает с использованием модели распределенного запаздывания, в которой алгоритмы используются для прогнозирования будущего на основе запаздывающих значений [2].

На основе проведенного анализа можно допустить вариант использования российских данных о заболеваемости COVID-19 в качестве исходных данных для дальнейшего моделирования распространения COVID-19 на территории Республики Беларусь [3]. С 31 января по 30 мая 2020 года российская ежедневная сводка COVID-19 сообщала о совокупных подтвержденных, умерших и выздоровевших данных, которые были получены из ресурсного центра Джона Хопкинса по COVID-19. Данные с 31 января по 20 мая используются в качестве обучающего набора для построения моделей ARIMA, а данные с 21 по 30 мая – в качестве проверочного набора для оценки точности этих моделей (таблица 1) [4].

Критерий оценки	Подтверждено	Умерло	Поправилось
Среднее значение	48 819	451	8 313
Среднеквадратичное отклонение	85 093	794	18 299
Максимум	308 705	2 972	85 392
Минимум	2	0	0
Асимметрия	1,73	1,74	2,55
Экцесс	1,70	1,77	5,76

Модель *ARIMA* – это метод прогнозирования временных рядов, предложенный Боксом и Дженкинсом в 1970-х годах. Модель включает *AR(p)*, *MA(q)*, *ARMA(p, q)*, *ARIMA(p, d, q)* и так далее. Среди них *ARIMA(p, d, q)* означает, что временной ряд дифференцируется *d* раз, и каждое наблюдение в последовательности представлено линейной комбинацией прошлых *p* наблюдений и *q* остатков. Основная идея модели состоит в том, чтобы рассматривать последовательность как набор зависящих от времени случайных величин и описывать ее с помощью математической модели, основанной на предыдущем значении последовательности для прогнозирования будущих значений. Прогнозируемое значение представляет собой линейную функцию, состоящую из самого последнего истинного значения и самой последней ошибки прогноза [5]. В общем виде модель *ARIMA(p, d, q)* может быть представлена в виде уравнения 1 [6]:

$$y_t = \mu + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} - \vartheta_1 \varepsilon_{t-1} - \vartheta_2 \varepsilon_{t-2} \dots - \vartheta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

где μ , β , ϑ – параметры модели, t – позиция параметра, ε – временный ряд.

Алгоритм построения *ARIMA*-модели:

1 Применяется расширенный тест Дики-Фуллера (*ADF*), чтобы определить, является ли временной ряд стационарным. Если исходная последовательность не является стационарной, ее можно преобразовать в стационарную, используя дифференциальную операцию и затем определяют значение *d*.

2 Строятся графики автокорреляции (*ACF*) и частичной автокорреляции (*PACF*) для определения значений *p* и *q*.

3 Разрабатывается модель соответствия и оценки. Проверяется, является ли остаточная последовательность белым шумом. Если $P > 0,05$, то остаточная последовательность представляет собой белый шум. Для проверки разработанной модели *ARIMA* определяются: средняя процентная ошибка (*MPE*), средняя абсолютная процентная ошибка (*MAPE*) и средняя абсолютная стандартизованная ошибка (*MASE*).

4 Проводится прогнозирование на следующие 30 дней. Для проверки точности модели сравниваются фактические значения с прогнозируемыми значениями от 5,21 до 5,30. При использовании *MAPE* принимают: чем меньше значение, тем выше точность модели. Это выражение можно представить уравнением 2:

$$MAPE = \frac{1}{n} \sum_0^n \left| \frac{e_t}{y_t} \right| \quad (2)$$

где n – количество дней, e_t – фактическое значение, y_t – значение модели.

59-я научная конференция аспирантов, магистрантов и студентов

Данный алгоритм использовался для прогнозирования и проверки точности модели (сравнение разницы между прогнозируемым значением и фактическим с 21.05.2020 по 30.05.2020). Значения, представленные в таблице 2, показывают, что MAPE подтвержденной модели составляет 0,60. Это означает, что модель обладает высокой точностью предсказания и надежностью.

Таблица 2 – Сравнение реальных и прогнозируемых значений 5.21 – 5.30

Дата	Подтверждено			Умерло			Поправилось		
	Реальные	Прогноз	$ e_t/y_t $	Реальные	Прогноз	$ e_t/y_t $	Реальные	Прогноз	$ e_t/y_t $
21.05.2020	317 554	317 466	0,00	3 099	3 096	0,00	92 681	92 731	0,00
22.05.2020	326 448	326 087	0,00	3 249	3 205	0,01	99 825	100 070	0,00
23.05.2020	335 882	334 621	0,00	3 388	3 316	0,02	107 936	107 409	0,00
24.05.2020	344 481	343 064	0,00	3 541	3 435	0,03	113 299	114 747	0,01
25.05.2020	353 427	351 427	0,01	3 633	3 557	0,02	118 798	122 086	0,03
26.05.2020	362 342	359 716	0,01	3 807	3 675	0,03	131 129	129 425	0,01
27.05.2020	370 680	367 937	0,01	3 968	3 789	0,05	142 208	136 764	0,04
28.05.2020	379 051	376 096	0,01	4 142	3 906	0,06	150 993	144 103	0,05
29.05.2020	387 623	384 199	0,01	4 374	4 024	0,08	159 257	151 442	0,05
30.05.2020	396 575	392 250	0,01	4 555	4 142	0,09	167 469	158 781	0,05
MAPE			0.60			3,90			2,40

Эпидемия в России достигла пика в мае, когда ежедневно в течение 10 дней подряд подтверждалось более 10 000 новых случаев. Несмотря на то, что по состоянию на 20 мая число подтвержденных случаев входило в топ-5 в мире, потенциал российской системы здравоохранения мог бы более эффективно отреагировать на пандемию COVID-19. Модель ARIMA имеет зрелую теоретическую базу и четкие критерии, что позволит эффективно прогнозировать эпидемическую тенденцию инфекционных заболеваний. Это поможет эффек-

тивно распределять медицинские ресурсы и будет иметь определяющее значение для научной профилактики и борьбы с COVID-19 [4]. Российские данные применимы и для Республики Беларусь, что позволяет также достаточно достоверно анализировать и прогнозировать заболеваемость на территории Беларуси.

Заключение. Возможности системы здравоохранения Республики Беларусь и России могут эффективно реагировать на пандемию COVID-19. Модели ARIMA показывают хороший эффект подгонки и могут быть использованы для краткосрочного прогнозирования тренда COVID-19, предоставляя Беларуси, России и схожим странам теоретическую основу для формирования новой политики в борьбе с COVID-19.

Список литературы

1. Tanujit Chakraborty, Indrajit Ghosh Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis // *Chaos, Solitons & Fractals*. – 2020. – №135. – С.228-239.
2. Модель ARIMA в машинном обучении [Электронный ресурс]. – Режим доступа: <https://www.biconsult.ru/products/model-arima-v-mashinnom-obucheniі>.
3. Ларькин, А. Д. Специфика данных о распространении COVID-19 на территории Республики Беларусь / А.Д. Ларькин // *Новые информационные технологии в научных исследованиях: материалы XXVII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов, Рязань, 7–9 декабря 2022 г.* / Рязанский государственный радиотехнический университет им. В. Ф. Уткина. – Рязань: РГРТУ им. В. Ф. Уткина, 2022. – С. 115-116.
4. Fang L, Wang D, Pan G Analysis and Estimation of COVID-19 Spreading in Russia Based on ARIMA Model // *SN Compr Clin Med*. - 2020. – №2(12). – С. 2521-2527.
5. Bahareh Fanoodi, Behnam Malmir, Farzad Firouzi Jahantigh Reducing demand uncertainty in the platelet supply chain through artificial neural networks and ARIMA models // *Comput Biol Med*. – 2019. – №113.
6. Как создать модель ARIMA для прогнозирования временных рядов [Электронный ресурс]. – Режим доступа: <http://surl.li/fojyd>.

UDC 004.852

ANALYSIS AND ASSESSMENT OF THE SPREAD OF COVID-19 BASED ON THE ARIMA MODEL

Larkin A.D

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Tonkovich I.N. – PhD, associate professor, associate professor of the Department of ICSD

Annotation. There is a necessary to assess and predict the epidemic situation in order to further improve the effectiveness of pandemic surveillance. The ARIMA model was designed to predict cumulative confirmed, dead, and recovered, respectively, and is therefore ideal for this analysis.

Keywords: COVID-19, ARIMA model, estimation, time series.