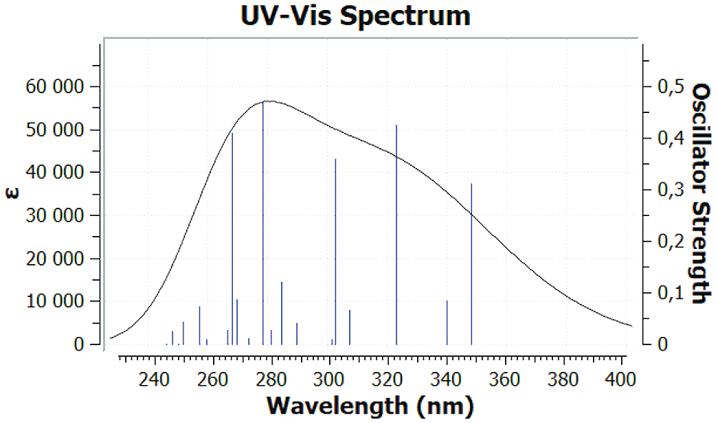

Рисунок 2 – Спектр поглощения молекулы

Вторая широкая и интенсивная полоса поглощения с максимумом при 323.19 нм относится к переходу в возбужденное синглетное состояние молекулы ($S_0 \rightarrow S_3$). Расчеты показывают, что данное возбужденное состояние описывается волновой функцией, отвечающей наложению двух функций. Возбуждение электрона с 106 МО на 108 МО дает главный вклад в полосу поглощения при 323.19 нм (табл. 1, рис. 2). Остальные переходы имеют маленькое значение f и запрещены по симметрии.

Теоретический спектр поглощения оптимизированной молекулы в среде растворителя рассчитан с помощью программного пакета Gaussian 16, используя уровень теории TD-DFT/RB3LYP/6-31++G. Усредненный масштабирующий коэффициент программы при расчете УФ спектров равен 0.99. Рассчитанный электронный спектр поглощения молекулы в среде растворителя представлен на рисунке 2.

ЛИТЕРАТУРА

1. *Siyamak Shahab, Masoome Sheikhi, Liudmila Filippovich, Evgenij Dikusar, Radwan A. Alnajjar, Mikhail Atroshko and Marina Drachilovskaya*, "Antitumor and Antioxidant Activities of the New Synthesized Azomethine Derivatives: Experimental and Theoretical Investigations", Letters in Organic Chemistry (2020) 17:1.

2. *Атрошко М.А., Шахаб С.Н.* Квантово-химический расчет и синтез новых азометиновых соединений, обладающих антиоксидантной активностью / Атрошко М.А., Шахаб С.Н. // Сахаровские чтения 2019 года: Экологические проблемы XXI века, Минск, 23–24 мая 2019 г. – С. 62–65.

ПРОГНОЗИРОВАНИЕ РАСПРОСТРАНЕНИЯ ИНФЕКЦИОННЫХ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

FORECASTING THE SPREAD OF INFECTIOUS DISEASES USING TIME SERIES ANALYSIS

***И. В. Лефанова*[1,2], *Т. В. Смирнова*[1,2]**

***I. V. Lefanova*[1,2], *T. V. Smirnova*[1,2]**

[1]Белорусский государственный университет, БГУ, г. Минск, Республика Беларусь
[2]Учреждение образования «Международный государственный экологический институт имени А. Д. Сахарова» Белорусского государственного университета, МГЭИ им. А. Д. Сахарова БГУ, г. Минск, Республика Беларусь
eis@iseu.by, irina.lefanova@gmail.com

[1]Belarusian State University, BSU, Minsk, Republic of Belarus
[2]International Sakharov Environmental Institute of Belarusian State University, ISEI BSU, Minsk, Republic of Belarus

В работе представлен обзор основных способов прогнозирования распространения эпидемий инфекционных заболеваний и формирования краткосрочных, так и долгосрочных проекций заболеваемости, а также рассматривается применение интегрированной модель авторегрессии – скользящего среднего ARIMA для прогнозирования распространения инфекционных заболеваний на примере данных о распространении COVID-2019 с web-ресурса https://ourworldindata.org/coronavirus

The paper presents an overview of the main methods for predicting the spread of epidemics of infectious diseases and the formation of short-term and long-term projections of morbidity, and also considers the use of an integrated

autoregression model – ARIMA moving average to predict the spread of infectious diseases using the data on the spread of COVID-2019 from the https web resource as an example. http://ourworldindata.org/coronavirus.

Ключевые слова: Математическое моделирование, временные ряды, прогнозирование, модель ARIMA.

Keywords: Mathematical modeling, time series, forecasting, ARIMA model.

https://doi.org/10.46646/SAKH-2022-2-399-402

Математические методы прогнозирования распространения инфекционных заболеваний активно исследуются с начала двадцатого века. В последние годы в связи с появлением и распространением коронавирусной болезни (CoViD19), которая может закончиться тяжелым острым респираторным синдромом (SARS-COV-2), а также большим объемом статистических данных по заболеваемости количество исследований по данной тематике стремительно растет. Первый случай нового коронавируса, известного как SARS-CoV-2 и обычно именуемого COVID-19, был зарегистрирован в декабре 2019 года в городе Ухань в Китае.

Несмотря на существование ряда известных математических моделей распространения инфекционных заболеваний, ни одна из них не может с достаточной долей вероятности спрогнозировать ситуацию развития пандемии вируса Covid-19.

Для того, чтобы модель с точностью отражала реальное положение распространения инфекционного заболевания необходимы точные статистические данные и грамотный анализ внешних факторов, которые влияют на распространение эпидемии как качественно, так и количественно. К таким факторам, кроме собственно показателей, характеризующих инфекцию, при рассмотрении инфекционных заболеваний обычно относят плотность населения региона, средний возраст населения и соотношение лиц разных возрастных категорий, карантинные меры, степень социальной активности групп населения и т.п. Вместе с тем, большинство данных факторов, способствующих формированию наиболее точного прогноза не являются очевидными и с большой долей неопределенности могут быть включены в математические модели.

Все математические модели распространения инфекционных заболеваний можно разделить на три большие категории:

1) Статистические модели, в основе которых лежат линейная и логистическая регрессия, а также ряд других методов машинного и глубокого обучения;

2) Компартментные математические модели, в основе которых лежит разделение населения определенного региона на ряд компартмент и описание состояний компартмент и соотношений между ними при помощи систем дифференциальных уравнений (SIR – sucseptible, infected, recovered, SEIR – susceptible, exposed, infectious, removed; SEIRS – susceptible, exposed, infected, recovered, succumbed);

3) Гибридные или смешанные модели, включающие в себя элементы статистических и компартментных моделей с целью повышения точности прогноза.

В указанных типах моделей могут применяться следующие вычислительные методы и алгоритмы:

1) метод наименьших квадратов;

2) нейронные сети разного типа;

3) алгоритм случайного леса;

4) метод ближайших соседей разных модификация;

5) методы опорных векторов;

6) кластерный анализ данных.

Согласно используемому математическому аппарату, все математические модели прогнозирования распространения эпидемий инфекционных заболеваний можно разделить на следующие категории:

1. Классические эпидемиологические модели переходов состояний, или компартментные модели. Данные модели предполагают большую степень неопределённости при построении первоначальной системы дифференциальных уравнений для описание переходов индивидов из одной компартменты в другую, причем при введении дополнительных компартмент для учета того или иного социального фактора, влияющего на распространение инфекционного заболевания степень неопределенности увеличивается и в систему включаются дополнительные уравнения.

2. Модели, основанные на анализе временных рядов, в частности ARIMA-модели. Данные модели при проведении полного анализа и составления прогноза достаточно сложно настроить, однако они дают обычно хороший результат в том случае, если требуется качественный прогноз на среднесрочный и краткосрочный период времени.

3. Адаптивные модели экспоненциального сглаживания стали достаточно популярным инструментом прогнозирования развития пандемии именно коронавирусной инфекции.

Главной задачей анализа временного ряда является получение исчерпывающей информации, выявление всех возможных зависимостей в анализируемых данных: общий тренд развития показателя временного ряда, циклические и сезонные колебания, а также стохастическую (случайную) составляющую временного ряда.

Под временным рядом понимают упорядоченную во времени последовательность значений вида

$$Y(t) = Y_1, Y_2, \dots, Y_t,$$

где t – момент времени.

Для определения модели временного ряда последовательно идентифицируют сезонные и циклические компоненты, далее детерминированную составляющую, и, в конечном итоге, обрабатывают остаточный ряд с помощью метода авторегрессионного проинтегрированного скользящего среднего (ARIMA).

Порядок модели ARIMA изображается как (p, d, q) со значениями для порядка или количества раз, когда функция встречается при запуске модели.

Модель ARIMA использует разностные данные, чтобы сделать их стационарными, что означает согласованность данных во времени. Эта функция устраняет влияние тенденций или сезонности, таких как рыночные или экономические данные.

Сезонность возникает, когда данные демонстрируют предсказуемые повторяющиеся закономерности. Крайне важно контролировать сезонность, поскольку она может повлиять на точность результатов.

Модели ARIMA могут быть построены с использованием сезонных и несезонных форматов. Сезонная модель должна учитывать количество событий в каждом сезоне в дополнение к авторегрессионным, разностным и средним условиям для каждого сезона.

Бокс и Дженкинс (1971) представили метод, который сочетает в себе авторегрессионные модели (AR) и модели скользящего среднего (MA). Модель ARMA (p, q) представляет собой комбинацию моделей AR (p) и MA (q) и лучше всего подходит для моделирования одномерных временных рядов.

Модель ARIMA (AutoregRessive Integrated Moving Average) представляет собой один из способов анализа и прогнозирования временных рядов. Модели ARIMA широко использовались для обнаружения вспышек инфекционных заболеваний. Модель ARIMA была создана Боксом и Дженкинсом в 1970-х годах для описания изменений временного ряда в математическом подходе. В прошлом ARIMA использовалась для прогнозирования нескольких вспышек заболеваний, таких как геморрагическая лихорадка с почечным синдромом, гепатит-В и др.

Модель ARIMA состоит из трех компонентов:

− AR (Авторегрессия) – означает авторегрессию, которая показывает изменяющуюся переменную, которая регрессирует на свои собственные предыдущие или запаздывающие значения. Другими словами, он предсказывает будущие значения на основе прошлых значений и определяется параметром p в модели авторегрессии

$$Y_t = \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + \varepsilon_t,$$

который определяется по PACF (partial auto-correlation function) – частной корреляционной функции, или частичной автокорреляции, между Y_t и Y_{t-k} при исключении влияния $Y_{t-1}, \ldots, Y_{t-k+1}$. $\theta_1, \ldots, \theta_p$ – оцениваемые коэффициенты, ε_t – случайное возмущение, описывающее влияние переменных, не учтенных в модели.

− MA (скользящее среднее) – используется для определения количества прошлых ошибок прогноза, используемых для прогнозирования будущих значений; представляет собой зависимость между наблюдаемым значением и остаточной ошибкой модели скользящего среднего, примененной к предыдущим наблюдениям определяется параметром q, получаемого из ACF (auto-correlation function, автокорреляционная функция)

$$\rho_k = \frac{cov\{Y_t, Y_{t-k}\}_t}{var\{Y_t\}_t},$$

$$y_t = \varepsilon_t + \alpha\varepsilon_{t-1} + \cdots + \alpha_q \varepsilon_{1-q},$$

где ε_t – белый шум, всегда являющийся стационарным процессом. Скользящее среднее показывает наличие колебаний в ряду. Чем выше значение скользящего среднего, тем выше вероятность колебаний.

− I (интегрирующий член) – он наблюдает разницу между статическими значениями данных и предыдущими значениями. Цель состоит в том, чтобы получить стационарные данные, не зависящие от сезонности. Это означает, что статистические свойства рядов данных, такие как среднее значение, дисперсия и автокорреляция, остаются постоянными во времени. Для проверки стационарности ряда используются расширенный тест Дики-Фулера (ADF) тест Квятковского-Филлипса-Шмидта-Шина (Kwiatkowski–Phillips–Schmidt–Shin, KPSS). Эти же тесты позволяют определить параметр d модели.

Модель авторегрессионного интегрированного скользящего среднего, или ARIMA, представляет собой подход к прогнозированию временных рядов, который используется для прогнозирования будущего значения переменной на основе ее собственных прошлых значений. Он использует авторегрессию и скользящее среднее, а также включает дифференциальный порядок для удаления тренда и/или сезонности. Модель ARIMA(p, d, q) для нестационарного временного ряда $T(n)$ имеет вид

$$\Delta^d T(n) = c + \sum_{i=1}^{p} a_i \Delta^d T(n-i) + \sum_{j=1}^{q} b_j \varepsilon(n-j) + \varepsilon(n)$$

В данном выражении $\varepsilon(n)$ – стационарный временной ряд белого шума, c, a_i, b_j – параметры модели, Δ^d – оператор разности временного ряда порядка d, гарантирующий стационарность ряда (последовательное взятие d раз разностей первого порядка – сначала от временного ряда, затем от полученных разностей первого порядка, затем от второго порядка и т. п.)

Идея анализа временного ряда заключается в нахождении зависимости текущего значения от показателей предыдущих временных отрезков. При зависимости только от предыдущего значения, говорят об авторегрессии

первого уровня AR (p=1) от двух предыдущих – AR (p=2) и т.д. Помимо самих значений временных рядов, зависимость может быть найдена и в ошибках значений каждой временной точки. В случае, если текущая ошибка зависит только от предыдущей, то это первый порядок скользящего среднего MA (q=1), если ошибка зависит от двух предыдущих ошибок – MA (q=2) и т. д. Таким образом, частным случаем модели ARIMA является модель ARMA с параметрами (p, q). Данной моделью можно пользоваться только если временной ряд является стационарным.

Для достижения стационарности необходимо брать разности ряда до тех пор, пока он не станет стационарным (часто также применяют логарифмическое преобразование для стабилизации дисперсии). Число разностей, которые были взяты, чтобы достичь стационарности, определяются параметром d. В итоге получаем единую модель ARIMA (p, d, q), которую принято считать процессом авторегрессии порядка p и d раз проинтегрированного скользящего среднего порядка q.

Подбор коэффициентов p и q модели ARIMA выполняется с помощью вычисления функции автокорреляции (ACF – autocorrelation function) и частичной автокорреляции (PACF – partial autocorrelation function) и анализа их графиков. После анализа графиков автокорреляции подбирается одна или несколько конфигураций ARIMA-моделей. Для оценки качества модели и выбора из ряда созданных моделей одной оптимальной может быть использован информационный критерий Акаике (AIC – Akaike information criteria), при этом наилучшей считается модель с наименьшим значением критерия.

Информационный критерий Акаике применяется для выбора оптимальной статистической модели из нескольких существующих. В общем случае данный критерий может быть определен как

$$AIC = 2k - 2ln(L)$$

где k – число параметров в статистической модели, а L – максимизированное значение функции правдоподобия модели.

Абсолютное значение информационного критерия Акаике не имеет смысла, он указывает только на относительный порядок сравниваемых моделей.

Затем с помощью ACF оцениваются остатки модели. Если отсутствует автокорреляция остатков, то ARIMA-модель конкретного показателя считается финальной и используется для прогнозирования, если же остатки модели автокоррелируют, то модель признается негодной для прогнозирования.

Наиболее популярными мерами точности прогноза в одномерных данных временного ряда являются среднеквадратическая ошибка (RMSE) и средняя абсолютная ошибка в процентах (MAPE). RMSE и MAPE вычисляются как

$$RSME = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \hat{x}_t)^2}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right| * 100,$$

где n – фактическое и прогнозируемое значения в момент времени соответственно, t – последовательность моментов времени. Меньшее значение RMSE и MAPE указывает на лучшую калибровку и, следовательно, на лучшую производительность.

ARIMA-метод является достаточно гибким, наиболее точным и дает наиболее правдоподобный прогноз.

Ежедневные данные о всех зарегистрированных случаях COVID-19 для обучения модели берутся с web-ресурса https://ourworldindata.org/coronavirus.

ЛИТЕРАТУРА

1. *Cryer J., Chan K.* Time Series Analysis/ J. Cryer, K. Chan. – New York: Springer, 2008. – 491 p.

2. *Azar A., Hassanien A.* Modeling, Control and Drug Development for COVID-19 Outbreak Prevention / A. Azar, A. Hassanien– New York: Springer, 2022. – 1122 p.

3. *Metcalfe A., Cowpertwait P.* Introductory Time Series with R/ A. Metcalfe, P. Cowpertwait – New York: Springer, 2009. – 246 p.

4. *Brockwell P., Davis R.* Introduction to Time Series and Forecasting/P. Brockwell, R. Davis – New York: Springer, 2016. – 425 p.