Innovativity in Modeling and Analytics Journal of Research
vol. 6, 2021, pp.15-22 ***iMAJOR***
http://imajor.info/JDA ISSN 2534-9619

SARIMA based Methodology for Forecast of COVID-19 Confirmed Accumulated Cases in Bulgaria

Danail Sandakchiev[1]

Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

[1] sandukchie@uni-sofia.bg

Abstract. Despite the development of vaccines in the fight against COVID-19, the continuing lack of their availability in many parts of the world, allows for the novel virus to persist circulating among communities and evolve. Most recently, a new strain of the virus (Delta variant) has caused waves of infections in many countries around the globe. Although Bulgaria has access to abundance of vaccines, the vaccine rate remains low, as little less than 20% of the population has been fully vaccinated. This situation challenges the health care system in the country once more with the Delta variant becoming the predominant strain of the virus. In this paper, a methodology is proposed on the basis of the seasonal-ARIMA technique that makes a 7-day forecast for the development of accumulated confirmed cases of COVID-19 in Bulgaria. The procedure can be run for n number of consecutive weeks. The methodology can be applied by authorities to support them in the planning of resources and measures in the attempt to contain the spread of the Delta variant.

Keywords: COVID-19 SARIMA Bulgaria Machine Learning.

1 Introduction

COVID-19 continues to be a major challenge for humanity 18 months after it was classified as pandemic by the World Health Organization (Cucinotta and Vanelli, 2020). Even though vaccines were approved towards the end of 2020, distribution and accessibility in poorer countries remain a significant challenge in the fight against the spread of the novel virus (Guidry et al., 2021). As COVID-19 keeps circulating among communities, it mutates in new strains that can potentially be more dangerous to human health. One such strain, the Delta variant, emerged in India as early as March 2021 (Kunal et al., 2021). The Delta variant transmits easier than other strains, it leads to more frequent admissions of infected people in hospitals and intensive care units and by end of June it had already spread in 88 different countries, becoming the predominant COVID-19 strain (Vaishya et al., 2021).

The Delta variant has caused a new wave of infection across many countries. Bulgaria has started to experience significant increase in cases as of August. In June and July there were respectively 3 477 and 3 121 confirmed cases, whereas in August alone close to 27 000 confirmed cases of COVID-19 were reported (https://coronavirus.bg, 2021). At the end of July, the average number of daily new cases (on a weekly basis) had reached a low number of 185 cases and at the end of August the number has reached more than 1 400 cases a day. By middle of August, the Delta variant had become the predominant COVID-19 cases with 99% of the observed cases being due to the Delta variant (https://coronavirus.bg/bg/news/2255, 2021). The vaccination campaign in Bulgaria is progressing at very unsatisfactory pace. After more than 8 months since vaccines became available in the country, less than 20% of the population is fully vaccinated (https://coronavirus.bg, 2021).

Although the Delta variant established itself as prominent quite recently, researchers have started analyzing and modelling its impact and some works on the topic are already published. Brereton and Pedercini (2021) considered the development of the COVID-19 cases in the UK where the emergence of the Delta variant coincided with the lifting of measure that was being planned. The authors created a simulation model based on the Susceptible, Infected, Exposed and Recovered (SIRD) method to forecast that the cases would peak at hundred of thousands, if the government would have lifted the measures in July. Shah et al. (2021) applied a spline model to predict the daily cases of COVID-19 in Scotland in light of the Delta variant. They also used a Cox regression model to estimate probabilities for hospitalization and death following infection with the Delta variant. Results showed that in the worst-case scenario daily cases could reach record levels, not seed in previous waves, but still hospitalization and death cases would not be as high as in the previous waves. Reingruber et al. (2021) developed a modeling framework based on data in France and concluded that the current vaccination rate would have prevented a new wave of infections from the original strain of the virus, but the same vaccination rate is not capable of preventing a new wave from the Delta variant. Their simulation support the argument that the Delta variant spreads much easier among communities.

Articles focused on the Delta variant in the field of data science will likely continue to be published in the near future. Even before the Delta variant became dominant, many researchers applied different machine learning techniques and methods to model various health, social, environmental aspects impacted by COVID-19. Romeo and Frontoni (2021) proposed a machine learning algorithm (Hierarchical Priority Classification eXtreme Gradient Boosting) that helps authorities prioritize administration of vaccines among people. Dabbah et al. (2021) estimated COVID-19 mortality risk using random forest classification model. Aljame et al. (2021) implemented ensemble-based method, named deep forest, to diagnose COVID-19 cases based on clinical and routing laboratory data. A very common method used by analysts to model COVID-19 cases is the Auto-Regressive Integrated Moving Average (ARIMA). ArunKumar et al. (2021) created a methodology based on ARIMA to forecast accumulated confirmed, recovered and death cases in the countries with the highest number of COVID-19 cases. To make 7-day forecast of newly confirmed cases in Japan and South Korea, Duan et al. (2021) utilized ARIMA model. In Nigeria, Aronu et al. (2020) modeled the survival rate of COVID-19 patients in the country with the help of ARIMA as well.

In light of the recent increase of COVID-19 cases in Bulgaria due to the emergence of the Delta variant in the country, the current paper applies a methodology based on seasonal ARIMA to make a 7 day forecast on the accumulated confirmed cases in Bulgaria for several consecutive weeks and evaluate the errors of the predictions using the actual reported cases for each 7 day period. The analysis of the data concludes that seasonality is present with 7 observations in a cycle. Hence, the forecast period of 7 days coincides with the length of one season. The objective is to provide estimates of confirmed cases with reasonable error levels. The methodology can help authorities to prepare and plan efficiently resources in the fight against the spread of the virus. The rest of the paper is structured as follows: Section 2 describes the methodology applied; Section 3 presents the results after applying the methodology for several consecutive 7 day periods; finally, Section 4 has conclusions drawn based on the results

2 Methodology

The methodology applied in this paper is based on Auto-Regressive Integrated Moving Average (ARIMA). A flow chart of the process can be viewed in Appendix A. In this section, the steps of the process are described in more details as well. The first steps refer to data exploration and this section are included also conclusions from the data analysis part. The Bulgarian government publishes data on the number of confirmed, infected, recovered and vaccinated cases on daily basis. The paper uses the data provided by the government, as it can be considered as credible and reliable. The process is implemented in python and Spyder served as integrated development environment.

Logically, the first step in the process includes data extraction from the data source, which in this case is https://data.egov.bg/. The data is being updated by the government on a daily basis and it is available since 6th of June 2020. The data set does not contain any missing values. After removal of unnecessary features, the final time series data set consists of dates and the respective accumulated confirmed cases. The research proceeds with the analysis phase. Since the ARIMA technique is being utilized, it is of high importance to transform the time series into stationary and determine whether there is seasonality in the series. The auto-correlation and partial auto-correlation show evidence of both non-stationary data and seasonality in the series. It is determined that each seasonal cycle includes 7 observations or 7 days. To make the time series non-stationary, the differencing technique is applied. After 2nd order differencing, evidences confidently suggest presence of stationary data.

Following the transformation of the time series to stationary and the conclusion that seasonality with 7 observations (days) per cycle is present, the methodology proceeds with split of the data into train and test subsets. The first 90 percent of the observations constitute the train sample and the remaining 10 percent are kept for testing. Using the auto_arima function from the pmdarima library available in python, the optimal parameters for seasonal-ARIMA (SARIMA) are determined. The Akaike Information Criterion (AIC) serves as criterion to decide the best configuration of the SARIMA parameters.

The model is trained and predictions are made for the test period. The predicted values and the actual values from the test subset are measured in terms of errors using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). As next step, the model is built this time by using all observations and finally, a 7-day forecast is created. When the actual data for the 7-day forecast period becomes available, the model is validated by estimating the errors between the actual and predicted values. The new daily confirmed cases are added to the time series and the methodology is repeated from the train/test split step. The model is calibrated and a new 7-day forecast can be made. This procedure can run for n weeks or until a significant change in the data occurs (for example, the seasonality pattern changes). In this paper, the procedure is performed for three seasons (three weeks)

3 Results

In this section are presented the results after running the procedure for two consecutive weeks. The weeks, subject of forecasting, cover the period 24.8.2021 - 13.9.2021. The parameters of the model for each week is included, along with evaluation of the train/test process. The forecast and actual values for the first two weeks are presented, and estimation of the error are provided. For the third week, only the forecast is available, as at the time of the writing of the paper, the actual values were not available yet

3.1. First week

The first week for which this study makes a forecast encompasses the period from 24th August until 30th August of 2021. Table 1 presents the forecast and actual values of accumulated confirmed cases for the corresponding date from the period. Following the methodology, the selected parameters for the SARIMA model are (p=2, d=2, q=1)(P=5, D=0, Q=0). As noted in the previous section, the analysis concluded that time series contain seasonality with 7 observations per seasonal cycle.The AIC of the model results in 6 286. During the test phase, the model evaluation showed MSE of 43 765 748, RMSE of 6 615 and MAE of 4 190. Using the actual accumulated cases for the forecast 7-day period, also a validation of the predictions is performed that shows MSE of 638 099, RMSE of 799 and MAE of 774. The test data subset has considerably more observations than the validation phase, which explains the significant difference in errors.

Table 1 - First Week

Date	Forecast	Actual
8/24/2021	442,785	443,186
8/25/2021	444,396	445,097
8/26/2021	445,945	446,698

8/27/2021	447,500	448,431
8/28/2021	449,100	450,144
8/29/2021	450,229	451,148
8/30/2021	450,931	451,599

3.2. Second week

Once the actual reported cases of the first forecast week become available, they are added to the time series data and the process is executed again. The second week in the study covers the period from 31st August until 6th September 2021. Table 2 shows the forecast and actual accumulated cases of COVID-19 for each corresponding day from the second week. The methodology suggest applying the same values of the parameters of SARIMA as in the first week, namely (p=2, d=2, q=1)(P=5, D=0, Q=0). The AIC of the model results in 6 382. The testing evaluation measures show MSE of 96 694 423, RMSE of 9 833 and MAE of 6 425. Validating the model using the actual values of accumulated cases for the 7-day forecast period results in MSE of 369 935, RMSE of 608 and MAE of 442.

Table 2 - Second week

Date	Forecast	Actual
8/31/2021	453,601	453,689
9/1/2021	455,676	455,742
9/2/2021	457,482	457,487
9/3/2021	459,365	459,051
9/4/2021	461,236	460,691
9/5/2021	462,467	461,545
9/6/2021	463,187	462,033

3.3. Third week

It is proceeded in same way, as in the second week, the actual cases for the second week are added to the time series and the procedure is ran once more. The third week covers the dates between 7th September and 13th September. The predicted values for each date are available in Table 3. Since the week has not concluded, at the time of the authoring of this paper, the actual values and validation errors are not available. This time the methodology leads to a different SARIMA model with parameters (p=0, d=2, q=3)(P=5, D=0, Q=0). The AIC is estimated to 6 466 and the performance measures from the test phase result in MSE of 160 333 959, RMSE of 12662 and MAE of 8 462.

Table 3 - Third week	
Date	Forecast
9/7/2021	464,059
9/8/2021	466,059
9/9/2021	467,753
9/10/2021	469,340
9/11/2021	470,958
9/12/2021	471,846
9/13/2021	472,330

The actual cases can be added to the time series data once they become available and the process can continue to be executed in this fashion n times, unless a change in the pattern of the source data is noticed.

4. Conclusions

In this paper, a methodology is proposed for predicting accumulated confirmed COVID-19 cases in Bulgaria. Analysis of the time series data concluded to a presence of seasonality in the time series where each seasonal cycle contains 7 observations. As expected, the data analysis also showed trend in the time series, which was overcome with the differencing technique. The methodology applies seasonal-ARIMA as a modeling technique to create a forecast for the next 7 days (in other words, next season). Once the actual confirmed cases for the forecast period become available, the procedure can be executed with the updated data to make a forecast for the next week.

As an example, the paper showcases the use of the methodology for three weeks during the a period of increase in COVID-19 cases due to the emergence of the Delta variant in the country. The results of the forecasts appear to contain reasonable errors based on the validation step. The methodology utilizes publicly available data and it is easy to be implemented. The Delta variant has already caused dramatic increase of confirmed cases in Bulgaria which suggests that the fight against COVID-19 is far from over, especially considering the slow vaccination rate in the country. Efficient planning of health care resources and measures is of vital importance, and the proposed procedure could serve as an assistant tool to interested institutions. As future work, the same methodology can be evaluated for predicting development of death and recovered cases.

Acknowledgements

The paper is supported by the project under the Sofia University St.Kl. Ohridski grant N80-10-93 FOR 2021

References

[1] ALJAME M., IMTIAZ A., AHMAD I., MOHAMMED A. (2021). DEEP FOREST MODEL FOR DIAGNOSING COVID-19 FROM ROUTINE BLOOD TESTS. SCIENTIFIC REPORTS, 11. HTTPS://DOI.ORG/10.1038/S41598-021-95957-W

[2] ARONU, C. O., EKWUEME, G. O., SOL-AKUBUDE, V. I., OKAFOR P. N. (2020). CORONAVIRUS (COVID-19) IN NIGERIA: SURVIVAL RATE. SCIENTIFIC AFRICAN, 11, E00689. HTTPS://DOI.ORG/10.1016/J.SCIAF.2020.E00689

[3] ARUNKUMAR, K. E., ET AL (2021). FORECASTING THE DYNAMICS OF CUMULATIVE COVID-19 CASES (CONFIRMED, RECOVERED AND DEATHS) FOR TOP-16 COUNTRIES USING STATISTICAL MACHINE LEARNING MODELS: AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) AND SEASONAL AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA). APPLIED SOFT COMPUTING JOURNAL, 103 (2021) 107161. HTTPS://DOI.ORG/10.1016/J.ASOC.2021.107161

[4] BRERETON C., PEDERCINI M. (2021). COVID-19 CASE RATES IN THE UK: MODELLING UNCERTAINTIES AS LOCKDOWN LIFTS. SYSTEMS, 9, 60. HTTPS://DOI.ORG/10.3390/SYSTEMS9030060

[5] COVID-19 UNIFIED INFORMATION PORTAL (2021). HTTPS://CORONAVIRUS.BG/, ACCESSED ON AUGUST 31, 2021.

[6] COVID-19 UNIFIED INFORMATION PORTAL (2021). HTTPS://CORONAVIRUS.BG/BG/NEWS/2255, ACCESSED ON AUGUST 31, 2021.

[7] CUCINOTTA, D., VANELLI, M. (2020). WHO DECLARES COVID-19 A PANDEMIC. ACTA BIOMED, VOL. 91, N. 1: 157-160. DOI: 10.23750/ABM.V91I1.9397

[8] DABBAH M. A. ET AL. (2021). MACHINE LEARNING APPROACH TO DYNAMIC RISK MODELING OF MORTALITY IN COVID-19: A UK BIOBANK STUDY. SCIENTIFIC REPORTS, 11. HTTPS://DOI.ORG/10.1038/S41598-021-95136-X

[9] DUAN, X., ZHANG, X. (2020). ARIMA MODELLING AND FORECASTING OF IRREGULARLY PATTERNED COVID-19 OUTBREAKS USING JAPANESE AND SOUTH KOREAN DATA. DATA IN BRIEF, 31, 105779. HTTPS://DOI.ORG/10.1016/J.DIB.2020.105779

[10] GUIDRY J., ET AL. (2021). U.S. PUBLIC SUPPORT FOR COVID-19 VACCINE DONATION TO LOW- AND MIDDLE INCOME COUNTRIES DURING THE COVID-19 PANDEMIC. VACCINE, 39 (2021) 2452- 2457. HTTPS://DOI.ORG/10.1016/J.VACCINE.2021.03.027

[11] KUNAL S., ADITI, GUPTA K., ISH P. (2021). COVID-19 VARIANTS IN INDIA:POTENTIAL ROLE IN SECOND WAVE AND IMPACT ON VACCINATION. HEART AND LUNG (2021). HTTPS://DOI.ORG/10.1016/J.HRTLNG.2021.05.008

[12] REINGRUBER J., PAPALE A., RUCKLY S., TIMSIT J., HOLCMAN D. (2021). MONITORING AND FORECASTING THE SARS-COVID-19 PANDEMIC IN FRANCE. MEDRXIV. HTTPS://DOI.ORG/10.1101/2021.07.28.21260870

[13] ROMEO L., FRONTONI E. (2021). A UNIFIED HIERARCHICAL XGBOOST MODEL FOR CLASSIFYING PRIORITIES FOR COVID-19 VACCINATION CAMPAIGN. PATTERN RECOGNITION, VOL. 121, 108197. HTTPS://DOI.ORG/10.1016/J.PATCOG.2021.108197

[14] SHAH S. A. ET AL. (2021). PREDICTED COVID-19 POSITIVE CASES, HOSPITALISATIONS, AND DEATHS ASSOCIATED WITH THE DELTA VARIANT OF CONCERN, JUNE–JULY, 2021. THE LANCET, VOL. 3, ISSUE 9. HTTPS://DOI.ORG/10.1016/S2589- 7500(21)00175-8

[15] VAISHYA R., SIBAL A., SHARMA H., SINGH S. K. (2021). LACK OF VACCINATION AND ASSOCIATED COMORBIDITIES PREDISPOSE TO THE NEED FOR INTENSIVE CARE IN INDIVIDUALS INFECTED WITH THE DELTA VARIANT - A CASE COHORT STUDY FROM A TERTIARY CARE HOSPITAL IN NEW DELHI, INDIA. DIABETES & METABOLIC SYNDROME: CLINICAL RESEARCH & REVIEWS, 15 (2021) 102203. HTTPS://DOI.ORG/10.1016/J.DSX.2021.102203