Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 583–586
DOI: 10.15439/2022F65
ISSN 2300-5963 ACSIS, Vol. 30

The effectiveness analysis of selected IT tools for predictions of the Covid-19 pandemic

Paweł Dymora, Mirosław Mazurek, Kamil Łyczko
Rzeszów University of Technology
al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland
Email: pawel.dymora@prz.edu.pl, mirekmaz@prz.edu.pl,
160773@stud.prz.edu.pl

***Abstract*—The article presents the problem of the complexity of prediction and the analysis of the effectiveness of selected IT tools in the example of the Covid-19 pandemic data in Poland. The study used a variety of tools and methods to obtain predictions of extinct infections and mortality for each wave of the Covid-19 pandemic. The results are presented for the 4th wave with a detailed description of selected models and methods implemented in the prognostic package of the statistical programming language R, as well as in the Statistica and Microsoft Excel programs. Naive methods, regression models, exponential smoothing methods (including ETS models), ARIMA models, and the method of artificial intelligence - autoregressive models built by neural networks (NNAR) were used. Detailed analysis was performed and the results for each of these methods were compared.**

I. Introduction

PREDICTION is the process of making certain anticipations with a definable probability of how phenomena will develop in the future. It derives from the field of statistics. It is rational in nature and uses data from the past in its course. Forecasting is used in various areas, such as predicting the weather, electricity consumption, product prices, etc. The resulting forecasts can provide valuable information and help in making decisions about future activities.

Predicts makes it possible to determine the possible future values of a time series. Various methods are available for their determination. They are based on mathematical models that describe the values of the series. The models may take into account many factors, e.g. historical values of the series, values of predictors, characteristics of the series, etc. A model is created by performing a series analysis and parameter estimation based on the data at hand.

Since the beginning of the Covid-19 pandemic, many disease prediction models have emerged, shaping the interest of the media, policymakers, and the broader public [1, 2]. However, forecasting the future development of a pandemic is challenged by the inherent uncertainty rooted in many "unknown unknowns," not only about the contagious virus itself, but also the human, social, and political factors that coevolve and keep the future of the pandemic open. Forecasting models have varying degrees of predictive accuracy. Researchers have attempted analyses during previous epidemics of the 21st century, namely SARS, H1N1, and Ebola. As reported in [1, 3], predictions of Ebola deaths have often been far from the ultimate reality, with a strong tendency to overestimate. It is therefore important to communicate the uncertainty of such analyses. The Covid-19 pandemic spread rapidly around the world, and researchers attempted to estimate the risk. Many researchers around the world have used various prediction techniques such as the Susceptible-Infected-Recovered model, Susceptible-Exposed-Infected-Recovered model, and Automatic Regressive Integrated Moving Average (ARIMA) model to predict the spread of this pandemic. The ARIMA technique has not been extensively used in Covid-19 forecasting by researchers due to the claim that it is not suitable for use in complex and dynamic contexts. However, in [4], the authors proposed the use of time series algorithms, Autoregressive Integrated Moving Average (ARIMA) and Autoregressive (AR). ARIMA-based models showed promising results compared to AR-based models. However, the most difficult challenge was parameter identification due to the sudden increase-decrease trend in coronavirus cases. The proposed work presents prediction quality scoring metrics for both models. In [5], verification was performed to see how accurate the best-fit predictions of the ARIMA model were with the actual values reported after the entire prediction period. The results showed that despite the dynamic nature of the disease and the continuous changes made by the Kuwaiti government, the actual values for most of the observed period were within the prediction limits of our chosen ARIMA model with a 95% confidence interval. Another direction taken by the researchers is to apply machine learning (ML) based forecasting mechanisms. ML models have long been used in many application domains that required the identification and prioritization of adverse threat factors. In the paper [6], four standard prediction models such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector

machine (SVM), and exponential smoothing (ES) were used to predict risk factors. The results proved that ES performs best among all the models used, followed by LR and LASSO, which perform well in predicting new confirmed cases, mortality, and recovery rate, while SVM performs poorly in all prediction scenarios given the available data set.

The author's paper [7] proposes an approach to forecasting the spread of pandemics based on a vector autoregression model. Time series of the number of new cases and the number of new deaths were combined to obtain a common prediction model. Test results based on data from the United Arab Emirates, Saudi Arabia, and Kuwait showed that the proposed model achieved a high level of accuracy, outperforming many existing methods. which can be a valuable tool in pandemic management.

The pandemic has shown that having knowledge and prediction of its spread will allow one to respond appropriately and attempt to undertake containment. Results obtained in [8] using a network model based on long-short-term memory (LSTM) have shown promise. The proposed model was used to predict the dates when other countries would be able to contain the spread of Covid-19.

In [9], the authors presented the results of a study on developing a neural network model for predicting the spread of Covid-19. They proposed a predictor based on a classical approach with a deep architecture that learns using the NAdam training model. Official data from government and open source repositories were used for training. The results of the proposed model showed high accuracy, which reached more than 99% in some cases.

The scope of this paper is to use different forecasting methods and to compare the results obtained for a given set of data from the course of the Covid-19 pandemic in Poland. The aim is to analyze the correctness and degree of fit of the forecasts obtained using different algorithms.

In the analysis, we used data on the daily incidence and death rates registered in Poland during the Covid-19 pandemic. On their basis, mathematical models were created, and then forecasts of subsequent values for time horizons of different lengths were carried out. In this way, possible scenarios for the course of the pandemic were presented.

The paper is divided into five chapters. The introduction provides a review of the literature and recent trends used in the prediction of Covid-19 pandemic trends. Chapter 2 characterizes the methods and IT tools used in the study such as regression and ARIMA models, methods based on neural networks, and main packages in R, Statistica, and Excel environments. Chapter 3 presents the selected issue of forecasting the evolution of a pandemic. Chapter 4 presents a description of the experiments and a comparative analysis of the results of prediction methods for the 4th wave of the pandemic in Poland. The summary, conclusion, and scope of future research are presented in Chapter 5.

II. Prediction Methods and IT Tools

The paper presents various techniques used in time series prediction - from simple ones, such as naive methods, through adaptive models and autoregressions to more advanced ones, such as ARIMA models or neuro-networks. The obtained models will be evaluated in terms of their fit to historical realizations of the series and the quality of generated forecasts. In addition, the use of selected software tools useful in forecastings such as Microsoft Excel, Statistica, and RStudio environment using R language developed for statistical purposes is presented. We use regression models which is a statistical method of describing utilizing a function the dependence of the values of some variables (explanatory) on the values of others (explanatory, predictors) [10], and autoregressive (AR) models describe the explanatory variable as a function of its lagged values [11-12]. Prediction is also possible based on neural network-based models. Artificial neural networks have a layered structure, which includes neurons that in a simplified way mimic the operation of cells found in the human brain. With the use of neural networks, it is possible to model the autoregression of time series. It consists in feeding the network input with delayed values of time series. An example of such solution implementation is nnetar() function from FORECAST package where one-way neural networks with one hidden layer (NNAR models) are used therefore forecasting [13-15].

III. Forecasting the Evolution of a Pandemic Using Various IT Tools and Methods

The subject of this study is data related to the course of the Covid-19 pandemic in Poland. The decision to use the results from their timeliness and availability at the time of work creation. The collected data describe the daily numbers of infections, deaths, and tests performed between 05.03.2020 and 25.10.2021. They form a time series that will be used to test the effectiveness of different prediction methods.

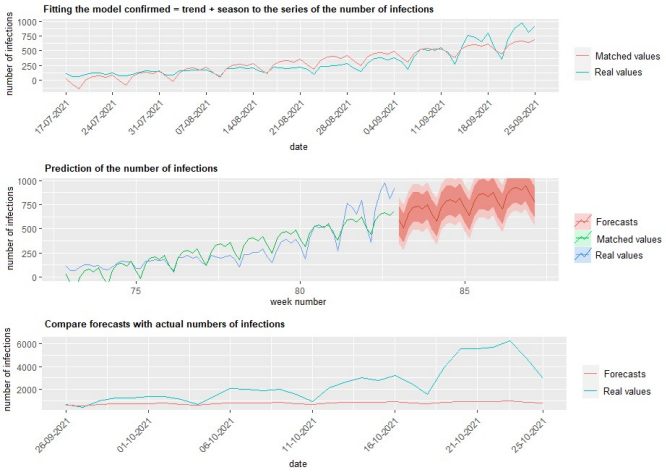

Fig. 1 Regression charts of a series of infection numbers for the 4th wave

TABLE I.
A SUMMARY OF THE VALUES OF THE MEASURES OF THE ACCURACY OF THE PREDICTIONS OF THE NUMBER OF INFECTIONS OBTAINED BY THE NAIVE METHODS

		ME	MAE	MSE	RMSE	MAPE
Forecasts for wave 1	simple naive method	61.03333	109.1	16914.433	130.0555	14.77851
	seasonal naive method	75.56667	108.2333	16411.967	128.1092	14.6654
	incremental naive method	-3.80327	111.4168	16287.586	127.6228	16.56339
Forecasts for wave 2	simple naive method	10850.97	10850.97	144349654	12014.56	49.82617
	seasonal naive method	12019.67	12019.67	170837763	13070.49	56.40277
	incremental naive method	10222.1	10222.1	128708885	11344.99	46.87582
Forecasts for wave 3	simple naive method	12524.6	12767.6	215657209	14685.27	48.94015
	seasonal naive method	7015.067	7351.933	72519578	8515.843	29.32963
	incremental naive method	12074.23	12400.53	205532085	14336.39	47.49426
Forecasts for wave 4	simple naive method	1525.667	1593.467	4984209.6	2232.534	53.85878
	seasonal naive method	1719.233	1719.233	5393828.5	2322.462	58.65325
	incremental naive method	1500.687	1571.495	4871069	2207.05	53.14973

To simplify the code of scripts, functions were prepared to create graphs using methods provided by the ggplot2 package of R language. To validate the quality of models and forecasts, functions were created to calculate the values of error measures (ME, MAE, MSE, RMSE, MAPE) and to select the objects of models and forecasts that are characterized by the most satisfactory features (Table I) [12-15].

The fitting of models to historical data and forecasts generated by them are presented in Fig. 1. Figure shows the fit of the finally selected models to the training series and a comparison of the generated forecasts for the 30-day horizon with the test series. The forecast charts also show the ranges for the 80% and 95% confidence levels.

IV. TIME SERIES ANALYSIS OF PANDEMIC WAVE 4 DEATH NUMBERS

The following methods were selected to forecast the values of death numbers for the 4th wave (time period from 26.09.2021 to 25.10.2021):

- R - simple naive method,
- R - seasonal naive method,
- R - incremental naive method,
- R - linear regression model including linear trend and seasonal variables built on training time series with observations from 17.07.2021 to 25.09.2021),
- R - ETS*(A,Ad,A)* model estimated using historical values of death numbers from 05.03.2020 to 25.09.2021,
- R - ARIMA*(2,1,2)(0,1,1)* model built on the basis of modified by adding constant 1 and undergoing Box-Cox transformation training time series with observations from 05.03.2020 to 25.09.2021,
- R - NNAR*(22,1,12)* model selected based on the training series with observations from 05.03.2020 to 25.09.2021,
- Statistica - Holt model estimated based on the training series with observations from 28.03.2020 to 25.09.2021,
- Statistica - Winters model estimated based on modified (addition of constant 1 and natural logarithmization) training time series with values from 28.03.2020 to 25.09.2021.
- Statistica - ARIMA*(3,1,3)(2,1,2))* model built based on modified by adding constant 1 and undergoing natural logarithmization training time series with observations from 28.03.2020 to 25.09.2021,
- Microsoft Excel - ETS*(A,A,A)* model estimated from the modified training series with death numbers from 28.03.2020 to 25.09.2021.

We concluded that predictions obtained using the Winters method in the Statistica program are characterized by very large values in comparison with predictions created by other methods. They differed significantly from the test (actual) values and were therefore not considered in further analyses. The predictions (excluding those mentioned) are shown in the graph (Fig. 2).

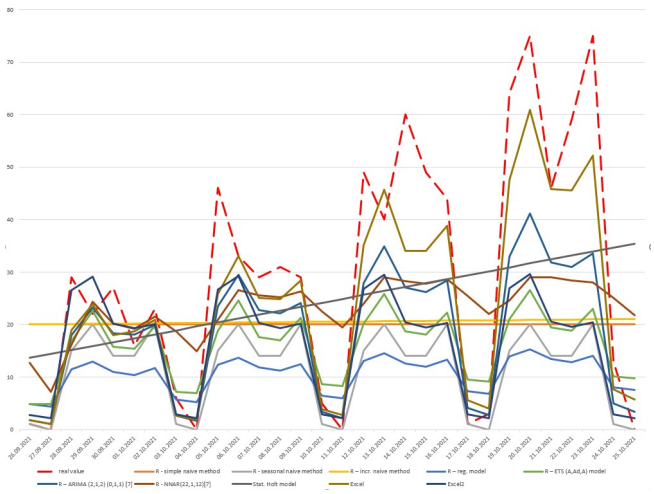

Fig. 2 Comparison of predictions with actual death counts for the 4th wave

Ex post forecast error measures were calculated. The calculation of MAPE values omitted cases for which the value of observations in the test series was equal to 0 (27.09.2021, 04.10.2021, 11.10.2021, and 25.10.2021). The most satisfactory predictions were obtained from the ARIMA*(3,1,3) (2,1,2)* model in Statistica software.

V. Conclusion

The study used a variety of tools and methods to obtain projections of expired infection and death rates for each wave of the Covid-19 pandemic. Forecasting was performed using models and methods implemented in the forecast package of the statistical programming language named R and the programs Statistica and Microsoft Excel. With the first tool, naive methods, regression models, exponential smoothing methods (including ETS models), ARIMA models, and artificial intelligence method - autoregressive models built by neural networks (NNAR) were applied for the examined time series. In Statistica software, modules were used to create predictions based on exponential smoothing methods and ARIMA models. In the case of Excel, predictions were obtained using prediction sheets.

MAE, RMSE, and MAPE error measures were used to verify the quality of the applied models and methods. Based on their values calculated for the training time series, a preliminary selection of the best variants of methods and models within one category was made (e.g. selection of the best regression model from among models taking into account various predictors). The final selection of the best solution for a particular time series was based on the values of the ex-post forecast error measures (calculated for the test set). Based on the analysis of the final results (forecasts obtained with the selected methods), based on the values of the test error measures, the solutions that allowed obtaining the most satisfactory predictions were indicated. For the time series of infection rates, a multiplicative variant of the Winters method from the family of exponential equalization methods was selected as the best forecasting method. For the time series of deaths, the use of ARIMA models was selected as the best approach.

References

[1] Le Ha Anh and Nguyen Minh Trang and Nguyen Thi Phuong Linh, The Influence of Work-from-home on job performance during COVID-19 pandemic: Empirical evidence Hanoi, Vietnam, Proceedings of the International Conference on Research in Management & Technovation, vol. 28, pp. 73-81, http://dx.doi.org/10.15439/2021KM59, 2021

[2] F. Grabowski, A. Paszkiewicz, M. Bolanowski: Wireless networks environment and complex networks, Lecture Notes in Electrical Engineering, Analysis and Simulation of Electrical and Computer Systems, Springer International Publishing Switzerland, ISBN 978-3-319-38545-7, vol. 324, str. 261-270, 2015.

[3] P. Nadella, A. Swaminathan, S. V. Subramanian, SV, "Forecasting efforts from prior epidemics and COVID-19 predictions". EUROPEAN JOURNAL OF EPIDEMIOLOGY, Vol. 35, Issue 8, Page 727-729, DOI: 10.1007/s10654-020-00661-0, 2020.

[4] D. Prajapati, M. Kanojia, "Forecasting of COVID-19 Cases in INDIA Using AReIMA and AR Time-Series Algorithm", Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SOCPAR 2021), Book Series Lecture Notes in Networks and Systems, Vol. 417, Page 361-370, DOI: 10.1007/978-3-030-96302-6_33, 2022.

[5] G. Alabdulrazzaq, M. N. Alenezi, Y. Rawajfih, B. A. Alghannam, A. A. Al-Hassan, F. S. Al-Anzi, "On the accuracy of ARIMA based prediction of COVID-19 spread ", RESULTS IN PHYSICS, Vol. 27, Article Number 104509, DOI: 10.1016/j.rinp.2021.104509, 2021.

[6] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. W. On, W. Aslam, G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models", IEEE ACCESS, Vol. 8, Page 101489-101499, DOI: 10.1109/ACCESS.2020.2997311, 2020.

[7] K. Rajab, F. Kamalov, A. K. Cherukuri, "Forecasting COVID-19: Vector Autoregression-Based Model", Arabian Journal for Science and Engineering, DOI: 10.1007/s13369-021-06526-2, 2022.

[8] S. Kumar, R. Sharma, T. Tsunoda, T. Kumarevel, A. Sharma, "Forecasting the spread of COVID-19 using LSTM network", BMC BIOINFORMATICS, Vol. 22, Issue SUPPL 6, Article Number 316, DOI: 10.1186/s12859-021-04224-2, 2021.

[9] M. Wieczorek, J. Silka, M. Wozniak, "Neural network powered COVID-19 spread forecasting model", CHAOS SOLITONS & FRACTALS, Vol. 140, Article Number 110203, DOI: 10.1016/j.chaos.2020.110203, 2020.

[10] M. Sobczyk, "Prognozowanie. Teoria, przykłady, zadania." Wydawnictwo Placet, 2008.

[11] R. J. Hyndman, G. Athanasopoulos, "Forecasting: Principles and Practise", https://otexts.com/fpp2/. Access 15.09.2021 r.

[12] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ Access 27.09.2021 r.

[13] https://www.rstudio.com/. Access 27.09.2021 r.

[14] https://www.rdocumentation.org/packages/forecast/versions/8.15. Access 27.09.2021 r.

[15] https://www.statsoft.pl/Programy/Architektura-STATISTICA/Programy-desktop/. Access 27.09.2021 r.