Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society

On the accuracy of Covid-19 forecasting methods in Russia for two years

I.A. Moloshnikov\textsuperscript{a}, A.G. Sboev\textsuperscript{a,b,}\textsuperscript{*}, A.V. Naumov\textsuperscript{a}, S.V. Zavertyaev\textsuperscript{a}, R.B. Rybka\textsuperscript{a}

\textsuperscript{a}NRC “Kurchatov Institute”, Academician Kurchatov sq., 1, Moscow, 123098, Russia
\textsuperscript{b}MEPhI National Research Nuclear University, Kashirskoye sh., 31, Moscow, 115409, Russia

Abstract

The effectiveness of predicting the dynamics of the coronavirus pandemic for Russia as a whole and for Moscow is studied for a two-year period beginning March 2020. The comparison includes well-proven population models and statistic methods along with a new data-driven model based on the LSTM neural network. The latter model is trained on a set of Russian regions simultaneously, and predicts the total number of cases on the 14-day forecast horizon. Prediction accuracy is estimated by the mean absolute percent error (MAPE). The results show that all the considered models, both simple and more complex, have similar efficiency. The lowest error achieved is 18% MAPE for Moscow and 8% MAPE for Russia.

© 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society.

Keywords: covid-19 forecasting; time series analysis; total cases prediction; machine learning; SIR

Introduction

Over the past two years, the coronavirus itself has mutated, the strength of its virulence has changed due to vaccination, which has led to a change in the dynamics of the pandemic and therefore requires additional study of the accuracy of prediction models.

The literature presents a number of methods to solve this problem based on SIR or SEIR models [1, 2, 3] and their modifications [4, 5, 6, 7], statistical models like ARIMA and Holt’s trend models [8, 9, 10], regression type of machine learning models [11, 12], and its combinations [13, 14, 15]. For a number of countries, such as USA and Japan, available data by regions, population composition, migration and quarantine decisions gave possibility to build complex models [16].

\textsuperscript{*} Corresponding author. Tel.: +7-926-253-7217.
E-mail address: Sboev\_AG@nrcki.ru

1877-0509 © 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society.
In our previous works [15, 17] on studying the accuracy level of predicting the dynamics of the pandemic for Russia as a whole and for individual regions, a comparative analysis of possible approaches was carried out using data for one year (2020-2021). It was shown that on the same initial data, machine learning models and population models have a relatively small advantage over statistical methods, and in cases of a limited amount of data, they are inferior to them. We therefore use statistical methods of time series forecasting as a baseline against which to compare. A number of papers explore different prediction depths: 7 days [18], 14 [17], 28 days [16]. Following the prior work of 2021 [17], the horizon of 14 days has been chosen for the forecast.

At the moment, time series of data on the Covid-19 pandemic dynamic for more than two years have been accumulated. This data allows for an extended assessment of the effectiveness of the predictive ability of models, taking into account the large variability in the dynamics of the pandemic. Based on the accumulated history, this paper analyzes the current level of accuracy of some newly-proposed models and some well-proven models for forecasting the dynamics of Covid-19 development for Russia [17] and Moscow [15].

1. Data

We used statistical data on the number of cases, recoveries and deaths taken from the Yandex DataLens project\(^1\) for the period from March 2021 to February 2022. The final data set contains values for 706 days from 03/12/2020 to 02/16/2022 for 85 regions (Federal subjects) of Russia and Russia as a whole.

In order to evaluate the efficiency of the prediction methods at various stages of the epidemic and on regions with different population and different scale of the epidemic, the data was preprocessed as follows:

- the data – time-series of the numbers of Covid-19 cases, recoveries, and deaths – was normalized per 100 thousand population as \( v_{\text{norm}} = v \cdot 10^5 / P_{\text{region}} \), where \( v \) is the value for the current day, \( P_{\text{region}} \) is the population of the region taken from the Yandex DataLens;
- the observation duration available in the dataset was divided into 5 parts (folds) of equal duration, out of which 4 splits into training and testing data are formed. The separation was performed in the data-accumulating manner for the training part, as shown in Figure 1: when some fold is chosen for testing, all preceding ones are used for training.

\(^{\text{1}}\) Yandex DataLens project: https://datalens.yandex.ru/
2. Methods

2.1. Statistical model

We used the Holt-Winters exponential smoothing method from the “statsmodels” library\(^2\) (with the following parameters: trend = additive, damped_trend = true, seasonal = None, seasonal_periods = None, use_boxcox = true), which proved itself efficient on the data from the first year of the pandemic [15, 17]. The output of prediction was the total number of new Covid-19 cases over 14 days. Two ways of applying exponential smoothing to obtain the prediction output have been tested:

- take the numbers of new cases for each day from the training part as input, predict the number of new cases for the 14 days ahead, and sum up the predicted values (further referred to as ES\(_{\text{daily}}\));
- take as input the total number of new cases over 14 days for each 14-day interval from the training part, and predict the total number of cases for the 14 days ahead (further referred to as ES\(_{\text{sum}}\)).

2.2. Population models

We use two population models, Susceptible-Infected-Removed (SIR) and Susceptible-Exposed-Infected-Removed (SEIR).

In the SIR model, the dynamics of the numbers of susceptible \(S\), infected \(I\), and recovered and dead \(R\) is defined by Equations 1.

\[
\begin{align*}
\frac{dS}{dt} &= -\frac{\beta IS}{N}, \\
\frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I, \\
\frac{dR}{dt} &= \gamma I
\end{align*}
\]

(1)

Here the total population \(N = S + I + R\) is taken equal to the population of the region, which equals \(10^5\) after the preprocessing described in Section 1. The propagation rate \(\beta\) and the recovery rate \(\gamma\) are calculated from the dynamics of \(S\), \(I\), and \(R\) on the last day \(t\) before the starting date of the prediction, by applying Equations 1 with the time step of 1 day (see Equations 2).

\[
\begin{align*}
S(t) - S(t-1) &= -\frac{\beta I(t-1)S(t-1)}{N}, \\
I(t) - I(t-1) &= \frac{\beta I(t-1)S(t-1)}{N} - \gamma I(t-1), \\
R(t) - R(t-1) &= \gamma I(t-1),
\end{align*}
\]

(2)

The values of \(S\), \(I\), and \(R\) for the days \(t\) and \(t-1\) are obtained from the statistical data (preprocessed as described in Section 1): \(R(t)\) is the sum of the cumulative numbers of recovered and deceased accumulated by the day \(t\), \(I(t)\) is the cumulative number of cases by the day \(t\) minus \(R(t)\), and \(S(t) = N - I(t) - R(t)\).

As a result of iterative application of Equation 2, the values of \(S\), \(I\), and \(R\) for the days following \(t\) are calculated. The predicted number of new cases over the next 14 days was defined as \(S(t+1) - S(t+15)\). This model is further referred to as SIR\(_{\text{static}}\).

As a basic model for comparison, we also use the modified SEIR model that was previously successfully applied to data from March 2020 to February 2021 [15]. In this model, \(\beta = R_0(t) * \gamma\), where \(R_0(t)\) is predicted by a regression

\(^2\) https://www.statsmodels.org/stable/
2.3. ML model

The proposed model was based on an architecture consisting of a combination of Long short-term memory (LSTM) and fully connected (FC) layers (see Figure 2). The input for the model was a vector of numbers of new daily cases during the preceding 28 days, additionally divided by 1000, and the region identifier, which was processed by the vector representation layer (Embedding). The output was the total number of cases in 14 days divided by 1000.

The model was trained on data from all regions of Russia and Russia as a whole, which was considered as a separate region. For validation, 20% samples randomly selected from the training dataset were used. The MSE was used as the loss function. Such an organization of the learning process has made it possible to build a single predictive model for all regions. However, for subsequent comparison, this model was adjusted for the target region, which determines the boundaries of the division into training and test parts when implementing the cross-validation procedure. This was due to the different start of data collection in different regions.

The best performance was shown by the model with the following combination of hyperparameters: 2 neurons of the LSTM layer, 6 values of embedding layer for vectorization of the region index, 8 neurons with a hyperbolic tangent activation function in a fully connected layer, 1 neuron with a linear activation function in the output layer, as well as parameters batch_size = 21, learning_rate = 0.003, optimizer: Adam. TensorFlow version 2.6 was used to implement the model.

3. Experiments and results

In this work, the forecast accuracy was estimated using the Mean Absolute Percentage Error (MAPE) metric:

\[
MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|,
\]

where \( n \) is the number of days in the test part, \( A_t \) and \( F_t \) are true and predicted values of the total 14-day number of cases (preprocessed as described in Section 1) for the day \( t \).

Tables 1 and 2 present the accuracies for Russia and Moscow by the MAPE metric. The “Fold” columns present accuracies calculated for the testing parts of the corresponding folds. The “Mean” columns present accuracies calculated for the concatenation of the predicted time series of the testing parts from all folds. For the column “Mean Mar 2020 - Feb 2022”, the models were trained and tested on the folds depicted in Figures 3 and 4, while for the column

---

3 https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
Table 1. MAPE of prediction of Covid-19 cases for Russia

| Method     | 2020-2022 | Mean (Mar 2020 - Feb 2021) | Mean (Mar 2020 - Feb 2022) |
|------------|-----------|---------------------------|---------------------------|
|            | Fold_1    | Fold_2        | Fold_3 | Fold_4 |                   |                   |
| SEIR SVR   | 9         | 6            | 8       | 17      | 9                 | 10                |
| SIR_static | 9         | 7            | 9       | 14      | 9                 | 10                |
| ES_daily   | 6         | 6            | 8       | 14      | 6                 | 8                 |
| ES_sum     | 6         | 5            | 8       | 14      | 6                 | 8                 |
| LSTM.FC    | 32        | 18           | 7       | 16      | 16                | 18                |

Table 2. MAPE of prediction of Covid-19 cases for Moscow

| Method     | 2020-2022 | Mean (Mar 2020 - Feb 2021) | Mean (Mar 2020 - Feb 2022) |
|------------|-----------|---------------------------|---------------------------|
|            | Fold_1    | Fold_2        | Fold_3 | Fold_4 |                   |                   |
| SEIR SVR   | 16        | 22           | 25     | 29      | 16                | 23                |
| SIR_static | 18        | 18           | 21     | 29      | 18                | 21                |
| ES_daily   | 13        | 15           | 20     | 26      | 15                | 18                |
| ES_sum     | 13        | 14           | 19     | 27      | 15                | 18                |
| LSTM.FC    | 36        | 16           | 16     | 23      | 30                | 23                |

“Mean Mar 2020 - Feb 2021”, the models were trained and tested on the data from March 2020 to February 2021 divided into 4 folds in the same way.

Examples of the predictive ability of models are presented in Figures 3 and 4.

The tables show that, though all models show similar accuracy, LSTM.FC is inferior to other models during the first fold, probably because of the lack of training data, but can compete with other models on later folds.

4. Conclusion

The calculations carried out on the data set accumulated by today demonstrate the stability of the prediction accuracy relative to the first year. The error level of forecasting the total number of new cases over 14 days for Moscow and Russia is 18% MAPE and 8% MAPE, respectively. The achieved values are in agreement with those shown in foreign sources: for example, the average forecast of cases in the US states for 28 days is 37.8%. The use of more complex models (SEIR SVR with time-varying parameters, LSTM.FC trained on data from all regions) does not lead to a meaningful increase in prediction accuracy compared to SIR with constant parameters and to exponential smoothing.
meaningful increase in prediction accuracy compared to SIR with constant parameters and to exponential smoothing.

Table 2. MAPE of prediction of Covid-19 cases for Moscow

Table 1. MAPE of prediction of Covid-19 cases for Russia

Russia is 18% MAPE and 8% MAPE, respectively. The achieved values are in agreement with those shown in foreign

Accomplishments

This study was supported by the Russian Foundation for Basic Research project №20-04-60528 and carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute”, http://ckp.nrcki.ru/.

References

[1] E. B. Postnikov, Estimation of covid-19 dynamics “on a back-of-envelope”: Does the simplest sir model provide quantitative parameters and predictions?, Chaos, Solitons & Fractals 135 (2020) 109841.
[2] N. A. Kudryashov, M. A. Chmykhov, M. Vigdorowitsch, Analytical features of the sir model and their applications to covid-19, Applied Mathematical Modelling 90 (2021) 466–473.
[3] G. Perakis, D. Singhvi, O. Skali Lami, L. Thayaparan, Covid-19: A multiwave sir-based model for learning waves, Production and Operations Management (2022).
[4] I. Rahimi, A. H. Gandomi, P. G. Asteris, F. Chen, Analysis and prediction of covid-19 using sir, seiqr, and machine learning models: Australia, Italy, and uk cases, Information 12 (3) (2021) 109.
[5] P. Singh, A. Gupta, Generalized sir (gsir) epidemic model: An improved framework for the predictive monitoring of covid-19 pandemic, ISA transactions 124 (2022) 31–40.
[6] J. L. Spouge, A comprehensive estimation of country-level basic reproduction numbers r 0 for covid-19: Regime regression can automatically estimate the end of the exponential phase in epidemic data, PloS one 16 (7) (2021) e0254145.
[7] N. A. Kudryashov, M. A. Chmykhov, M. Vigdorowitsch, Comparison of some covid-19 data with solutions of the sir-model, in: AIP Conference Proceedings, Vol. 2425, AIP Publishing LLC, 2022, p. 340009.
[8] M. Abotaleb, T. Makarovskikh, System for forecasting covid-19 cases using time-series and neural networks models, Engineering Proceedings 5 (1) (2021) 46.
[9] H. Swapnarekha, H. S. Behera, J. Nayak, B. Naik, P. S. Kumar, Multiplicative holts winter model for trend analysis and forecasting of covid-19 spread in india, SN Computer Science 2 (5) (2021) 1–15.
[10] F. A. Awwad, M. A. Mohamoud, M. R. Abonazel, Estimating covid-19 cases in makkah region of saudi arabia: Space-time arima modeling, PLoS One 16 (4) (2021) e0250149.
[11] E. Gambhir, R. Jain, A. Gupta, U. Tomer, Regression analysis of covid-19 using machine learning algorithms, in: 2020 International conference on smart electronics and communication (ICOSEC), IEEE, 2020, pp. 65–71.
[12] B. Zhou, G. Yang, Z. Shi, S. Ma, Interpretable temporal attention network for covid-19 forecasting, Applied Soft Computing 120 (2022) 108691.
[13] R. Vega, L. Flores, R. Greiner, Simlr: Machine learning inside the sir model for covid-19 forecasting, Forecasting 4 (1) (2022) 72–94.
[14] M. A. Bakar, N. Mohamed, D. A. Pratama, M. F. A. Yusran, N. A. Aleng, Z. Yanuar, L. Niken, Modelling lock-down strictness for covid-19 pandemic in asean countries by using hybrid arima-svr and hybrid sir-ann, Arab Journal of Basic and Applied Sciences 28 (1) (2021) 204–224.
[15] A. Shoev, N. Kudryashov, I. Moloshnikov, A. Serenko, A. Naumov, R. Rybka, Baseline accuracy of forecasting covid-19 cases in moscow region on a year in retrospect using basic statistical and machine learning methods, in: Journal of Physics: Conference Series, Vol. 2036, IOP Publishing, 2021, p. 012029.
[16] S. ¨O. Arık, J. Shor, R. Sinha, J. Yoon, J. R. Ledsam, L. T. Le, M. W. Dusenberry, N. C. Yoder, K. Popendorf, A. Epshteyn, et al., A prospective evaluation of ai-augmented epidemiology to forecast covid-19 in the usa and japan, NPJ digital medicine 4 (1) (2021) 1–18.
[17] A. Naumov, I. Moloshnikov, A. Serenko, A. Sboev, R. Rybka, Baseline accuracies of forecasting covid-19 cases in russian regions on a year in retrospect using basic statistical and machine learning methods, Procedia Computer Science 193 (2021) 276–284.

[18] O. Istaiteh, T. Owais, N. Al-Madi, S. Abu-Soud, Machine learning approaches for covid-19 forecasting, in: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), IEEE, 2020, pp. 50–57.