Baseline accuracy of forecasting COVID-19 cases in Moscow region on a year in retrospect using basic statistical and machine learning methods

A G Sboev¹², N A Kudryshov¹, I A Moloshnikov¹, A V Serenko¹, A V Naumov¹, R B Rybka¹
¹ NRC "Kurchatov Institute", Academician Kurchatov sq., 1, Moscow, 123098, Russia
² MEPhI National Research Nuclear University, Kashirskoye sh., 31, Moscow, 115409, Russia
E-mail: sag111@mail.ru

Abstract. The large amount of data that has accumulated so far on the dynamics of the COVID-19 outbreak has allowed to assess the accuracy of forecasting methods in retrospect. This work is devoted to comparing a set of basic time series analysis methods for forecasting the number of confirmed cases for 14 days ahead: machine learning methods, exponential smoothing, autoregressive methods, along with variants of SIR and SEIR. On the year-long data for Moscow, the best basic model is showed to be SEIR within which the basic reproduction number $R_0$ is predicted using a regression model, achieving the mean error of 16% by the MAPE metric. The resulting accuracy can be considered a baseline for a more complex prospective model that would be based on the presented approach.

Keywords: COVID forecasting, Machine learning, Time series analysis

1. Introduction

Currently, despite many publications devoted to the problem of describing the spread of the COVID-19 epidemic, models of this process that would be actually efficient and useful for practice are still lacking [1]. The need for these models for planning policies to contain the outbreak is undeniable. So, the task of assessing accuracy of different forecasting methods in retrospect seems extremely relevant.

In the literature, forecasting the dynamic of the epidemic is generally considered as the task of forecasting the numbers of confirmed cases [2–4], recoveries [2; 4] or deaths [2; 4; 5]. The forecast is usually based on the input data for a few previous days, and predicts the respective time series for some days ahead. The input data vectors for a day may contain additional features indirectly characterising the evolving situation: mitigation measures [6; 7], statistic of search queries containing keywords relevant to the...
Baseline accuracy of forecasting COVID-19 cases in Moscow region

epidemic [8], data inferred from social network texts [9], and so on. Also, different works consider different forecasting horizons: 1 day ahead [3; 10], 7 days [11], 10 days [4; 5; 12] or more: note the competitions “COVID-19 Data Challenge” by SberBank‡ and the Zindi§ competition.

A vast variety of forecasting methods is used in different works:

- the tomorrow-like-today approach, also referred to as the naive approach [13].
- exponential growth models [14].
- population models such as Susceptible-Infected-Removed (SIR), Susceptible-Exposed-Infected-Removed (SEIR) and their modifications [1; 2; 5; 10; 15–19].
- machine learning methods for regression [4; 11–13; 20], including SVM [4], LSTM [3; 10; 11; 20; 21], and deep neural network models [22].

However, the works devoted to comparing different approaches [4; 11; 13; 20] show that the more sophisticated methods sometimes show no significant increase in accuracy compared to the simplistic ones: for instance, complex machine learning methods do not outperform single ones when forecasting for 10 days [4].

Due to that, a task of practical importance considered in this paper is to assess accuracy levels achieved by basic forecasting methods on the available retrospective data. This will prepare the basis for further research and improvement of models for forecasting the dynamics of the epidemic.

2. Data

We use data from the Yandex DataLens∥ service for a period of one year (from March 2020 to March 2021) for the Moscow region, which contains daily numbers of confirmed cases, recoveries and deaths.

We normalize the data per 100 thousand of population, and preprocess it by discarding the points before the day with the first detected cases and replacing all values less than or equal to 0 with interpolated values for their non-zero neighboring values.

In this work, we focus on predicting the number of confirmed cases, considering it to be the most real-time indicator of the dynamics of the outbreak. The target value for forecasting is the total increase in the number of confirmed cases over a certain period of time (14 days), because we believe it to be less dependent on stochastic factors unrelated to the epidemic spread (delays in testing, in reporting, etc.).

‡ https://ods.ai/competitions/sberbank-covid19-forecast/
§ https://zindi.africa/competitions/predict-the-global-spread-of-covid-19
∥ https://datalens.yandex/covid19
Baseline accuracy of forecasting COVID-19 cases in Moscow region

3. Models

3.1. Dummy

As the simplest basic time series prediction model we use the tomorrow-like-today approach, in which the future value of the time series is taken equal to its current value. In spite of the simplicity of such approach, it often proves hard to outperform, and thus is here used for comparison purposes.

3.2. Statistical models

Out of this class of models, we use the Holt-Winters exponential smoothing model namely Holt Winter’s method from the statsmodels library with the “trend” parameter set to “additive” and other parameters kept default (denoted further as Smoothing); and the ARIMA autoregressive model with automatic selection of optimal model parameters from the pmdarima library (denoted as AutoARIMA).

The input data for statistical models is the entire time series (preprocessed as described in Section 2), from its very beginning and up to the day for which prediction is being made. The output is the values of the preprocessed time series for the 14 days following the current day.

3.3. Machine learning methods

We use the following popular machine learning methods: Support Vector Regression with linear kernel and max_iter=5000 (further referred to as LinearSVR); least-squares linear regression with the parameter normalize=True (LR); an ensemble of decision tree models with n_estimators=100 (RandomForest); gradient boosting with n_estimators=100 (GBR).

All of the above models are taken from the sklearn library [23], with all hyperparameters kept default except those specified above. The input data for the models is a vector consisting of the values of the preprocessed series for the last 14 days before the current day, while the output is the value of the series for the day that is 14 days after the current day.

3.4. Population models

3.4.1. SIR with parameters adjusted independently for each day

In the SIR model [24; 25] considered here (further referred to as SIR), 14-day forecast is performed each day with fitting the model parameters to the past 14 days. For the parameters to fit we consider the population size $N$ potentially susceptible to infection, the susceptible-to-infected transition rate $\beta$ and infected-to-removed (recovered or deceased) transition rate $\gamma$. The compartment sizes of the susceptible $S$, the infected $I$ and the removed
Baseline accuracy of forecasting COVID-19 cases in Moscow region

$R$ for the day before the period under forecast are set on base of the known values of the Confirmed, Recovered and Deceased series: $R = \text{Recovered} + \text{Deceased}$, $I = \text{Confirmed} - R$, $S = N - I - R$. Fitting the parameters is performed using the least-squares algorithm minimizing the mean square error for the $I$ and $R$ series on 14 past days.

3.4.2. SIR and SEIR with parameters changing in expert-preset dates The model parameters ($\beta$ and $\gamma$ for SIR; $\beta$, $\gamma$, and $\delta$ for SEIR) are constant in the intervals delimited by preset dates:

$$\beta(t) = \begin{cases} 
\beta_1 & \text{if } t <= date_1 \\
\beta_2 & \text{if } date_1 < t <= date_2 \\
... \\
\beta_n & \text{if } date_{n-1} < t <= date_n \\
\beta_{n+1} & \text{if } t > date_n 
\end{cases} \quad (1)$$

The dynamics of $\gamma(t)$ and $\delta(t)$ are defined analogously. Dates $date_i$ correspond to significant events that affect the evolving outbreak. Dates for the Moscow region are obtained from the Wikipedia webpage*: 

(i) 31-03-2020 – the start of lockdown (stay-at-home order);
(ii) 12-04-2020 – closure of almost all enterprises and organizations stops, introduction of mandatory permits for leaving home;
(iii) 12-05-2020 – imposition of mandatory wearing of masks in public transport, resumption of the work of industrial organizations and construction facilities;
(iv) 09-06-2020 – the end of lockdown;
(v) 01-09-2020 – end of summer vacation and the beginning of the school year;
(vi) 29-09-2020 – imposition of remote learning;
(vii) 31-12-2020 – the new year holidays; 
(viii) 15-01-2021 – lifting of most of restrictive measures except masks;

The 14-day forecast for a day $t$ involves fitting the parameters $\beta_i$, $\gamma_i$, and $\delta_i$ by the non-linear least-squares algorithm minimizing the mean square error of the $I$ series over all last data up to $t$. Here, $i$ includes dates no later than $t$ minus 7 days; if $t$ is less than 7 days from $date_i$, then $\beta_i$, $\gamma_i$, and $\delta_i$ do not take new values at $date_i$ because such interval is considered too short for parameter fitting.

Further, the SIR and SEIR models with such parameter fitting are referred to as SIR\_expert and SEIR\_expert.

* Wikipedia: The spreading of COVID-19 in Moscow (in Russian): https://ru.wikipedia.org/wiki/%D0%9A%D0%80%D1%81%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D0%B6%D0%BD% D0%B8%D0%B5_COVID-19_%D0%B2_%D0%9C%D0%BE%D1%81%D0%BA%D0%B2%D0%B5
Baseline accuracy of forecasting COVID-19 cases in Moscow region

3.4.3. SEIR and SIR with regression model for \( R_0(t) \) In this model, the parameter \( \beta(t) = \gamma \cdot R_0(t) \) is obtained on base of the basic reproduction number \( R_0(t) \) of how many people an average infected person infects in a day. It is calculated as

\[
R(t) = \frac{\sum_{k=t-w}^{t} \text{Daily\_cases}(k)}{\sum_{i=t-2w}^{t-1} \text{Daily\_cases}(i)},
\]

(2)

where \( \text{Daily\_cases}(t) \) is the number of daily confirmed cases for day \( t \), \( t \) is the number of the current day, \( w \) is the window size (we use \( w = 4 \) in accordance with the approach of Yandex\(^\#\)).

\( \gamma \) is searched for using the least-squares algorithm minimizing the mean square error of \( I \) on 14 past days. Initial compartment states are obtained from the known daily values of confirmed cases, recoveries and deaths. \( N \) is set fixed at 100,000 in accordance with the prior normalizing of the series of confirmed and recovered.

The future values of \( R_0(t) \) are forecast using a regression model based on Epsilon Support Vector Regression (SVR)\(^\dagger\dagger\) with the parameters \( C=100, \text{kernel}="\text{rbf}" \). During fitting, SVR receives the number of the day \( t \) at its input, and outputs \( R_0(t) \). Fitting SVR and finding \( \gamma \) is carried out for every day using past values of \( R_0(t) \) calculated by Eq. 2. The 14-day forecast is obtained with the help of SVR predicting \( R_0 \) for each of the 14 days ahead, while \( \gamma \) is fixed in its value obtained for the current day.

The SEIR model is employed in the same fashion. For the initial parameter for Exposed we take its value obtained for the previous day. The \( \delta \) parameter is searched for analogously to \( \gamma \). Parameter search is carried out by minimizing the mean square error of either the \( I \) series (further referred to as “SEIR SVR \( R(t) \) v1”), or of both \( I \) and \( R \) series (“SEIR SVR \( R(t) \) v2”).

4. Results

The forecasting accuracy is assessed by the mean absolute percentage error (MAPE) metric:

\[
MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|, \tag{3}
\]

where \( n \) is the number of days in the series, \( A_t \) and \( F_t \) are true and predicted values of the preprocessed time series respectively for the day \( t \).

In order to evaluate each model on different stages of epidemic, we split the time series into 5 equal parts (folds). When using machine learning models, the first part (Fold_0) is used as their training set, therefore, for the sake of consistency, the results of the models are only compared for the 4 folds (Fold_1 - Fold_4).

Table 1 contains the MAPE metric of the forecasting models on different folds. As it follows from the table data, machine learning methods give poor accuracy, essentially

\(^\#\) https://yandex.ru/covid19/stat?utm_source=main_graph&utm_source=main_notif&geoId=225 (in Russian)

\(^\dagger\dagger\) https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
REFERENCES

Table 1. Prediction error (by the MAPE metric) of different models on the 4 folds of the series of confirmed cases. "History days" denote the length of past series a model uses.

| Model                | History days | Fold_1 | Fold_2 | Fold_3 | Fold_4 | Mean |
|----------------------|--------------|--------|--------|--------|--------|------|
| Dummy                | all          | 28     | 16     | 12     | 28     | 21   |
| Smoothing all        | 16           | 12     | 28     | 21     | 20     |
| AutoARIMA            | 14           | 432    | 81     | 23     | 18     | 139  |
| LR                   | 14           | 337    | 18     | 45     | 22     | 113  |
| LinearSVR            | 14           | 354    | 73     | 44     | 20     | 123  |
| RandomForest         | 14           | 34     | 38     | 29     | 33     | 34   |
| GBR                  | 14           | 22     | 23     | 10     | 38     | 23   |
| SIR                  | 14           | 16     | 28     | 18     | 18     | 20   |
| SIR,expert           | all          | 34     | 38     | 29     | 33     | 34   |
| SEIR,expert          | all          | 22     | 23     | 10     | 38     | 23   |
| SIR,SVR R(t)         | 14           | 16     | 28     | 18     | 18     | 20   |
| SIR,SVR R(t) v1      | 14           | 13     | 18     | 14     | 21     | 17   |
| SIR,SVR R(t) v2      | 14           | 13     | 19     | 15     | 18     | 16   |

worse than that of the Dummy model. Results higher than the Dummy model are achieved by the following models: Smoothing, AutoARIMA, SEIR,expert, and SEIR and SIR models with regression model for $R_0(t)$.

5. Conclusion

The main conclusion of the obtained results is that the most effective instrument for predicting the dynamics of the coronavirus epidemic among basic methods is SEIR with regression model for $R_0(t)$. In future work, we plan to use this approach with more advanced models for predicting time-dependent parameters of the SEIR and SIR models.

Acknowledgements

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] Roda W C, Varughese M B, Han D and Li M Y 2020 Infectious Disease Modelling 5 271 – 281 ISSN 2468-0427 URL http://www.sciencedirect.com/science/article/pii/S2468042720300075
REFERENCES

[2] Hamzah F B, Lau C, Nazri H, Ligot D, Lee G, Tan C, Shaib M, Zaidon U, Abdullah A, Chung M et al. 2020 Bull World Health Organ 1 32

[3] Wang P, Zheng X, Ai G, Liu D and Zhu B 2020 Chaos, Solitons & Fractals 140 110214 ISSN 0960-0779 URL http://www.sciencedirect.com/science/article/pii/S096007792030610X

[4] Rustam F, Reshi A A, Mehmood A, Ullah S, On B, Aslam W and Choi G S 2020 IEEE Access 8 101489–101499 ISSN 2169-3536

[5] Anastassopoulou C, Russo L, Tsakris A and Siettos C 2020 PLOS ONE 15 1–21 URL https://doi.org/10.1371/journal.pone.0230405

[6] Shvetsova O, Adeel A B, Catalano M, Catalano O, Giannelli F, Muftuoglu E, Riggs T, Sezgin M H, Tahir N, VanDusky-Allen J, Zhao T and Zhirmov A 2020 Institutional origins of protective COVID-19 policies dataset URL https://orb.binghamton.edu/working_paper_series/7/

[7] Hale T, Angrist N, Cameron-Blake E, Hallas L, Kira B, Majumdar S, Petherick A, Phillips T, Tatlow H and Webster S 2020 Oxford COVID-19 government response tracker URL https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker

[8] Derval G, François-Lavet V and Schaus P 2020 Nowcasting COVID-19 hospitalizations using Google Trends and LSTM URL https://crcs.seas.harvard.edu/files/crcs/files/ai4sg_2020_paper_57.pdf

[9] Qazi U, Imran M and Ofli F 2020 SIGSPATIAL Special 12 6–15

[10] Yang Z, Zeng Z, Wang K, Wong S S, Liang W, Zanin M, Liu P, Cao X, Gao Z, Mai Z et al. 2020 Journal of Thoracic Disease 12 165 URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7139011/

[11] İsmail Kirbaş, Sözen A, Tuncer A D and Şinasi Kazancıoğlu F 2020 Chaos, Solitons & Fractals 138 110015 ISSN 0960-0779 URL https://www.sciencedirect.com/science/article/pii/S0960077920304136

[12] Petropoulos F and Makridakis S 2020 PLOS ONE 15 1–8 URL https://doi.org/10.1371/journal.pone.0231236

[13] Chaurasia V and Pal S 2020 Research on Biomedical Engineering 1–13 URL https://link.springer.com/article/10.1007/s42600-020-00105-4

[14] Grasselli G, Pesenti A and Cecconi M 2020 JAMA 323 1545–1546 ISSN 0098-7484 URL https://jamanetwork.com/journals/jama/articlepdf/2763188/jama_grasselli_2020_vp_200049_1597263971.14383.pdf

[15] Wu J T, Leung K and Leung G M 2020 The Lancet 395 689–697

[16] Piovella N 2020 Chaos, Solitons & Fractals 140 110243 ISSN 0960-0779 URL http://www.sciencedirect.com/science/article/pii/S0960077920306391

[17] Kröger M and Schlickeiser R 2020 Journal of Physics A: Mathematical and Theoretical 53 505601 URL https://doi.org/10.1088/1751-8121/abc65d
REFERENCES

[18] Prodanov D 2020 arXiv preprint arXiv:2010.07000

[19] Overton C E, Stage H B, Ahmad S, Curran-Sebastian J, Dark P, Das R, Fearon E, Felton T, Fyles M, Gent N, Hall I, House T, Lewkowicz H, Pang X, Pellis L, Sawko R, Ustianowski A, Vekaria B and Webb L 2020 Infectious Disease Modelling 5 409–441 ISSN 2468-0427 URL https://www.sciencedirect.com/science/article/pii/S2468042720300245

[20] Zeroual A, Harrou F, Dairi A and Sun Y 2020 Chaos, Solitons & Fractals 140 110121 ISSN 0960-0779 URL http://www.sciencedirect.com/science/article/pii/S096007792030518X

[21] Shastri S, Singh K, Kumar S, Kour P and Mansotra V 2020 Chaos, Solitons & Fractals 140 110227 ISSN 0960-0779 URL http://www.sciencedirect.com/science/article/pii/S0960077920306238

[22] Rodriguez A, Tabassum A, Cui J, Xie J, Ho J, Agarwal P, Adhikari B and Prakash B A 2020 medRxiv (Preprint https://www.medrxiv.org/content/early/2020/09/29/2020.09.28.20203109.full.pdf) URL https://www.medrxiv.org/content/early/2020/09/29/2020.09.28.20203109

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E 2011 Journal of Machine Learning Research 12 2825–2830

[24] Kudryashov N A, Chmykhov M A and Vigdorowitsch M 2021 Applied Mathematical Modelling 90 466 – 473 ISSN 0307-904X URL http://www.sciencedirect.com/science/article/pii/S0307904X20304984

[25] Kermack W O and McKendrick A G 1927 Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115 700–721