Modeling of the COVID-19 Epidemic in the Russian regions Based on Deep Learning

Olga Krivorotko
Sobolev Institute of Mathematics SB RAS,
Novosibirsk, Russia
Moscow Institute of Physics and Technology,
Dolgoprudny, Russia
0000-0003-0125-4988

Nikolay Zyatkov
Sobolev Institute of Mathematics SB RAS,
Novosibirsk, Russia
0000-0001-5847-4194

***Abstract*—The neural network of COVID-19 5 days forecasting in Russian Federation region based on epidemic and social data from 2020 to 2023 is constructed and analyzed. The structure of neural network consists in recurrent and full-connected layers. In addition to training the neural network, its hyperparameters were optimized, such as the optimal number of neurons in each layer, regularization parameters, and optimizer parameters. It is shown that the mean squared error on the test period from 07.2022 to 05.2023 is approximately 5% for new diagnosed of COVID-19 and hospitalized ones in Moscow, Saint Petersburg and Novosibirsk region. The proposed approach makes it possible to refine mathematical models in epidemiology.**

Keywords— epidemic, data processing, machine learning, deep learning, LSTM, short-term forecasting

I. Introduction

The COVID-19 pandemic became a reason to construct new mathematical models in epidemiology. Classical approaches such as SIR-type models [1]–[3] and agent-based models [4]–[7] should be modified taking into account socio-economic processes that influence on the epidemic situation as well as region features (medical system, ecology, population ages and so on) and population behavior for better forecasting. These models are characterized by their coefficients (infection transmission, reproduction number, probability of hospitalized, mortality rate, etc.) that should be identified using epidemic data (inverse problem) [8], that could contain gaps and noise. The inverse problems are ill-posed and its numerical solution should be regularized [9]. The regularization algorithm and mathematical models should be flexible to new strains of infectious disease, restrictive measures in the region and other seasonable epidemy for more suitable forecasting.

New mean-field game (MFG) models of epidemic propagation use social behavior of population and could get forecasting maps closer to the reality [10], [11]. But the numerical solution of the MFG problem is usually unstable for a big-time modeling as well as the non-smooth data.

Machine learning is one of the most promising approaches for modeling and forecasting the epidemic situation using the world-wide statistics.

Deep learning approaches, such as data driven as well as physics informed neural networks (PINNs) are widely used in different applications as well as in epidemiology. In paper [12] the recurrent neural networks for disease prediction are developed for determination the future flues using real flu's data. In paper [13] authors apply two approaches of deep learning: support vector regression and long short-term memory networks usually called LSTM, are a special kind of recurrent neural networks (RNN), capable of learning long-term dependencies. Simulation results show that LSTM provides more realistic results in the Indian Scenario. The combination of different LSTM and residual neural networks are capable of learning dynamical systems and shows that the vaccination rate with higher efficacy lowers the infectiousness and basic reproduction number based on COVID-19 data for the state of Tennessee in USA [14].

In papers [15]–[17] PINNs that the alternative to traditional numerical methods for solving system of differential equations used to describe dynamics of infectious diseases are proposed. PINNs are included the classical SIR models described by systems of ordinary differential equations (ODEs). The system of ODEs and its time derivative are included in the residual loss function of PINNs in addition to the data error between the current network output and the time series data of the compartment sizes. The results show that the proposed PINNs approach is a reliable candidate for both solving such systems and for helping identify important parameters that control the disease dynamics.

This paper proposes short-term forecasting COVID-19 epidemic in the Russian Federation regions, i.e. expected number of new diagnosed and hospitalized cases of COVID-19 in Moscow, Saint-Petersburg and Novosibirsk region, using deep learning approach for epidemic and social data of considering regions and world-wide since 2020 (see Section II). The neural network based on recurrent and full-connected layers is described Section III. Deep optimization is applied on epidemic and social data by dividing them into training, validation, and test samples, as well as the use of the roll forward cross-validation technique (see Section III-B). The results of forecasting for considered regions are analyzed in Section IV.

II. Feature Engineering and Data Processing

The infectious disease is described by statistical information of epidemic, social and economic situation in the region. Every outbreak of new epidemic is characterized by changes in epidemiological and socioeconomic parameters. The statistical information should be prepared and proceeded for usage in machine learning. In the next sections epidemic and social data are described and analyzed.

A. Epidemic Data

The epidemic data of COVID-19 propagation in Moscow, Saint Petersburg and Novosibirsk region was collected from open sources (websites) such as number of tested and diagnosed people using polymerase chain reaction (PCR) analysis, recovered, hospitalized, critical and mortality people, self-isolation index by Yandex and the rate of

antibody IgG of investigated population. The detailed burials with COVID-19 diagnose are used for Novosibirsk region as well [3]. Also, we use world-wide and Europe epidemic data about new diagnosed and mortality cases of COVID-19. All COVID-19 epidemic data for considered regions are collected at website https://covid19-modeling.ru/data.

The most of data are expressed as time series and processed as follows:

- Filling in missing values in the data using linear interpolation.
- Replacement of extreme values (negative, too large, etc.) by interpolation (elimination of outliers in the data).
- Data smoothing (exponential moving average). The noisier data were smoothed with a 14-day exponential moving average, the less noisy data were smoothed with a 7-day exponential moving average.

The Fig. 1 demonstrates the epidemic data for Moscow that are used in feature engineering.

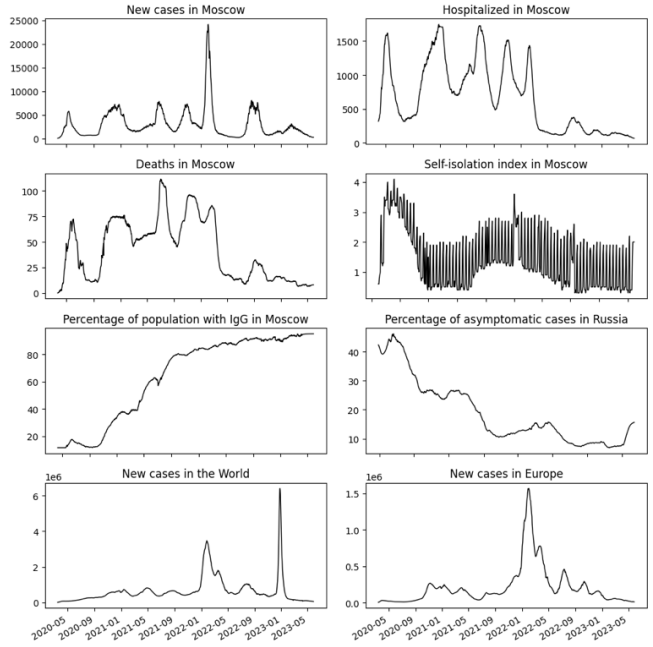

Fig. 1. Statistical data of COVID-19 propagation in Moscow, World and Europe from 05.2020 to 05.2023 after exponential moving average: new diagnosed and hospitalized people in Moscow (1 row), mortality and Yandex self-isolation index in Moscow (2 row), percentage of people with IgG antibodies in Moscow and asymptomatic cases in Russia (3 row) and new diagnosed in the World and Europe (4 row).

Note, that the emergence of new strains of SARS-CoV-2 (Delta, Omicron, Centaurus, Cerberus, etc.) is accompanied by an outbreak, making the data non-uniformly distributed. For normalization of time series (for example, data in the first row in Fig. 1) the logarithm transform is applied.

Additional data were generated to train the neural network: logarithmic increments for $\delta = 3$, $\delta = 7$ and $\delta = 14$ days of new diagnosed $f_1(t)$ and hospitalized $f_2(t)$ people in considered regions as follows:

$$f_i^{\delta}(t) = \ln(f_i(t) + 1) - \ln(f_i(t - \delta) + 1), \quad i = 1,2.$$

Fig. 2 shows the obtained time series (features) for new diagnosed and hospitalized cases due to COVID-19 in Moscow.

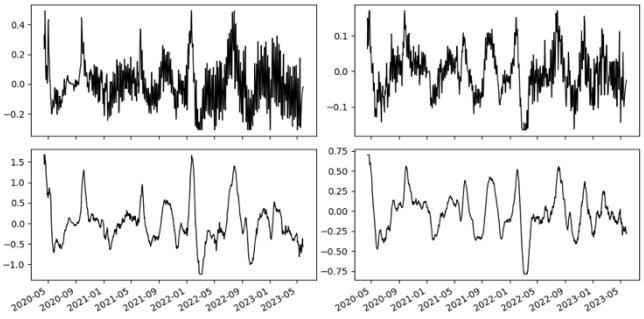

Fig. 2. The logarithm of new diagnosed $f_1^{\delta}(t)$ (1 column) and hospitalized $f_2^{\delta}(t)$ (2 column) people with COVID-19 in Moscow with $\delta = 3$ (1 row) and $\delta = 14$ days (2 row).

B. Social Data

In additional to the epidemic data, we use restriction measures (social data) that have big influence on the epidemic propagation [18]. Such social data (masks wearing, open/close workplaces, schools, and public places) are available at regional websites. The main dates and measures in three regions are demonstrated in Table I. These measures were collected in binary time series using following rules:

- Holidays and weekends. Binary time series filled with values 0 or 1 (1 is a holiday or weekend,) 0 - other days.
- Seasonality. There are 7 binary time series. First row for Monday: 1 if the current day is Monday, 0 for the other days. Similarly for the other 6 days of the week.
- • Restrictions in the region (see Table I). There are two binary time series. For the first row: 1 means that restrictions are imposed on the current day, 0 - other days. For the second row 1 means that some restrictions are cancelled, 0 - other days.

TABLE I. THE MAIN RESTRICTION MEASURES IN MOSCOW, SAINT PETERSBURG AND NOVOSIBIRSK REGION FROM MARCH 2020 TO MARCH 2022

Restriction measures	**Moscow**	**Saint Petersburg**	**Novosibirsk region**
2020			
Closed of schools	16.03 – 12.04		18.03 – 01.04
Public places are closed, self-isolation and quarantine measures are controlled by government	28.03 – 31.05		
Mandatory wearing of masks	27.04 (to the March 2022)		
Restrictions access to public places, 30% of workplaces are switched to remote mode	13.11 – 15.01.2021	05.10 – 02.11	–
2021			
Public, workplaces and schools are temporary closed	12.06 – 20.06		–
Public places, 30% workplaces and schools are mandatory closed	28.10 – 07.11		

QR codes are applied in public places	08.11 – 31.12	–
2022		
All restriction measures are cancelled	14.03	

III. Neural Network Model

Supervised machine learning methods are a class of mathematical methods that are characterized not by a direct problem solution, but by training and identifying empirical patterns on a set of N experiments (historical data) $X = \{x_i\}$, $i = 1, \dots, N$, with previously known results $Y = \{y_j\}$. In this paper we use $j = 1,2$ for Y set where y_1 and y_2 represents the new diagnosed and hospitalized people respectively. Each object x_i from X is characterized by M features, i.e. $x_i = \{x_{i1}, . x_{iM}\}$. In this paper x_i represents the part $(t_i - L, t_i)$, $i = 1, \dots, N$, of time series of statistical data (see Fig. 3). In this paper we use the follows $M = 20$ features for each region:

- daily diagnosed $f_1(t)$ and its logarithm functions $f_1^3(t)$, $f_1^7(t)$, $f_1^{14}(t)$,
- hospitalized $f_2(t)$ and its logarithm functions $f_2^3(t)$, $f_2^7(t)$, $f_2^{14}(t)$,
- the logarithm functions of daily mortality for $\delta = 3$ and $\delta = 7$ days,
- the percentage of people with IgG antibodies and its logarithm functions for $\delta = 3$, $\delta = 7$ and $\delta = 14$ days,
- Yandex self-isolation index

and for all regions

- daily diagnosed $f_1(t)$ in the World and Europe and its logarithm functions $f_1^3(t)$ and $f_1^7(t)$,
- the percentage of asymptomatic cases in Russia and the logarithm function with $\delta = 7$.

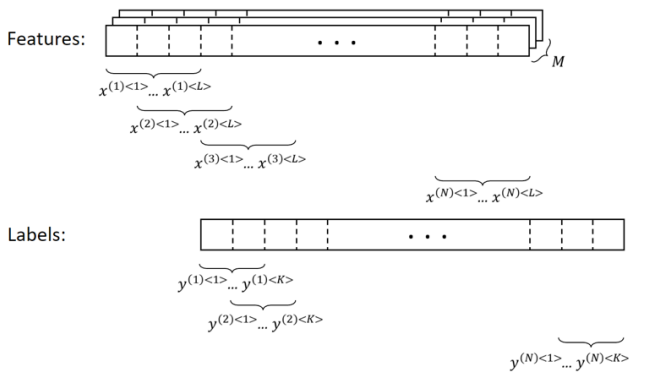

Fig. 3. The scheme of data processing in neural network.

For forecasting by the neural network, we chose two target functions:

$$F_i(t+k) = \ln(f_i(t+k)+1) - \ln(\mathrm{SMA}(f_i(t),7)+1),$$

where $\mathrm{SMA}(f_i(t),7)$ is 7-days smooth moving average of f_i.

A. Model Structure

The structure of neural network is based on the data type (time series) with long short-term memory and described by recurrent and full connected layers. The input layer consists in $M = 20$ time series with time window L days (see Fig. 3). On the next step input data x_{im}, $i = 1, \dots, N$, $m = 1, \dots, M$, transform through the two recurrent layers of LSTM type with dropout regularization [19-20]. Such layers could analyze features and time characteristics in data.

The output data of the second LSTM layer are processed by batch normalization [21]. This data is transformed using full connected layer with ReLU activation function.

The neural network has two outputs. The first and the second outputs form the 5 days forecasting of time series of sequences of new diagnosed $F_1(t+k)$, and hospitalized $F_2(t+k)$, $k = 1, \dots, 5$, people of COVID-19. The forecasting process is based on full connected layer.

B. Parameter Optimization and Cross-Validation

We divided the data into training (from 03.2020 to 01.2022), validation (from 01.2022 to 07.2022), and test (from 07.2022 to 05.2023 for Moscow and Saint Peterburg and from 07.2022 to 12.2022 for Novosibirsk region) periods. This breakdown was chosen for the following reasons: the training data include Alpha, Beta, Delta strains of SARS-CoV-2, the validation sample has Omicron strain, and the test sample has Centaur strain for investigation.

On the training sample we optimized the weights of the neural network using stochastic gradient method with backpropagation and regularization approaches.

Using validation sampling we optimized the neural network's hyperparameters by minimizing the mean square error of the neural network prediction with real data on the validation period. After obtaining the optimal set of hyperparameters (see Table II) the neural network was trained for the period from 03.2020 to 07.2022 and the forecast for the test period was performed.

TABLE II. OPTIMAL NEURAL NETWORK HYPERPARAMETERS FOR MOSCOW, ST. PETERSBURG AND NOVOSIBIRSK REGION

Hyperparameter	**Moscow**	**St. Petersburg**	**Novosibirsk region**
Number of forecast days	$K = 5$		
Time window look back period (number of days used to get the forecast)	$L = 5$ days	$L = 17$ days	$L = 5$ days
Number of training features	$M = 20$		
Optimization method	Adam	RMSprop	RMSprop
Descent coefficient in the gradient method (learning rate)	0.0016	0.015	0.006
Batch size (number of training samples for the gradient descent method)	35	5	45
Number of epochs (number of complete passes through the training dataset)	100		
LSTM, layer 1: dimension of the output	12	24	8
LSTM, layer 2: dimension of the output	6	17	17
LSTM layers: dropout regularization ratio	0.03	0.04	0.03
LSTM layers: recurrent dropout regularization	0.19	0.45	0.66
Dense layer: dimension of the output	11	49	13

IV. Forecasting Results

The forecasting results at day $t+5$ for t from the test period for new diagnosed of COVID-19 (Fig. 4) and hospitalized with COVID-19 (Fig. 5) are presented for Moscow (1st row), Saint Petersburg (2nd row) and Novosibirsk region (3rd row). The scenario of COVID-19 propagation (red line for left pictures) is compared with the real data at the day $t+5$ (black dots for left pictures). The middle pictures in Fig. 4 and 5 describe the difference between predicted (blue dots) and true (red line) values, i.e. if blue dots are far from the red line then the prediction is worse. The histogram of the model prediction error (right column of Fig. 4, 5) calculated as the difference between the true values and the obtained predictions. The mean absolute error (MAE) on the test period is less then 5% (Table III). The higher accuracy for Novosibirsk region relates to more detailed data processing for the neural network [2], [3], [22].

TABLE III. MAE OF THE NEURAL NETWORK ON TEST PERIOD (PEOPLE)

Output data	Moscow	Saint Petersburg	Novosibirsk region
New diagnosed f_1	342	240	58
Hospitalized f_2	10	11	35

Conclusion

The constructed neural network, based on the processing of real data on the spread of COVID-19 in the regions of the Russian Federation, showed high accuracy of 5-day short-term forecasts for new diagnosed and hospitalized people on the test period from July 2022 with real data. The presented approach is suitable for modeling of COVID-19 propagation in any Russian region for any period.

The proposed approach makes it possible to refine other mathematical models of the spread of infectious diseases, such as compartmental or agent-based models [3], [6]. It is possible to combine deep learning with classical models for learning of new disease characteristics and better forecasting.

Acknowledgment

The research is supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, project No. 0714-2020-0005.

References

[1] W.O. Kermack, A.G. McKendrick. “A contribution to the mathematical theory of epidemics,” Proceedings of the Royal Society, vol. 115, pp. 700–721, August 1927.

[2] O.I. Krivorotko, S.I. Kabanikhin, N.Yu. Zyatkov, A.Yu. Prikhodko, N.M. Prokhoshin, M.A. Shishlenin. “Mathematical modeling and forecasting of COVID-19 in Moscow and Novosibirsk region,” Num. Anal. Appl., vol. 13, no. 4, pp. 332–348, October 2020.

[3] O.I. Krivorotko, N.Y. Zyatkov. “Data-driven regularization of inverse problem for SEIR-HCD model of COVID-19 propagation in Novosibirsk region,” Eurasian Journal of Mathematical and Computer Applications, vol. 10, iss. 1, pp. 51–68, 2022.

[4] A.I. Vlad, T.E. Sannikova, A.A. Romanyukha. “Transmission of acute respiratory infections in a city: agent-based approach,” Mathematical Biology and Bioinformatics, vol. 15, no 2, pp. 338–356, December 2020.

[5] C.C. Kerr, R.M. Stuart, D. Mistry et al. “Covasim: An agent-based model of COVID-19 dynamics and interventions,” PLoS Comput. Biol., vol. 17, no. 7, ID e1009149, July 2021.

[6] O. Krivorotko, M. Sosnovskaia, I. Vashchenko, C. Kerr, D. Lesnic. “Agent-based modeling of COVID-19 outbreaks for New York state and UK: parameter identification algorithm,” Infectious Disease Modelling, vol. 7, pp. 30–44, March 2022.

[7] O.I. Krivorotko, S.I. Kabanikhin, M.A. Bektemesov, M.I. Sosnovskaya, A.V. Neverov. “Simulation of COVID-19 propagation scenarios in the Republic of Kazakhstan based on regularization of agent model,” Diskretn. Anal. Issled. Oper., vol. 30, no 1, pp. 40–66, May 2023.

[8] S.I. Kabanikhin, Inverse and Ill-Posed Problems, Walter de Gruyter, 2011.

[9] B. Kaltenbacher, A. Neubauer, O. Scherzer, Iterative Regularization Methods for Nonlinear Ill-Posed Problems, Walter de Gruyter, 2008.

[10] V. Petrakova, O. Krivorotko. “Mean field game for modeling of COVID-19 spread,” Journal of Mathematical Analysis and Application, vol. 514, ID 126271, October 2022.

[11] Y.T. Chow, S.W. Fung, S. Liu, L. Nurbekyan, S. Osher. “A numerical algorithm for inverse problem from partial boundary measurement arising from mean field game problem (Version 1),” unpublished.

[12] A. Puleio. “Recurrent neural network ensemble, a new instrument for the prediction of infectious diseases,” The European Physical Journal Plus, vol. 136, no 3, pp. 319–334, March 2021.

[13] S. Dash, S. Chakravarty, S.N. Mohanty, C.R. Pattanaik, S. Jain. “A Deep Learning Method to Forecast COVID-19 Outbreak,” New Gener. Comput., vol. 39, pp. 515–539, July 2021.

[14] T.K. Torku, A.Q.M. Khaliq, K.M. Furati. “Deep-Data-Driven Neural Networks for COVID-19 Vaccine Efficacy,” Epidemiologia, vol. 2, iss. 4, pp. 564–586, November 2021.

[15] L. Nguyen, M. Raissi, P. Seshaiyer. “Modeling, Analysis and Physics Informed Neural Network approaches for studying the dynamics of COVID-19 involving human-human and human-pathogen interaction,” Computational and Mathematical Biophysics, vol. 10, iss. 1, pp. 1–17, February 2022.

[16] J. Malinzi, S. Gwebu, S. Motsa. “Determining COVID-19 Dynamics Using Physics Informed Neural Networks,” Axioms, vol. 11, iss. 3, pp. 121, March 2022.

[17] S. Berkhahn, M. Ehrhardt. “A physics-informed neural network to model COVID-19 infection and hospitalization scenarios,” Advances in Continuous and Discrete Models, vol. 2022, iss. 1, pp. 61–88, October 2022.

[18] Q. Cao, B. Heydari. “Micro-level social structures and the success of COVID-19 national policies,” Nat. Comput. Sci., vol. 2, pp. 595–604, September 2022.

[19] S. Hochreiter, J. Schmidhuber. “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, November 1997.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov. “Dropout: A simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] S. Ioffe, C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” International Conference on Machine Learning, vol. 37, pp. 448–456, July 2015.

[22] O. Krivorotko, M. Sosnovskaia, S. Kabanikhin. “Agent-based mathematical model of COVID-19 spread in Novosibirsk region: Identifiability, optimization and forecasting,” Journal of Inverse and Ill-posed Problems, April 2023 (online).

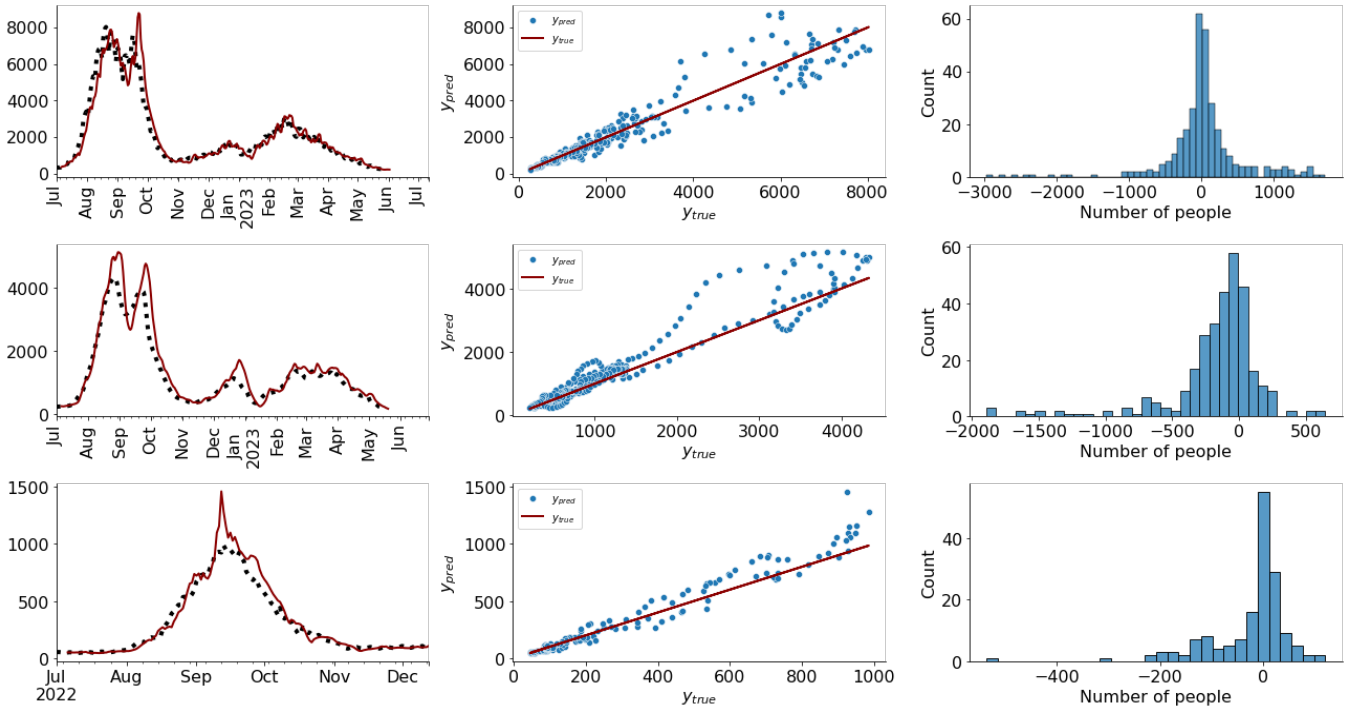

Fig. 4. The forecasting of new diagnosed people (left, red curve) in Moscow (1st row), Saint Petersburg (2nd row) and Novosibirsk region (3rd row) to 5 days based on test data (black dots). The central picture is the comparison of prediction data (blue dots) with true (red line), the right picture is the error histogram.

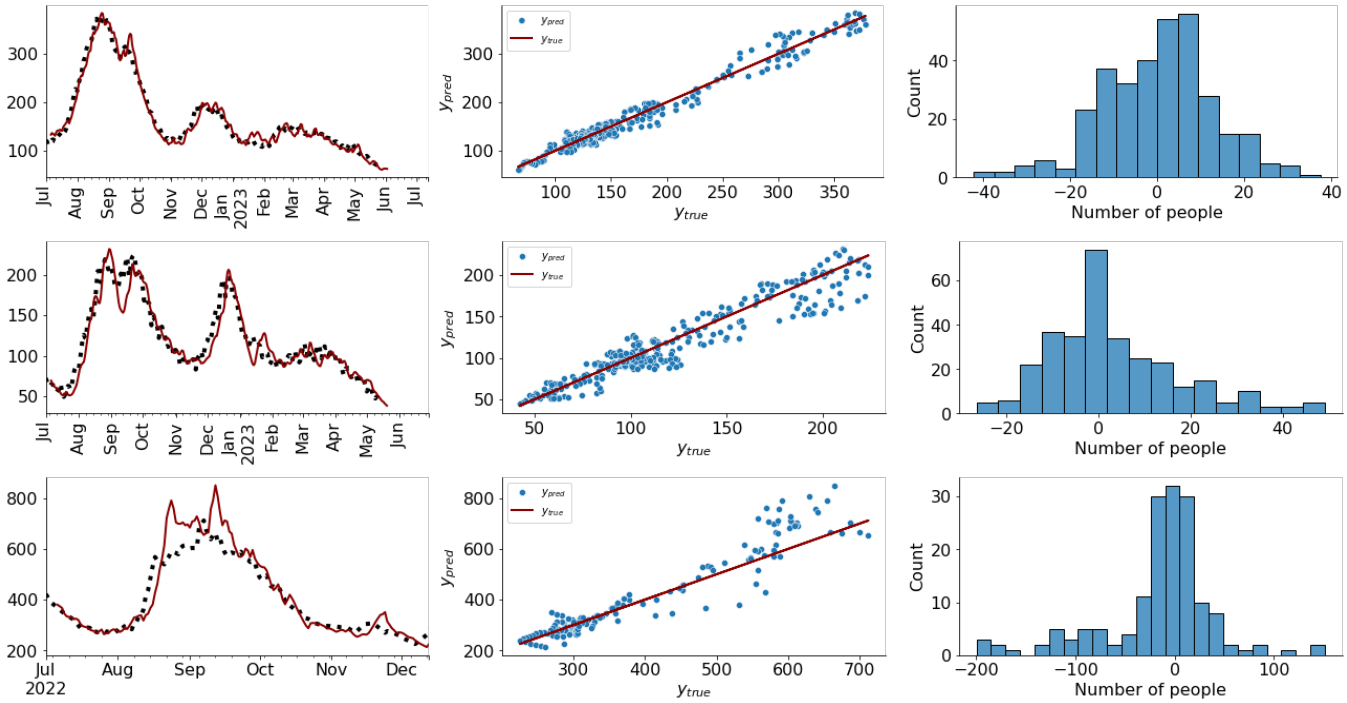

Fig. 5. The forecasting of hospitalized people (left, red curve) in Moscow (1st row), Saint Petersburg (2nd row) and Novosibirsk region (3rd row) to 5 days based on test data (black dots). The central picture is the comparison prediction data (blue dots) with true (red line), the right picture is the error histogram.