Start of Epidemy in a City: Short-Term Forecast of Covid-19 with GMDH-Based Algorithms and Official Medical Statistics

2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT) | 978-1-7281-7443-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSIT49958.2020.9322033

Anna Boldyreva
Moscow Institute of Physics and Technology
Moscow, Russia
anna.boldyreva@phystech.edu

Mikhail Alexandrov
RANEPA
Moscow, Russia
Autonomous University of Barcelona
Barcelona, Spain
malexandrov@mail.ru

Olexiy Koshulko
Glushkov Institute of Cybernetics
Kyiv, Ukraine
koshulko@gmail.com

Svetlana Popova
Technological University Dublin
Dublin, Ireland
spbu.svp@gmail.com

***Abstract*—The sudden onset and quick development of an unknown epidemic may lead to tragic consequences: panic of population due to victims and unpreparedness of authorities for effectively help to population. These circumstances define extremely high requirements to the tools for short-term operational forecast. Namely, such tools should provide reliable results when model of phenomenon is unknown (factors of disease spreading) and data are limited (time series of observations). GMDH-based algorithms just meet these requirements unlike modern differential or advanced statistical models. In this study we test different algorithms from GMDH Shell platform on the example of Covid-19 epidemic in Moscow during the period March 30-April 12, 2020. The forecast horizon is from 1 to 7 days, the initial information is only the official dynamics of diseased patients. Our model is autoregression with variables of different powers. The results of forecast are compared with the accuracy of popular statistical autoregression using exponential smoothing with trend. We suppose that the proposed approach will be useful for short-term forecast at the start of epidemic due to its simplicity and reliability.**

Keywords— epidemic, Covid-19, short-term forecast, GMDH, GMDH Shell

I. INTRODUCTION

A. Motivation

The COVID-19 epidemic was declared as a pandemic by the World Health Organization at March 11, 2020 [1]. The first patient died in Moscow at March 19, and we had 1226 diseased patients, 11 the dead ones and 28 the recovered ones by March 30. This day in Moscow, an isolation regime was introduced.

The first subjective predictions of doctors-infectologists about the development of epidemic in Moscow appeared in mass media at March 27 [2]. They predicted a peak in 1-2 weeks just before the middle of April. Later it proved that it was only the wish but not reality. Meantime the quantity of diseased persons was growing rapidly and just that time one applied mathematician performed modeling development of epidemic in Moscow with and without isolation regime. This mathematician used well-known differential model of epidemic SEIR modified by Richard Neyer. The Neyer's program is in open access for anybody [3]. The results were published at the mentioned date March 30 and then it was repeated on several sites during week [4]. These results showed that without the strong isolation regime, that is quarantine, Moscow would have more than 50000 heavy diseased patients at the beginning of May and more than 100000 dead patients during the epidemic. Moscow government knew these results and began to consider the regime of quarantine at April 6. The quarantine was introduced in a week on April 13.

April 1, we received the request from the administration of the Russian Presidential Academy of National Economy and Public Administration (RANEPA) to propose any tool (i.e. some method or even technology) which could provide reliable short-term forecasts of spreading epidemic in Moscow. The difficulties consisted in 2 circumstances:

1. The model of phenomenon was unknown (factors of disease spreading);
2. Data were limited (time series of observations).

The first circumstance means that we have no any a priori information to use it in predictive model. The second circumstance means that we can't build a model for long-term forecast..

That moment Russian segment of Internet was full of different opinions concerning both predictions themselves and tools for these predictions. It looked as an Information warfare that could be the subject of research like [5, 6].

We stopped on the technology GMDH, which was known us by our previous experience. The reason is: this technology allows to build reliable models under circumstances mentioned above unlike modern differential or advanced statistical models. GMDH-based models have optimal complexity providing a balance between the accuracy of complex models and the noise immunity of simple models [7, 8]. This property is very important at the start of the epidemic.

B. Problem Setting

To build the models, we use 4 algorithms from the GMDH Shell platform [9]: combinatorial (Combi), neural-

like with relaxation (RIA, Neuro), stepwise with additions (Forward), and stepwise mixed (Mixed). The first two are the traditional algorithms of GMDH. The second two are the algorithms of regression analysis but they use model generation based on the GMDH principles.

In our study we use the official statistics of diseased patients in Moscow during the period March 30 - April 12. The training interval is only one week March 30 - April 5. The forecast interval is the next week April 6-12.

The forecast horizon varies from 1 to 7 days. Here we do not build a forecast model for 7 days with interpolation of the result on intermediate 1, 2, ... 6 days. On the contrary, we build models separately for each day of the week.

The quality of all the results is compared with the forecast of the traditional exponential smoothing with a linear trend. This model is also titled as double exponential smoothing [10]. Its advantages are well-known: simplicity of tuning and adaptability to data. The forecast with this popular model is considered as a Baseline.

Our data and results refer to Moscow. However, we suppose that our study will be useful for other local homogeneous territories with a large number of dwellers. Here the term "homogeneous" means the equal population density within a given territory, similar behaviour of dwellers regarding contacts and isolation, and equal readiness of healthcare system to aid the population. For this reason, we used the word "city" in the title of the paper.

C. Related work

There are no any specific models and corresponding algorithms related to short-term forecast of epidemic without any substantial information about the process and/or large number of data. Here, one can use well-known econometric models, in particularly, the mentioned popular model of double exponential smoothing. It has a special name as Holt-Winters model [10]. This model has 2 parameters reflecting inertial properties of time series under consideration, which can be tuned to provide the best result. We use this approach in our research as a baseline. The practice proposes many other smarter models for successful short-term forecast but they can be built on enough long time series [11].

In the framework of hypothesis about an ideal development of epidemic one can use the logistic equation [12]. Such models are commonly used to get a qualitative forecast of the development of epidemic with respect to diseased patients and to evaluate approximately the moment of reaching peak (or vicinity of the peak). The equation depends on 3 parameters: initial number of diseased patients, maximum possible number of the diseased ones, and velocity of growth of diseased patients. These parameters can be easy recovered on existing experimental data with the least square method. But the reality is slightly far from the ideal model and short-term forecast of diseased patients with logistic equation proves far from the real process. We could make sure in such a result on the stage of preliminary experiments.

Apart from the simplest models described above there is a large class of theoretical models of mathematical epidemiology [13]. They include deterministic populational models, stochastic populational models, and agent-oriented models. All these models need many a priori given parameters, which are unknown in advance or are known very approximately. So, these models are good for long-term qualitative forecast but not for high accuracy short-term forecast.

II. Data and Tools

A. Dynamics of Diseased, Dead, and Recovered Patients

In our research we use only official medical statistics about diseased patients. It covers the period March 30 – April 12. The data are presented in Table 1 and on Fig. 1. We also show the dynamics of the diseased patients during the period March 30 – May 9 on Fig. 2 to demonstrate that we are still far from the peak of epidemic. It's easy to see that in a month after our experiments the number of diseased patients has grown approximately 10 times.

B. Group Method of Data Handling and GMDH Shell

Group Method of Data Handling (GMDH) is a technology of machine learning (ML) for creating noise immunity models. The ideas and perspectives of GMDH are presented in many publications; see, for example, [7] Theoretical bases of GMDH are described the most completely in the well-known paper [8]. GMDH does not orient on certain class of functions, but the most popular GMDH-based tools use polynomial functions of many variables [9,14] This fact has the simple explanation: any continuous functions of many variables on hypercube can be presented in the form of uniformly-convergent polynomial series.

TABLE I. Official Medical Data

Data	Diseased	Dead	Recovered
30.03.2020	1226	11	28
31.03.2020	1613	11	70
01.04.2020	1880	16	115
02.04.2020	2475	19	140
03.04.2020	2923	20	168
04.04.2020	3357	27	194
05.04.2020	3893	29	198
06.04.2020	4494	29	206
07.04.2020	5181	31	222
08.04.2020	5841	31	270
09.04.2020	6698	38	313
10.04.2020	7822	50	350
11.04.2020	8852	58	499
12.04.2020	10158	72	687

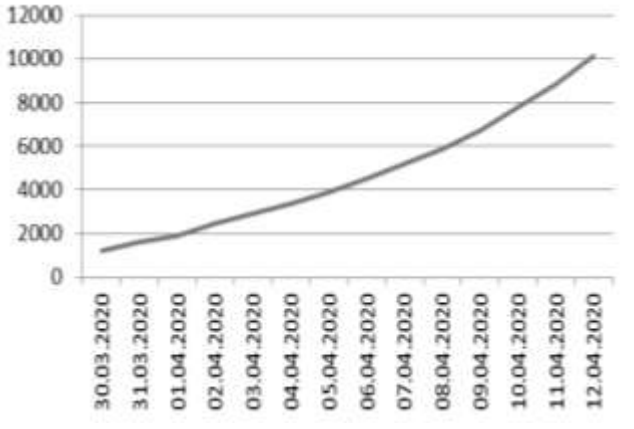

Fig. 1. Dynamics of diseased patients, 30.03-12.04 (persons/days)

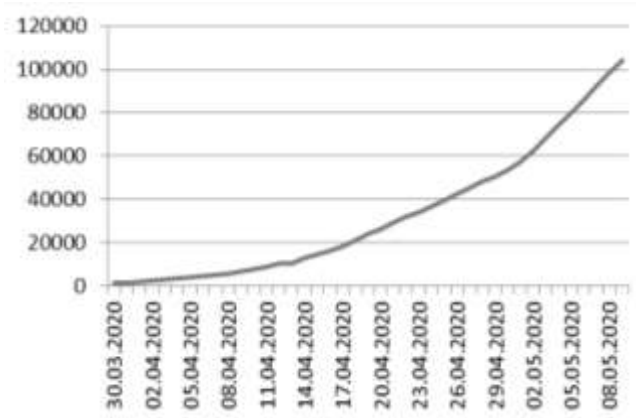

Fig. 2. Dynamics of diseased patients, 30.03-09.05 (persons/days)

GMDH as ML has numerous natural-scientific, technical, and humanitarian applications presented in papers and thematic monographs. For the first detail acquaintance of English-speaking readers we could recommend two useful books [15, 16].

In our experimental research we use algorithms from the available platform GMDH Shell, which include 3 modes: Extrapolation (Forecast), Approximation (Regression), and Classification [9]. The mode "Forecast" offers the following 4 algorithms:

Combi. It is a classical GMDH-based sort out algorithm, which considers all possible combinations of variables;

Neuro. It is also a GMDH-based neuro-similar relaxation algorithm, where generated variables are used together with the initial ones;

Forward. It is similar to the stepwise regression, where the procedure adds new member to a current model having tested it according principles of GMDH;

Mixed. It is similar to the stepwise regression, where the procedure may add successful members to a current model and also delete the unsuccessful ones from a current model having tested them according principles of GMDH.

A user has an opportunity to define and limit the class of polynomial models. For example, he/she can specify:

- Regression, autoregression, or hybrid model;
- Form of variables and the maximum number of members in a model being constructed (Combi, Forward, Mixed);
- Form of generative function and width of neuron layer (Neuro);
- Etc.

A user has possibility to apply different criteria for assessment of the quality of forecast as training-testing, k-fold cross validation, and also different measures of error. GMDH Shell before building a model analyzes the given data and proposes the best options. A user can agree or not and make his/her own choice.

III. Experiments

A. Data Preprocessing and Tuning

Preprocessing consists in transformation of time series to their logarithms. Such a procedure proves to be useful when we deal with time series of integral numbers having exponential or almost exponential growth or fall. It allows to get data with less variability (almost similar to linear trend) and therefore to build more accurate models. Our previous numerous experiments with GMDH Shell justify such a procedure. It should say that in the mentioned experiments we used along with logarithms also square roots and cubic roots. The further results were very close.

Tuning of algorithms:

- Combi and Forward use variable, their squares and their pairwise products;
- Neuro uses maximum 6 layers; its width is equal 5;
- Mixed uses variables, their products and division.

These options were proposed by the GMDH Shell supporting means and corrected. We used here some recommendations from [17].

The model quality is evaluated with 2-fold cross validation. It is the same as the symmetric criterion of regularity [7]. The errors are measured by mean absolute percentage error (MAPE). Time lag equals 2 weeks, which is equal to incubation period. But due to our short time series this period proved to be less.

Basic algorithm has two 2 parameters reflecting inertial properties of time series: one refers to random process (its average is equal zero) and the second one refers trend. Both parameters are tuned separately for each day of forecast. This operation was completed automatically by testing on the grid of the mentioned 2 parameters. Therefore, baseline proves to be relatively high.

B. Experiments

The experiments aim to determine the averaged error of forecast with respect to the whole week. Speaking "averaged error" we mean the average values of forecast errors for 1 day, 3 days, 5 days, and 7 days. It is necessary to make 3 steps to get these errors:

1. A model is built for 1-day forecast and then this model is used to make 1-day forecasts for each day April 6-12;
2. These operations are repeated for 3-days forecast, 5-days forecast, and 7-days forecast;
3. We calculate average MAPE for these 4x7=28 forecasts.

The results of calculations are shown in Table 2 for each algorithm including the Basic one.

TABLE II. Averaged Error of Forecast [%]

Algorithm	1 day	3 days	5 days	7 days
Combi	0,78	3,85	8,20	7,00
Neuro	1,45	5,85	6,10	5,73
Forward	0,78	3,85	8,20	7,75
Mixed	1.05	5,93	9,38	6,80
Basic	2,36	8,93	16,55	20,25

IV. Conclusion

A. Results

The results in graphical form are shown on Fig. 3. It is easy to see the following:

- All GMDH-based algorithms for all days of week show better results than the Basic algorithm, this advantage is approximately 2-3 times;
- Combi and Forward are the best algorithms for 1-day and 3-days forecasts, and Neuro is the best one for 5-days and 7-days forecasts;
- Combi and Forward demonstrate practically the same results for all days of week, it is the expected result because both algorithms have almost similar way of model generation;

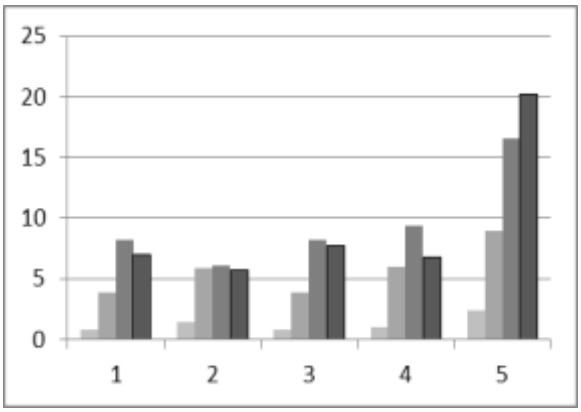

Fig. 3. Errors (MAPE %) of all algorithms for different days of week; here 1 - Combi, 2 - Neuro, 3 - Forward, 4 - Mixed, 5 - Basic

B. Future Work

In future we suppose:

- To consider forecast of dynamics of the recovered and dead patients using GMDH-based algorithms;
- To test GMDH-based algorithms not only at the beginning of the epidemic but also on other periods of epidemic development including the peak of its growth;
- To consider middle-term forecast with GMDH-algorithms including options of automatic switching of algorithm, the latter is similar to so-called intelligent modeling [18].

References

[1] World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means; https://time.com/5791661/who-coronavirus-pandemic-declaration/

[2] Predictions of Moscow infectologists [rus]; https://yandex.ru/turbo/s/mr-7.ru/articles/216465/

[3] Neyer's platform for COVID-19 modeling; https://covid19-scenarios.org/

[4] Modeling epidemic in Moscow on Neyer's software platform [rus]; https://meduza.io/feature/2020/03/30/v-moskve-vveli-zhestkie-karantinnye-mery-pohozhe-eto-pravilno-matematicheskaya-model-pokazyvaet-chto-inache-mogli-by-pogibnut-bolshe-100-tysyach-chelovek?utm_source=facebook.com&utm_medium=share_fb&utm_campaign=share&fbclid=IwAR2EfWfWOu2PLFrLPwIJl2NebZE00T0GZBdgtm_xP-n5vL6hC0H6mk-NnIw

[5] A. Petrov, O. Proncheva, "Propaganda battle with two-component agenda," In: Proc. of the MACSPro (Workshop 2019), Austria, CEUR, vol. 478, 2019, pp. 28–38; http://ceur-ws.org/vol-2478/

[6] P. Petrov, O. Proncheva, "Modeling position selection by individuals during informational warfare with a two-component agenda," In: Mathematical Models and Computer Simulations, vol.12, no.2, 2020, pp. 154–163.

[7] V. Stepashko, "Developments and prospects of GMDH-based inductive modeling," Advances in Intelligent Systems and Computing II; Springer, AISC book series, vol. 689, 2017, pp. 346–360.

[8] V. Stepashko, "Method of critical variances as analytical tool of theory of inductive modeling," J. Autom. Inf. Sci., vol. 40, no. 3, 2008, pp. 4–22.

[9] Platform GMDH Shell; http//www.gmdhshell.com

[10] C. Holt, "Forecasting trends and seasonals by exponentially weighted averages," Intern. J. of Forecasting, vol. 20, no. 1, 2004, pp. 5–10.

[11] Y. Lukashin, Adaptive methods for short-term forecast of time series. Textbook, M: Finance&Statistics, 2003 [rus]; https://www.studmed.ru/lukashin-yup-adaptivnye-metody-kratkosrochnogo-prognozirovaniya-vremennyh-ryadov_39c4e89988b.html

[12] Bygu's learning applications: logistic function; https://byjus.com/maths/logistic-function/

[13] V. Leonenko, Mathematic epidemiology, Handbook, St.Petersburg: ITMO University, 2018 [rus]; https: // books.ifmo.ru/file/pdf/2383.pdf

[14] Resource GMDH in IRTC ITS of the NAS of Ukraine; http//mgua.irtc.org.ua/

[15] S.J. Farlow, Self-Organizing methods in modeling: GMDH type algorithms, Statistics: A series of textbooks and monographs, Book 54, 1st edition. Marcel Decker Inc., New York, Basel, 1984.

[16] H.R. Madala, A.G. Ivakhnenko, Inductive learning algorithms for complex systems modelling, CRC Press, New York, 1994.

[17] O. Koshulko, G. Koshulko, "Validation strategy selection in combinatorial and multilayered iterative GMDH algorithms," Proc. 4th Intern. Workshop on Inductive Modeling (IWIM–2011), IRTC ITS of the NAS of Ukraine, Kyiv, 2011, pp. 51–54.

[18] V. Stepashko, "On the self-organizing inductive-based intelligent modeling," Advances in Intelligent Systems and Computing III. Springer, AISC book series, vol. 871, 2018, pp. 433–448; DOI: 10.1007/978-3-030-01069-0_31.