Case Study of the Ukraine Covid Epidemy Process Using Combinatorial-Genetic Method

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/CSIT49958.2020.9322000

Olha Moroz
Department for Information Technologies of Inductive Modelling
International Research and Training Centre
for Information Technologies and Systems of the NASU
Kyiv, Ukraine
olhahryhmoroz@gmail.com

Volodymyr Stepashko
Department for Information Technologies of Inductive Modelling
International Research and Training Centre
for Information Technologies and Systems of the NASU
Kyiv, Ukraine
stepashko@irtc.org.ua

***Abstract*—The article presents the modelling results of the Ukraine Covid-19 epidemy process using the combinatorial-genetic method based on official statistical data. A comparison with some other known methods for model construction and the process prediction is also given. This research is important for defining tendency of coronavirus evolvement in time and predicting its future activity in order to take some protective measures.**

Keywords—inductive modelling, GMDH, combinatorial algorithm COMBI, genetic algorithm GA, combinatorial-genetic method, COMBI-GA, coronavirus, covid-19 epidemy.

I. INTRODUCTION

Firstly, a new variation of coronavirus SARS-CoV-2 was identified in December 2019 in Wuhan (Hubei), China [1]. Despite all protection measures, the COVID-19 pandemic was spread in the following months both in the China and the whole world. For example, the World Health Organization reported by 6 May 2020 on 6.66 million + 119 thousand total cases and 393 thousand + 4288 deaths worldwide.

The coronavirus COVID-19 pandemic defines the global health crisis of our time and the greatest challenge we have faced since World War Two. Since its emergence in Asia late last year, the virus has spread to every continent except Antarctica. New cases are detecting daily in Europe, the Americas, and Africa. Countries are trying to slow the virus spread by testing and treating patients, carrying out tracing contacts, limiting travel, quarantining citizens, and cancelling large events like sports, concerts, and schools. The pandemic is spreading like a wave that floods more and more regions. But the virus is much more than a threat for health. By putting a strain on regions, countries and peoples, it may potentially to cause destructive social, economic and political consequences.

Relevant and precise modeling and prediction of the pandemic indicators on the basis of official statistical data is an urgent current task to improve security measures and help to cope with economy problems in every country.

To build such predictive models, it is reasonable to use the inductive modeling tools based on the Group Method of Data Handling (GMDH) [2] as one of the most effective methods for the analysis, modeling and forecasting of complex processes from experimental data under conditions of incompleteness of a priori information and short data samples.

Among the sorting-out methods of inductive modelling based on GMDH principles, the combinatorial-genetic method COMBI-GA [3] is very promising. In [4], the application results of the COMBI-GA for building optimal linear and nonlinear models in the class of models linear in parameters from data samples with both small and large number of observations are presented.

In this research, we use the COMBI-GA for finding best autoregression or polynomial models to forecast the coronavirus evolution in Ukraine. A comparison of obtained results with three other known methods is also given.

Section II of this paper describes the modelling task. Section III considers briefly five methods being compared and their features for solving this task. Section IV presents comparative research methodology. In Section V, results of modelling and prediction by the COMBI-GA are presented and compared. Section VI presents discussion, and Section VII does conclusive remarks.

II. MODELING TASK

In Ukraine coronavirus was firstly identified on March 01, 2020. Quarantine limitations have three main periods: strong quarantine from 18.03 till 11.05 (firstly imposed till 03.04, then prolonged twice till 24.04 and 11.05); first quarantine weakening from 12.05 till 22.05 (open parks and squares, dentistry, shops, hairdressers); for 22.05 to 1.06, it is just announced the next quarantine easing (public transport operation resumed).

We were given time series data of the number of people with confirmed coronavirus in Ukraine from the World Data Center (WDC) [5] for the period from March 01 to May 20, 2020. The goal is to construct best model for prediction of the confirmed cases taking into account different quarantine periods.

To achieve this goal, a comparison of models obtained by four different methods is considered: COMBI-GA in autoregression class with optimizing the model structure; Back Propagation Neural Network from WDC; standard autoregression (without sample division) and Lasso method.

III. APPLIED METHODS

The hybrid sorting-out COMBI-GA algorithm that comprises the combinatorial algorithm COMBI GMDH [2] and genetic algorithm GA [6] has the following general structure: 1) generation of a random set of partial model structures of a given size as an initial population of the

COMBI-GA; 2) estimating coefficients of every partial model using least squares method (LSM); 3) calculation of an external criterion values (fitness function of the GA) for each model, for example, the regularity criterion as typical for GMDH; 4) current selection of the best partial models (elite selection in GA) or reduction/rejection of worst individuals from the parent and offspring populations, and then formation of new population of the same size; 5) check a stop criterion, for example, achieving a given accuracy or number of iterations; stop if it is fulfilled, otherwise go to the next step; 6) use of genetic operators (crossover and mutation) with a given probability to selected individuals of the population and forming a set of partial model structures for the next generation; go to step 2.

Now the COMBI-GA method can be used to select best model in polynomial and/or autoregressive classes [7]. To search for the optimal model in these classes, the given data vector z [h×1] of the time series of the confirmed cases is transformed into the input matrix X [n×m] and the output vector y [n×1]. And then the standard algorithm COMBI-GA is working.

The originality of the construction of models according to this algorithm in the autoregression class is that there is chosen not only the optimal order of autoregression, but also the optimal composition of the delayed (lag) terms of this model. Thus, the hidden internal patterns of inertial effects of the afteraction in this process are detected which increases the predictive capabilities of a built model.

The method of Least Absolute Shrinkage and Selection Operator (LASSO) for estimation of coefficients of a linear regression model was first formulated in 1996 [8]. A proper algorithm is available in the Matlab statistics toolbox.

The method introduces an additional regularization component into the optimization functionality of the model to obtain a more stable solution. The condition of minimizing the square error when estimating the parameters is expressed by the formula:

$$\hat{\theta} = \arg\min\left(\|y - X\theta\|^2 + \lambda \|\theta\|\right),$$

where $\hat{\theta}$ is the vector of model parameters and λ is the regularization parameter which makes sense of the penalty for complexity. A cross-validation procedure is used to find the desired value of λ. In the course of minimization, some coefficients become equal to zero, i.e. at the same time there is a selection of informative variables.

There are many advantages in using LASSO method, first of all it can provide very good prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase of the bias, this is especially useful when there is small number of observations and large number of features. In terms of the parameter λ tuning, the bias increases and the variance decreases when λ increases, hence a trade-off between bias and variance has to be found. Moreover, the LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable, this way also overfitting is avoided.

The Backpropagation neural network (BPNN) is a multilayered, feedforward network and the most extensively used one due to its excellent function approximation ability [9, 10]. It is considered as one of the simplest and most general methods used for supervised learning of multilayered neural networks [9]. Backpropagation works by approximating a non-linear relationship between inputs and the output by adjusting the values of neuron weights in hidden layers. It can further be generalized for inputs that are not included in the training patterns (predictive ability).

A typical BPNN usually contains three kinds of layers including input layer, hidden layer, and output layer. Input layer is the entrance of the algorithm. It inputs one instance of the data into the network. The dimension of the instance determines the number of inputs in the input layer. Hidden layer contains one or several layers. It transfers intermediate data to the output layer that generates the final output of the neural network. The number of outputs is determined by the encoding of the classification results. In BPNN each layer consists of a number of neurons. The linear or nonlinear functions in each neuron are frequently controlled by two kinds of parameters, weight and bias.

IV. Research Technique

To apply the algorithms when solving this task, the data interval from 25.04 to 20.05 2020 (totally 26 days or data points) was taken as the period of active increase in the intensity of the infection process, see Table 1 and Fig. 1.

TABLE I. DATA OF THE CONFIRMED CASES IN UKRAINE 25.04 TO 20.05 [5]

Date	Real data	Date	Real data	Date	Real data
25.04	8125	04.04	12331	13.05	16425
26.04	8617	05.04	12697	14.05	16847
27.04	9009	06.05	13184	15.05	17330
28.04	9410	07.05	13691	16.05	17858
29.04	9866	08.05	14195	17.05	18291
30.04	10406	09.05	14710	18.05	18616
01.05	10861	10.05	15232	19.05	18876
02.05	11411	11.05	15648	20.05	19230
03.05	11913	12.05	16023		

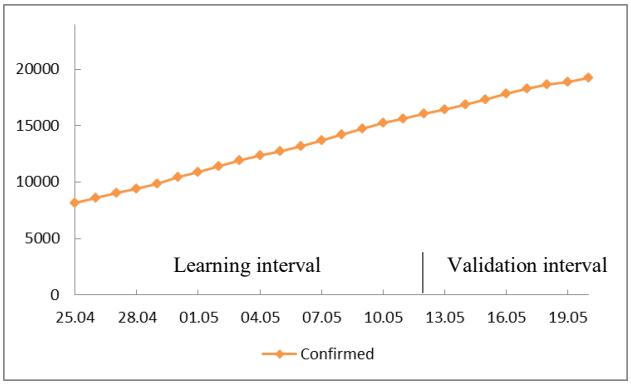

Fig. 1. Number of confirmed cases of people infected in Ukraine

When using the COMBI-GA, the whole data set was divided as follows: 17 days from 25.04 to 11.05.2020 (the strong quarantine period) were allocated for the model construction as the model *learning* set: 11 days for the *training* subset A to estimate model parameters and 6 days for the *checking* subset B to select the optimal autoregression model by the regularity criterion:

$$AR_B = \frac{1}{n_B}\left\|y_B - X_B\hat{\theta}_A\right\|^2, \hat{\theta}_A = (X_A^T X_A)X_A^T y.$$

To validate the forecast accuracy of the built optimal model, we use data of the *validation* subset: 9 days from 12.05 to 20.05.2020. As it was the period of the first quarantine relaxation, we have the possibility to find out if these new conditions change the patterns inherent to this process. The further "pure" forecast was done for the period from 21.05 to 31.05.2020 (11 days).

The same learning and validation sets were applied when using other methods to solve this modeling task. Generally, we compare the performance of the following four methods:

1. COMBI-GA for building optimal *autoregression* model as explained above;
2. *Backpropagation neural network* (BPNN) generating a nonlinear transformation of input data to an output response, namely we used results of its tuning and application for modeling and prediction presented on the WDC site [5]; as it can be supposed, the BPNN build here some kind of a nonlinear autoregression;
3. Lasso method we used for obtaining an *autoregression* model as well, applying the algorithm from the Matlab toolbox working in an automatic mode without any external division of the dataset;
4. Standard or *ordinal autoregression* when we simply estimate all the given lag parameters (one-time parametric identification without any structural optimization);

Note that the Backpropagation model is evidently not available because it is an implicit nonlinear neural network which cannot be explicitly presented in any mathematical form but can be used to calculate approximation and prediction. And these calculated results we have simply taken in the ready numerical form from the WDC site [5].

V. Modeling Results

A. Results Obtained Using COMBI-GA

First experiment was aimed to test modeling results for various time lags. Table 2 shows models have been built by COMBI-GA for maximum 3, 7, 10 and 14 lags and the respective values of the Mean Absolute Percentage Error:

$$\mathrm{MAPE} = \frac{100\%}{n_v} \sum_{k=1}^{n_v} \left| \frac{z_k - \hat{z}_k}{z_k} \right|$$

on the validation subset (for dates 12.05 to 20.05, $n_v = 9$). Here k is the discrete time of the process or the ordinal numeration of days from the beginning of the period.

TABLE II. BEST MODELS OBTAINED BY COMBI-GA IN AUTOREGRESSION CLASS WITH DIFFERENT LAG NUMBERS

Lags	Model	MAPE %
3	z(k) = – 0,441z(k–1) + 1.445z(k–3)	1.93
7	z(k) = – **0.043z(k–1) – 0.168z(k–3) + 1.213z(k–7)**	**0.44**
10	z(k) = – 0.387z(k–3) + 0.761z(k-5) – 0.584z(k–6) + + 1.201z(k–10)	0.83
14	z(k) = 0.561z(k–2) – 0.503z(k–3) – 0.301z(k–6) + + 0.195z(k–12) – 0.064z(k–13) + 1.111z(k–14)	1.62

As it follows from these results, optimal autoregression model found by the COMBI-GA algorithm contains only 3 lags out of the max their number 7:

$$z(k) = -0.043z(k-1) - 0.168z(k-3) + 1.213z(k-7).$$

These 3 lags may be interpreted as the most informative ones for predicting the process under consideration. In fact, this result can be explained on the basis of on the already known observation that the symptoms of this disease appear in 5 to 7 days after infection. Hence the lag 7 appropriate to 7th day plays a key role in this relationship which is reflected in that the coefficient at the term z(k–7) has maximal value among others and is positive. At the same time the terms z(k–1) and z(k–3) are less significant and has negative coefficients, at that the first one is minimal.

B. Comparative Results

Here we present the results of comparing performance of models obtained by the 4 methods announced above:

1) optimal COMBI-GA autoregression;

2) Back Propagation Neural Network;

3) Lasso method;

4) standard autoregression (with complete set of 7 lags);

Table 3 shows best models built by the 5 methods and the appropriate MAPE values. Figs 2 to 5 illustrate comparisons of real data values with the modeling and prediction results for each of the 4 used methods.

To build the classical autoregression model, we use 7 lag numbers as in the optimal model synthesized by the COMBI-GA.

TABLE III. MODELS AND MAPE VALUES WHEN PREDICTING BY THE 4 METHODS THE CONFIRMED CASES IN UKRAINE ON THE VALIDATION SET

Method	Model	MAPE %
1	z(k) = – 0.043z(k–1) – 0.168z(k–3) + 1.213z(k–7)	0.44
2	Back Propagation NN, WDC, n/a	1.32
3	z(k) = – 0.422z(k–1) +0.634z(k–2) – 0.313z(k–3) – – 0.402z(k–6) + 1.503z(k–7)	1.59
4	z(k) = – 0.523z(k–1) +0.822z(k–2) – 0.179z(k–3) – – 0,424z(k–4) + 0,281z(k–5) – 0.687z(k–6) + + 1.713z(k–7)	5.73

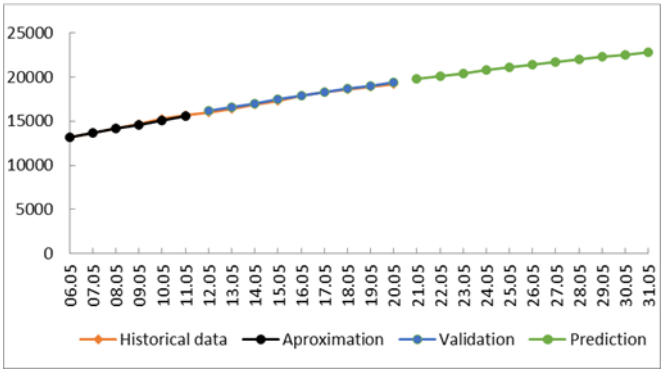

Fig. 2. Confirmed cases modeled and predicted using COMBI-GA

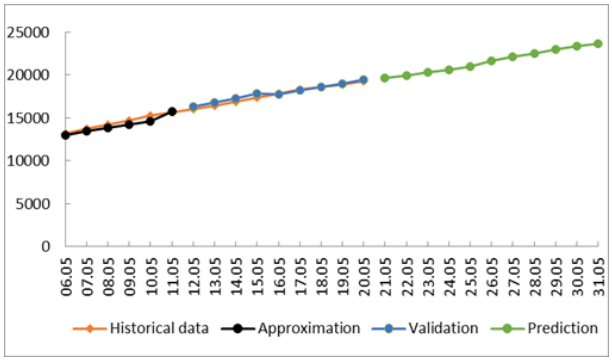

Fig. 3. Approximation and prediction of confirmed cases in Ukraine obtained using BPNN by the WDC

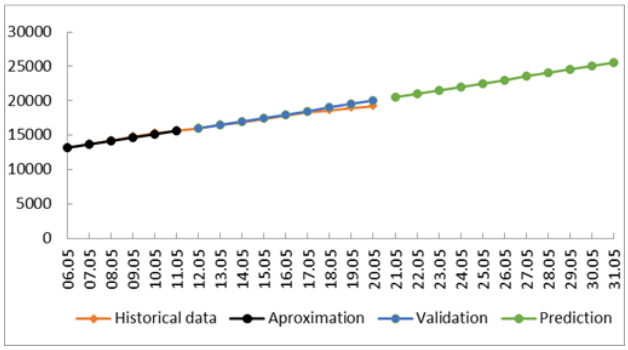

Fig. 4. Confirmed cases modeled and predicted using LASSO

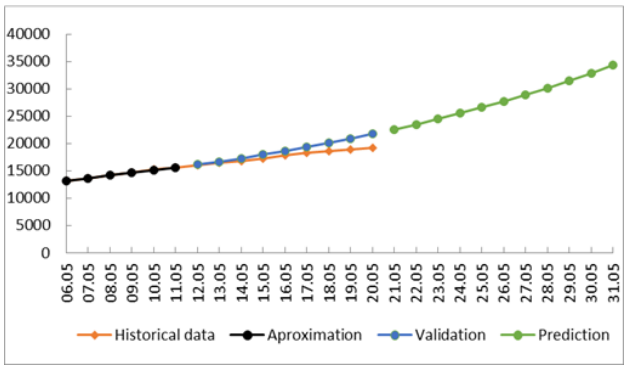

Fig. 5. Confirmed cases modeled and predicted using autoregression

The obtained results show that COMBI-GA algorithm has built the optimal model in the autoregression class with minimum error on the validation subset.

VI. Discussion

Firstly, it is worth to stress the fact that the nonlinear BPNN model have appeared to be not fully relevant to this process of raising the number of confirmed cases of infections in Ukraine. Namely, the WDC results presented in the Fig. 3 demonstrate sufficient approximation but an inadequate behavior of the model because of nonmonotonic growth as compared to the historical data.

The rest three models 1, 3, 4 were constructed in the class of autoregressive dependences and are very serviceable to predict monotonic processes. As it can be observed, all these models include the lag variables (inputs) z(k–1), z(k–3) and z(k–7) that indicates their significant influence on the evolution of coronavirus. It has logical confirmation because, as it is known, the symptoms of the coronavirus disease appear in 5 to 7 days after infection.

But when taking into account the validation indicators of these models, the quality of them is quite different. Model 1 built by COMBI-GA contains only the three lags mentioned above and has lowest validation error. Two other models 3 and 4 contain some additional terms. Namely, the standard (complete) autoregression overestimates the real tendency to grow the process. The Lasso method gives somewhat optimized structure and shows better validation accuracy than the ordinary autoregression.

As for the prediction results of these 4 methods on the prediction interval 21.05 to 31.05 (11 days), one can point out that the model 1 displays a weak tendency to stabilizing the process (Fig. 2), models 2 and 3 gives quasi linear growth (Fig. 3 and 4), whereas model 4 tends evidently to an exponential growth (Fig. 5).

VII. Conclusion

The modeling results of the Ukraine Covid-19 epidemy process were presented in the paper. These results for the indicator of confirmed cases dynamics were obtained using five different methods. The best optimal model of the coronavirus infection growth was built by the COMBI-GA in an optimized autoregression class with 7 lags. We can conclude that this algorithm has detected the hidden internal patterns of inertial dynamic effects of the process due to some level of intelligence inherent to this specific kind of GMDH algorithms. It is worth to stress that this result was generated in the linear class of dynamic models.

Besides that algorithm, we tested in this task 3 other approaches: nonlinear method BPNN as well as 2 other ones in the autoregression class (Lasso-based optimal model and standard autoregression with complete set of 7 lags). These 3 methods in general correctly reproduce the growth tendency of the process but they showed worse validation accuracy than COMBI-GA.

References

[1] E. Volz, H. Fu, H. Wang et al. Genomic epidemiology of a densely sampled COVID19 outbreak in China. medRxiv; 19-03-2020. https://doi.org/10.1101/2020.03.09.20033365

[2] H.R. Madala, A.G. Ivakhnenko, Inductive Learning Algorithms for Complex Systems Modeling. New York: CRC Press, 1994.

[3] O.H. Moroz, "Sorting-Out GMDH algorithm with genetic search of optimal model," Control Systems and Machines, no. 6, pp. 73-79, 2016. (In Russian)

[4] O. Moroz, V. Stepashko, "Hybrid Sorting-Out Algorithm COMBI-GA with Evolutionary Growth of Model Complexity," Advances in Intelligent Systems and Computing II / N. Shakhovska, V. Stepashko, Editors, AISC book series, Berlin: Springer Verlag, vol. 689, pp. 346-360, 2017.

[5] phttp://wdc.org.ua/

[6] J. Holland, Adaptation in natural and artificial systems: An introductory analysis with application to biology, control, and artificial intelligence, University of Michigan, Computers, 1975, 183 p.

[7] L. Ljung, System Identification. Theory for the User, PTR Prentice Hall, Upper Saddle River, 1999, 609 p.

[8] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” J. Royal. Statist. Soc. B., 1996, vol. 58, no. 1, pp. 267-288.

[9] Wei Lu, “Neural Network Model for Distortion Buckling Behaviour of Cold-Formed Steel Compression Members,” Helsinki University of Technology, Laboratory of Steel Structures Publications 2000, no. 16, http://www.hut.fi/Yksikot/Rakennus/Teras/TKK-TER-16.pdf

[10] M.H. Hagan, H.B. Demuth, and M.H. Beale, *Neural Network Design*, PWS Publishing Company, 1996.