Short-Term Covid-19 Incidence Prediction in Countries Using Clustering and Regression Analysis

Fuad Aleskerov[1,2], Sergey Demin[1,2], Alexey Myachin[1,2](✉), and Vyacheslav Yakuba[1,2]

[1] HSE University, 20 Myasnitskaya Street, Moscow 101000, Russia
{alesk,sdemin,amyachin,vyakuba}@hse.ru

[2] Institute of Control Science of Russian Academy of Science, 65 Profsoyuznay Street, Moscow 117997, Russia

Abstract. As of February 07, 2022, more than 395 million cases of COVID-19 had been identified in the world, with 5.74 million deaths. The paper considers methodology for predicting the number of cases in the short term using a preliminary assessment of countries based on three indicators: expert assessments of the law-abiding population, the level of education and restrictive measures taken in the country. The description and composition of the groups obtained are given. An assessment of the accuracy of the forecast results is made. A comparison of the considered models of 2020 with 2022 is given.

Keywords: Covid-19 · Clustering · Pandemic · Incidence prediction · Regression analysis

1 Introduction

In connection with the problems caused by the COVID-19 pandemic that began in 2019, the number of various studies on this topic has grown rapidly. A query on the Google Academy for the keyword "COVID-19" showed about 4,700,000 results, which is a significant number given the relatively short study period. These works deal with the various aspects of the study of this disease: data collection and visualization [3], clinical symptoms [13], and the description of specific cases of patients [11]. Many articles study the impact of COVID-19 on human health, and possible concomitant diseases [5, 7, 12]. However, despite the extensive statistical data, a relatively small number of works, as far as the authors know, apply the quantitative methods for making short and medium term forecasts.

This study focuses on two main aspects. First, the methodology of dividing 66 countries into groups with similar structures according to the following basic system of indicators is proposed: the aggregated value of weekly quarantine measures and restrictions adopted in the country, the general level of education, and expert opinion on the law-abidingness of citizens. Naturally, much attention is paid to the choice of

S. Dzitac et al. (Eds.): ICCCC 2022, AISC 1435, pp. 333–342, 2023.
https://doi.org/10.1007/978-3-031-16684-6_29

the methodology for splitting the sample of countries under study and the basic system of indicators. Second, the compilation of the models for predicting the increase in the number of cases of COVID-19 is performed. Since it is very difficult to draw up a single model for all countries, identifying clusters can significantly increase the final forecast accuracy. Significance of the model is testing took place on data for the period of February 01, 2020–October 01, 2020, and compared with the result for January 2022. In most cases, the determination coefficient exceeds 0.8.

2 Initial Data

Since the ultimate goal of the study is to build a model that allows predicting of the increase in the number of cases in certain countries of the world in the short term, it is assumed that a basic system of indicators is used to determine differences in the behavior of citizens in the presence of restrictive measures, and the trend of the disease itself. To determine the behavior of citizens, it is assumed to use both the general level of education and law-abidingness, and the severity of the restrictive measures adopted in the country. Working with data on education is not particularly difficult: there are publicly available databases on the number of citizens with different levels of education [10]. However, the big question is the operationalization of the concept of "law abiding" in this task and the choice of characterizing indicators. Among the possible options, it is proposed to use various aggregated indices, including [4], or the partial use of indicators from some studies (e.g. Public Trust in Politicians from [9]). However, these indices characterize the level of trust rather than law-abidingness. In this regard, it is decided to switch to expert assessments conducted by specialists from the HSE University and the Trapeznikov Institute of Control Science of Russian Academy of Sciences. The peer review required certain limitations for the original sample, and therefore 66 countries were included. We consider working with the selected scorecard in more detail:

1. Weekly data on the number of cases of COVID-19 in the country, taken from [6]. Since the comparison of the absolute number of values in some cases is not always representative, in this study, a weekly gain is used. All values are logarithms.
2. Aggregate indicator characterizing the level of restrictive measures adopted in the country for the population due to the worsening of the epidemiological situation [2]. The measures taken are normalized on a scale of [0; 1].
3. An aggregate indicator characterizing the level of education, based on the data presented in [10]. The initial data is the percentage of the population with no education, incomplete primary, primary, lower secondary, upper secondary and post secondary. The level of education is normalized on a scale of [0; 1].
4. The general level of law-abidingness in the country. The indicator is evaluated on the basis of expert estimates, in the scale of [0; 10]

3 Finding Clusters

3.1 The Results of 2020

The study assumes an increase in forecast accuracy by compiling not a single model for all countries, but after preliminarily dividing them into clusters. For this purpose, methods of pattern analysis [1, 8], classification and clustering is considered. The results based on pattern analysis is difficult to interpret, and it is very difficult to compose a test sample for classification. In this regard, classical clustering methods are used for further analysis. The big question remains is both the choice of a particular methodology and the validity of this choice. In the study, the selection criterion is the interpretability of the final results and the accuracy of the prediction.

Results that improve the accuracy of predictive models are obtained using the k-means method. The criteria for assessing the quality of clustering to determine the number of clusters are silhouette estimates (in conjunction with the coefficient of the determination of regression models to determine the forecast accuracy). However, due to the use of dynamic data, the resulting clusters may differ in composition and in number, depending on the week studied, in which the measures taken in the country are considered, and a forecast is made about the increase in the number of cases. Consider an example for July 18, 2020. The clustering is shown in Fig. 1.

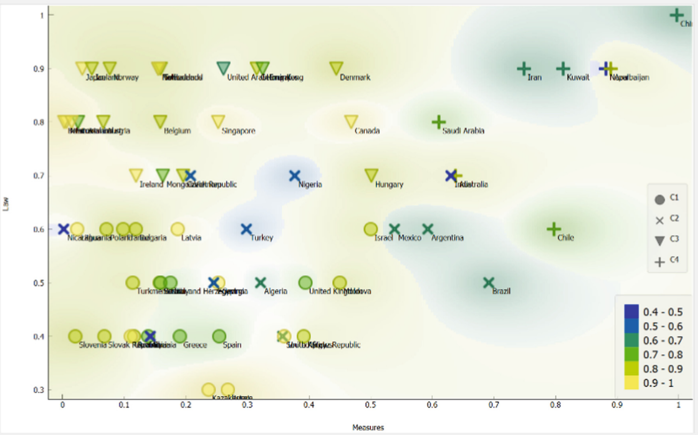

Fig. 1. Splitting the original sample of countries into groups (as of July 18, 2020).

The X axis characterizes the measures taken as of July 18, 2020, the Y axis is the law-abiding population in different countries. The color scale reflects the normalized indicator characterizing the level of education.

The composition of the groups differs somewhat depending on the date in question (although, in some cases, not significantly). For example, consider an alternative breakdown presented on September 19, 2020.

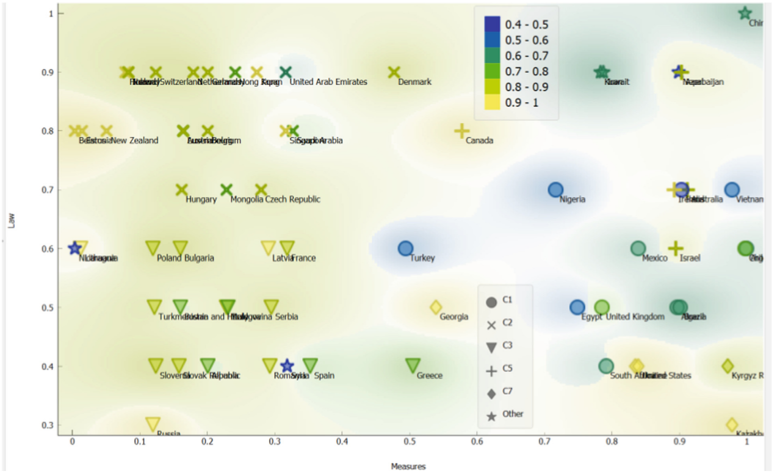

Fig. 2. Splitting the original sample of countries into groups (as of September 19, 2020).

The above figures clearly show the difference between the resulting clusters. In some cases, not only the number of clusters differs, but also their composition. However, such a result does not contradict the original problem, namely, the compilation of regression models to predict the increase in morbidity. This approach, on the contrary, allows an adjustment of the results over time, providing higher accuracy when making short-term forecasts.

3.2 The Results of 2022

The clustering has been carried out for the entire period from the beginning of 2020 to 2022. To compare the results, as well as to demonstrate the effectiveness of the proposed approach, an example of short-term forecasting of the increase in incidence on January 15, 2022 is given. Measures taken are based on data as of January 8, 2022.

Clustering is also based on the k-means algorithm, with the number of clusters adjusted when using silhouette estimates. Note that, for the purpose of comparison, analysis has been also carried out based on other methods, including: ordinal-fixed, ordinal-invariant and diffusion-invariant pattern clustering, hierarchical clustering, density-based spatial clustering of applications with noise (DBSCAN). The efficiency criterion is the interpretability of the final results (in our case, the accuracy of regression analysis models obtained based on clustering for short-term forecasting of the number of cases of COVID-19).

Figure 3 shows the results of the clustering. Designations are similar to Fig. 1 and 2. The X axis characterizes the measures taken as of January 08, 2022, the Y axis is the law-abiding population in different countries. Colors reflect the normalized indicator characterizing the level of education.

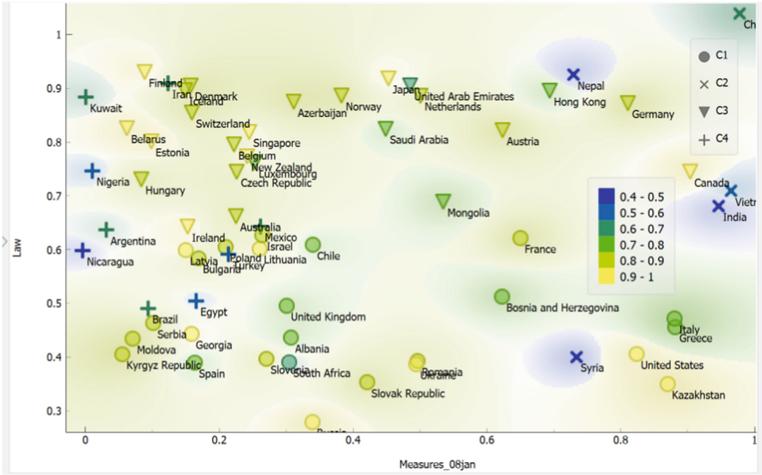

Fig. 3. Splitting the original sample of countries into groups (as of January 08, 2022).

4 The Regression Analysis

4.1 Regression Models for 2020

The next stage of the study is the construction of regression models in order to predict the increase in the number of number of new cases in the short term. Such models cannot take into account all possible factors, but they are very useful in drawing up general recommendations for the population and in making the necessary decisions. As mentioned, the general model gives an insufficiently accurate forecast (the comparison is shown in Table 1), and therefore separate models for each cluster are presented (Fig. 4).

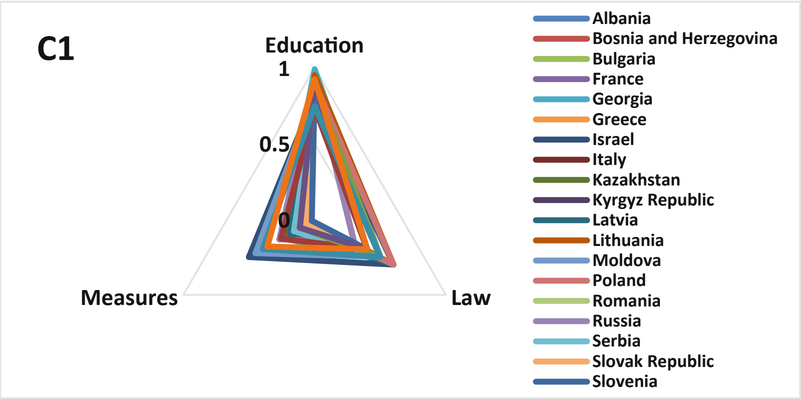

Fig. 4. Visualization of Cluster 1 (as of July 18, 2020).

Cluster 1 consists of 24 countries: Albania, Bosnia and Herzegovina, Bulgaria, France, Georgia, Greece, Israel, Italy, Kazakhstan, Kyrgyz Republic, Latvia, Lithuania, Moldova, Poland, Romania, Russia, Serbia, Slovak Republic, Slovenia, Spain, Turkmenistan, Ukraine, United Kingdom, United States. The relatively average indicators of both the measures taken as of July 18, 2020, and the values of expert assessments regarding law-abidingness, and high education indicators are typical. The multiple regression model in this cluster looks like this:

$$N_t(Cluster\,1) = 0{,}015 + 0{,}523N_{t-1} - 0{,}715N_{t-2} - 0{,}009N_{t-3} + 0{,}868N_{t-4} + 0{,}078N_{t-5}$$

Thus, the increase in the number of incidences of COVID-19 on July 25, 2020 is projected based on clustering carried out on July 18, 2020. For an accurate forecast, the previous 5 weeks have been shown. Moreover, the coefficient of determination is 0.911. For comparison, when compiling a single regression model for 66 countries on the same data (and for the same period), the coefficient of determination is 0.678. Thus, there is a significant improvement in the quality of the forecast due to the division of countries into clusters (a full comparison is shown in Table 1 and 2) (Fig. 5).

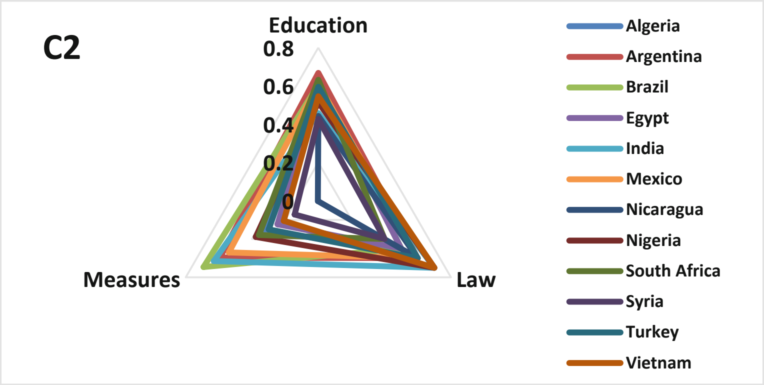

Fig. 5. Visualization of Cluster 2 (as of July 18, 2020).

The second cluster includes 12 countries: Algeria, Argentina, Brazil, Egypt, India, Mexico, Nicaragua, Nigeria, South Africa, Syria, Turkey, Vietnam. In this case, a relatively high indicator of the level of education and law-abidingness, is characteristic. The regression equation is:

$$N_t(Cluster\,2) = 0{,}036 - 0{,}356N_{t-1} - 0{,}328N_{t-2} - 0{,}461N_{t-3} - 0{,}44N_{t-4} + 1{,}737N_{t-5}$$

The coefficient of determination is the smallest among all the clusters obtained: at 5 weeks it is 0.812. However, in this cluster the R-squared value is relatively high and at 2 weeks, and further changes insignificantly.

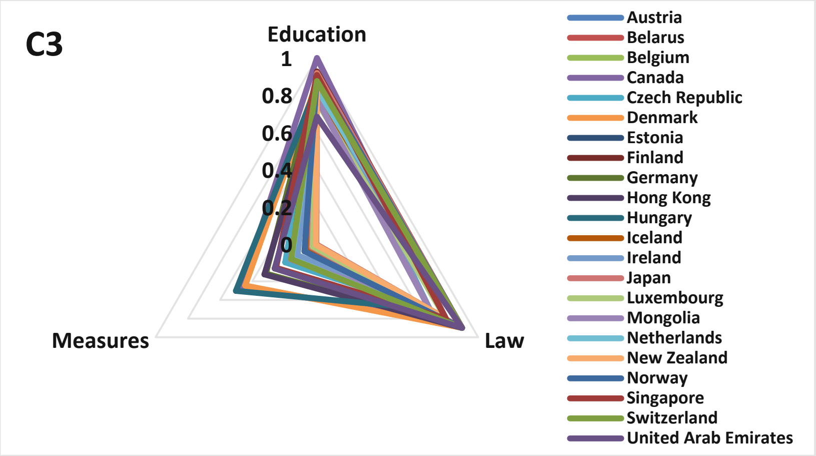

Fig. 6. Visualization of Cluster 3 (as of July 18, 2020).

The third cluster is typical for 22 countries: Austria, Belarus, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, Germany, Hong Kong, Hungary, Iceland, Ireland, Japan, Luxembourg, Mongolia, Netherland, New Zealand, Norway, Singapore, Switzerland, United Arab Emirates. In this cluster, relatively high indicators of education, the level of law-abidingness, and relatively low values of the aggregate indicator of measures taken as of July 18, 2020, are visible. Regression equation for this cluster is:

$$N_t(Cluster\ 3) = 0{,}001 - 1{,}209N_{t-1} + 0{,}587N_{t-2} - 0{,}923N_{t-3} + 3{,}363N_{t-4} - 0{,}259N_{t-5}$$

An interesting feature of Cluster 3 is the significant increase in the value of the determination coefficient from 1 to 2 weeks: 0.355 and 0.776, respectively. For 5 weeks, R-squared is 0.857, which, despite its high value, only surpasses the value of Cluster 2. However, we note that this value can be increased (although not very significantly) with an increase in the number of periods considered (Fig. 7).

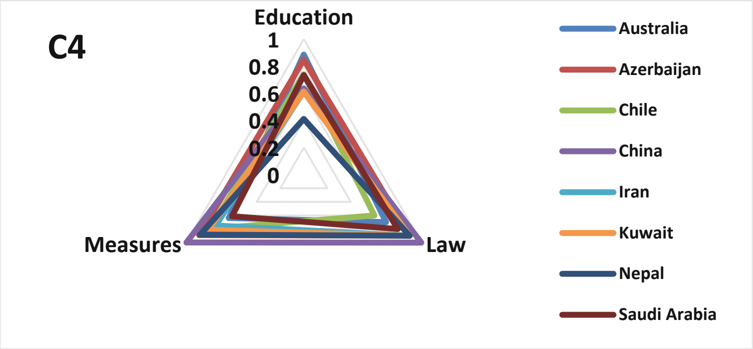

Fig. 7. Visualization of Cluster 4 (as of July 18, 2020).

Cluster 4 is described by 8 countries: Australia, Azerbaijan, Chile, China, Iran, Kuwait, Nepal, Saudi Arabia. Figure 6 shows the features characteristic of this cluster: average and relatively high values for all considered indicators: education level, law-abidingness, and measures taken as of July 18, 2020. The regression equation for the cluster is:

$$N_t(\mathit{Cluster}\,4) = 0{,}014 + 0{,}048N_{t-1} - 0{,}529N_{t-2} + 0{,}848N_{t-3} - 0{,}922N_{t-4} + 1{,}293N_{t-5}$$

The very high value of R-squared is characteristic, as it is gradual insignificant increase when a larger number of weeks are included in the model: from 0.955 at 2 weeks to 0.991 at 5. Since the value of the determination coefficient in this cluster is initially very high, the increase is the least significant (when compared with other obtained clusters).

Below is a general comparison of the results. As an example, the regression model for Cluster 1 on July 18, 2020 is built followed by the forecast of an increase the number of cases on July 25, 2020. The accuracy of the model, as before, is assessed based on the coefficient of determination. The results are shown in the Table 1.

Table 1. Comparison of the results as of July 25, 2020 for Cluster 1 with the entire sample

Number of weeks	R-squared for cluster 1	R-squared for all countries
1	0.487	0.346
2	0.893	0.644
3	0.896	0.655
4	0.896	0.679
5	0.911	0.678

In Table 1, the advantage of using data from one cluster from the entire sample is obvious, with the exception of the short-term forecast based on one-week data. Next, we present the results of other clusters.

Table 2. Comparison of results as of July 25, 2020 for Clusters 2–4

Number of weeks	Cluster 2	Cluster 3	Cluster 4
1	0.763	0.355	0.883
2	0.763	0.776	0.955
3	0.766	0.832	0.977
4	0.789	0.844	0.985
5	0.812	0.857	0.991

The results for the other periods under consideration shows similar results: with the preliminary division of countries into clusters (for a specific date), the forecast accuracy of increases. The composition of clusters and their number may change over different periods, which is due to the use of a dynamic indicator (measures).

4.2 Regression Models for 2022

To compare the results, we present regression models for predicting the number of cases on January 15, 2022. Models are built for 4 clusters obtained on the basis of k-means.

For all obtained clusters, the coefficient of determination is not lower than 0.95 (for logarithmic data). Below are the resulting models:

$$
\begin{aligned}
&N_t(Cluster\,1) = 0{,}001 + 0{,}075N_{t-1} - 0,108N_{t-2} + 0{,}248N_{t-3} - 1{,}3N_{t-4} + 1{,}99N_{t-5} \\
&N_t(Cluster\,2) = 0{,}025 + 0{,}523N_{t-1} - 1{,}75N_{t-3} + 1{,}25N_{t-5} \\
&N_t(Cluster\,3) = 0{,}009 - 0{,}46N_{t-1} + 1{,}59N_{t-2} - 0{,}97N_{t-3} - 1{,}05N_{t-4} + 1{,}68N_{t-5} \\
&N_t(Cluster\,4) = 0{,}001 + 0,149N_{t-1} - 0{,}239N_{t-2} - 0{,}007N_{t-3} - 0{,}319N_{t-4} + 1{,}217N_{t-5}
\end{aligned}
$$

5 Conclusion

COVID-19 poses a serious threat to the health of the entire world. Restrictive measures have been introduced in many countries, and the consequences of this disease remain to be studied. In this regard, it is highly relevant to create mathematical models for predicting the spread of the disease.

The article provides a methodology for improving the accuracy of predicting the increase in the number of cases in 66 countries of the world with a preliminary division of countries into clusters using cluster analysis methods based on 3 indicators: the level of education in the country, the law-abidingness of citizens and the restrictive measures taken. The use of dynamic data allows an adjustment of the number of clusters and their composition, which also increases the forecast quality.

Acknowledgements. The article was prepared within the framework of the HSE University Basic Research Program.

References

1. Aleskerov, F., Egorova, L., Gokhberg, L., Myachin, A., Sagieva, G.: Pattern analysis in the study of science, education and innovative activity in Russian regions. Proc. Comput. Sci. **17**, 687–694 (2013)
2. Coronavirus Government Response Tracker. https://www.bsg.ox.ac.uk/research/research-projects/coronavirus-government-response-tracker
3. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet. Infect. Dis **20**(5), 533–534 (2020)
4. Edelman Trust Barometer (2020). https://www.edelman.com/trustbarometer

5. Fang, L., Karakiulakis, G., Roth, M.: Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir. Med. **8**(4), e21 (2020)
6. Johns Hopkins University. https://coronavirus.jhu.edu/map.html
7. Li, G., et al.: Coronavirus infections and immune responses. J. Med. Virol. **92**(4), 424–432 (2020)
8. Myachin, A.L.: Pattern analysis in parallel coordinates based on pairwise comparison of parameters. Autom. Remote. Control. **80**(1), 112–123 (2019)
9. The Global Competitiveness Report 2017–2018. http://reports.weforum.org/global-competitiveness-index-2017-2018/
10. The World Bank. Education Statistics – All Indicators. https://databank.worldbank.org/databases/education
11. Zhang, Y., et al.: Coagulopathy and antiphospholipid antibodies in patients with Covid-19. New Engl. J. Med. **382**(17), e38 (2020)
12. Zheng, Y.Y., et al.: COVID-19 and the cardiovascular system. Nat. Rev. Cardiol. **17**(5), 259–260 (2020)
13. Zu, Z.Y., et al.: Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology **296**(2), E15–E25 (2020)