УДК 519.237:330.43

STATISTICAL ANALYSIS AND ECONOMETRIC MODELING OF THE COVID-19 PANDEMIC

V.I. Malugin, A.K. Kornievich, V.A. Potapovich

Belarusian State University, Nezalezhnosti Ave., 4, 220030, Minsk, Belarus
Malugin@bsu.by

This paper presents the results of solving the following tasks: development of statistical methods for classifying countries in the European Region by the intensity of the COVID-19 pandemic; building country ratings that characterize the intensity of the pandemic; assessment of the relationship of country ratings with the economic indicators of countries; development of econometric models of the epidemic process in the Republic of Belarus.

***Keywords*:** COVID-19 typology; cluster analysis; econometric modeling; statistical ratings; economic indicators.

Introduction

The problem of analyzing the COVID-19 pandemic in various aspects is given considerable attention in the world scientific literature [1]. An important direction in the ongoing research is the development of methods for statistical analysis of the COVID-19 pandemic based on the data available in the mode of regular updating. Both analytic simulation [2] and statistical models [3] are used to analysis and predict the epidemic process at the level of individual countries. Considerable attention is paid to the tasks of analyzing the COVID-19 pandemic in a multi-country aspect [1, 4].

This research has the following objects: 1) statistical multi-country analysis of the pandemic COVID-19 typology and evaluation of its influence on the economic indicators of countries by means the machine learning algorithms; 2) econometric modeling and forecasting of the epidemic process in the Republic of Belarus.

1. The problems and used data

Multy-country COVID analysis problem. It is assumed that the available panel statistical data include the values of N indicators of epidemic process obtained for some sample of countries of volume n at time $t\,(t=1,...,T)$:

$$x_{i,t}=(x_{i1,t},...,x_{iN,t})'\in\Re^N\ (i=1,...,n,\ t=1,...,T).$$

In the context of the COVID-19 analysis, panel data have a heterogeneous

cluster structure. It is supposed that the most important factor of heterogeneity is the difference between countries by a latent feature, which characterizes the intensity of the COVID-19 epidemic process. According to this property, countries can be assigned to one of the L classes. This property is expressed by a discrete random variable $d_{it} \in \{1,...,L\}$, indicating the class number for country i at time t. Class numbers $\{d_{it}\}$ are interpreted as country ratings of the intensity of the epidemic process.

The problem of statistical classification: to divide the sample $\{x_{i,t}\}$, heterogeneous in terms of latent feature, into L homogeneous subsamples (classes) that differ in the space of classification features by the degree of intensity of the epidemiological process. The solution to this problem is the classification matrix $D = \{d_{i,t}\} (i = 1,...,n,\ t = 1,...,T)$.

Single-country COVID analysis problem. The purpose of statistical analysis of the epidemic process within a single country is to solve the following tasks based on available statistical data: assessment and short-term forecasting of the growth of new infections; building long-term forecasts, the purpose of which is to assess the turning point of the epidemic wave and the moment of its completion.

To solve these problems, we use the daily and weekly data for 30 countries of the European region (Armenia, Austria, Azerbaijan, Belarus, Bulgaria, Croatia, Czech, Denmark, Estonia, Finland, France, Georgia, Germany, Great Britain, Greece, Hungary, Ireland, Italy, Kazakhstan, Latvia, Lithuania, Moldova, Netherlands, Poland, Romania, Russia, Slovenia, Spain, Switzerland, Turkey) from March 1, 2020 to April 18, 2022.

The list of available indicators includes: total number of infections (Total – $T(t)$), number of active cases of infection (*A*ctive – I(t)), number of recovered (*R*ecovered – $R(t)$), number of deaths (Deceased – $D(t)$) [5].

2. Used approaches and algorithms

Statistical analysis of the pandemic COVID-19 typology. The following classification features are constructed and used:

- the ratio of the number of closed cases to the total number of infected (Closed to Total);
- the ratio of the number of closed cases to the number of active cases (Closed to Active);
- daily growth rate of the total number of infections or the ratio of the current value of cases to the previous one (Total Infections Daily Rate);

- mortality rate – the proportion of deaths from the total number of officially registered cases of COVID-19 (Death Rate).

Since there is no training sample and the number of classes L is not known, it is necessary to use panel data classification algorithms in the self-learning mode. To solve this problem, it is proposed the approach to the analysis of panel data with a cluster structure [6].

This approach includes the following steps:

- preliminary statistical analysis of sample and outlier detection;
- censoring and scale transformation of features to interval (0,1) in such a way that values close to zero correspond to a more favorable course of the epidemic process and vice versa;
- cluster analysis of initial non-classified sample in cross-sectional representation by means of hierarchical cluster analysis and L-means algorithm;
- calculation and analysis of pandemic statistics at country and multi-country levels.

Discriminatory abilities of classification features in the L-means algorithm are shown in Figure 1.

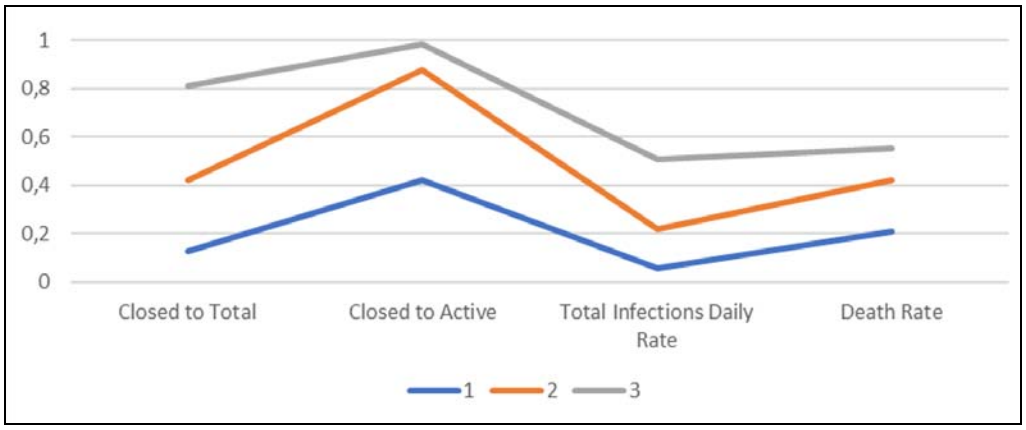


Figure 1 – Values of classification features for cluster centers 1, 2, 3

Based on the estimated classification matrix $D = \{d_{i,t}\}(i = 1,...,n,\ t = 1,...,T)$ the following indicators of the COVID-19 pandemic are constructed:

$d_{it} \in \{1, \ldots, L\}$ – *daily country rating* (*DCR*), which characterizes the degree of intensity of the epidemic for country i at time t: rating values 1 and L correspond to the lowest and highest degree of intensity of the epidemic process;

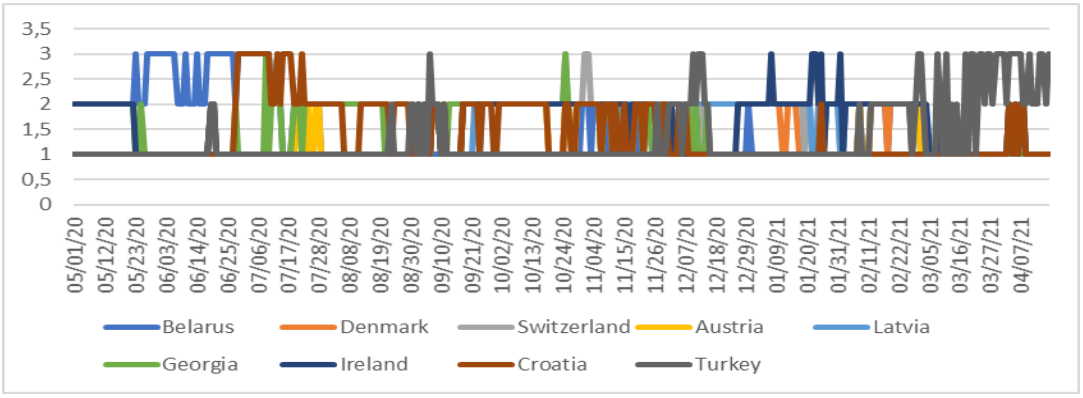
ACR_i – *Average Country Rating for the entire time interval*:

$$ACR_i = \frac{1}{T}\sum_{t=1}^{T} d_{it} \in (1,\ L),\ i = 1,...,n;$$

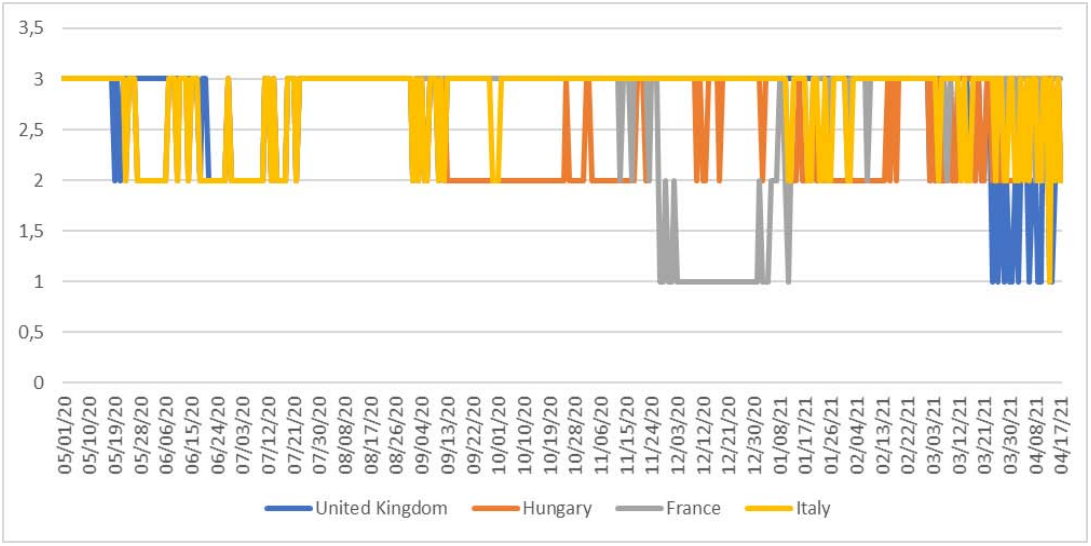
IMI_t – *Integral Multicounty Indicator of COVID-19* at time $t = 1,...,T$:

$$IMI_t = \frac{1}{n}\sum_{i=1}^{n} d_{it} \in (1,\ L),\ t = 1,...,T.$$

Figure 2 illustrates the daily DCR rating for countries from classes 1 (panel *a*) and 3 (panel *b*) with lowest and highest degree of intensity of the epidemic process respectively.



a)



b)

Figure 2 – DCR rating for countries of class 1 (a) and 3 (b) up to April 18, 2021

Table shows the average values (at the end of 2020) for GDP Annual Growth Rate and Unemployment Rate [7] for classes 1, 2, 3.

Table – Country rankings with GDP growth rates and Unemployment rate

Class (rating)	GDP Annual Growth Rate	Unemployment Rate
1	-3,610	8,395
2	-4,063	7,141
3	-7,716	6,630

The results of ranking all countries according to the ACR rating are presented on Figure 3 for two term intervals of COVID-19, including: 1) March, 2020 – April, 2021; 2) March, 2020 – April, 2022. It can be concluded that the typology of the epidemiological process in countries as a whole is preserved for new waves of the epidemic.

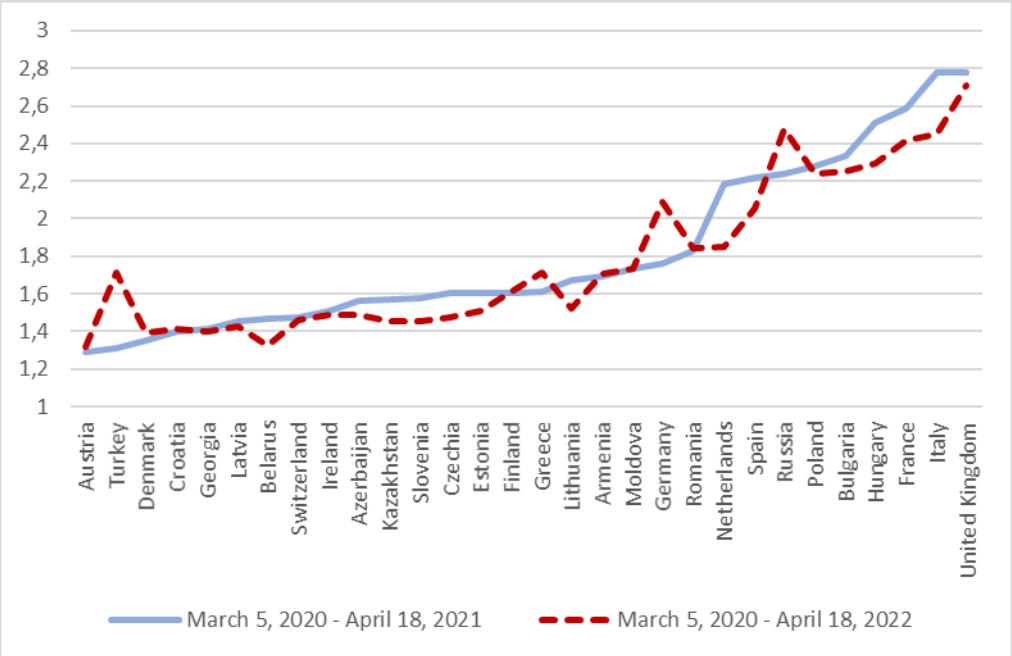


Figure 3 – The results of ranking all countries according to the ACR rating for the first two waves and entire observation period

Econometric modeling and analysis of COVID-19 in the Republic of Belarus. To analyze and predict the main indicators of the COVID-19 epidemic process two types of econometric models have been developed:

1) vector error correction model (Vector Error Correction Model – VECM COVID-19 RB) for analysis and forecasting within a single wave;

2) Markov-switching models for estimation the turning points of rise and fall of the epidemic process for the entire observation period [8].

Both models are based on assumptions close to the SIR (*Susceptible-Infectious-Recovered*) model. The main one is the assumption of the existence of a long-term equilibrium dependence for a steady state of the epidemic [2]:

$$I(t) + C(t) + S(t) = N \text{ or } I(t) + C(t) = N - S(t) = T(t),$$

where for the moment of time N – the size of the entire population; $S(t)$ – the number of persons susceptible to infection; $C(t)$ – the number of closed cases of infection, including those who recovered $R(t)$ and died $D(t)$.

Weekly time series $I(t)$, $C(t)$ are used to build a linear regression model MS-LR-AR with Markov switching of states, that allows autocorrelation of residuals. For two classes of states of the epidemic process, "rising" and "recession", the first differences of the time series $DI(t)$ and $DC(t)$ are used, that is, weekly changes in the variables $I(t)$ в зависимости от $C(t)$. The constructed model is based on the established long-term cointegration relationship between these time series and take the form:

$$DI_t = c_{d(t)} + \beta_{d(t),1} t + \beta_{d(t),2} DC_t + \eta_t,$$

where the values of the variable $d(t)$ indicate the class of epidemic states: $d(t)=1$ for the class "rising" and $d(t)=2$ for the class "decline".

All parameters of the models are unknown and are estimated using the EM (*Expectation-Maximization*) machine learning algorithm [8]. To correct the autocorrelation of residuals of the model, the algorithm proposed [9] is used.

Conclusions

Based on the obtained results of typology analysis (Figures 1–3 and Table), it can be concluded that in countries with the highest degree of intensity of the epidemic process, there is a greater decline in economic growth. It may be also supposed that the intensity of the epidemic process in each country is largely due to the ongoing anti-COVID state policy and effectiveness of anti-COVID measures.

The constructed econometric models are recommended to be used for modeling and forecasting the number of active infections within a single wave (VECM COVID-19 RB model) and for the entire period of observation of the epidemic (MS-LR-AR COVID-19 RB). The necessary condition for building these models is the stability and controllability of the epidemic process.

References

1. Cao Longbing, Liu Qing. COVID-19 Modeling: A Review // SSRN papers, 2021. URL: https://ssrn.com/abstract=3899127.

2. Kermack W., McKendrick A. Contributions to the mathematical theory of epidemics // Bulletin of Mathematical Biology. 1991. №53(1–2), pp. 33–55.
3. Kharin Yu.S, Valoshka V.A, Dernakova O.V, Malugin V.I, Kharin A.Yu. Statistical forecasting of the dynamics of epidemiological indicators for COVID-19 incidence in the Republic of Belarus // Journal of the Belarusian State University. Mathematics and Informatics. 2020. №3, pp. 36–50 (in Russian).
4. Jang S.Y., Hussain-Alkhateeb L., Rivera Ramirez, T. et al. Factors shaping the COVID-19 epidemic curve: a multi-country analysis // BMC Infect Dis 21, 2020. URL: https://doi.org/10.1186/s12879-021-06714-3
5. Worldometers.info [Electronic resource]. URL: https://www.worldometers.info/coronavirus. Date of access: 27.02.2022.
6. Malugin V.I., Hryn N.V., Novopoltsev A.Yu. Statistical analysis and econometric modelling of the creditworthiness of non-financial companies // Int. J. Computational Economics and Econometrics. 2014. Vol. 4(1/2), pp. 130-147.
7. The World Bank Group [Electronic resource]. URL: https://data.worldbank.org/. Date of access: 02.03.2022.
8. Malugin V. Statistical Estimation and Classification Algorithms for Regime-Switching VAR Model with Exogenous Variables / V. Malugin, A. Novopoltsev // Austrian Journal of Statistics. 2017. Vol. 46, pp. 47–56.
9. Malugin V.I. Discriminant analysis of multivariate autocorrelated regression observations under conditions of parametric heterogeneity of models // Informatics. 2008. №3(19), pp. 17–28 (in Russian).