



Research article

Modelling and predicting the spread of COVID-19 cases depending on restriction policy based on mined recommendation rules

Vitaliy Yakovyna¹ and Natalya Shakhovska^{2,*}

¹ University of Warmia and Mazury in Olsztyn, Poland

² Lviv Polytechnic National University, Ukraine

* **Correspondence:** Email: nataliya.b.shakhovska@lpnu.ua.

Abstract: This paper is an extended and supplemented version of the paper "Recommendation Rules Mining for Reducing the Spread of COVID-19 Cases", presented by the authors at the 3rd International Conference on Informatics & Data-Driven Medicine in November 2020. The paper examines the impact of government restrictive measures on the spread and effects of COVID-19. The work is devoted to the improvement of recommendation rules based on novel ensemble of machine learning methods such as regression tree and clustering. The dynamics of migration between countries in clusters, and their relationship with the number of confirmed cases and the percentage of deaths caused by COVID-19, were studied on the example of Poland, Italy and Germany. It is shown that there is a clear relationship between the cluster number and the number of new cases of diseases and death. It has also been shown that different countries' policies to prevent the disease, in particular the timing of restrictive measures, correlate with the dynamics of the spread of COVID-19 and the consequences of the disease. For example, the results show a clear proactive tactic of restrictive measures by example of Germany, and catching up on the spread of the disease by example of Italy. A regression tree and guidelines about influence of features on the spreading of COVID-19 and mortality due to this infection have been constructed. The paper predicts the number of deaths due to COVID-19 on a 21-day interval using the obtained guidelines on the example of Sweden. Such forecasting was carried out for two potential government action options: with existing precautionary actions and the same precautionary actions, if they had been taken 20 days earlier (following the example of Germany). The RMSE of the mortality forecast does not exceed 4.2, which shows a good prognostic ability of the developed model. At the same time, the simulation based on the strategy of anticipatory introduction of restrictions gives 2–6% lower values of the forecast of the number of new cases. Thus, the results of this study provide an opportunity to assess the impact of decisions about restrictive measures and predict, simulate the consequences of restrictions policy.

Keywords: clustering, classification, COVID-19, prognostic data, machine learning

1. Introduction

One of the largest epidemics in a hundred years is coronary disease COVID-19, caused by SARS-CoV-2: A single-stranded RNA-containing strain of SARS-CoV of the genus betacoronavirus, first genetically detected in 2019 in a sample of a patient with SARS pneumonia in Wuhan, China. According to the World Health Organization (WHO), the transmission of this virus from person to person happens by airborne droplets within the families of patients, which causes a fairly rapid spread of outbreaks. As of early 2021, the coronavirus COVID-19 is affecting 218 countries and territories around the world and 2 international conveyances [1]. Due to the rapid spread of cases, their wide territorial coverage and potentially serious consequences that threaten human life, the World Health Organization has declared the spread of COVID-19 a pandemic [2]. Thus, during the year of pandemic spread of the disease, more than 86 million cases of the disease were confirmed, which took lives of 1,864,996 people worldwide as of January 5, 2021 [1]. The first year of the pandemic passed without any specific treatment for the disease [3], and the main measures to reduce the spread of the epidemic were restrictive measures by governments to limit human contact and thus slow down the spread of this virus infection. Despite the start of COVID-19 vaccination, which began in various countries in the last days of 2020, severe restrictive measures and lockdowns are still widely used around the world, including most EU countries, due to the increase in new infections and the spread of new strains of the virus. However, despite a year of active efforts and measures to slow the spread of the epidemic, we still do not have a reliable answer to the question of what restrictive measures are needed to effectively reduce the spread of infection, what should be the scope and timing of these restrictions. Some measures that seemed to be effective in the first half of 2020 were found to be insufficient in the autumn of the same year and unable to effectively stop the epidemic (for example, the experience of Sweden).

In addition to medical issues, such as: which populations are most at risk, what is an effective treatment protocol, creating an effective and safe vaccine etc., there are a number of socially oriented questions, such as: how long will the epidemic last, what restrictive measures are most effective, what will be the dynamics of the disease, and how many people it will affect etc. [4]. The latter are particularly important for governments and health systems, as they enable them to plan restrictive measures with minimal harm to the economy and well-being of citizens, to plan and prepare health system to provide timely and effective care to those who are ill and thus reduce the number of victims of the pandemic. Therefore, governments and WHO face the difficult task of finding a fine line between preserving the economic and social life of a society and the life and health of each member of that society. And the numerous protests against restrictive measures that occur periodically in various parts of the world, despite the significant spending of governments to support sectors of the economy affected by the epidemic, are an example of this.

In addition to medical and biological research, modern information technology and algorithms can make a significant contribution to the fight against the epidemic by analyzing, modeling and predicting the spread of disease. Methods and tools of Artificial Intelligence [5], Machine Learning [6,7], and Big Data [8,9] can significantly accelerate progress in these areas (see, e.g., [5–9]). In addition, AI can make a significant contribution to the development of vaccines and specific drugs against COVID-19, as well as to the development of various expert systems that will be suitable for

combating the epidemic and its consequences.

As vaccination of the population is just beginning and the creation of effective collective immunity will take some time, quite a long time, at the moment "social distancing" and quarantine restrictions remain the only effective means of curbing the rate of infection. In such a situation, the role of modeling and predicting the COVID-19 spread, the maximum position and its duration is growing. The results of such modeling and predictions are the basis for sound governmental decisions concerning appropriate measures and resources allocation, including pulmonary ventilation systems or dedicated hospitals.

This article is an expanded and supplemented version of "Recommendation Rules Mining for Reducing the Spread of COVID-19 Cases", presented by the authors at the 3rd International Conference on Informatics & Data-Driven Medicine in November 2020, and published in CEUR Workshop Proceedings [10]. In [10] we built recommendation rules based on novel ensemble of machine learning methods such as regression tree and clustering. The aim of this work is to further improve the accuracy and predictive power of these production rules, to study them in Poland, Italy and Germany, to study the dynamics of movement between clusters, and their relation to number of confirmed cases and mortality rates caused by COVID-19.

2. Related work

The scale and consequences of the COVID-19 pandemic have led to a rapid increase in publications on both medical and biological aspects, as well as research on the spread, monitoring and control of the disease. A special service "Covid-19 Data Science" was created [11] which aims to report and balanced evaluation of research results, data sets, applications and models, and comments related to Covid-19 and its underlying virus, SARS-Cov-2. The volume of publications on this topic is so large that separate articles have appeared on the means of processing articles on the subject of the COVID-19 pandemic. For example, A. Doanvo et al. [12] proposed a fast and scalable machine-based framework that automatically parses literature related to a new disease and can help identify research overlap between COVID-19 and other coronaviruses, research hotspots, and areas warranting exploration. Case study performed by the authors of [12] indicates that COVID-19 publications have focused on public health, clinical care, and testing for coronaviruses, as opposed to the more limited number focused on basic microbiology, including pathogenesis and transmission.

The development of new effective models for predicting the number and dynamics of new COVID-19 infections [13–15], as well as epidemiological forecasting in general [16–18], is extremely important for the health care system, as it enables effective planning to eliminate or reduce possible epidemics. The main requirement for such models is the maximum possible accuracy and reliability of the forecast of epidemiological time series [15]. To meet these requirements, a number of AI-based models have been used for a number of years [19,20]. The methods used to solve the problem of constructing the COVID-19 prediction model have ranged from statistical autoregression like ARIMA to more robust machine learning methods [21–23].

A large group of articles is devoted to modeling and predicting the spread of coronavirus using various mathematical models and machine learning methods. Thus, in [24] an intelligent, adaptive analytical model of COVID-19 disease was built, and, based on deep learning, a data stream guided real-time incremental learning algorithm based on neural networks was proposed to estimate parameters of this model. Modeling of such a disease on such a scale is significantly complicated by the fact that data on the disease and its factors is constantly obtained and updated, and thus the

parameters of the model changeover time depending on external factors. The authors [24] developed a non-intrusive algorithm for adaptive incremental learning, which eliminates the need to re-learn or rebuild the model with each update of test dataset. This paper describes modeling of spread of the disease taking vaccination into account, which is based on Susceptible, Infected, Recovered, Vaccinated and Deceased states of individuals in a population going through an epidemic. A model validated on the basis of COVID-19 incidence in India [24] was used to study various epidemic control strategies. The authors analyzed different disease prevention strategies using a control theory approach. The authors considered three simple strategies to prevent the spread of the virus, including Herd Immunity, Complete Vaccination, and Complete Lockdown, as well as their proposed strategy Controlled Natural Immunization through risk based Population Compartmentalization. The authors of this study consider a strategy of controlled natural immunization through risk-based population compartmentalization an effective way to reduce the number of deaths in the risk group. This strategy involves isolating high-risk populations, while low-risk populations will develop immunity. It is believed that after the end of preventive isolation, the high-risk population will be surrounded by a sufficient number of immunized persons to prevent the spread of infection [24]. However, as we can see from the experience of Sweden, where the government introduced similar measures in the first half of 2020, such a strategy did not work. Another model, excluding vaccination (SIR - Susceptible, Infected, Recovered model), was considered in the article [25] for modeling the development of a pandemic in the first 100 days of the disease. E. Kaxiras, G. Neofotistos, and E. Angelaki consider a simple analytical model of the spread of infectious diseases, which is based on the numerical solution of the widely used SIR model. The model described in [25] contains three parameters: the position of the mid-point of the sigmoid representing the behavior of the susceptible population; the lag between the sigmoid that describes the recovered population and the sigmoid of the susceptible population; and the number of daily cases near the peak of the curve. The authors modified the original model and presented an approximate numerical solution of the differential equations of the model. In [25], they used the built model with their defined parameters to predict the total number of confirmed cases of infection in 50 countries during the first 100 days (until April 28, 2020), also they used it to predict the date of reaching the maximum number of infected. However, the proposed model could not describe precisely enough the time series for other countries contained in the database used by the authors, which in [25] is explained by the fact that either the data was too noisy or did not reach the maximum curve, and therefore this model was not able to provide a good fit. As will be shown later, in the main part of the article, our proposed approach based on machine learning methods does not have such shortcomings and satisfactorily describes the behavior of the spread of infection in different countries. Among the important limitations of the studied SIR-model, the authors [25] also include the impossibility of describing and predicting the mortality rate, which is explained by the need to take into account additional information such as capabilities of health care system of each country. Another limitation of such model is the prediction of only the lower limit of cases, which can potentially be eliminated by using nonlinear terms in the differential equations describing the model [26].

In [29] is devoted to predict of mortality of patients with severe COVID-19. In this work, using multivariate regression analysis, the predictors and factors influencing the mortality and clinical condition of patients with COVID-19, as well as the necessary conditions for the development of treatment algorithms for such patients were investigated. In this study, it was found that factors as the presence of any comorbid disease, elevated C-reactive protein levels, and the presence of dyspnea significantly increase mortality in moderate to severely ill Covid-19 patients.

Thus, Ribeiro et al. [21] consider different regression-based models like cubist, ridge, reference vector regressions, ARIMA along with random forest, and ensemble overlay training for one-, three-, and six-days prediction of cumulative value of COVID-19 cases in Brazil. They demonstrate that, in most cases, support vector regression and stack ensemble training perform better against the accepted criteria than the other models studied. In general, the models developed in [21] can generate accurate forecasts with an error of less than 6.90%. In [21], it was concluded that further improvement in the development of COVID-19 prediction models can be achieved by combining deep learning and ensemble learning accumulation, by using multi-objective optimization for tuning the hyperparameters of prediction models and adopting a set of features that allows to explain the dependencies of future COVID-19 cases.

Artificial neural networks and neuro-like structures are successfully used in prediction problems, including multifactorial and time series prediction, the dynamics of COVID-19 propagation in particular. For example, in [27] a method of increasing the accuracy of model prediction by creating a non-iterative supervised learning predictor is presented. Such a predictor is based on the neural-like structure of the successful geometric transformations model and Ito decomposition. The use of this method to predict the cost of health insurance has allowed the authors [27] to significantly improve the accuracy of forecasting compared to other methods, including linear regression, multilayer perceptron or SVM. In [27] it was concluded that the developed method, due to the improved properties of generalization, can be successfully used in various subject areas to improve the accuracy of processing large amounts of data. While most publications are on modeling and predicting the spread of the coronavirus pandemic, not many articles explore hybrid methods. One such approach was used by Najmul Hasan in [28] to predict the COVID-19 epidemic using a hybrid model that combines empirical mode decomposition and artificial neural networks. This work used COVID-19 time series data for the period from January 22 to May 18, 2020. The first step of the study involved pre-processing and noise reduction of the input data using ensemble empirical mode decomposition, after which the resulting noiseless dataset was used to train an artificial neural network with one hidden layer. In [28], ANN configuration trained using the Levenberg-Marquardt algorithm with 20 hidden neurons was used. The obtained results were compared with the results obtained by the methods of traditional statistical analysis, such as regression analysis and moving average review with three trend categories (linear, exponential and polynomial). The author [31] shows a significant advantage of his approach compared to traditional statistical analysis, which once again emphasizes the importance and prospects of using AI methods and tools to analyze and predict the dynamics of complex processes, such as the spread of COVID-19. The results and forecasts obtained by such methods give governments and health care organizations an opportunity to take preventive measures by understanding the upcoming COVID-19 situation for better healthcare management.

More complex multifactor COVID-19 prediction models are considered by K.C. Santosh in his article "COVID-19 Prediction Models and Unexploited Data" [29]. This paper argues that existing models based on both the widely used SEIR/SIR or curve-fitting epidemiological models and the machine-based models are unable to describe some key issues that have never been used before, such as hospital resources, the number of daily tests for coronavirus, demographic and social characteristics of certain areas, population density, population distribution by risk groups, etc. Santosh proposes to introduce time-dependent mathematical models with dynamic self-adjusting parameters. One of the most common approaches to modeling and predicting the spread of the COVID-19 pandemic is to use a variety of time series forecasting techniques ranging from classical ARMA/ARIMA models to

machine learning techniques from simple regression models to deep learning technologies [30]. Sengupta et al. conducted data analysis on pandemic data at the global level, and then used the simulation results to analyze the situation in India. The focus of their study was to predict the position and duration of the maximum incidence. At the same time, their model took into account demographic data, such as gender, geographical location of the region, age, etc. Time series forecasting techniques were used for modeling and forecasting, which included linear and polynomial regression, SVM, and LSTM (Long short-term memory). Using the k-means method in [30], the data of confirmed cases of COVID-19 were clustered by individual countries and three clusters were identified: the first with a high mortality rate and a low recovery rate; the second with a low mortality rate and a high recovery rate; and a third with average levels of both mortality rate and recovery rate. According to Sengupta et al., classical forecasting methods such as ARIMA, Facebook Prophet forecasting model and polynomial regression show better results compared to models based on machine and deep learning methods. In [30], it is argued that temporal data on the spread of coronavirus infection in India are best described by a Sigmoid model, which makes it possible to predict the peak incidence.

Basu and Campbell [13] propose a model based on Long Short-Term Memory, which was trained on cumulative number of COVID-19 cases and deaths separately. They train the developed model using the dataset containing number of confirmed COVID-19 cases during the period of more than four months. The parameters of the proposed model can be adjusted in such a way as to provide the required forecasting accuracy. In the article [13] using the developed LSTM-model, a quantitative assessment and analysis of the impact of measures taken by various US counties on the spread of the disease has been carried out.

Hu et al. [16] built and investigated a machine learning model based on composite autoencoders to describe the spread of coronavirus in China in April 2020 based on data from previous periods. They clustered cities and provinces based on the characteristics obtained from the developed auto-coding model. Similar LSTM-based models have been used to study patient-related statistics [17], as well as to identify trends in the spread of the epidemic in China [18].

The study by S. Tuli et al. [15] is based on an improved machine learning model for analyzing and predicting the spread of the COVID-19 epidemic in different countries. The authors of [15] have developed and studied the Robust Weibull model with iterative weighting. They conclude that the developed Weibull model is statistically better than the Gaussian baseline model for prediction of the COVID-19 outbreak. It was shown [15] that simple Gaussian model results in an overly optimistic COVID-19 spreading scenario. The developed model was deployed on a cloud computing platform using the FogBus framework for more accurate and realistic forecasting of the epidemic's growth dynamics.

3. Materials and methods

3.1. Dataset description

In this study we have used data from the Oxford COVID-19 Government Response Tracker as well as European Centre for Disease Prevention and Control (ECDC) Covid-19 Cases [31]. Dataset WPP2019_TotalPopulationBySex was taken from [32]. It contains information on population by year and population density. All data is filtered for 2020.

The Oxford COVID-19 Government Response Tracker (OxCGRT) dataset provides information

on which governments took action, what action they did, and when they took it. The OxCGRt systematically gathers information on a range of common policy responses by a given government and determines the extent to which the government is implementing these measures. The respective scores are combined into a set of policy indicators.

The second dataset contains new public data on the geography of COVID-19 cases worldwide from European Centre for Disease Prevention and Control. Each line or record contains the number of new cases per day, by country or region. These data have been reduced to the number of cases per 10,000 habitants using the United Nations population dataset [32]. Unfortunately, available datasets did not contain information about number of tests taken, only the number of new cases per day. The mortality rate was used as the relative value. Datasets were grouped by country name, and the number of confirmed cases was normalized by dividing the population by country. Dates were changed to days from the beginning of the disease, and the analysis in all countries began on the same day, when the normalized number of confirmed cases became higher than 0.001. The part of characteristics obtained from these datasets are shown in Table 1. This table presents the list of restrictions. Totally dataset consists of 38 features.

Table 1. Variables and their meaning used as attributes in the study.

Variable	Meaning
c1	School closing
c2	Workplace closing
c3	Cancel public events
c4	Restrictions on gatherings
c5	Close public transport
c6	Stay at home requirements
c7	Restrictions on internal movement
c8	International travel controls
e1	Income support
h1	Public information campaigns
h2	Testing policy
h3	Contact tracing

For the performed analysis we have merged the mentioned datasets and considered data from January 01, to May 31 for building the models, while the data from June 01 to August 31 have been used for verification of the forecasting based on the developed recommendation rules.

3.2. Data preprocessing

Dataset consists of attributes with different nature. Because obtained dataset contains categorical and numerical variables at the stage of preprocessing one hot encoding is applied to categorical variables. The noise data and outliers are presented too. That is why at the first stage the following steps are required: feature selection, empty data analysis, data normalization and scaling.

Feature selection is made based on theory of information. The joint mutual information between each feature and target attribute ConfirmedCases is calculated as:

$$JI(f_i) = \sum_{f_j \in S} (I(f_i; c) + I(f_i; c \vee f_j)),$$

where c is target class, $I(f, c)$ is mutual information, f_j is already selected feature, f_i is processed feature.

The Cross-Validated Committees Filter (CVCF) based on information gain and decision tree algorithm was used for feature selection. CVCF is implemented in CRAN package in RStudio. The authors of CVCF place special emphasis on using ensembles of decision trees [33]. The main steps of the algorithm are the following: (1) split the training dataset using an nf -fold cross validation scheme, (2) for each of these nf parts, a learning algorithm is trained on the other $n-1$ parts, resulting in n different classifiers. Here, C4.5 is used as the learning algorithm, (3) these n resulting classifiers are then used to find the most common attributes in trees' roots.

The list of attributes after features selection consists of all restrictions (Table 1), `country_name`, `confirmed_cases`, `confirmed_death`, `e1_income_support`, `h1_public_information_campaigns`, `h2_testing_policy`, `h3_contact_tracing`. As result, 19 from 38 features were selected.

The next step is empty data analysis. The dataset consists of 66,998 rows, 26,003 of them have empty values in selected attributes. Due to small quantity of rows with empty data these rows were eliminated. In addition, binning was made as well.

Although various nonlinear models such as artificial neural networks or SVMs have recently been widely used in various subject areas, many questions remain about their use with noisy or incomplete data (see e.g. [34,35]). One of common methods of overcoming noisy data is the use of decomposition methods that improve the accuracy of neural network prediction by creating clear input signals [35,36]. Among the effective methods of fighting missing data is the original approach described in [34]. Izonin et al. [34] proposed to use an ensemble of neural networks to solve the problem of missing data, in which the first network predicted the required value, and the second improved the error of the first neural network. Improving the quality of missing data in this work was achieved due to the additional use of an extended-input SGTm neural-like structure to form a polynomial response surface of the predicting value.

The authors propose to use Multivariate imputation by chained equations (MICE) algorithm for missing data imputation (package `Mice` in RStudio) [37]. This is sequential regression multiple imputation technic. MICE works well on missing at random data. But it is possible to use this method for not random data imputation.

The results of MICE are given on Figure 1. Totally 89 rows were omitted the rest of missed data was imputed.

It is desirable to bring all input variables to a single range and normalize (the maximum absolute value of input variables should not exceed one). Otherwise, errors due to variables varying over a wide range will be more influential than errors due to variables varying over a narrow range. By ensuring that each feature changes within the same range, we ensure that each has an equal effect. Therefore, the input variables, as a rule, are scaled so that the variables change in the range of the function, as a rule, $[0,1]$ or $[-1,1]$. The Softmax scaling is used. The distribution of dataset for each feature is shown in the Figure 2.

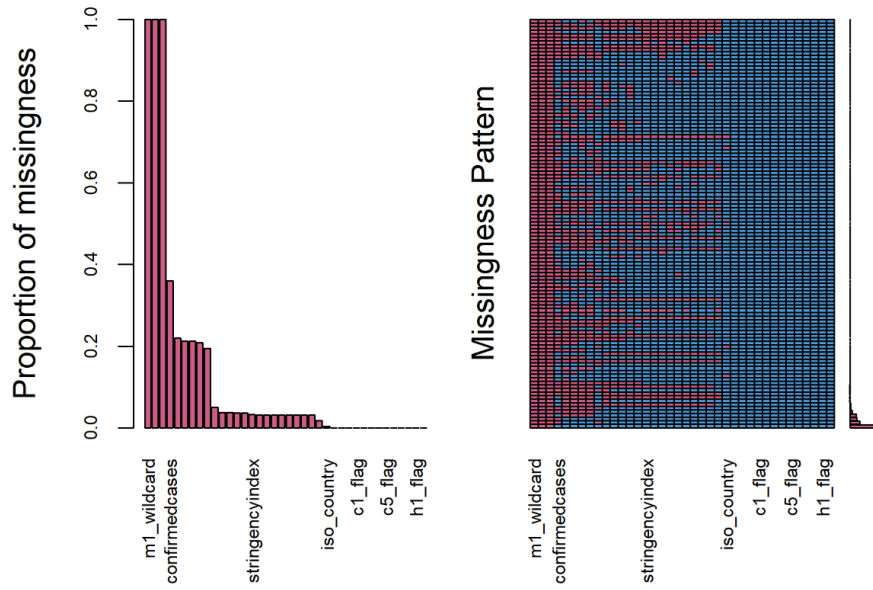
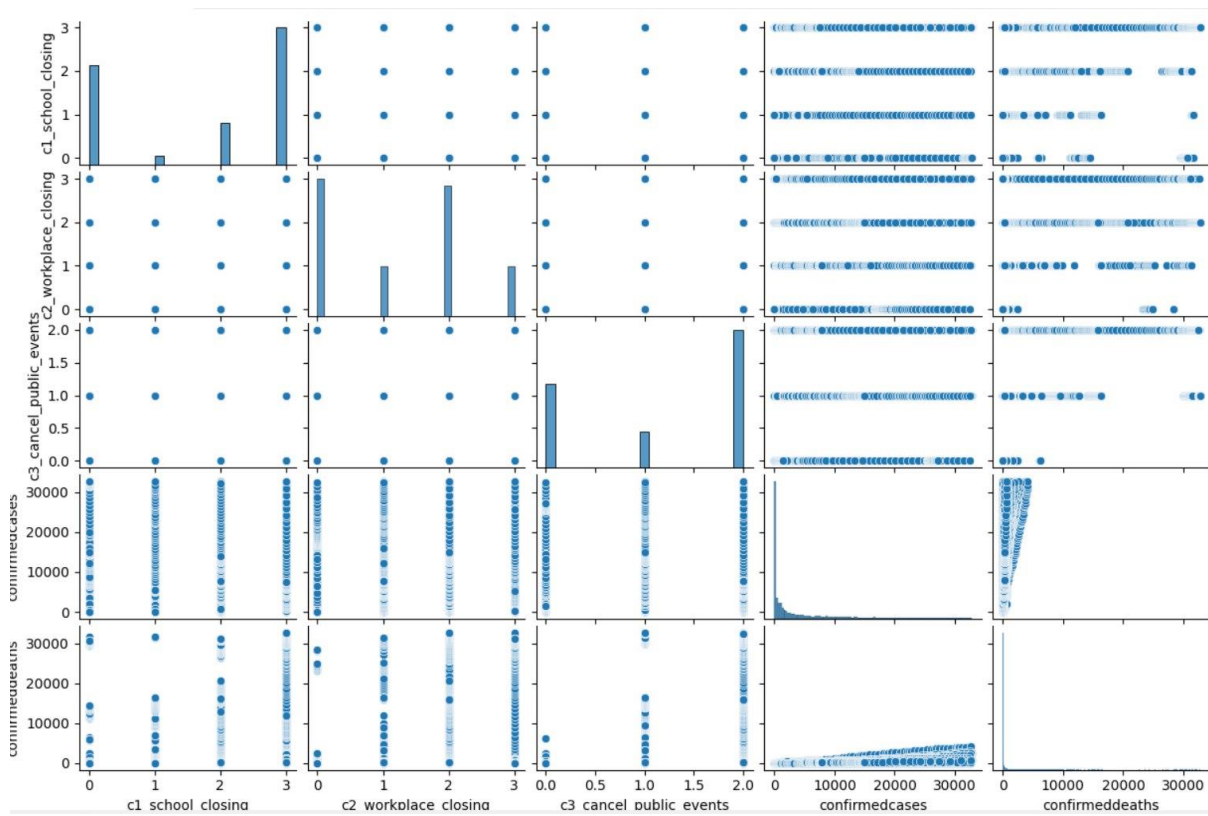


Figure 1. The results of MICE.



(a)

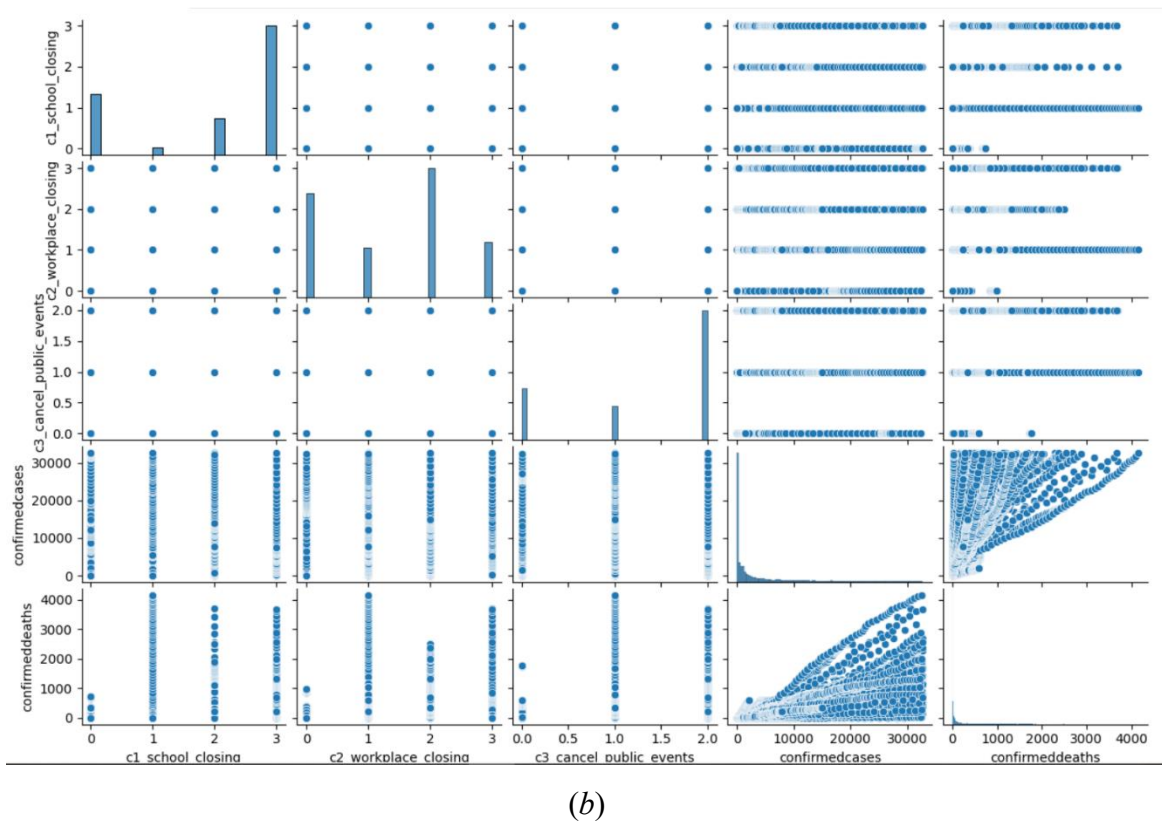


Figure 2. Data distribution before (a) and after (b) preprocessing.

4. Results and discussion

In the first stage of our study, we clustered the data and built a regression decision tree [10]. K-means method was used for clustering. The gap-statistics allows to find the appropriate number of clusters. Cluster centroids allow to find “average” object in each group and to create the regularization rules. The cluster # 2 and #3 shows countries under restriction, cluster #1 consists of countries without almost any restriction (see Table 2, Table 3, and Table 4) [10].

Table 2. Feature weights for the three clusters obtained.

Cluster	School closing	Workplace closing	Cancel public events	Close public transport
1	0.04368254	0.02120635	0.07822222	0.005587302
2	2.89865433	2.26230025	1.94790791	1.109703532
3	1.88113534	1.02194531	1.30058844	0.256905504

Table 3. Feature weights for the three clusters obtained cd.

Cluster	Stay at home req.	Restrictions on internal movement	International travel controls
1	0.0351746	0.01853968	1.137778
2	1.7232443	1.68355761	3.438709
3	0.5112496	0.50931118	2.850190

Table 4. Feature weights for the three clusters obtained cd.

Cluster	Income support	Public information campaigns	Testing policy	Contact tracing
1	0.02933333	0.7869206	0.6651429	0.6673016
2	1.00667578	1.9615749	1.5702271	1.3380467
3	1.17223953	1.9540325	1.8985808	1.6339218

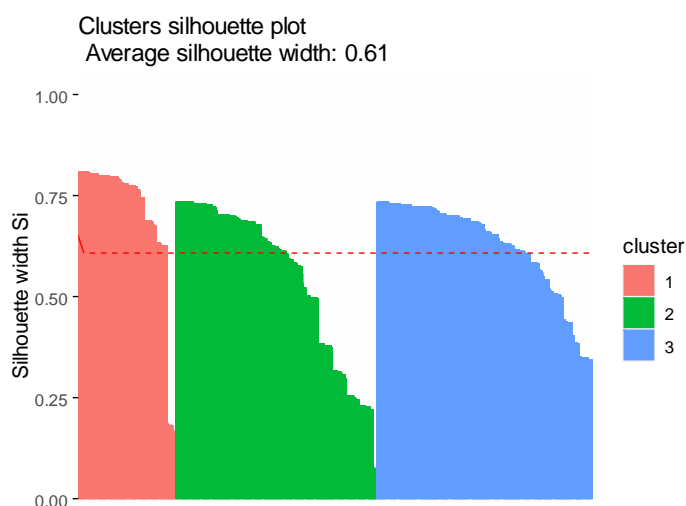
The clusters obtained differ by the most distinct recommendations, which can be summarized as follows [10]:

- Cluster #1: control international travels,
- Cluster #2: recommended to close schools, recommended to close workplace recommended to stay at home, Restrictions on internal movement, International travel controls?
- Cluster #3: income support, testing policy, International travel controls, recommended to stay at home and cancel public events.

Given above conclusions are formed on the statistical analysis of each restriction in separated cluster. For example, Stay_at_home restriction is frequently appeared in Cluster 2 (Table 3).

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The Silhouette parameters of the clusters are given below (Figure 3).

cluster	size	ave.sil.width
1	1400	0.70
2	2897	0.54
3	3071	0.62

**Figure 3.** Silhouette plot.

To build the recommendation rules, the regression decision tree was developed in [10]. Three circumstances are key here:

1. the rules make it possible to implement sequential dichotomous data segmentation,
2. two objects are considered similar if they appear in the same segment of the partition,

3. at each step of the partition, the amount of information about the variable under investigation (response) increases.

The main feature of the algorithm is its k -arc structure. Branching on a chosen trait x splits training objects on k subsamples, where k is the number of different characteristic values.

Without loss of generality, we assume that the feature x has values from $\{0, 1, \dots, k-1\}$, $k \geq 2$. In this case, when constructing of the decision tree, from the vertex x there are k arcs labeled with numbers from $\{0, 1, \dots, k-1\}$.

Let σ be the label of one of the arcs leaving the vertex x , $\sigma \in \{0, 1, \dots, k-1\}$. To form a new current subset of objects and a new current set of features, those objects from T are deleted for which the value of feature x is not equal to σ , and also feature x itself is removed from the set of features.

Let v be a hanging vertex generated by a branch of a tree with the inner vertices x_{j_1}, \dots, x_{j_r} and let the arc outgoing from the vertex x_{j_i} , $i \in \{1, \dots, r\}$, be labeled σ_i . Further let $T(v)$ is the current set of objects that hit the vertex v . The vertex v is associated with a pair $(B, w(v))$, where $w(v)$ is equal to the mean value of the target variable over all objects from $T(v)$, and B is an elementary conjunction of the form $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$. If the vertex v is not pendant, then we assign to it the conjunction

$$B = x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r} \dots$$

Truth interval of elementary conjunction B denote by NB . Let S be a recognizable object. For each hanging vertex $(B, w(v))$, a check is performed that the description of the test object belongs to the truth interval NB . If the description S belongs to NB , then the object S is associated with the value of the target variable $w(v)$. Object S is assigned the value of the target variable

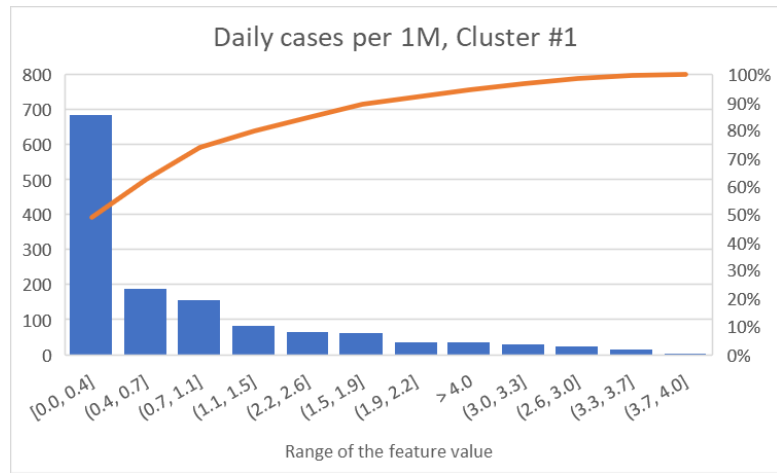
$$W = \frac{\sum_{i=1}^l w(v_i) * I_{B_i}}{\sum_{i=1}^l I_{B_i}},$$

where

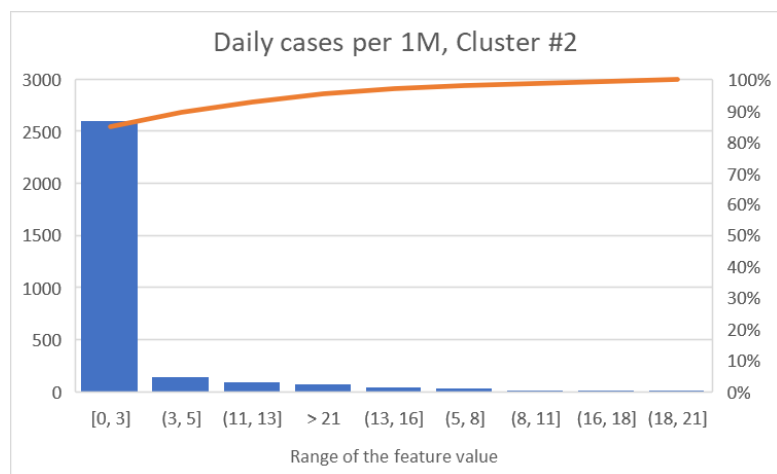
$$I_{B_i} = \begin{cases} 1, & \text{if } S \in N_{B_i}, \\ 0, & \text{otherwise.} \end{cases}$$

The influence of such clustering on COVID-19 spreading, peak position and duration as well as mortality rate will be the subject of our further study. The same country can be joined to different clusters in different time slots. So, the clustering by country will not be so unambiguous. That is why time series for separated country can be interesting and will be the subject of the further study. Besides, the influence of such clustering on COVID-19 spreading, peak position and duration as well as mortality rate will also be the subject of our further study.

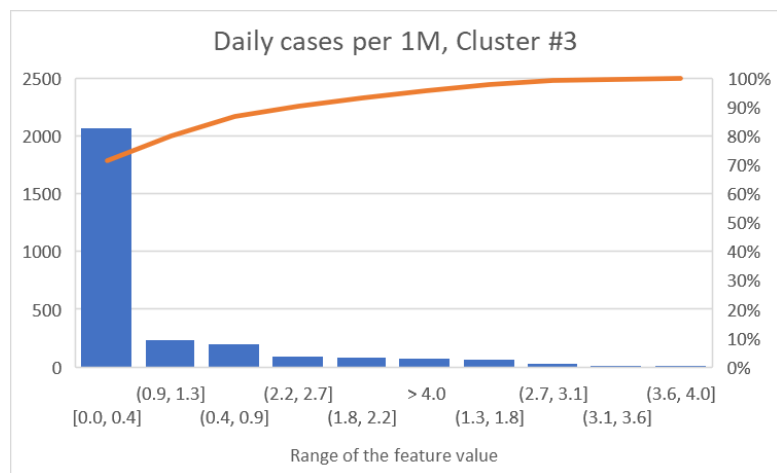
As further research actually deals with time series that reflect the dynamics of confirmed cases of COVID-19 for different countries and the dynamics of changes in the membership of certain countries in clusters, common methods were used to analyze time series. At this stage, forecasting was performed for individual clusters. The histogram of the frequencies of confirmed cases of COVID-19 per 1 million populations for different clusters is shown in Figure 4, and the distribution of mortality values for different clusters – in Figure 5 respectively.



(a)



(b)

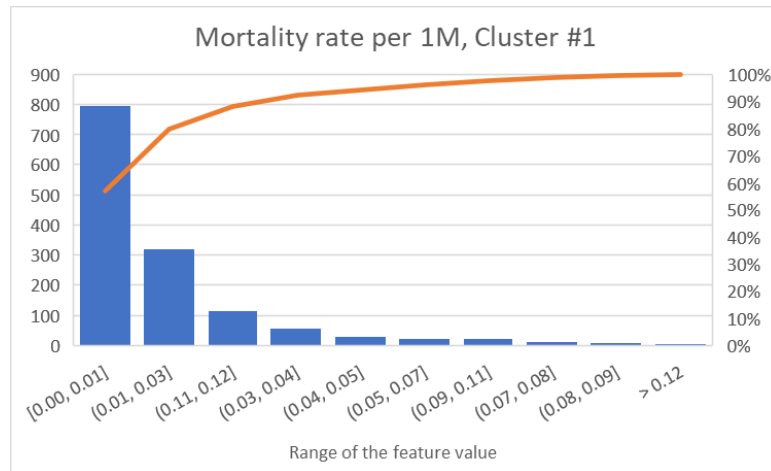


(c)

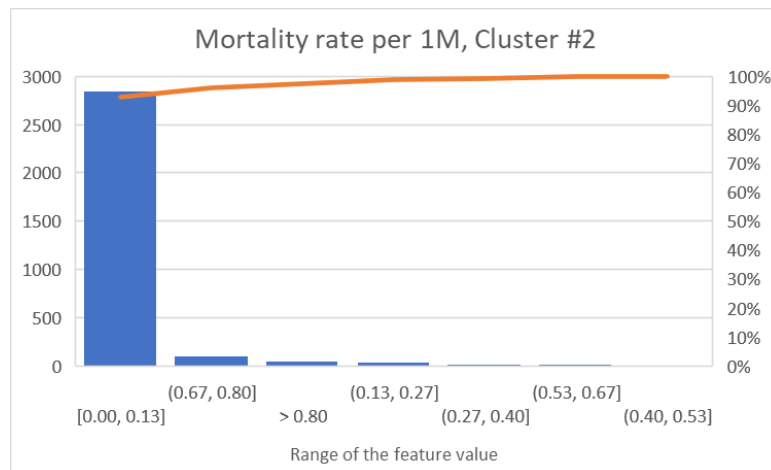
Figure 4. Histogram of frequencies of confirmed cases of COVID-19 by clusters: (a) the first cluster; (b) the second one; (c) the third.

In the analysis of Figures 4 and 5, it should be kept in mind that the total amount of data

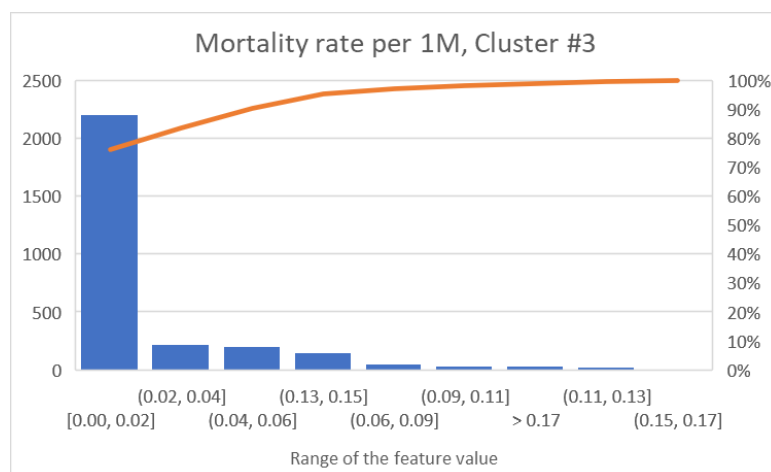
corresponds to the number of days, not the number of countries, and countries could migrate from one cluster to another during the history of observations.



(a)



(b)



(c)

Figure 5. Histogram of COVID-19 mortality rates by clusters: (a) the first cluster; (b) the second one; (c) the third.

The median values for each cluster were 0.393, 0.267, and 0.114 new daily confirmed cases of infection per 1M population, respectively, and 0.00861, 0.00714, and 0.00306 daily deaths due to COVID-19 per 1M population. Thus, for countries belonging to the third cluster, the spread of the epidemic is much slower than for the first and second clusters, although the reduction in mortality rate from the first cluster to the third is smaller than the reduction in the number of newly confirmed cases of COVID-19 (2.8 times vs. 3.4 times) and such behavior requires further research. From the analysis of Figure 3, we can distinguish the second cluster in terms of the distribution of values of new cases of COVID-19. Although the median daily diurnal infection is lower than for the first cluster, more than 80% of new diurnal cases are in the range of up to 3 per 1 million populations, while for the first and third clusters this figure is 0.4 new diurnal cases per cluster. 1 million populations. That is, in the case of the second cluster, the variability of this variable is higher, which makes it possible to consider this cluster as a transitional stage between the first and second, which are probably quite stable. The analysis of Figure 5 does not give such a clear relationship: for all clusters, 90% of the values are within 0.13 deaths per 1 million populations, with, however, significantly different median values (see above).

All countries were divided into 2 categories: countries that were mostly in the same cluster all the time, and countries whose clusters changed over time. From the group of countries that were in different clusters during the study, 3 countries were randomly selected (Germany, Italy, Poland) similar in demographic, geographical and social characteristics, but different in terms of epidemic development and investigated the impact of government measures (see Table 1) on change of the cluster of the country (Figures 6–8, respectively), which corresponds to the change in the dynamics and consequences of the incidence of COVID-19 (see Figures 4 and 5).

From the analysis of Figures 4, 5 it can be concluded that there is a clear relationship between the cluster number and the number of new cases of morbidity and death. There are also different countries' policies to prevent the disease, which have different effects on the spread of COVID-19 (Fig. 6–8). For example, in Germany, proactive tactics are pronounced, while in Italy – catching up on the nature of the spread of the disease. The next step in the study is to predict the number of new cases of COVID-19 on the example of Sweden (the country not studied at the previous step). Such forecasting was carried out for two potential options for government action: with existing preventive actions and the same preventive actions, if they were taken 20 days earlier (following the example of Germany). Thus, we get the answer to two questions:

- (i) whether our model has prognostic power, if the prognosis corresponds to the real picture of the disease;
- (ii) simulating whether the incidence could have been reduced if the same measures had been introduced earlier.

If both questions are answered in the affirmative, the model and conclusions of this work can be used by different governments as one method of making informed decisions about imposing certain restrictions in order to prevent the spread of the COVID-19 pandemic and reduce the severity of the disease.

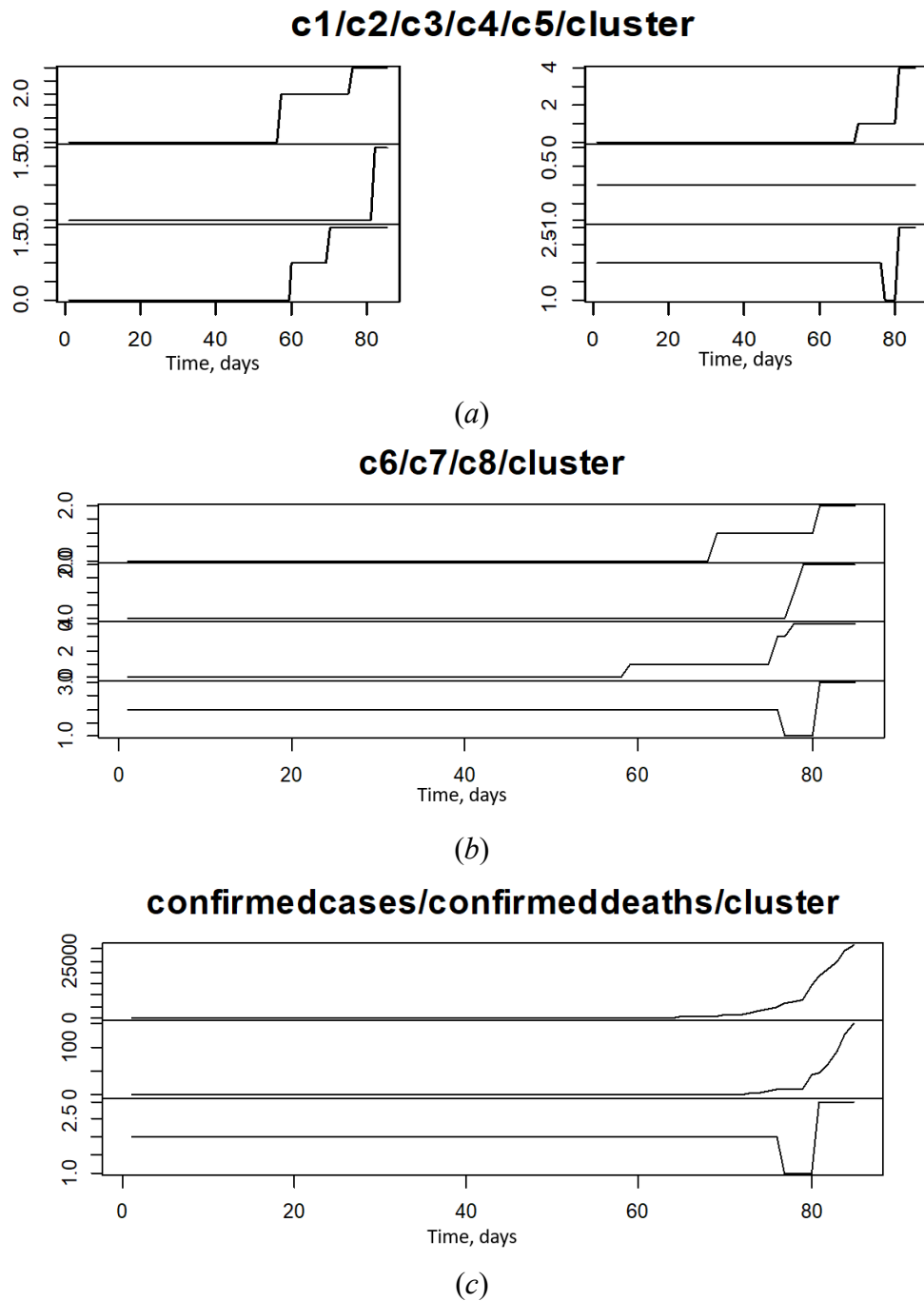


Figure 6. Dynamics of cluster change with respect to different features for the case of Germany: (a) features c1–c5 and cluster number; (b) features c6–c8 and cluster number; (c) confirmed number of diseases and mortality rate.

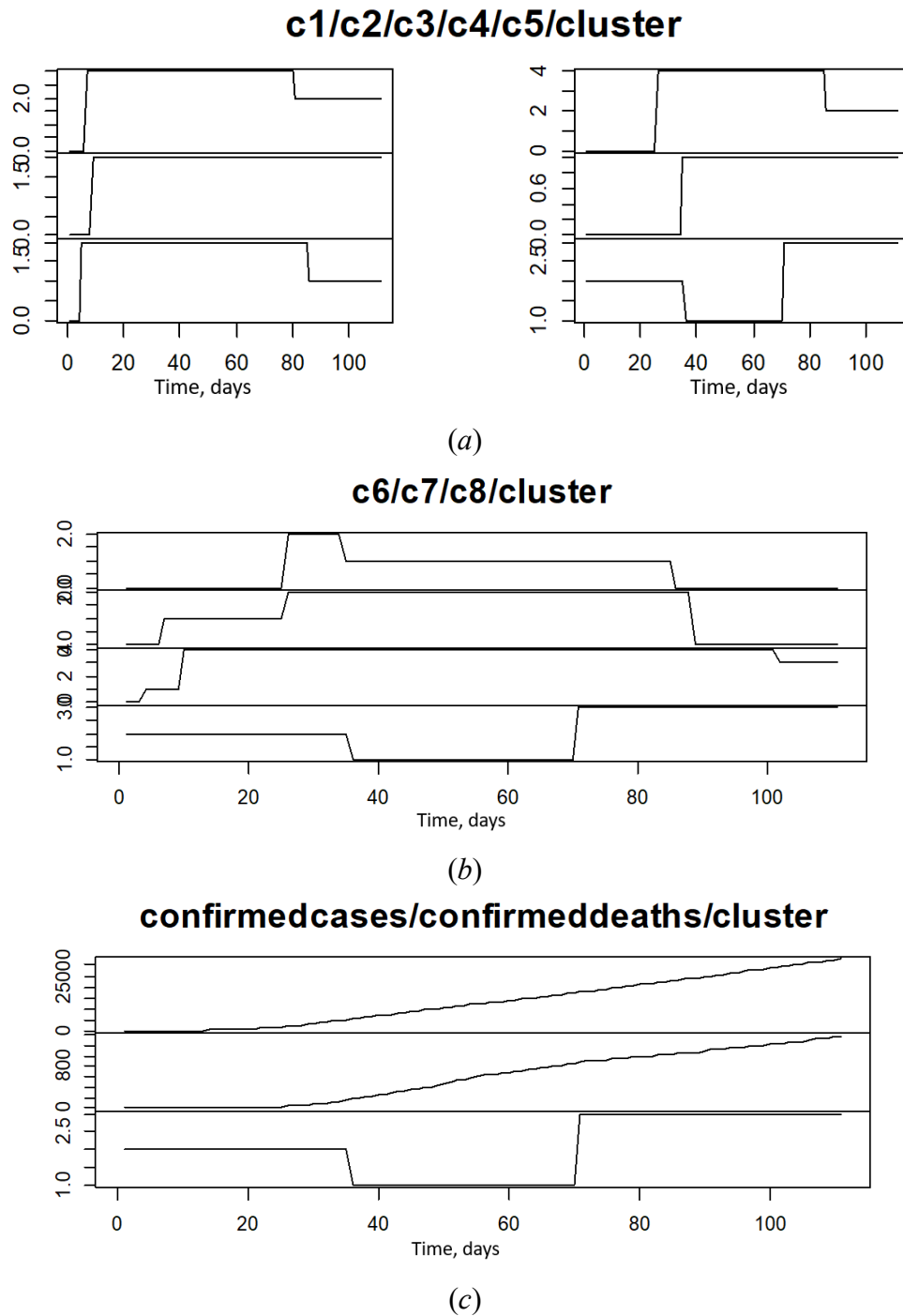


Figure 7. Dynamics of cluster change in relation to different features for the case of Poland: (a) features c1–c5 and cluster number; (b) features c6–c8 and cluster number; (c) confirmed number of diseases and mortality rate.

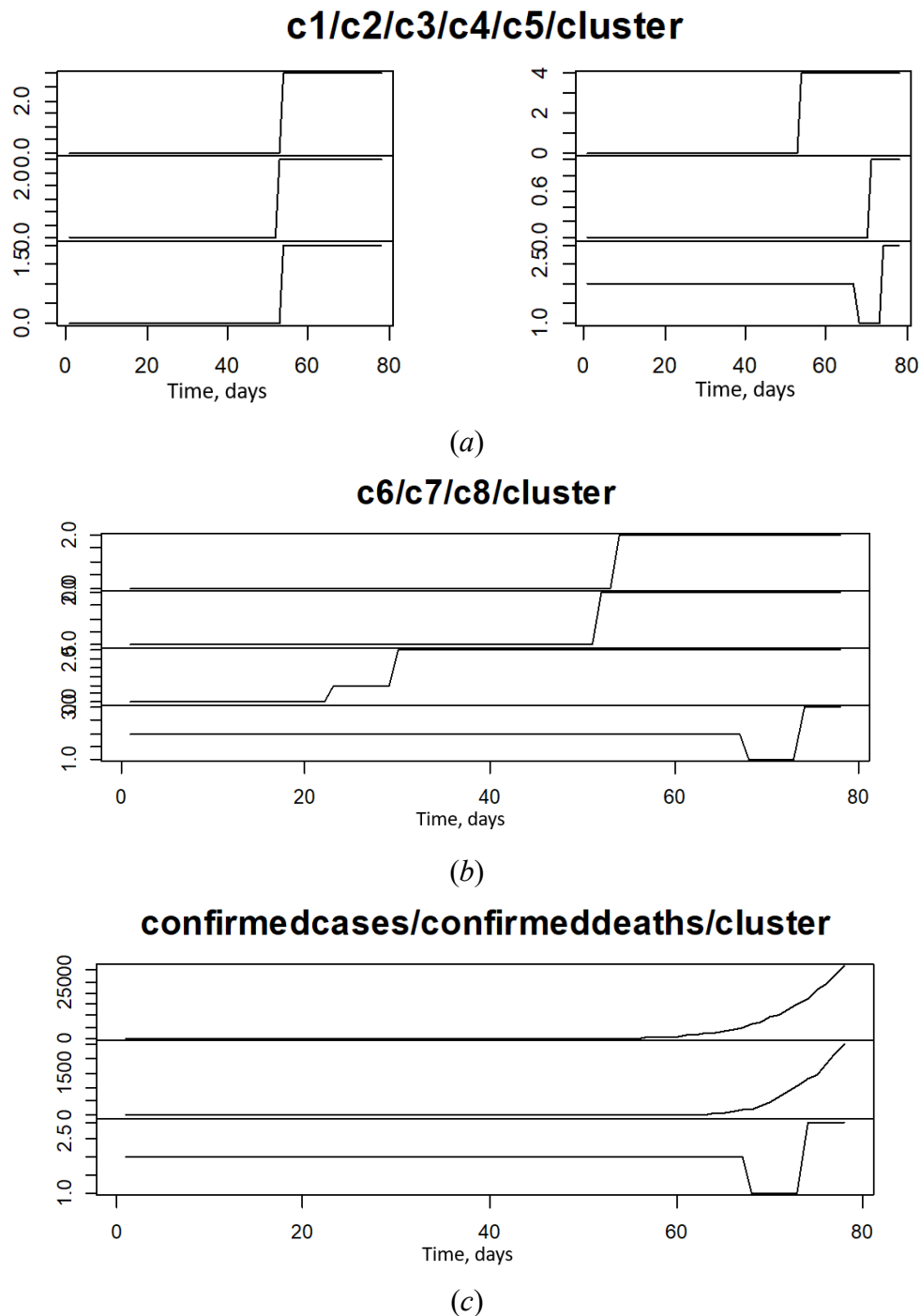


Figure 8. Dynamics of cluster change with respect to different features for the case of Italy: (a) features c1–c5 and cluster number; (b) features c6–c8 and cluster number; (c) confirmed number of diseases and mortality rate.

To build a forecast model, we construct a regression tree. Incidence data for the period from January 1 to October 31, 2020 were used to construct the tree, while data for November and December 2020 were used to check the quality of the forecast. Table 5 shows the importance of variables for random forest prediction. %IncMSE is the most robust and informative measure. It is the increase in MSE of predictions (estimated with out-of-bag-CV) as a result of variable j being permuted (values

randomly shuffled). IncNodePurity relates to the loss function which by best splits are chosen. The loss function is MSE for regression and gini-impurity for classification. More useful variables achieve higher increases in node purities, that is to find a split which has a high inter node 'variance' and a small intra node 'variance'. IncNodePurity is biased and should only be used if the extra computation time of calculating %IncMSE is unacceptable.

Table 5. Degree of importance of features for random forest method.

Feature	%IncMSE	IncNodePurity
date	1762329	60371343113
c1_school_closing	20312295	112539124311
c2_workplace_closing	22704695	155417997013
c3_cancel_public_events	17357035	84284697645
c4_restrictions_on_gatherings	21327188	133060453250
c5_close_public_transport	19265937	114212676197
c6_stay_at_home_requirements	19246615	100860861469
c7_restrictions_on_internal_movement	22388811	95209240436
c8_international_travel_controls	30728123	152711818057
e1_income_support	31968524	195651691579
h1_public_information_campaigns	3670697	19353642670
h2_testing_policy	30246687	155123828287
h3_contact_tracing	16762839	91517681970

The constructed rules have the following form, and the resulting regression tree is shown in Figure 9.

- 1) root 41602 2.144700e+12 4344.2440
- 2) e1_income_support < 0.5 18045 3.451650e+11 1577.5170
 - 4) c2_workplace_closing < 0.5 9754 4.310477e+10 504.8472 *
 - 5) c2_workplace_closing >= 0.5 8291 2.776335e+11 2839.4650 *
- 3) e1_income_support >= 0.5 23557 1.555594e+12 6463.5960
 - 6) c8_international_travel_controls >= 3.5 8874 4.392391e+11 4550.0060
 - 12) c5_close_public_transport < 0.5 3428 8.620474e+10 2583.1110 *
 - 13) c5_close_public_transport >= 0.5 5446 3.314248e+11 5788.0730
 - 26) h2_testing_policy < 1.5 2573 7.956870e+10 3627.6140 *
 - 27) h2_testing_policy >= 1.5 2873 2.290908e+11 7722.9360 *
- 7) c8_international_travel_controls < 3.5 14683 1.064221e+12 7620.1170
 - 14) c2_workplace_closing < 0.5 2049 5.259635e+10 3436.5260 *
 - 15) c2_workplace_closing >= 0.5 12634 9.699460e+11 8298.6180
 - 30) e1_income_support < 1.5 3462 1.935803e+11 5760.5320 *
 - 31) e1_income_support >= 1.5 9172 7.456460e+11 9256.6270
 - 62) h2_testing_policy < 1.5 3278 2.321877e+11 7107.6480 *
 - 63) h2_testing_policy >= 1.5 5894 4.899009e+11 10451.8000 *

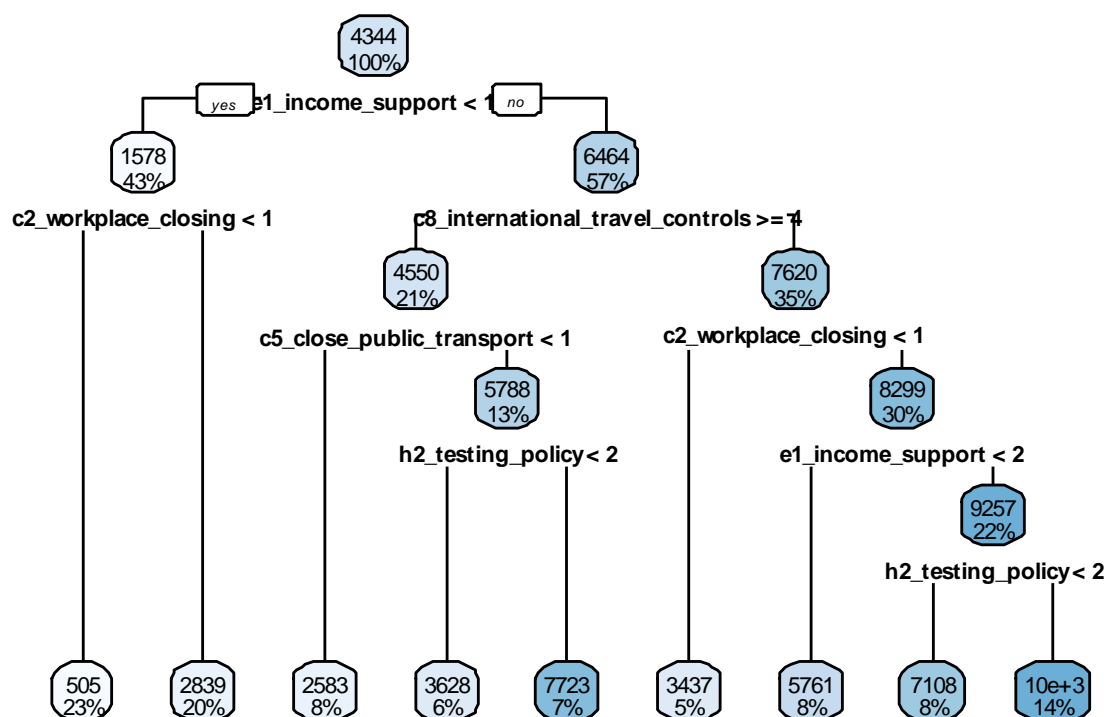


Figure 9. The resulting regression tree for predicting the incidence of COVID-19 depending on government restrictive measures.

The number of new cases due to COVID-19 was predicted for an interval of 21 days. To verify the prognosis, data on new cases from March 1, 2020 to March 21, 2020 were used. Table 6 shows the actual data projected taking into account the current strategy of imposing restrictions, as well as a simulation for the case of anticipatory restrictive measures.

We used Root Mean Square Error (RMSE) implemented in RStudio (<https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/>). RMSE of the new cases forecast in Sweden for 21 consecutive days from March 1 to March 21, 2020, made on the basis of production rules obtained from the constructed regression tree. The larger the RMSE, the larger the difference between the predicted and observed values, which means the worse a regression model fits the data. Conversely, the smaller the RMSE, the better a model is able to fit the data. RMSE is calculated for renormalised confirmed cases. It means the actual value range is thousands. As can be seen from Table 6, the biggest absolute difference between real data and predicted data for 16-03-20 is 17 or relative error is 0.0006. RMSE is equal 4.2.

At the same time, the simulation based on the strategy of anticipating the introduction of restrictions gives the value of the new cases forecast less than 2% (forecast for 7 days) to 6% (forecast for 21 days). Thus, the results of our clustering of data, the study of the dynamics of transition between clusters, and its impact on the spread and consequences of the COVID-19 pandemic, as well as a forecast model based on production rules based on the regression tree, allow us to assess the impact of decisions on restrictive measures and to predict and simulate the consequences of a policy of imposing restrictions.

Table 6. Forecast and actual number of new cases due to COVID-19 for Sweden in the period from March 1 to March 21, 2020

Date	Real data	Forecast for the current strategy of imposing restrictions	Simulation for the case of anticipatory restrictive measures	Deviation of simulation from real data, %
01-03-20	21603	21612	21603	0
02-03-20	22135	22142	22135	0
03-03-20	22434	22438	22210	1
04-03-20	22695	22697	22468	1
05-03-20	23172	23178	22940	1
06-03-20	23829	23832	23591	1
07-03-20	24574	24579	24083	2
08-03-20	25358	25355	24851	2
09-03-20	26058	26049	25537	2
10-03-20	26567	26568	26036	2
11-03-20	26846	26852	26309	2
12-03-20	27301	27316	26482	3
13-03-20	28055	28061	27213	3
14-03-20	28753	28742	27890	3
15-03-20	29410	29421	28234	4
16-03-20	30098	30081	28894	4
17-03-20	30456	30448	28933	5
18-03-20	30715	30714	28872	6
19-03-20	31145	31139	29276	6
20-03-20	31811	31819	29902	6
21-03-20	32619	32625	30662	6

5. Conclusions and future work

This work continues research [10] aimed at building models for the spread of COVID-19 depending on restrictive measures taken by governments around the world, and identifying signs that have high prognostic power to predict pandemic behavior and its consequences depending on government action. In the work on the basis of data clustering, three clusters have been identified, which differ in restrictive measures and, accordingly, in the dynamics of morbidity and mortality, in the scale of each country. At this stage of the study, factors such as country size, population heterogeneity, heterogeneity of social, demographic or geographical characteristics within one country and between different countries were not taken into account. The influence of these factors will be the subject of further research. In addition, the rationing of the number of new cases per day on the number of tests conducted in each country, in the presence of such information, will increase the reliability of modeling and accuracy of the forecast.

Data clustering showed a different distribution of variables between clusters, as well as a change in a country's membership in a particular cluster over time. Therefore, further research was conducted for each cluster separately, and the dynamics of migration between countries between clusters depending on changes in restrictive measures and the dynamics of the epidemic in each country were

studied. All countries were divided into 2 categories: countries that were mostly in the same cluster all the time, and countries whose clusters changed over time. From the group of countries that were in different clusters during the study, 3 countries (Germany, Italy, Poland) were selected similar in demographic, geographical and social characteristics, but different in terms of epidemic development and investigated the impact of government measures on cluster change. The paper shows that there is a clear relationship between the cluster number and the number of new cases of morbidity and death. There are also different countries' policies to prevent the disease, which have different effects on the spread of COVID-19. For example, in Germany, proactive tactics are pronounced, while in Italy – catching up on the nature of the spread of the disease.

A regression tree and guidelines for the effect of symptoms on the prevalence of COVID-19 morbidity and mortality rates due to this infection were constructed. The paper predicts the number of deaths due to COVID-19 on the example of Italy. Such forecasting was carried out for two potential options for government action: with existing preventive actions and the same preventive actions, if they were taken 20 days earlier (following the example of Germany). Predictions of the number of new deaths in Sweden due to COVID-19 were made for an interval of 35 days. To verify the prognosis, data on morbidity from March 1, 2020 to March 21, 2020 were used. The RMSE of the mortality forecast does not exceed the value of 4.2, which shows a good prognostic ability of the developed model. At the same time, the simulation based on the strategy of anticipatory introduction of restrictions gives 2–6% lower values of the forecast of the number of new cases. Thus, the results of our clustering of data, the study of the dynamics of transition between clusters, and its impact on the spread and consequences of the COVID-19 pandemic, as well as a forecast model based on production rules based on the regression tree, allow us to assess the impact of decisions on restrictive measures and to predict and simulate the consequences of a policy of imposing restrictions.

The limitations of the study are the following:

- It is impossible to change the scenario automatically based on restrictions. For example, we have modeled the ConfirmedCases number (Table 6) based on the set of restrictions changes for manually chosen countries (Figures 6 – 8). However, the result of modelling depends on the list of chosen countries. The automatic changes can be done, for example, based on multi-agent system based on mobile cellular automata [38];
- The modelling of confirmed cases and confirmed death should be provided for separated cluster more detailed. As it is shown in [39], the quality of classifier model is different for different clusters.

Conflict of interest

The authors declare no conflict of interest.

References

1. *Worldometer*: Coronavirus Update (Live): 119,292,681 Cases and 2,644,627 Deaths from COVID-19 Virus pandemic, 2020. Available from: <https://www.worldometers.info/coronavirus/>.
2. *WHO Director-General*: Opening remarks at the media briefing on COVID-19, 11 March 2020. Available from: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
3. *World Health Organization*: Coronavirus disease (COVID-19), 2020. Available from:

<https://www.who.int/emergencies/diseases/novel-coronavirus-2019>.

4. X. Zhang, R. Ma, L. Wang, Predicting turning point, duration and attack rate of COVID-19 outbreaks in major Western countries, *Chaos Soliton. Fract.*, **135** (2020), 109829.
5. K. E. Asnaoui, Y. Chawki, A. Idri, Automated methods for detection and classification pneumonia based on X-ray Images using deep learning, preprint, arXiv:2003.14363.
6. Y. Kryvenchuk, I. Helzynskyy, T. Helzhynska, N. Boyko, R. Danel, Synthesis control system physiological state of a soldier on the battlefield, *CEUR-WS.org*, **2488** (2019), 297–306.
7. N. Melnykova, V. Melnykov, E. Vasilevskis, The personalized approach to the processing and analysis of patients' medical data, *CEUR-WS.org*, **2255** (2018), 103–112.
8. N. Shakhovska, S. Fedushko, M. Greguš ml., N. Melnykova, I. Shvorob, Y. Syerov, Big data analysis in development of personalized medical system, *Proced. Computer Sci.*, **160** (2019), 229–234.
9. Z. Wang, K. Tang, Combating COVID-19: Health equity matters, *Nat. Med.*, **26** (2020), 458–458.
10. V. Yakovyna, N. Shakhovska, K. Shakhovska, J. Campos, Recommendation rules mining for reducing the spread of COVID-19 cases, *CEUR-WS.org*, **2753** (2020), 219–229.
11. *Covid-19 Data Science*: Home, 2020. Available from: <https://www.covid-datascience.com/>.
12. A. Doanvo, X. Qian, D. Ramjee, H. Piontkivska, A. Desai, M. Majumder, Machine learning maps research needs in COVID-19 literature, *Patterns*, **1** (2020), 100123.
13. S. Basu, R. H. Campbell, Going by the numbers: Learning and modeling COVID-19 disease dynamics, *Chaos Soliton. Fract.*, **138** (2020), 110140.
14. R. A. Neher, R. Dyrdak, V. Druelle, E. B. Hodcroft, J. Albert, Potential impact of seasonal forcing on a SARS-CoV-2 pandemic, *Swiss. Med. Wkly.*, **150** (2020), 20224.
15. S. Tuli, S. Tuli, R. Tuli, S. S. Gill, Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, *Int. Things*, **11** (2020), 100222.
16. Z. Hu Z, Q. Ge, L. Jin, M. Xiong, Artificial intelligence forecasting of Covid-19 in China, preprint, arXiv:200207112.
17. S. K. Bandyopadhyay, S. Dutta, Machine learning approach for confirmation of COVID-19 cases: Positive, negative, death and release, preprint, medRxiv/2020.03.25.20043505.
18. Z. Yang, Z. Zeng, K. Wang, S. S. Wong, W. Liang, M. Zanin, et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, *J. Thorac. Dis.*, **12** (2020), 165–174.
19. J. K. Davis, T. Gebrehiwot, M. Worku, W. Awoke, A. Mihretie, D. Nekorchuk, et al., A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model, *Environ. Model. Software*, **119** (2019), 275–284.
20. J. M. Scavuzzo, F. Trucco, M. Espinosa, C. B. Tauro, M. Abril, C. M. Scavuzzo, et al., Modeling Dengue vector population using remotely sensed data and machine learning, *Acta Tropica*, **185** (2018), 167–175.
21. M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, L. dos Santos Coelho, Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil, *Chaos Soliton. Fract.*, **135** (2020), 109853.
22. D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos Soliton. Fract.*, **134** (2020), 109761.
23. K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. Hyman, et al., Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020, *Infect. Dis. Model.*, **5** (2020), 256–263.

24. J. Farooq, M. A. Bazaz, A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies, *Chaos Soliton. Fract.*, **138** (2020), 110148.
25. E. Kaxiras, G. Neofotistos, E. Angelaki, The first 100 days: Modeling the evolution of the COVID-19 pandemic, *Chaos Soliton. Fract.*, **138** (2020), 110114.
26. A. S. Fokas, N. Dikaios, G. A. Kastis, Predictive mathematical models for the number of individuals infected with COVID-19, preprint, medRxiv/2020.05.02.20088591.
27. R. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, O. Pavlyuk, Development of the non-iterative supervised learning predictor based on the ito decomposition and SGTm neural-like structure for managing medical insurance costs, *Data*, **3** (2018), 46.
28. N. Hasan, A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model, *Int. Things*, **11** (2020), 100228.
29. K. C. Santosh, COVID-19 prediction models and unexploited data, *J. Med. Syst.*, **44** (2020), 170.
30. S. Sengupta, S. Mugde, G. Sharma, Covid-19 pandemic data analysis and forecasting using machine learning algorithms, preprint, medRxiv/2020.06.25.20140004.
31. *Azure Open Datasets Catalog: COVID-19 Data Lake*, 2020. Available from: <https://azure.microsoft.com/en-us/services/open-datasets/catalog/covid-19-data-lake/>.
32. *United Nations Department of Economic and Social Affairs: Population Dynamics*, 2016. Available from: <https://www.un.org/development/desa/en/key-issues/population.html>.
33. D. Gamberger, N. Lavrac, C. Groseelj, Experiments with noise filtering in a medical domain, in *Proc. of the Sixteenth Intl. Conf. on Machine Learning*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., (1999), 143–151.
34. I. Izonin, R. Tkachenko, V. Verhun, K. Zub, An approach towards missing data management using improved GRNN-SGTm ensemble method, *JESTECH*, paper in press.
35. E. Meng, S. Huang, Q. Huang, W. Fang, L. Wu, L. Wang, A robust method for non-stationary streamflow prediction based on improved EMD-SVM model, *J. Hydrol.*, **568** (2019), 462–478.
36. M. El-Hendawi, Z. Wang, An ensemble method of full wavelet packet transform and neural network for short term electrical load forecasting, *Electr. Power Systems Res.*, **182** (2020), 106265.
37. S. van Buuren, K. Groothuis-Oudshoorn, Mice: Multivariate imputation by chained equations in R, *J. Stat. Soft.*, **45** (2011), 1–67.
38. Y. Vyklyuk, M. Manylich, M. Škoda, M. M. Radovanović, M. D. Petrović, Modeling and analysis of different scenarios for the spread of COVID-19 by using the modified multi-agent systems—evidence from the selected countries, *Results Phys.*, **20** (2021), 103662.
39. N. Shakhovska, I. Izonin, N. Melnykova, The hierarchical classifier for COVID-19 resistance evaluation, *Data*, **6** (2021), 6.



AIMS Press

©2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)