

Attention-Based and Time Series Models for Short-Term Forecasting of COVID-19 Spread

Jurgita Markevičiūtė^{1,*}, Jolita Bernatavičienė², Rūta Levulienė¹, Viktor Medvedev², Povilas Treigys² and Julius Venskus²

¹Institute of Applied Mathematics, Vilnius University, Vilnius, 03225, Lithuania

²Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, 08412, Lithuania

*Corresponding Author: Jurgita Markevičiūtė. Email: jurgita.markeviciute@mif.vu.lt

Received: 19 March 2021; Accepted: 02 May 2021

Abstract: The growing number of COVID-19 cases puts pressure on healthcare services and public institutions worldwide. The pandemic has brought much uncertainty to the global economy and the situation in general. Forecasting methods and modeling techniques are important tools for governments to manage critical situations caused by pandemics, which have negative impact on public health. The main purpose of this study is to obtain short-term forecasts of disease epidemiology that could be useful for policymakers and public institutions to make necessary short-term decisions. To evaluate the effectiveness of the proposed attention-based method combining certain data mining algorithms and the classical ARIMA model for short-term forecasts, data on the spread of the COVID-19 virus in Lithuania is used, the forecasts of epidemic dynamics were examined, and the results were presented in the study. Nevertheless, the approach presented might be applied to any country and other pandemic situations. The COVID-19 outbreak started at different times in different countries, hence some countries have a longer history of the disease with more historical data than others. The paper proposes a novel approach to data registration and machine learning-based analysis using data from attention-based countries for forecast validation to predict trends of the spread of COVID-19 and assess risks.

Keywords: COVID-19 spread modeling; attention-based forecasting; machine learning; data registration; data analysis; ARIMA

1 Introduction

The COVID-19 pandemic has added an extremely high element of unpredictability to the global economy and the situation in general. Governments are trying to overcome the infection by taking serious measures in an effort to stabilize the situation. Experts are trying to predict how the situation may change and how it will look when the coronavirus can be restrained and which states will be the first to come out of the economic recession. At the moment, the priority is to



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

solve urgent health care problems and maintain economic stability. Experts are already trying to look into the near future and understand how the disease rate can progress.

Short-term and long-term forecasting models are generally used to forecast certain situations and to alert us to events in the future so that we are better prepared. Short-term forecasting models [1], which predict for no more than a few days, react and reflect more up-to-date data than long-term models. The latest new data in long-term models has a much less impact on the forecast growth curves as the more recent data form a much smaller proportion of the total. Therefore, to create accurate short-term forecasts, it is necessary to be especially careful and pay attention to the up-to-date data on new cases. Practice shows that stochastic methods are most effective for short-term forecasts, while deterministic methods, like SIR or SEIR [2,3], are most suitable for describing long-term scenarios.

Effective short-term prediction models are needed to predict the number of new cases. In this regard, it is important to develop strategic planning methods in the public health care system to avoid further increases in incidence of infection, as well as to introduce special measures to reduce the scope of infection. Various methods based on mathematical modeling and data mining are powerful tools for understanding the COVID-19 transmission [4], as well as for modeling and studying different scenarios for short-term pandemic forecasting. The accuracy of forecasting methods largely depends on the quality and quantity of available data used to make forecasts and assess the situation. The problems of forecasting the dynamics of confirmed cases of COVID-19 are being widely discussed in many recent scientific publications [5–13].

Despite the limitations associated with medical data-based forecasting and the specific nature of the data being analyzed, forecasting plays an important role as it enables a better understanding of the current situation and makes plans for the future. Mathematical modeling and disease prediction are powerful tools for understanding the spread of COVID-19 and studying different scenarios. Various methods and time series analysis are currently being used for short-term forecasting of COVID-19 epidemic disease dynamics: linear forecasting models, including autoregressive integrated moving average (ARIMA) model [5,14–16], artificial intelligence inspired models [10,13,17–19], hybrid forecasting models [20–22]. Because of the limited amount of data available, models inspired by artificial intelligence were not used in this study. The authors propose an attention-based method and compare short-term forecasting results with the simple but effective ARIMA model for forecasting disease spread trends. By means of machine learning clustering, the authors of this paper draw attention to countries with a longer COVID-19 disease spread history and, based on the data obtained, calculates the forecast for the country under study. According to the recent research [5], two criteria are emphasized: the accuracy of forecasts and the simplicity of the models used. A variety of applicable strategies make it possible to combine several methods, such as data preprocessing, data mining methods, and common forecasting models. ARIMA model has the advantage of a simple structure and a strong ability to interpret the data. Rather than focusing on which model is more correct, we need more models that answer additional questions and the use of which may affect the spread of COVID-19 [23].

The main purpose of this article is to provide short-term forecasting that could produce a reliable forecast for policymakers to make the necessary decisions and to provide useful guidance. In this paper, the authors provide the results of statistical forecasting of confirmed cases of COVID-19 in Lithuania using the attention-based approach and ARIMA models. The proposed methods might be applied to any country and other pandemic situations. The paper is organized as follows: Section 2 presents the analyzed data and methodology for short-term forecasting of confirmed cases of COVID-19 using the ARIMA models and the attention-based machine

learning method. The experimental results of the proposed method as well as the comparison with Lithuanian forecasting results obtained using the ARIMA models are presented in Section 3. The results of the study are concluded in Section 4.

2 Materials and Methods

2.1 Data in the Context of Lithuania

During the pandemic, the source of Lithuanian COVID-19 data changed, which made the task of spread forecasting challenging. The main data provider in Lithuania is the National Public Health Center (NPHC) under the Ministry of Health. Since the beginning of the spread of COVID-19 in Lithuania, the data has been announced on the website of the Ministry of Health. The problem with such an announcement was that no historical data was available, only the daily statistics, thus the authors had to collect, process, and store data by creating their own database. However, because of problems with data collection in the NPHC—delays by various institutions, errors caused by human factors, etc.—historical data has also been revised, but this has not been announced. At first, the revisions were not substantial and did not affect the short-term forecasts. However, corrections became substantial before the autumn of 2020, when the number of confirmed cases increased. Since the end of August, the National Public Health Center under the Ministry of Health started sharing files with the time series data. Although there were still problems with the quality of the data, but historical data became available. Since November, the only institution announcing the COVID-19 data is Statistics Lithuania.¹ Although the quality problems have not been completely resolved, detailed daily datasets with revised historical data are now available.

Nevertheless, it has been noticed that some time series, such as recoveries or active cases, still suffer from quality problems due to delays in reporting recoveries by hospitals. Thus, in the middle of February 2021, Statistics Lithuania started announcing two time series for recovered and active cases: a *de jure* series, which announces registered cases and *statistical* series, which announces statistical estimates. These time series are very different from each other and have differences with the earlier recovered and active cases time series.

The authors use data² from the Coronavirus Resource Centre (CRC) at Johns Hopkins University as a data source for other countries. It was noticed that Lithuanian historical data of the disease was not correctly updated in the CRC source. Thus, while analyzing Lithuania in the context of the European countries, the authors used the data from Lithuanian data sources rather than from the CRC. For the clustering and attention-based forecasting, European countries and COVID-19 data of confirmed, recovered cases, deaths from COVID-19 per 100,000 people, and population density from the United Nations database on 2019 mid-year period was used in the research. The motivation for including population density was based on the assumption that more densely populated countries have a higher risk of more rapid spread of COVID-19.

Despite the fact that the beginning of confirmed cases of COVID-19 in Lithuania is February 28, 2020, the authors in this research use data on the spread of the disease in Lithuania during the period from March 12, 2020, to February 1, 2021, because only one case was registered until March 12, 2020. Moreover, we take the earlier definition of recovered cases since new definitions have appeared after the period we are investigating. Fig. 1 shows the dynamics of the daily

¹ <https://osp.maps.arcgis.com/apps/MapSeries/index.html?appid=c6bc9659a00449239eb3bde062d23caa>.

² <https://github.com/CSSEGISandData/COVID-19>.

confirmed cases and cumulative confirmed cases per 100,000 people. There was a huge increase in the number of confirmed cases per day in the autumn of 2020 and January 2021, while the spring and summer of 2020 were relatively stable periods. There are several reasons for this. In spring 2020, Lithuania announced a very strict lockdown—schools, kindergartens, and universities switched to online learning, all non-food shops and services were closed, only online trade was allowed, the number of imported cases was higher than inside the country; thus, the closure of borders helped to reduce the spread of COVID-19. During the summer period, a small number of confirmed cases might be possibly related to the reduced extent of traveling abroad and weather conditions. During the autumn there was no strict lockdown, mostly recommendations, schools and kindergartens operate in normal or mixed mode, universities operate in online or mixed mode. Thus, the spread becomes very intensive, and the prevention becomes belated and unresponsive to the actual situation of COVID-19. Therefore, as most cases occur within the country, border closures are not used to prevent the spread of the COVID-19 virus.

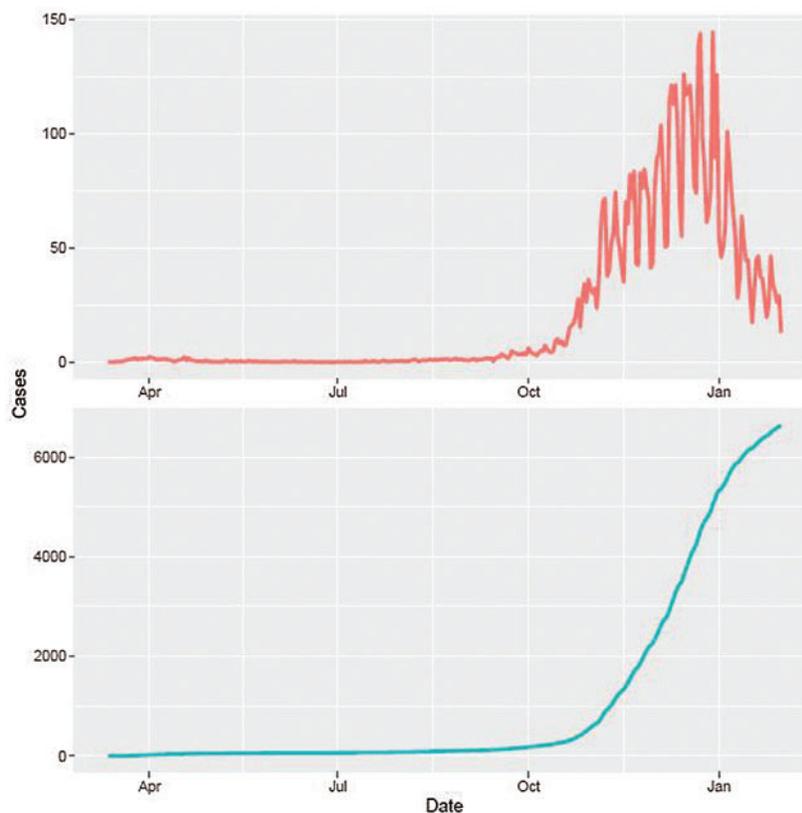


Figure 1: Confirmed (top) and cumulative (bottom) confirmed cases per 100,000 people

Every day by January 26, 2021, forecast models were built based on 35 subsequent observations, and forecasts were performed for 5 steps ahead. As mentioned above, there were different trends in virus spread during the spring and autumn periods, so the authors built models for the spring sub-period by analyzing data for the period from March 12 to June 30, 2020 (the forecast of five steps ahead is included), and for the autumn sub-period from October 1, 2020 to January 31, 2021. We refer to these periods as “first wave” and “second wave” though in

epidemiological terms these waves may have different time stamps. The authors do not investigate other sub-periods, since the summer was fairly stable and calm in terms of the spread of COVID-19.

Two approaches were used for the short-term forecasting of confirmed cases of COVID-19 in Lithuania: ARIMA models and the attention-based forecasting method. The results of ARIMA models and all the data used for modeling and forecasting are available at www.covid19.projektas.vu.lt.

2.2 ARIMA Model

ARIMA models are frequently used to forecast time series in a short period of time. Although such type of models is simple and easy to apply, these models show good performance for short-term forecasting. ARIMA models are best fitted to stationary or differentiated data if the data is stationary after the differentiation. In this paper, we apply the non-seasonal ARIMA(p, d, q) model:

$$y_t = \beta + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}, \quad (1)$$

where t denotes the time; p denotes the order of the autoregressive part (AR(p)); q denotes the order of the moving average part (MA(q)); d denotes the order of the differences of the data; y_t are the observations that are differentiated d times, usually $d = 0, 1, 2$; ε_t are model innovations.

To estimate the ARIMA model, the Hyndman-Khandakar algorithm [24] is used, which combines unit root tests, minimization of the AICc criterion (Akaike information criterion modified for the small samples), and the maximum likelihood estimator. With the view to choose the best model, the AICc criterion has been used. The motivation is related to the fact that a quite small sample of only 35 observations is used in the estimation procedure. The criterion is defined as:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} = 2k - 2 \ln(\hat{L}) + \frac{2k^2 + 2k}{n - k - 1},$$

where \hat{L} is the maximum value of the likelihood function of the model, n is a sample size, and k is a number of estimated parameters. More information on the ARIMA models can be found in [25].

2.3 Attention-Based Forecasting Method

The idea of the attention-based approach is to use a mechanism that selects specific factors from the data available. The idea can be accomplished by focusing attention on small regions of multidimensional information rather than on the data as a whole. In this way, dimensionality reduction techniques are used to draw attention to similar countries where the spread of the virus is similar and to further analyze the data for countries that fall into the attention cluster. The attention-based mechanism acts as an extractor of information while inferring the similarity and minimizing the number of countries used for spread forecasting.

The authors propose an attention-based forecasting method that consists of three steps:

1. Data registration regarding the first confirmed case of COVID-19 for the first wave and the day after July 1, 2020, when the number of confirmed cases per 100,000 is greater than or equal to 3 for the second wave.
2. Selection of countries most similar to Lithuania, using data mining and machine learning techniques.

3. Forecasting based on the use of trends in confirmed cases in the selected countries.

Each step of the method is described in more detail in the subsections below.

2.3.1 Data Registration

The outbreak of COVID-19 started at different time periods in different countries. Thus, some countries have a longer history and more historical data of the COVID-19 spread than others. The novelty of the proposed method is to consider the onset of the spread of COVID-19 in different European countries, and to compare the dynamics of the virus and integrate this knowledge into the forecast. The idea is to use data from countries with more historical disease data to forecast trends in Lithuania. For this purpose, we have recorded the data in such a way that the time series starts from the first confirmed case of COVID-19, i.e., we use an artificial time scale (number of days from the first confirmed case) rather than an actual calendar date for the spring sub-period (the first wave). For the autumn sub-period (the second wave), we took data from July 1, 2020, and registered data in such a way that the first day is the day when the number of confirmed cases per 100,000 population is greater than or equal to three. We chose a threshold value of three since the increase in confirmed cases has begun at this time point in the second sub-period for many countries. For the short-term forecasting for the first wave, countries that have more historical data, compared to Lithuania, on the disease from the first confirmed case of COVID-19 were selected: Austria, Belgium, Croatia, Denmark, Estonia, Finland, France, Germany, Greece, Iceland, Italy, Netherlands, North Macedonia, Norway, Romania, Spain, Sweden, Switzerland, and the United Kingdom. Accordingly, the following countries were identified in the same way for short-term forecasting in the case of the second wave: Austria, Belgium, Bulgaria, Croatia, Czechia, Denmark, France, Greece, Hungary, Ireland, Italy, Liechtenstein, Luxembourg, Malta, Montenegro, Netherlands, North Macedonia, Norway, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom. Lithuanian data was also included for both waves.

2.3.2 Selection of Countries

The multidimensional data describe complex objects or phenomena characterized by many features. For better comprehension, it is useful to provide data in an easy-to-understand form: to define the structure of the data, relationships, and clusters. Multidimensional data visualization methods are used to provide data mining results in a more comprehensive form by drawing attention to similarities. The attention-based selection of the European Union countries for forecasting is performed by integrating multidimensional data clustering and data dimensionality reduction methods: self-organizing neural network (SOM), multidimensional scaling (MDS), and t-distributed stochastic neighbor embedding (t-SNE). The data was first clustered using the SOM neural network. For the clustering result inspection, visualization techniques such as MDS or t-SNE methods can be used. Different visualization techniques were chosen to validate the clustering results obtained by the SOM, using methods based on different operating principles. Dimensionality reduction methods transform the analyzed dataset from the l -dimensional space R^l into a lower-dimensional space R^d ($d < l$), where the distribution of points can be observed visually, e.g., $d = 2$. In this study, we consider multidimensional data of the following form:

$$Y_i = (y_{i1}, y_{i2}, \dots, y_{in}, x_{i1}, x_{i2}, \dots, x_{in}, z_{i1}, z_{i2}, \dots, z_{in}, p_{-d_i}), \quad i = 1, \dots, m,$$

where $n = 1, \dots, 35$ (number of observations), m —number of countries analyzed, $y_{i1}, y_{i2}, \dots, y_{in}$ —number of cumulative confirmed cases, $x_{i1}, x_{i2}, \dots, x_{in}$ —number of cumulative deaths, $z_{i1}, z_{i2}, \dots, z_{in}$ —number of cumulative recovered cases, $p_{-}d_i$ —population density for i -th country, respectively, i.e., the study deals with data described by 106 features.

Typically, $Y_i \in R^l$, $i = 1, \dots, m$ are interpreted as points in the multidimensional space R^l , where l defines the dimensionality of the space (in our case, 106 features). If the dimensionality of the projection space is two, then the MDS and t-SNE methods can be applied to represent the multidimensional data in a 2D space.

The self-organizing neural network SOM was used for clustering of multidimensional data. SOM is a neural network-based method that is trained in an unsupervised way using competitive learning [26,27]. Self-organizing maps use a neighboring function to preserve the topological properties of the input space. Typically, SOM represents a set of interconnected neurons according to some topology, e.g., the rectangular SOM is a two-dimensional array of neurons. Each element of the input data set is connected to every individual neuron in the rectangular structure. Every neuron is entirely defined by its location on the grid by its specific index at the row and the column and by its weight (so-called code book vector). After SOM training, the data are presented to SOM, and the winning neuron for each input data is found. The winning neuron is the one to which the Euclidean distance of the input data vector is the shortest. In such a way, the input data are distributed on SOM, and resulting data clusters can be observed. t-SNE is a nonlinear dimensionality reduction technique based on Stochastic Neighbor Embedding [28]. The method minimizes the divergence between two distributions: a distribution that measures pairwise similarities of the high-dimensional objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding.

The MDS method is used to find a configuration of points in a space, usually Euclidean, where each point represents one of the objects or individuals, and the distances between pairs of points in the configuration match as well as possible the original dissimilarities between the pairs of objects or individuals [29]. The MDS method represents (dis)similarity data as distances in a low-dimensional space to make these data accessible to visual inspection and exploration [30].

2.3.3 Forecasting

The trends of the number of confirmed cases in the countries which belong to the same cluster as Lithuania and have more historical data on the disease from the first confirmed case of COVID-19 are used. The regression models with countries as covariates are considered. The forecasting is done for such a number of days ahead as is the history of the confirmed number of cases in the countries belonging to the same cluster as Lithuania. However, some countries, which belong to the same cluster as Lithuania, do not have a much longer history of confirmed cases than Lithuania. Thus, to obtain a forecast for required steps ahead, ARIMA models (see Section 2.2) were used to forecast the number of the confirmed cases in each country, and then these forecasts were employed in regression analysis to get the forecast of the confirmed cases in Lithuania.

To achieve the goal above, the linear regression with ARMA errors was used:

$$y_t = \beta_0 + \beta_1 x_t + \eta_t, \quad \eta_t = \sum_{i=1}^p \phi_i \eta_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}, \quad (2)$$

where p and q are AR and MA orders respectively; $\beta_0, \beta_1, \phi_i, \theta_j$ are unknown parameters and ε_t are random errors.

After the linear regression models were obtained for each country in the cluster, the forecast was calculated by taking the average of forecasts from these models:

$$\hat{y}_{LT;t} = \frac{1}{k} \sum_{l=1}^k \hat{y}_{l;t}, \quad (3)$$

where $\hat{y}_{l;t}$ is the forecast from each regression performed.

2.4 Forecast Accuracy

The comparison of the accuracy of the forecasts over the considered time period was made by choosing models for every interval of 35 days and forecasting five steps ahead. The following measures of the forecast accuracy were used:

$$RMSE = \sqrt{\frac{1}{T} \sum_{i=1}^T (\hat{y}_t - y_t)^2}, \quad MAPE = \frac{100\%}{T} \sum_{i=1}^T \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \quad MAE = \frac{1}{T} \sum_{i=1}^T |y_t - \hat{y}_t| \quad (4)$$

where T is the number of time points, y_t and \hat{y}_t are observed and fitted values of the dependent variable, respectively.

3 Results

3.1 ARIMA Forecasting Results

We estimate the ARIMA model for cumulative confirmed cases every day with 35 recent observations for the period March 12, 2020–January 26, 2021, dividing data into two sub-periods as follows: March 12–June 25, 2020, and October 1, 2020–January 26, 2021. Earlier data and the summer period are truncated due to a very small number of confirmed cases.

We would like to point out that modeling started at the end of March, and different types of models and lengths of data at first were used. Here we present the final version of the forecasting approach, thus the historical forecast in the paper might differ from the forecast announced for the public.

Note that the model was built based on the cumulative number of confirmed cases per 100,000 people. In addition, the authors analyze the non-seasonal ARIMA model for the daily data, though a slight seasonality might be observed because of the weekend effect, when fewer tests are performed. However, testing with the seasonality models does not have better performance than with the non-seasonal models.

A new ARIMA model is fitted for every day. The orders of each ARIMA model are shown in [Fig. 2](#).

For every day, not only a new model is built, but forecasts are performed for five steps ahead as well as prediction intervals with 80% and 95% prediction probabilities are computed.

The complete algorithm for every day model is given in Algorithm 1 presented in [Fig. 3](#).

The result of Algorithm 1 is the graph where the black line and black dots indicate predicted values, the red dots indicate true values, the dark blue band indicates the 80% prediction interval, and the light blue band indicates the 95% prediction interval (see [Fig. 4](#)).

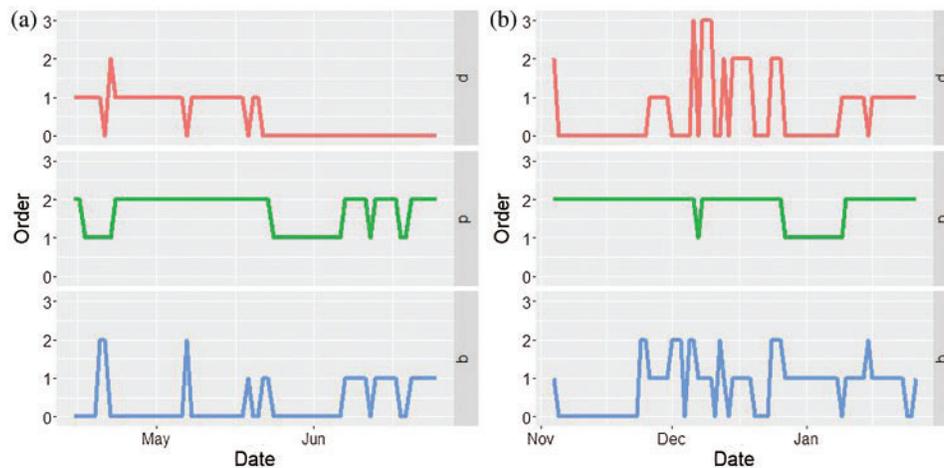


Figure 2: ARIMA orders for every day models. (a) March 12–June 25, 2020 (b) October 1, 2020–January 26, 2021

Algorithm 1 ARIMA short-term forecast

Require: Cumulative confirmed cases

Population size

Range of values for parameters p , d and q

Number of prediction steps ahead h

Ensure: Short-term prediction

- 1: Obtain COVID-19 data from Statistics Lithuania
 - 2: Select cumulative confirmed cases
 - 3: Compute number of cumulative confirmed cases per 100,000 people
 - 4: **for** i, j, k in range of values p , d and q **do**
 - 5: Estimate ARIMA(i, j, k) model for 35 days observations using maximum likelihood estimator
 - 6: **end for**
 - 7: Select ARIMA model with the smallest AICc value
 - 8: Perform the forecast for h steps ahead
 - 9: Restore values of the forecast from the relative scale to the absolute values of cumulative confirmed cases
 - 10: **if** the values of the prediction interval is below the last observed cumulative confirmed cases value **then**
 - 11: Set these values to the last observed cumulative confirmed cases value
 - 12: **end if**
 - 13: Draw a graph with last two values of the period, h days predictions, 80% and 95% prediction intervals
-

Figure 3: Algorithm 1. ARIMA short-term forecast

As it was mentioned earlier, the historical data has been changing over time, and the prediction scheme has also slightly changed over the pandemic period. Thus, a retrospective analysis of the goodness of the prediction of the ARIMA models has been accomplished. To achieve this goal, training set errors and prediction errors computed over the whole period, taking into account all historical models, were investigated (see Fig. 5: Algorithm 2).

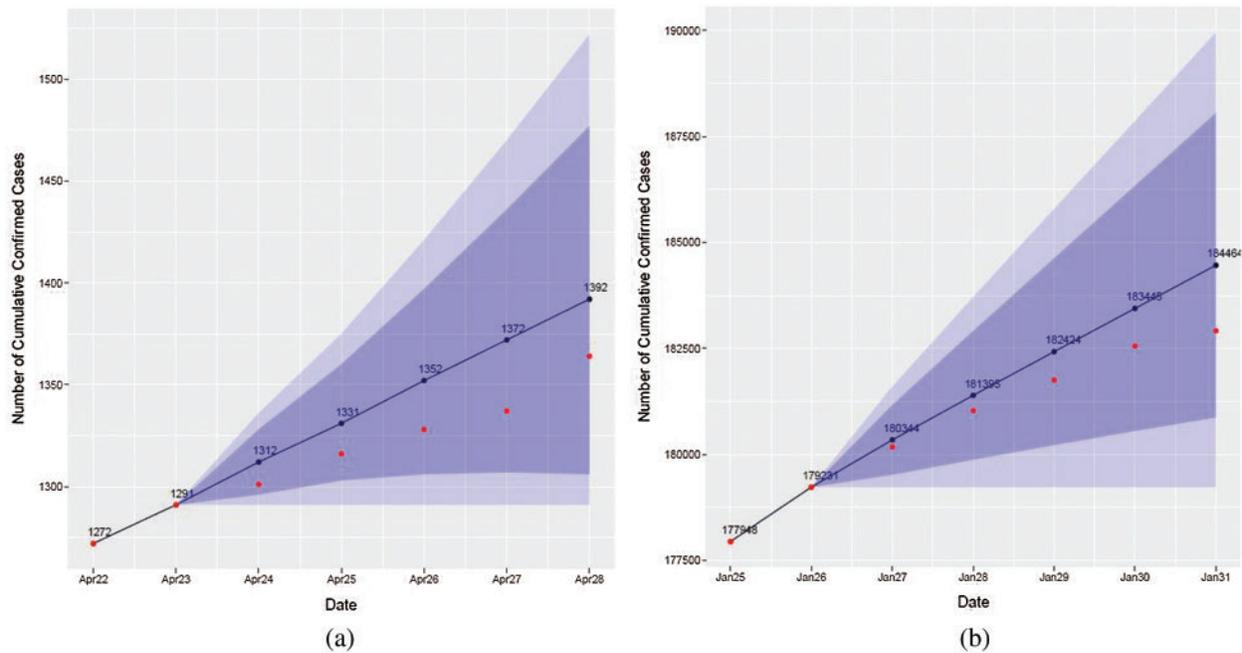


Figure 4: Visualization of forecasting results obtained by ARIMA. Red dots represent actual cases, blue line and dots represent ARIMA forecasts, and the blue and light blue bands are prediction intervals. (a) April 22–April 28, 2020 (b) January 25–January 31, 2021

Algorithm 2 Training and prediction

Require: Cumulative confirmed cases

Population size

Values range for parameters p , d and q

Number of prediction steps ahead h

Ensure: TrainingError table, PredictionValues table

1: **for** the data in a selected period **do**

2: Execute Algorithm 1

3: Save training set error measures RMSE, MAE, MAPE in the TrainingError table

4: Save forecast and values of prediction intervals in each step ahead separately in the

 PredictionValues table

5: Compute cumulative prediction errors for each forecast step ahead using Algorithm 3 and save it in the PredictionValues table

6: Compute indicator values

7: **if** true value is in the prediction interval **then**

8: indicator = 1

9: **else**

 indicator = 0

10: **end if**

11: and save them in the PredictionValues table

12: **end for**

Figure 5: Algorithm 2. Training and prediction

In total, 156 models (72 for the first wave and 84 for the second wave) were estimated, and training sets and prediction errors were saved and the final output of Algorithm 2 consists of two tables. Tab. 1 illustrates several *TrainingError* values.

Algorithm 3 Prediction error

Require: Observed values from the periods: March 12 – June 30, 2020, and October 1, 2020 – January 31, 2021
 Predicted values from the periods: March 12 – June 30, 2020, and October 1, 2020 – January 31, 2021, for each prediction step

Ensure: Cumulative RMSE, MAE and MAPE

```

1: for i in 1 to 156 do
2:   filter input data from the first value to the ith value
3:   compute RMSE, MAE, MAPE for the each prediction step separately
4: end for
    
```

Figure 6: Algorithm 3. Prediction error

Table 1: Training set errors

Model	RMSE	MAE2	MAPE
1	0.32	0.24	3.53
2	0.32	0.24	3.15
3	0.34	0.26	6.65
...

The result of a *Prediction Values* table is a table with 52 columns and 77 rows for the first wave and 89 rows for the second wave. The columns are as follows:

- *Date*: the date for the true and predicted values;
- *Observed*: the observed values;
- *S1_point, ..., S5_point*: point forecasts for each step ahead separately computed with Algorithm 1;
- *S1_lo80, ..., S5_lo80, S1_hi80, ..., S5_hi80, S1_lo95, ..., S5_lo95, S1_hi95, ..., S5_hi95*: prediction intervals lower and upper values computed with Algorithm 1;
- *RMSE_S1, ..., RMSE_S5, MAE_S1, ..., MAE_S5, MAPE_S1, ..., MAPE_S5*: cumulative prediction errors computed with Algorithm 3;
- *S1_int_80, ..., S5_int_80, S1_int_95, ..., S5_int_95*: indicator values, if true value is in the prediction interval.

The cumulative RMSE, MAE, and MAPE errors showed that the values of the median and variance increased with each prediction step (see Fig. 7, e.g., with MAPE error).

The same algorithms have been applied for the second period of data, as mentioned in SubSection 2.1: October 1, 2020–January 31, 2021.

The results obtained are very similar, but we have slightly larger errors. The bigger difference appears in the empirical probability for the true value to be in the prediction interval. Note that the authors have computed 80% and 95% prediction intervals, but retrospectively it is the indicator values if the true value is in this interval (Algorithm 2). Empirical prediction probabilities were computed as follows:

$$p = \frac{\text{\#number of true values in a prediction region}}{\text{\#number of total predictions performed}}.$$

The results are presented in Tab. 2. It can be seen that the results of the forecast are promising. For the further prediction steps, the probability is getting lower but is still above the

value of 0.89 for the step 5 with the 95% prediction interval over the first wave and greater than 0.80 for the second wave.

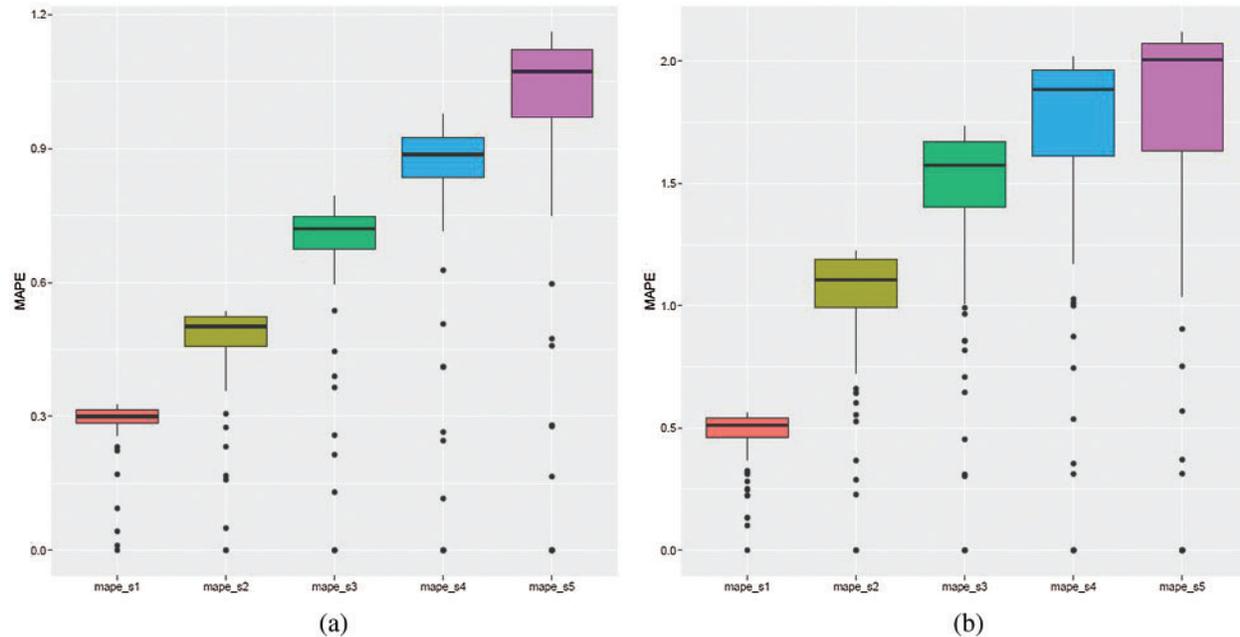


Figure 7: ARIMA models: cumulative MAPE errors. The figure shows boxplots for each prediction step, from one to five, ahead, that is *mape_s1* denotes MAPE values for the first step ahead forecast, *mape_s2* denotes MAPE values for the second step ahead forecast, etc. (a) March 12–June 30, 2020 (b) October 1, 2020–January 31, 2021

Table 2: ARIMA models: empirical prediction probabilities (P.I.–prediction interval)

Prediction step	First wave		Second wave	
	80% P.I.	95% P.I.	80% P.I.	95% P.I.
1	0.90	0.93	0.67	0.86
2	0.83	0.99	0.58	0.80
3	0.79	0.99	0.60	0.81
4	0.79	0.93	0.68	0.80
5	0.75	0.89	0.69	0.83

3.2 Attention-Based Forecasting Results

The results obtained using attention-based forecasting are presented in this section. The complete scheme for attention-based forecasting is given in Algorithm 4 (see Fig. 8). In an effort to compare the resulting forecasting accuracy of the two methods considered, the time periods, measures of the forecast accuracy, and the study of performance over the time kept the same as in Algorithms 2 and 3 (see Figs. 5 and 6).

Algorithm 4 Attention-Based short-term forecast

Require: cumulative new confirmed cases, deaths, recovered cases, population density for each analysed country.

Ensure: Short-term prediction

- 1: Obtain COVID-19 data from Statistics Lithuania and Johns Hopkins University Coronavirus Resource Center (CRC).
- 2: Select cumulative confirmed cases, deaths, recovered cases
- 3: Compute number of cumulative confirmed cases, deaths, recovered cases per 100,000 people
- 4: Add population density
- 5: Perform data registration from the first confirmed case of the COVID-19
- 6: Select countries with a longer history of the disease
- 7: Perform clustering of selected countries using the SOM neural network
- 8: Select the six countries closest to Lithuania based on dimensionality reduction (using MDS and t-SNE methods) and clustering results
- 9: **for each country do**
- 10: **if** history m of the number of confirmed cases in this country is smaller than h **then**
- 11: **for** i, j, k in range of values p, d and q **do**
- 12: Estimate forecasting model using observations of 35 days applying the maximum likelihood method; Choose the ARIMA model with the smallest AICc value; perform the forecast $h - m$ steps ahead
- 13: **end for**
- 14: **end if**
- 15: **end for**
- 16: **for each country do**
- 17: Obtain linear regression model with ARMA errors taking a country data as a covariate
- 18: Compute forecasts h steps ahead using regression model
- 19: **end for**
- 20: Compute an average of forecasts
- 21: Restore values of the forecast from the relative scale to the absolute values of cumulative confirmed cases
- 22: **if** the values of the prediction interval is below the last observed cumulative confirmed cases value **then**
- 23: Set these values to the last observed cumulative confirmed cases value
- 24: **end if**
- 25: Draw a graph with last two values of the period, h days predictions and 80% and 95% prediction intervals

Figure 8: Algorithm 4. Attention-based short-term forecast

3.2.1 Registration and Clustering Results

Following the data registration described in Section 2.3.1, 20 European countries with a longer history of the disease and virus spread than Lithuania were selected for the study: 17 European Union countries, the United Kingdom, Norway, and Northern Macedonia. Daily cumulative relative data (per 100,000 people per day) for the last 35 days of the research is used: cumulative new confirmed cases, deaths, recovered cases, and population density. As mention above, the multidimensional data consists of 106 features per country. Following the registration of data for the first case (see Section 2.3.1), a cluster and visual analysis of multidimensional data were carried out (see Section 2.3.2). To forecast the number of cases and disease trends in Lithuania, six countries were identified with more historical disease data, and in which the registered data for the past 35 days have a similar trend to Lithuania. [Tab. 3](#) presents the six countries selected for regression analysis for each month from March to June (registered for the first wave), which have been identified as most similar countries to Lithuania in terms of trend and similarity of the virus spread. The countries were ranked based on the summarized results of visual and cluster analysis (see Section 2.3.2). All six selected countries had a longer history from the first confirmed case of COVID-19 compared to Lithuania based on SOM output. In some cases, SOM output resulted in less than six countries in the same cluster as Lithuania. To select the lacking data to

form the cluster of six countries, the points closest to Lithuania, representing the countries, were identified based on Euclidean distances in the multidimensional space.

Table 3: Ranked first wave countries (most similar to Lithuania in terms of trends and similarities in the spread of the virus from March to June: the first wave)

Rank	March	April	May	June
1	Croatia	Croatia	Croatia	Croatia
2	North Macedonia	Romania	Finland	Greece
3	Greece	North Macedonia	Romania	Finland
4	Romania	Sweden	North Macedonia	Romania
5	Sweden	Greece	Greece	Estonia
6	Finland	Finland	Estonia	Norway

Tab. 3 shows that Croatia, Northern Macedonia, Romania, and Greece were the most similar countries to Lithuania in terms of virus spread trends from March to June. Since the visualization results obtained by the MDS and t-SNE methods are quite similar, we present only the results obtained by the t-SNE method. The visualization results of the analyzed data from the period from March 12 to April 15, 2020 are presented in Fig. 9a. When visualizing data, one color is allocated to countries in a single cluster, formed using the SOM. Similar trends in the spread of the virus during this period were also observed in Sweden, Finland, Estonia, and Norway, based on the results of visualization of the corresponding multidimensional data (see Fig. 9). As the situation in Romania and Macedonia changed already in May, these countries began to disappear from the cluster with Lithuania. The summer period is a stable and calm time in the spread of COVID-19, the situation changes slightly. Lithuania belongs to the same cluster as countries with a small number of confirmed cases. From August to September, the same cluster with Lithuania included Greece, Finland, Estonia, Norway, and Germany.

However, since November, the situation begins to change. Tab. 4 presents six countries from September to January (data registration for the second wave). From November, Lithuania was in the same cluster as Slovenia, Czechia, Slovakia, Austria, Italy, Montenegro, and Hungary. The number of cases in these countries has been increasing. As an example, the results of the visualization of the analyzed data in the period of December 23, 2020–January 26, 2021, obtained using the t-SNE method, are presented in Fig. 9b. The same clustering results were observed by applying the MDS method for clustering results inspection. However, MDS relies on Euclidean distances in multidimensional space, and when the figures are produced in 2D space, the points representing countries form quite dense clusters on the plane that are hardly readable but comparable to those obtained by t-SNE. Thus, only t-SNE results are being presented. Lithuania has been a part of the same cluster as Czechia, Slovenia, and Luxembourg. Other countries (Austria, Slovakia, Croatia, Montenegro, Hungary, Italy) have been identified as similar to Lithuania, although they belong to other clusters.

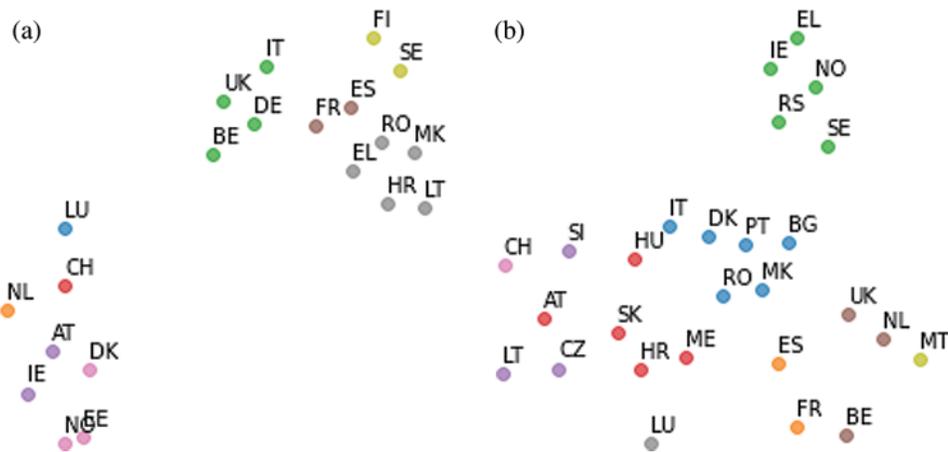


Figure 9: Visualization results obtained using t-SNE. Countries with the same color depict the same cluster meaning a similar COVID-19 spread. European Union and other countries have been assigned a two-letter country code.³ (a) March 12–April 15, 2020 (b) December 23, 2020–January 26, 2021

Table 4: Ranked second wave countries (most similar to Lithuania in terms of trends and similarities in the spread of the virus from September 2020 to the end of January 2021: second wave)

Rank	September	October	November	December	January
1	Bulgaria	Spain	Slovakia	Slovenia	Czechia
2	Sweden	North Macedonia	Italia	Austria	Slovenia
3	Ireland	Montenegro	Czechia	Belgium	Slovakia
4	Croatia	Romania	Spain	Hungary	Montenegro
5	Norway	Hungary	Hungary	Slovakia	Croatia
6	Romania	Slovakia	Montenegro	Italy	Austria

3.2.2 Forecasting Results

The approach outlined in Section 2.3.3 and the results presented in Section 3.2.1 were used to obtain the forecasts. Six countries from the same cluster and closest to Lithuania were used to make forecasts based on the number of cumulative confirmed cases. The same time intervals were considered to compare the results with those of the ARIMA models: the first is March 12–June 30, 2020 (corresponds to the first wave of COVID-19), and the second is October 1, 2020–January 31, 2021 (corresponds to the second wave).

Algorithm 4 was applied to obtain results, and the graph presents the final result (see Fig. 10): the black line and black dots represent the predicted values, the red dots represent true values, and the dark and light blue bands represent 80% and 95% prediction intervals, respectively.

³https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Country_codes.

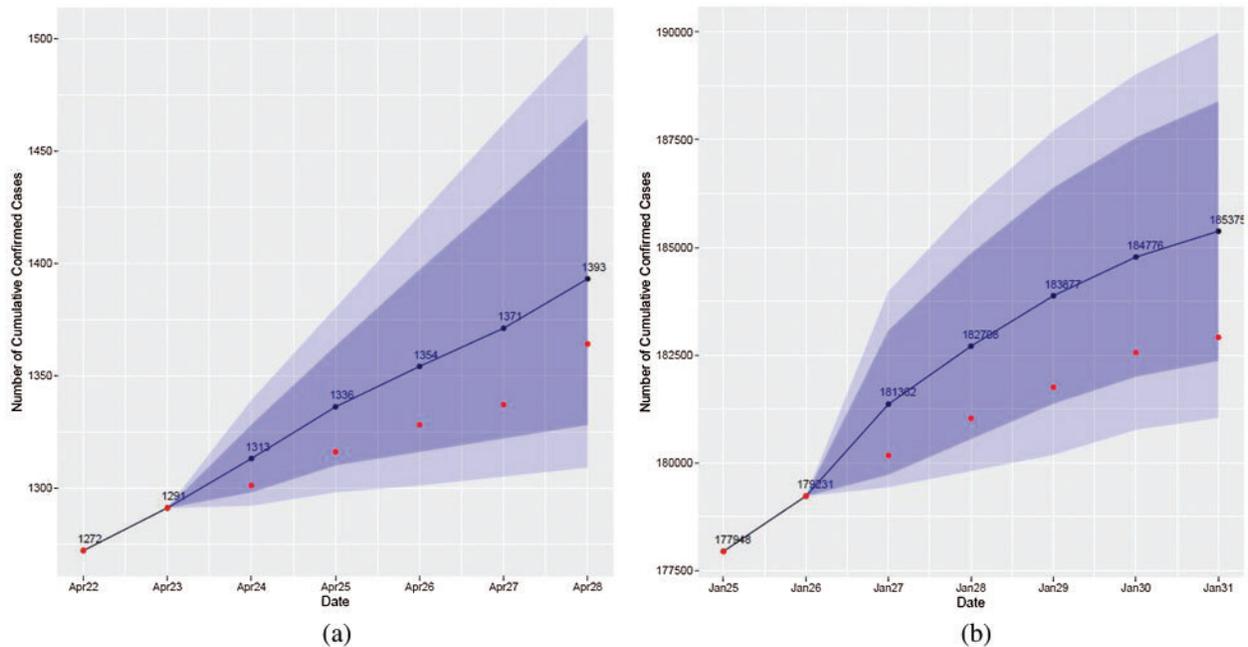


Figure 10: Visualization of forecasting results obtained by the attention-based method. The red dots represent actual cases, the blue line and dots represent forecasts using the attention-based method, and the blue and light blue bands are prediction intervals. (a) April 22–April 28, 2020 (b) January 25–January 31, 2021

Comparing the results of ARIMA models and attention-based forecasting (see, for example, Figs. 3 and 10), it can be concluded that one of these methods shows a better fit for some periods and the other is better for other periods. A higher forecast accuracy is achieved in periods when countries in the same cluster are closer to Lithuania.

All models, as in the case of ARIMA models, were estimated historically (overall 156 models were fitted: 72 in the case of the first wave and 84 in the case of the second wave), training set errors were saved and prediction errors over the two considered sub-periods were obtained (see Fig. 5: Algorithm 2). The cumulative values of RMSE, MAE and MAPE (see Fig. 6: Algorithm 3) show that the median and variance become larger with each prediction step (see Figs. 11a and 11b for MAPE values). The medians are similar, as in the case of ARIMA, however, the variance is slightly larger.

The comparison of the results obtained by the attention-based method for the two sub-periods (waves) shows that the forecast errors in the case of the first wave (see Fig. 11a for MAPE values) are smaller than in the case of the second wave (see Fig. 11b). This is due to the reason that the situation in countries differs more in the autumn than in the spring when countries from the same cluster as Lithuania were closer to each other.

The empirical probabilities that the true value is in the prediction region were obtained (see Tab. 5). The results show that the accuracy of the estimation is similar to that of ARIMA in the case of the first sub-period (wave) considered, but it is slightly worse in the case of the second wave.

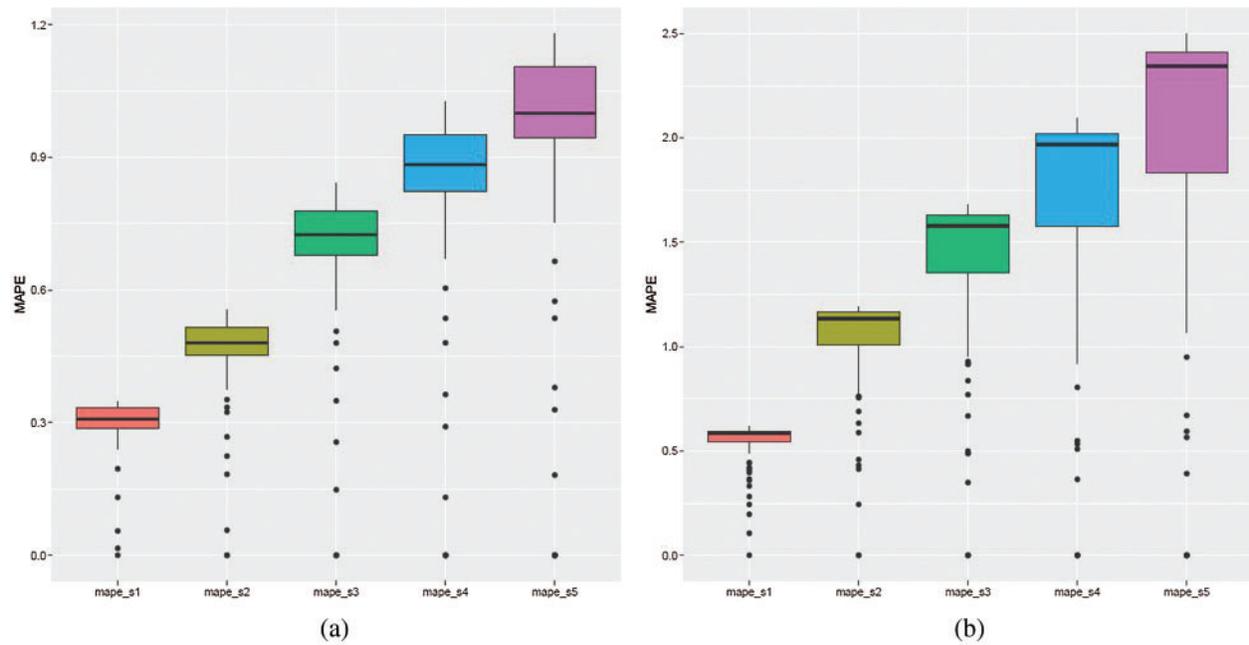


Figure 11: Attention-based method: cumulative MAPE. The figure shows boxplots for each prediction step one through five, ahead, that is *mape_s1* denotes MAPE values for the first step ahead forecast, *mape_s2* denotes MAPE values for the second step ahead forecast, etc. (a) March 12–June 30, 2020 (b) October 1, 2020–January 31, 2021

Table 5: Attention-based method: empirical prediction probabilities (P.I.–prediction interval)

Prediction step	First wave		Second wave	
	80% P.I.	95% P.I.	80% P.I.	95% P.I.
1	0.86	0.96	0.62	0.81
2	0.89	0.99	0.61	0.78
3	0.85	0.99	0.58	0.77
4	0.88	0.97	0.56	0.75
5	0.88	0.97	0.51	0.70

4 Discussion

This study investigates methods for obtaining short-term forecasts of the COVID-19 virus spread that can be useful for policymakers and public institutions in making necessary decisions and providing useful guidance as to what might happen in the coming week. By borrowing the idea of the attention-based approach from Long Short-Term Memory deep neural networks and combining this approach with the mathematical modeling and prediction methods, we obtain a powerful tool for understanding the spread of COVID-19 and exploring different short-term spread scenarios.

Two approaches were used for short-term forecasting of the confirmed COVID-19 cases in Lithuania: the ARIMA model and a new attention-based forecasting method, which combines machine learning techniques and statistical methods. The novelty of the approach presented above is the use of data from countries with a longer history of the disease to forecast trends in Lithuania. To this end, the authors introduce the data registration from the first confirmed case of COVID-19. Such a way of data registration and integral data analysis using techniques for clustering and multidimensional data dimensionality reduction allows to assess trends in the spread of the virus in different countries and to group them according to similarity, i.e., to draw attention to those countries where the spread of COVID-19 behaves in a similar way. Moreover, the proposed approach allows to assess the dynamics of the spread of the virus and changes in the situation over time. The clustering analysis shows the specificity of the virus spread and enables to review the measures applied in the countries of the same cluster to control the virus and assess the impact (effectiveness) of the measures applied on the increase in the number of newly confirmed cases of the disease. The attention-based focus and identification of countries that are similar to the investigative one, i.e., Lithuania, with the ability to have a longer history of virus spread analysis, as well as the forecast based on their trends, allows to create and foresee the virus spread scenarios based on the historical data of other countries.

Summarizing the results of the forecasting, it can be concluded that both methods demonstrate similar accuracy in forecasting of the so-called first wave time period COVID-19 cases (March 12–June 30, 2020), none of the methods outperforms the other. The forecast accuracy obtained using the attention-based forecasting, taking the second wave (October 1, 2020–January 31, 2021), is slightly lower compared to the results obtained by ARIMA. The explanation can be related to the fact that the situation and trends of confirmed cases in the countries being rather different, the number of countries in the same cluster as Lithuania is not large, and the distances from Lithuania in the cluster are varied. However, the attention-based forecasting approach gives promising results. Higher forecast accuracy is achieved in periods when the countries in the same cluster are closer to Lithuania. The two approaches discussed above complement each other and provide insights for the short-term forecasting of COVID-19 spread and enable to validate the forecasting results. The dimensionality reduction techniques, viewed as an attention-based method for similar COVID-19 spread country or region selection, combined with regression analysis, provide a means to validate the forecasting results. The approach presented in the paper can be applied to any country with the view to analyze other pandemic situations.

Acknowledgement: The authors are thankful for the high-performance computing resources provided by the Information Technology Open Access Center at the Faculty of Mathematics and Informatics of Vilnius University Information Technology Research Center.

Funding Statement: This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-COV-20-4.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] J. A. Doornik, J. L. Castle and D. F. Hendry, “Short-term forecasting of the coronavirus pandemic,” *International Journal of Forecasting*, 2020. <https://www.sciencedirect.com/science/article/pii/S0169207020301412>.

- [2] S. He, Y. Peng and K. Sun, "SEIR modeling of the COVID-19 and its dynamics," *Nonlinear Dynamics*, vol. 101, no. 3, pp. 1667–1680, 2020.
- [3] E. L. Piccolomini and F. Zama, "Monitoring Italian COVID-19 spread by an adaptive SEIRD model," *MedRxiv*, vol. 15, no. 8, pp. e0237417, 2020.
- [4] R. Ünlü and E. Namli, "Machine learning and classical forecasting methods based decision support systems for covid-19," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1383–1399, 2020.
- [5] T. Kufel, "ARIMA-Based forecasting of the dynamics of confirmed covid-19 cases for selected european countries," *Equilibrium*, vol. 15, no. 2, pp. 181–204, 2020.
- [6] J. C. Mora, S. Perez, I. Rodriguez, A. Nunez and A. Dvorzhak, "Application of a semi-empirical dynamic model to forecast the propagation of the COVID-19 epidemics in Spain," *MedRxiv*, vol. 2, no. 4, pp. 452–469, 2020.
- [7] M. Perc, N. Gorišek Miksić, M. Slavinec and A. Stožer, "Forecasting covid-19," *Frontiers in Physics*, vol. 8, pp. 127, 2020.
- [8] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani and L. dos S. Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil," *Chaos, Solitons and Fractals*, vol. 135, pp. 109853, 2020.
- [9] K. Sarkar, S. Khajanchi and J. J. Nieto, "Modeling and forecasting the COVID-19 pandemic in India," *Chaos, Solitons and Fractals*, vol. 139, pp. 110049, 2020.
- [10] İ Kırbaş, A. Sözen, A. D. Tuncer and FŞ Kazancıoğlu, "Comparative analysis and forecasting of COVID-19 cases in various european countries with ARIMA, NARNN and LSTM approaches," *Chaos, Solitons and Fractals*, vol. 138, pp. 110015, 2020.
- [11] F. Rahmadani and H. Lee, "Hybrid deep learning-based epidemic prediction framework of covid-19: South Korea case," *Applied Sciences (Switzerland)*, vol. 10, no. 23, pp. 1–21, 2020.
- [12] B. S. Gill, V. J. Jayaraj, S. Singh, S. M. Ghazali, Y. L. Cheong *et al.*, "Modelling the effectiveness of epidemic control measures in preventing the transmission of COVID-19 in Malaysia," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, pp. 1–13, 2020.
- [13] R. Majhi, R. Thangeda, R. P. Sugasi and N. Kumar, "Analysis and prediction of COVID-19 trajectory: A machine learning approach," *Journal of Public Affairs*, pp. e2537, 2020. <https://onlinelibrary.wiley.com/doi/10.1002/pa.2537>.
- [14] S. K. Sharma, S. Bhardwaj, R. Bhardwaj and M. Alowaidi, "Nonlinear time series analysis of pathogenesis of covid-19 pandemic spread in Saudi Arabia," *Computers, Materials and Continua*, vol. 66, no. 1, pp. 805–825, 2021.
- [15] V. Papastefanopoulos, P. Linardatos and S. Kotsiantis, "COVID-19: A comparison of time series methods to forecast percentage of active cases per population," *Applied Sciences (Switzerland)*, vol. 10, no. 11, pp. 3880, 2020.
- [16] O. D. Ilie, R. O. Cojocariu, A. Ciobica, S. I. Timofte, I. Mavroudis *et al.*, "Forecasting the spreading of COVID-19 across nine countries from Europe, Asia, and the American continents using the arima models," *Microorganisms*, vol. 8, no. 8, pp. 1–19, 2020.
- [17] B. Yan, X. Tang, B. Liu, J. Wang, Y. Zhou *et al.*, "An improved method for the fitting and prediction of the number of covid-19 confirmed cases based on LSTM," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1473–1490, 2020.
- [18] R. G. da Silva, M. H. Dal Molin Ribeiro, V. C. Mariani and L. dos Santos Coelho, "Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables," *ArXiv*, vol. 139, pp. 110027, 2020.
- [19] M. H. Tayarani N., "Applications of artificial intelligence in battling against covid-19: A literature review," *Chaos, Solitons and Fractals*, vol. 142, pp. 110338, 2021.
- [20] O. Castillo and P. Melin, "Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic," *Chaos, Solitons and Fractals*, vol. 140, pp. 110242, 2020.

- [21] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi and R. Gloaguen, "COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach," *MedRxiv*, vol. 8, no. 6, pp. 890, 2020.
- [22] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis," *Chaos, Solitons and Fractals*, vol. 135, pp. 109850, 2020.
- [23] J. Panovska-Griffiths, "Can mathematical modelling solve the current covid-19 crisis?," *BMC Public Health*, vol. 20, no. 1, pp. 551, 2020.
- [24] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp. 1–22, 2008.
- [25] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods: Theory and Methods*, 2nd ed., New York, NY: Springer New York, 1991.
- [26] T. Kohonen, *Self-Organizing Maps*, vol. 30. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.
- [27] J. Venskus, P. Treigys, J. Bernatavičienė, V. Medvedev, M. Voznak *et al.*, "Integration of a self-organizing Map and a virtual pheromone for real-time abnormal movement detection in marine traffic," *Informatica (Netherlands)*, vol. 28, no. 2, pp. 359–374, 2017.
- [28] L. Van Der Maaten and G. Hinton, "Visualizing data using t-sNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2625, 2008.
- [29] G. Dzemyda, O. Kurasova and J. Žilinskas, *Multidimensional Data Visualization*, vol. 75, New York, NY: Springer New York, 2013.
- [30] J. Bernatavičienė, G. Dzemyda, G. Bazilevičius, V. Medvedev, V. Marcinkevičius *et al.*, "Method for visual detection of similarities in medical streaming data," *International Journal of Computers, Communications and Control*, vol. 10, no. 1, pp. 8–21, 2015.