Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Expert Systems With Applications 214 (2023) 119034

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

COVID-19 forecasting using shifted Gaussian Mixture Model with similarity-based estimation

Emre Külah *, Yusuf Mücahit Çetinkaya, Arif Görkem Özer, Hande Alemdar

Department of Computer Engineering, Middle East Technical University, Cankaya 06800, Ankara, Turkey

ARTICLE INFO

Keywords:
COVID-19
Gaussian mixture models
Time-series data
Similarity-based estimation
Trend similarity score

ABSTRACT

The COVID-19 pandemic has caused a pronounced disturbance in the social environments and economies of many countries worldwide. Credible forecasting methods to predict the pandemic's progress can allow countries to control the disease's spread and decrease the number of severe cases. This study presents a novel approach, called the Shifted Gaussian Mixture Model with Similarity-based Estimation (SGSE), that forecasts the future of a specific country's daily new case values by examining similar behavior in other countries. The model uses daily new case values collected since the pandemic began and finds countries with similar trends using a specific time offset. The daily new case values data between the first day and $(today-N)$th day are transformed by employing the Gaussian Mixture Model (GMM) and, subsequently, a new vector of features is obtained for each country. Using these feature vectors, countries that show similar statistics in the past are found for any forecasted country. The future of the corresponding country is forecasted by taking the mean of the time-series plots after the offset points of similar countries are calculated. A brand new metric called a trend similarity score, which calculates the similarity between forecasted and actual values is also presented in this study. While the SGSE trend similarity score median varies between 0.903–0.947, based on the selection of the distance metric, the ARIMA model yields only 0.642. The performance of the SGSE was compared in seven European countries using four different public projects submitted to The European COVID-19 Forecast Hub. The SGSE gives the most accurate forecasts compared to all other models. The test sets' results show that trends and plateaus are predicted accurately for many countries.

1. Introduction

At present, the world continues to struggle with endless new waves of coronavirus (COVID-19) variants. The pandemic has greatly affected the economies, industries, business sectors, educational realms, and social environments of many countries. It is therefore essential to apply vaccination and social distancing rules that aim to "flatten the curve" and reduce daily new case values. Another important consideration is analysis of newly developing waves of the pandemic. With the help of credible predictions, countries can more effectively manage restrictions and take better precautions in the face of future waves of COVID and of similar diseases.

Since the first incidence of COVID-19, a significant number of studies have been conducted that forecast the number of future cases (Shinde et al., 2020). Considering the incubation period and transmission of the disease, the short-term future of the pandemic in a specific country mainly depends on the recent past (Satrio, Darmawan, Nadia, & Hanafiah, 2021; Tandon, Ranjan, Chakraborty, & Suhag, 2020; Zheng et al., 2020). On this basis, many recent studies have considered the historical data of the corresponding country (Hu, Ge, Li, Jin, & Xiong, 2020; Pham, Nguyen, Huynh-The, Hwang, & Pathirana, 2021). In general, the methods used in the literature mainly derive from auto-regressive and deep learning models (Benvenuto, Giovanetti, Vassallo, Angeletti, & Ciccozzi, 2020; Dehesh, Mardani-Fard, & Dehesh, 2020; Hu et al., 2020; Ketu & Mishra, 2022; Liao et al., 2021; Naeem et al., 2022; Saba & Elsheikh, 2020; Tandon et al., 2020).

Auto-regressive models are based on a country's history, while exponential models may not exactly depend on a country's historical information. However, geographic location and condition affect the spread of the virus, both within countries and across countries.

Traveling to and from countries was and is one of the key reasons why COVID-19 spread so widely (Farzanegan, Gholipour, Feizi, Nunkoo, & Andargoli, 2021; Mousavi et al., 2020). Consequently, although it seems that the disease has spread across countries in different ways, in most places, the fluctuations in the number of new cases look alike (Gautam, 2022; Hu et al., 2020).

* Corresponding author.
E-mail addresses: kulah@ceng.metu.edu.tr (E. Külah), yusufc@ceng.metu.edu.tr (Y.M. Çetinkaya), gorkem@ceng.metu.edu.tr (A.G. Özer), alemdar@ceng.metu.edu.tr (H. Alemdar).

https://doi.org/10.1016/j.eswa.2022.119034
Received 10 May 2022; Received in revised form 9 October 2022; Accepted 11 October 2022
Available online 18 October 2022
0957-4174/© 2022 Elsevier Ltd. All rights reserved.

It starts with a single incident; then, the number of cases increases rapidly, and the isolated case turns into a pandemic, resulting in the first wave. Therefore, some countries followed stricter border policies to further prevent the disease from spreading into their respective territories (Imtyaz, Haleem, & Javaid, 2020). Depending on the exact precautions taken, some countries managed to decrease the number of new cases (Sahoo & Sapra, 2020; Sun, Zhang, Yang, Wan, & Wang, 2020). However, almost every country experienced a second and third wave (Fisayo & Tsukagoshi, 2021; Seong et al., 2021). The behavioral similarities in the number of new cases, which differ on the time axis, lead to a search of a specific country's progress in relation to other countries' past time-series data. When time-series data of countries' daily new cases are examined, it can be observed that the data show similar patterns for many countries.

Given the necessity of forecasting, several methods have already been presented in the literature. The objective and fair performance comparison among these proposed methods is another crucial requirement. Several general-purpose, widely-used performance metrics exist such as mean absolute error (MAE), root means squared error (RMSE), and mean absolute percentage error (MAPE); these can be used to compare the performance of different approaches. These metrics consider the exact numerical difference between actual data and forecast results; however, the accurate estimation of the trend is more important than the net difference in forecasting the course of the pandemic.

In order to address these issues, the following contributions are made in this study: first, a novel method, the Shifted Gaussian Mixture Model with Similarity-based Estimation (SGSE), is proposed. The SGSE locates countries that have faced similar past patterns to forecast the future of the forecasted country by imitating the progress of similar countries starting from the end of the pattern found. Waves in the time-series data of the countries are represented by Gaussian distributions that make up a probabilistic model called the Gaussian Mixture Model (GMM). The time-series data of the countries are cropped according to various shifted time offsets, and GMM representations are constructed for all country-time offset pairs. Subsequently, GMM representations are used to find which countries' pasts are similar to the progress of the forecasted country (Eirola & Lendasse, 2013; Povinelli, Johnson, Lindgren, & Ye, 2004). For each similar country-time offset pair, the time-series data that follows the offset point is used while estimating the future new case values of the forecasted country. Second, a completely new performance evaluation metric, the trend similarity score (TSS), is presented. TSS is calculated using the cosine similarity of actual and forecasted trends. By using TSS, we can observe how much the forecasted values deviate from the actual trend—in contrast to metrics that only employ numeric differences such as MAE, RMSE, and MAPE. This new approach is validated using a publicly available COVID-19 dataset published by Roser, Ritchie, Ortiz-Ospina, and Hasell (2020) that uses 20 countries from different continents. The results demonstrate that the SGSE outperforms the baseline model by a large margin, considering all evaluation metrics. Third, the SGSE is an explainable model and can assist decision-makers while they take action. As the SGSE finds similar country-time interval pairs, decision-makers are able to pursue actions that result in a flattened curve or avoid actions that can potentially cause a new wave of disease for a similar country.

The remaining sections in this study are organized as follows: the next section provides a list of related studies in the literature. Next, details about the SGSE, along with the dataset, preprocessing operations, and data transformations are presented in Section 3. In Section 4, the results of the experimental evaluation are presented using RMSE, MAPE, and TSS metrics. The study concludes in Section 5.

2. Related work

Since the COVID-19 pandemic began, numerous studies have been conducted on forecasting models; in these, researchers have looked at the forecasting of various statistics such as mortality rate, pandemic end date, and the number of daily new cases. This section examines studies that have focused on the number of daily new cases. A majority of these specifically observed auto-regressive and deep learning models.

In their study, Dehesh et al. (2020) used a naive auto-regressive integrated moving average (ARIMA) model with different p (order of auto-regression), q (order of moving average), and d (degree of non-seasonal difference) parameters. They applied the ARIMA training process in the simplest form. Using auto-correlation function (ACF) and partial auto-correlation function (PACF) graphics, the best p, d, q parameters are observed. ARIMA models were trained using these parameters on five different countries from different continents. The data used cover 41 days for these five countries. The graphs show the forecasting results, and the statistical behavior of these five countries in the near future can be observed. Predictions for the next 17 days were as follows: a stable trend for China, unstable trends for Italy and Iran, a stationary trend for South Korea, and mostly controlled daily new cases for Thailand. With the exception of China, these countries had an uptrend between the specified dates. While numbers show a slight increase in China, the country showed a more stable trend as figures had already risen a lot by then.

Saba and Elsheikh (2020) integrated two commonly used approaches: the autoregressive model and the neural network. In these approaches, the model used is the nonlinear autoregressive neural network (NARANN); it predicts a time-series stemming from that series' past values. In this study, the authors compared the performances of the ARIMA and NARANN models using accumulated daily new cases in Egypt. The capabilities of the NARANN model are verified using test data for both short-term and long-term forecasts. Different statistical indices were used to measure the performance of the model. In the first process, the researchers compared the 7-day forecasts of the ARIMA and NARANN models. The ARIMA model's absolute percentage error increased with time, starting at 3.08% and eventually reaching 29.48%. A longer forecasting process using the NARANN model resulted in 7.75 mean absolute error and 10.4 root mean squared error.

Hu et al. (2020) conducted a pioneering study that used auto-encoders to forecast the number of daily new cases. The authors used a modified auto-encoder in their work; this model has a different structure than the classical auto-encoder because the number of nodes in layers is not the same. Input data is formed in time segments containing a particular day and its seven successive days. A total of 128 segments of time-series with a length of eight were used for training. The previous day is used as an observational input in forecasting the number of daily new cases for the next day. This process is repeated in forecasting the future with different n-step models. In the study, results were visualized for China and its provinces. The next two months for China were forecasted, but the first eight days, which can be used for the test, were used to measure model performance. For different steps from 6 to 10, 1.64%, 2.27%, 2.14%, 2.08%, 0.73% absolute percentage errors were obtained, respectively.

Singhal, Singh, Lall, and Joshi (2020) made COVID-19 predictions using two different models. The first one is a mathematical model accounting for various parameters relating to the spread of the virus. The second is a nonparametric model. First, the Fourier Decomposition Method decomposed time-series data into a desired set of frequency bands. The spread size is estimated by fitting the mixture of Gaussian functions on the trends obtained in the previous step. The performance of the proposed GMM model for the number of daily new cases was measured, and 1842.5 mean absolute error for the world, 731 for the USA, 102.38 for Italy, and 53.53 for India were obtained. Ayoobi et al. (2021) used a variety of deep learning models such as GRU, LSTM, and convolutional LSTM to predict new cases and mortality rates in Iran and Australia — one, three, and seven-day ahead. They reported that the bidirectional models' performances worked better than non-bidirectional.

Table 1
Summary statistics of smoothed daily new cases per million during the first and second waves with the structural view of the waves' daily cases patterns between given dates for several countries.

Team	Model	Description
Epiforecasts	EpiNow2	EpiNow2 implements a Bayesian latent variable to create an exponential growth model. The incidence of time steps is estimated using the trajectory of time-varying R_t calculated for each subsequent time step (Abbott et al., 2020).
IEM Health	CovidProjections	It uses SEIR model projections for daily incident confirmed COVID-19 cases and deaths by using AI to fit actual cases observed (Suchoski, Stage, Gurung, & Baccam, 2022).
Masaryk University	VAR	It uses a vector auto-regression model fitted to outlier-corrected transformed weekly aggregated series (Pavlik et al., 2020).
European COVID-19 Forecast Hub	ensemble	It is an ensemble model of all projects submitted to the European COVID-19 Forecast Hub (Sherratt et al., 2022).

Xu, Magar, and Barati Farimani (2022) combined deep learning models to forecast the subsequent 14 days of new case totals for Brazil, India, and Russia only. They reported that the LSTM model performed comparably well and exhibited the best performance considering the evaluation metrics MAE, R^2, and EV. The model is not capable of forecasting daily new cases totals for countries with rapidly changing numbers, whereas it forecasts cumulative cases more successfully. The reported MAE results of the CNN-LSTM model for Brazil, India, and Russia are 15563, 5245, and 986, respectively.

In a more recent study, Gautam (2022) used transfer learning and a long short-term memory (LSTM) network to forecast the smoothed number of daily new cases. Approximately 20 days of past data were given to the model as input, and five consecutive future days were predicted. The authors transferred the nonlinear patterns of certain countries to other countries' training processes. Models trained and tested with the same country and models trained with a country and tested with another country are visualized and compared. 5-day predictions were made with the Italy and America models for five countries, and an average of 0.99 and 1.51 root mean squared errors were obtained, respectively.

It can be misleading to compare the performances of different studies due to the differences in experimental settings, i.e., the countries reported and the time frame used in the evaluations. Besides, the data used differ in the studies. Some researchers use new case values daily, whereas others use cumulative daily new case values. The baseline and proposed models must be applied to the same dataset. For these reasons, the results found in this study are compared against a baseline method (ARIMA) widely used in the literature. Additionally, the performance of the SGSE is compared with four chosen models that are publicly available in The European COVID-19 Forecast Hub, which is a collaborative project that aims to collect, evaluate, and combine the forecasts of weekly COVID-19 cases in European countries (Sherratt et al., 2022). The models that the SGSE is compared with are listed in Table 1 along with short descriptions.

3. Shifted GMM with similarity-based estimation

In this section, as a first step, the input data that is fed into the model is described. The preprocessing steps performed on the data are then explained, including the Gaussian mixture representation construction. Finally, the details of the similarity-based forecasting approach are presented.

3.1. Dataset & preprocessing

The data have been taken from an open dataset of Our World In Data (Roser et al., 2020), which shows the number of daily new cases, the number of new deaths, the accumulated number of new cases, the accumulated number of deaths, and numerous other statistics from all countries linked to the COVID-19 pandemic. This dataset is available from December 31, 2019 to the present time; it is updated daily and made publicly available. The present study uses data up to December 16, 2021. The smoothed daily new cases per million fields are used within the scope of this study. The data are obtained by dividing the number of daily new cases (7-day smoothed) by the total population, then multiplying by 1 million. A total of 36 countries were excluded from the dataset since their statistics were corrupted or noisy.

Table 2 shows the statistical summaries of smoothed daily new cases per million during the first two waves some countries experienced during the COVID-19 pandemic. The pattern column reveals that some countries experienced the pandemic at similar structural patterns but different magnitude and dates. For example, during the first wave, Albania and Bosnia and Herzegovina had a similar pattern and normalized mean and standard values, but the intensity of the new case rate signal they experienced was different. Similarly, the Germany–Japan–United Kingdom and Poland–Romania plots represent groups with similar daily new case plot appearances with different amplitude and width.

Countries do not always have to accompany a similar country as time varies. During the second wave, Bosnia and Herzegovina's plot is not similar to Albania's but coincides with Poland and Romania. France–Belgium and Albania–Sweden have analogous plots. Moreover, as shown in Table 2, the duration and start–end dates of the experienced waves may differ. When the dataset is examined, it can be observed that the date difference was not more than two months to find the similarity.

When the dynamics in the data are examined, detecting similar patterns seems to be a beneficial approach in estimating the number of future new cases in a country because some countries have experienced the pandemic in a similar manner before others. Although it is easy to detect pattern similarity between the two signals upon initial observation, in order to express it mathematically, a representation for times-series data is required. The GMM is leveraged to represent and calculate the similarity.

3.2. GMM representation of time-series data

The GMM is a weighted sum of Gaussian densities, which consists of mean μ and covariance Σ values (Reynolds, 2009). It is possible to use one-dimensional GMM, $p(x)$, to represent time-series data (Kumar, Patel, & Woo, 2002):

$$p(x) = \sum_{k=1}^{K} \omega_k (x \mid \mu_k, \Sigma_k) \tag{1}$$

where K represents the number of mixture components, ω_k is the prior probability of the component, μ_k is the mean, Σ_k is the covariance of the component.

For a given dataset, the standard approach to train a GMM is the EM algorithm (Zhou, Lim, Kwon, et al., 2014). It is essential to first decide how many components – mean-covariance pairs – will represent

Table 2
Summary statistics of smoothed daily new cases per million during the first and second waves with the structural view of the waves' daily cases patterns between given dates for several countries.

Country	Wave 1						Wave 2					
	Dates	Norm. Mean	Norm. Std.	Min	Max	Pattern	Dates	Norm. Mean	Norm. Std.	Min	Max	Pattern
Albania	2020-03-14 to 2020-05-13	0.478	0.292	1.85	8.21		2020-10-01 to 2021-06-04	0.427	0.287	4.35	389.48	
Belgium	2020-03-04 to 2020-06-21	0.376	0.339	0.27	125.12		2020-09-03 to 2021-06-30	0.167	0.184	36.27	1533.14	
Bosnia and H.	2020-03-10 to 2020-06-07	0.540	0.302	0.22	15.99		2020-10-01 to 2021-06-01	0.413845	0.299	25.33	499.90	
France	2020-03-01 to 2020-05-04	0.209	0.264	0.379	196.66		2020-09-01 to 2021-07-01	0.315	0.189	27.05	906.66	
Germany	2020-03-01 to 2020-06-16	0.308	0.304	0.230	67.08		2020-10-01 to 2021-07-01	0.474	0.266	7.24	309.93	
Japan	2020-03-01 to 2020-06-01	0.306	0.326	0.14	4.37		2020-07-01 to 2021-07-01	0.316	0.267	0.88	52.12	
Poland	2020-03-09 to 2020-07-09	0.648	0.263	0.06	11.81		2020-09-15 to 2021-06-15	0.348	0.272	7.79	753.85	
Romania	2020-03-02 to 2020-06-03	0.551	0.339	0.02	19.22		2020-09-25 to 2021-07-01	0.401	0.272	2.71	440.87	
Sweden	2020-03-02 to 2020-07-22	0.476	0.271	0.25	104.44		2020-09-18 to 2021-06-01	0.496	0.258	23.64	710.96	
United Kingdom	2020-03-02 to 2020-07-04	0.467	0.343	0.22	72.17		2020-08-15 to 2021-04-14	0.280	0.245	14.77	887.05	

the series. Bayesian information criterion (BIC) is a model selection criterion commonly used to decide on the number of components of the GMM (Schwarz, 1978).

Several values for selecting the ideal number of GMM components for all countries are attempted, and by using the elbow method, $K = 5$ is set. Hence, five GMM components are used to represent the time-series data ($p \leq .001$). Instead of using the time-series data for a particular country, GMM is fitted on the data with five components and mean-covariance pairs; a vector with ten elements is then created. Following creation of the vector, it is updated by ordering the components according to their mean values. In Eq. (2), the vector of the ith country denoted by v_i is shown. In this equation, μ_{ij} and σ_{ij} represent the mean and covariance of jth component of ith country, respectively.

$$v_i = \{\mu_{i1}, \sigma_{i1}, \mu_{i2}, \sigma_{i2}, \mu_{i3}, \sigma_{i3}, \mu_{i4}, \sigma_{i4}, \mu_{i5}, \sigma_{i5}\} \tag{2}$$

In the remaining sections, the term 'GMM representation' will describe this vector. Fig. 1 shows Germany's time-series data and Gaussian components when GMM is fitted to the data:

3.3. Forecasting using similarity-based estimation

Before moving on to the training process, it is necessary to extract GMM representations from each country based on the different cutoff dates to learn which countries' history shows a similar time interval to the forecasted country. After examining the countries' data, it was decided that the time difference between countries can be two months at most. For this reason, five different replicas of the time-series data were extracted for each country, and each replica goes back 14 days from the previous replica. GMM representations are then created for each of the replicas. The GMM representations obtained for each country formed the dataset for the rest of the processes. For 182 countries with five different GMM representations, 910 GMM representations are obtained for the dataset. The dataset grows incrementally, and past days are not recalculated for upcoming days. Since the recorded data is used as a lookup table, the system effectively adapts to the new data. Fig. 2 shows the process in detail.

For 182 countries with five replicas, the structure of the created dataset D is given in Eq. (3):

$$D = \begin{Bmatrix} \mu_{111} & \sigma_{111} & \cdots & \mu_{11k} & \sigma_{11k} \\ \mu_{121} & \sigma_{121} & \cdots & \mu_{12k} & \sigma_{12k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{i(j-1)1} & \sigma_{i(j-1)1} & \cdots & \mu_{i(j-1)k} & \sigma_{i(j-1)k} \\ \mu_{ij1} & \sigma_{ij1} & \cdots & \mu_{ijk} & \sigma_{ijk} \end{Bmatrix} \tag{3}$$

where i represents the country index ($i \in [1, 182]$), j represents replica index ($j \in [1, 5]$), and k represents the GMM component index ($k \in [1, 5]$), respectively. For instance, $(\mu_{123}, \sigma_{123})$ represents properties of third GMM component of the second replica of the first country. The last row of the matrix includes the GMM components of the 5th replica of the 182nd country.

The ten most similar samples with regards to each country are found in the normalized version of the created D. These samples can be different replicas of any country or different replicas of different countries. Euclidean distance, Jensen–Shannon distance (Lin, 1991), and Wasserstein distance (Vaserstein, 1969) measures are used to calculate the similarity, and their results are compared in terms of prediction errors.

The Jensen–Shannon distance between two distributions p and q is defined as:

$$D_{JS}(p \parallel q) = \sqrt{\frac{1}{2} D_{KL}(p \| \frac{p+q}{2}) + \frac{1}{2} D_{KL}(q \| \frac{p+q}{2})} \tag{4}$$

The Kullback–Leibler Divergence D_{KL} is defined as:

$$D_{KL}(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{5}$$

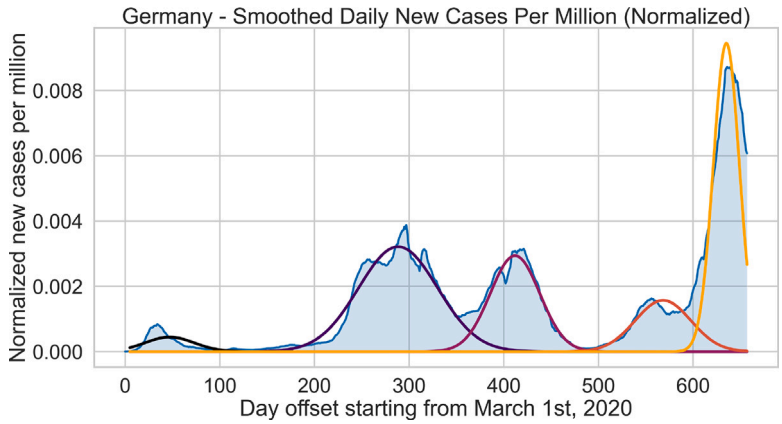


Fig. 1. Normalized smoothed number of daily new cases per million for Germany, between March 1, 2020–December 14, 2021 and fitted Gaussian components.

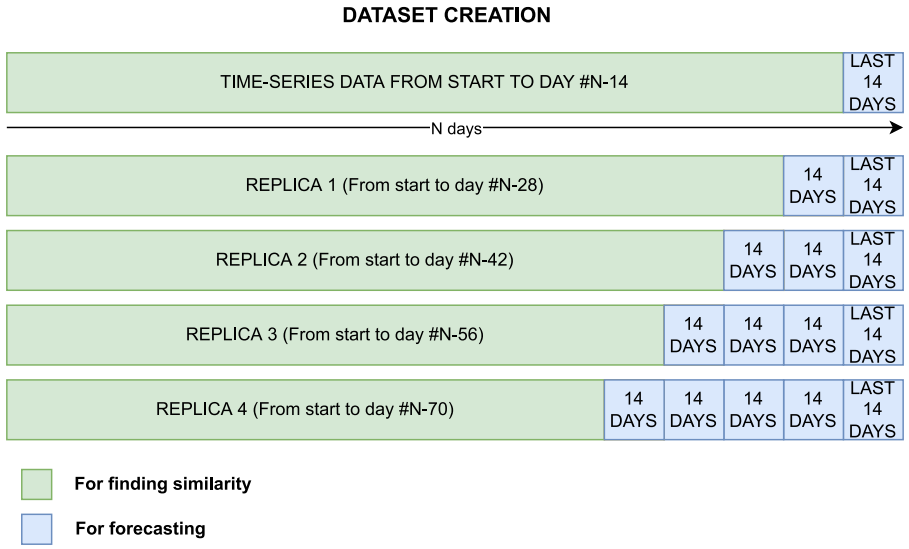


Fig. 2. Dataset creation by replicating original time-series data with an M-day shifted cutoff date.

Kullback–Leibler divergence is not symmetric and does not satisfy triangular inequality. Therefore, Jensen–Shannon distance is preferred to determine similar countries.

Wasserstein distance measures how much cost is required to transform one distribution to another, and this is formulated as:

$$D_W(p,q) = \inf_{\pi \in \prod(p,q)} \int_{\mathbb{R}x\mathbb{R}} |x-y| d\pi(x,y) \tag{6}$$

When the COVID-19 plot for each country is examined, it can be said that some countries have visually similar plots; however, they have dealt with the pandemic in a slower manner than others. The fact that similar samples have different lengths also demonstrates this. The velocity ratio should also be considered when forecasting the future of the forecasted country. Therefore, while forecasting the future of countries, the near future of each similar sample is interpolated (Akima, 1974) to coincide with the next 14 days length of the forecasted country using the velocity ratio between the forecasted country and similar samples. Let r be the velocity ratio between the forecasted country and a similar sample,

$$r = \frac{|x_i|}{|y_i|} \tag{7}$$

where x_i is the time-series data of the forecasted country and y_i is the corresponding time-series data of a similar sample. The value l gives us the vector length obtained from a similar sample's corresponding future.

$$l = 14 * r \tag{8}$$

By interpolating the first l days of the sample after the cutoff date, a vector of length 14 which will be an element of the final forecast vector, is then obtained.

To demonstrate, using the dataset D, with GMM components, the time series of country B with a shift amount of 28 days is found to be similar to the target country A. In other words, country B has a similar plot at the $(N-28)$th day to the plot of country A at the Nth day (see Fig. 3, $M = N - 28$). To put it more simply, country B experiences the pandemic $r = \frac{N}{N-28}$ times more quickly than country A. The velocity ratio r is used for "stretching" the data of country B. Upon stretching, interpolation is needed to "soften" the stretched data to prevent gaps among values. Later, since the forecasted number of days is 14, the first 14 values of the interpolated data of country B, starting from the $(N-27)$th day, are used while forecasting the future of country A.

Collecting interpolated future elements from the ten most similar samples, a set of vectors of length 14 is obtained. The behavior of these vectors shows how the forecasted country's estimated future will behave; however, their values do not match the number of daily new cases of the corresponding country. For this reason, as a last step in the process, these vectors are scaled to match the end of the time-series data of the country with the beginning of the forecasted element. The equally weighted mean of these forecast elements forms the estimated 14-day results of the forecasted country.

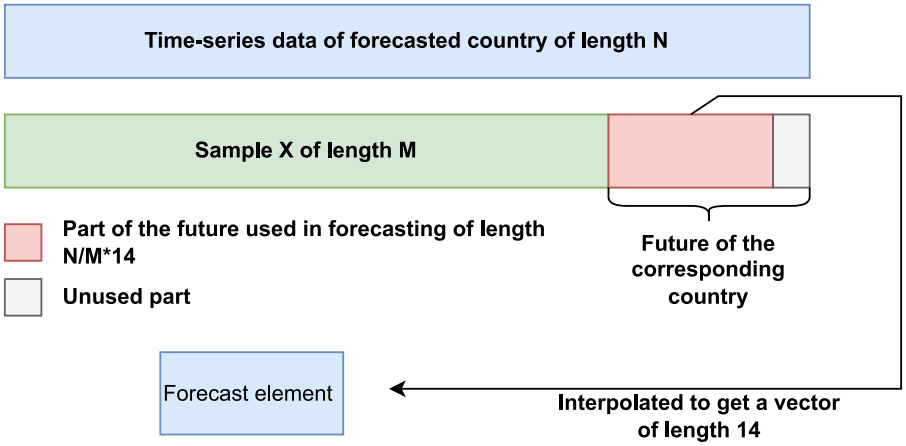


Fig. 3. Process of obtaining a forecast element by interpolating the corresponding part of the future of the sampled country.

To illustrate, the first value of the interpolated data (the value at $(N-27)$th day of country B) is not equal to the last value (the value at Nth day) of country A. Therefore, the scale factor s should be applied to the values of the interpolated data; Algorithm 1 shows the calculation and application of the scale factor s:

Algorithm 1: Scaling the forecast elements.

```
procedure SCALE(x, y, N)
  x ← time-series data of the target country
  N ← length(x)
  y ← the forecast values of length 14 after interpolation
    x_N ← x[N]
    y_0 ← y[0]
    s ← x_N / y_0
    for i ← 0 to 14 do
        y[i] ← y[i] * s
    end for
end procedure
```

Finally, after scaling, the average of forecast vectors from the ten most similar samples is taken to provide the estimated future vector for the target country. Eq. (9) shows the calculation of the estimated future:

$$f_i = \frac{\sum_{j=1}^{10} e_{ji}}{10} \tag{9}$$

where f_i is ith day of the estimated future, for $i \in [1, 14]$, and e_{ji} is the ith day of the jth forecast element

4. Experimental results

4.1. Overview

The EM algorithm and interpolation operations are implemented using Python's *scikit-learn* API. First, in this section – for a specific country – the closest samples and their time-series data are displayed to show the similarities. Next, the forecast data and the actual, confirmed, smoothed daily case numbers per million for some selected countries are visualized. Next, the forecast data and the actual, confirmed, and smoothed daily case numbers per million for several selected countries are visually presented. Finally, MAE, RMSE, MAPE, and the ACF1 score are listed to measure the performance of the created model. For all results, the used and presented COVID-19 data start on March 1, 2020, and end on December 15, 2021. The ACF1 is the autocorrelation of errors at lag 1. Other error functions are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{11}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - y_i}{x_i}\right| \tag{12}$$

4.2. Trend similarity score

MAE, RMSE, and MAPE are more efficient than other metrics for observation of the exact difference between actual data and forecast results. However, while estimating how COVID-19 will continue, accurate trend estimation is more important than the net difference. For this reason, it is more meaningful for us to fit linear lines on real data and estimation results and look at their cosine similarity. At this point, a new error metric, the TSS, is presented and Eq. (13) shows how it is calculated:

$$TSS = \frac{T_1 \cdot T_2}{\|T_1\| \times \|T_2\|} \tag{13}$$

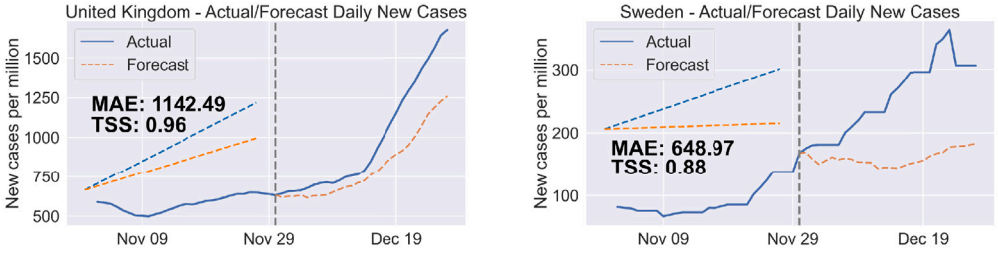
where T_1 and T_2 are the lines fitted on normalized actual and predicted data using linear regression. In Fig. 4, the estimations and errors for two different countries, the United Kingdom and Sweden are shown:

MAE and TSS error values for the UK's forecast are given in Fig. 4(a); and Sweden's can be found in Fig. 4(b). Although the MAE error for the UK is much higher than for Sweden, the TSS error value is much lower than Sweden's as the forecast trend is quite similar to the actual trend. One of the crucial reasons for this is that daily new cases per million values of countries are in very different ranges, and this problem can be solved using the MAPE error. Nevertheless, when TSS is compared with MAPE, it may result in a numerical amount different from the actual values; the comparison may also not display real performance in cases where the trend is correctly estimated.

4.3. Results

Two different experiments were conducted to measure model performance. The first experiment compares the model with an ARIMA model trained using the same dataset. The models' performances that make 14-day forecasts are measured with MAE and TSS errors. The second experiment compares the SGSE with four models submitted to the European COVID-19 Forecast Hub. In this organization, the submitted models provide weekly forecasts. Daily forecasts produced by the SGSE were aggregated, and 4-week forecasts were used in the evaluation.

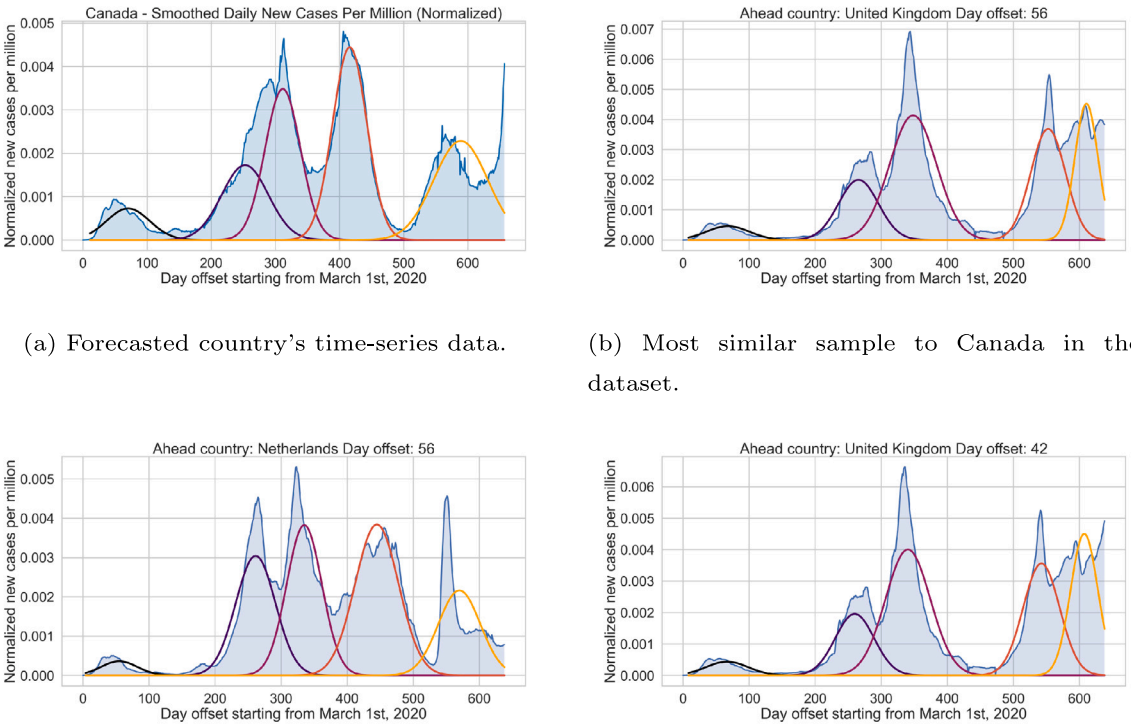
In the first experiment, forecasts are made for the 14 days following December 1, 2021, for 20 selected countries on different continents. The created prediction data covers the dates between December 1



(a) Forecast for the United Kingdom

(b) Forecast for Sweden

Fig. 4. MAE and TSS errors for the United Kingdom and Sweden using the SGSE with Wasserstein distance.



(a) Forecasted country's time-series data.

(b) Most similar sample to Canada in the dataset.

(c) Second most similar sample to Canada in the dataset.

(d) Third most similar sample to Canada in the dataset.

Fig. 5. Visualization of Canada's time-series data and the three most similar samples found using the GMM representation of Canada using Euclidean distance.

and December 14. For instance, in Fig. 5, the GMM representations of Canada and the top three samples most similar to Canada are visualized.

Table 3 lists all of the ten most similar samples to Canada's data. Each row represents a similar country and the shift offset days from the prediction day.

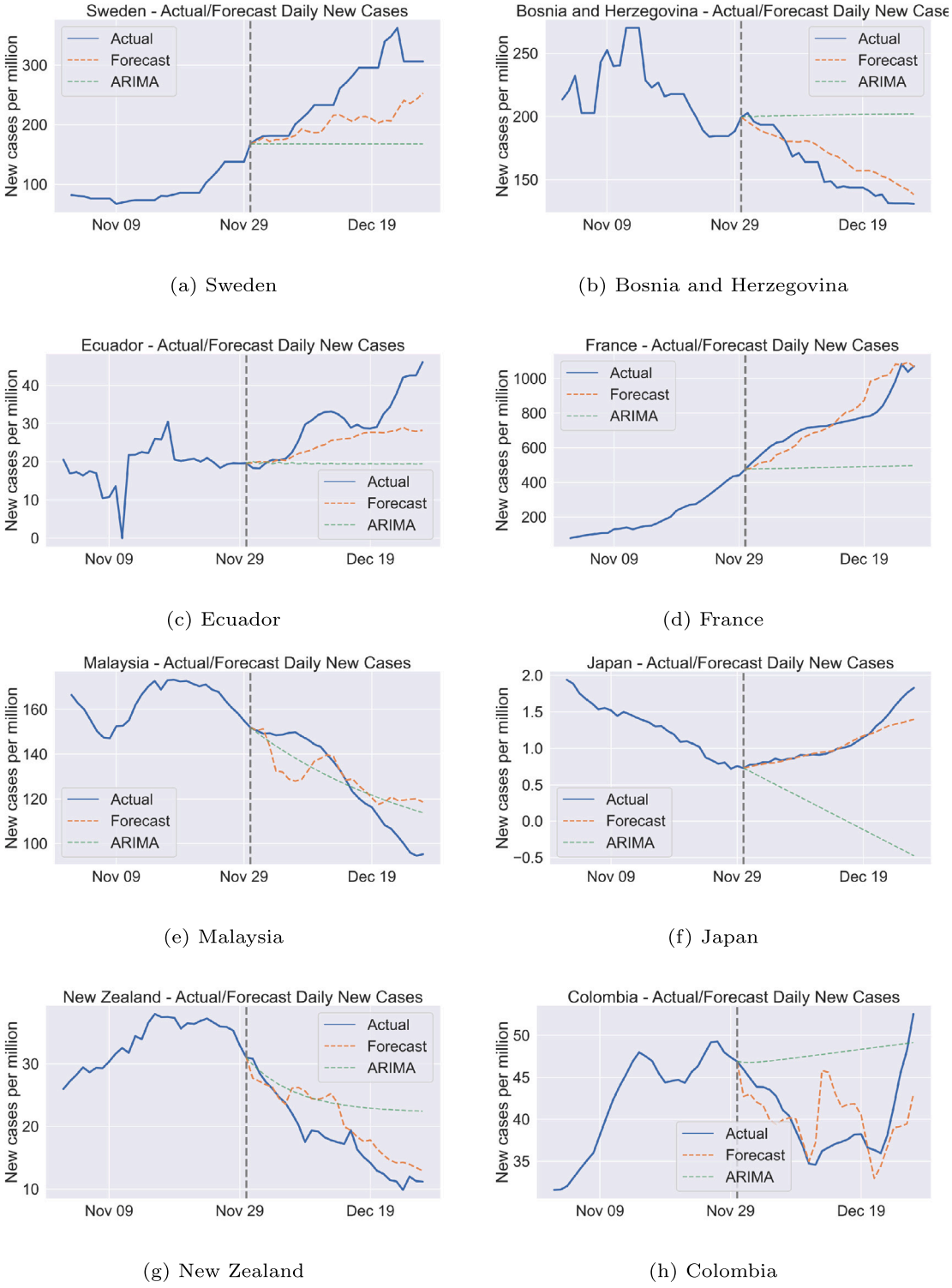
Table 3 shows that the GMM representation of the UK's time series data up to 56 days before is the most similar representation to the GMM representation of Canada's time-series data on the predicted day, meaning that the UK has dealt with the pandemic in a faster way than Canada. Therefore, the UK's pandemic experience can give insight about how the pandemic will continue in Canada.

Later, the UK's time-series data between 56 days before the prediction day and the prediction day itself is scaled taking into account the magnitude of the time-series data of Canada and used while calculating its country-specific forecast values. Since ten most similar samples have been chosen, 10 scaled future plots are obtained. In the end, the mean of the scaled future plots is taken to provide the final forecast values for Canada.

Table 3
Top 10 samples most resembling Canada's GMM representation.

Country	Offset
United Kingdom	56 days
Netherlands	56 days
United Kingdom	42 days
Cape Verde	14 days
Libya	70 days
Colombia	42 days
Netherlands	42 days
United Kingdom	70 days
Netherlands	70 days
Colombia	56 days

In Fig. 6, the estimation results of eight countries from different continents are visualized. The results indicate that the SGSE model is more accurate at forecasting the future than ARIMA's baseline model.



(a) Sweden (b) Bosnia and Herzegovina

(c) Ecuador (d) France

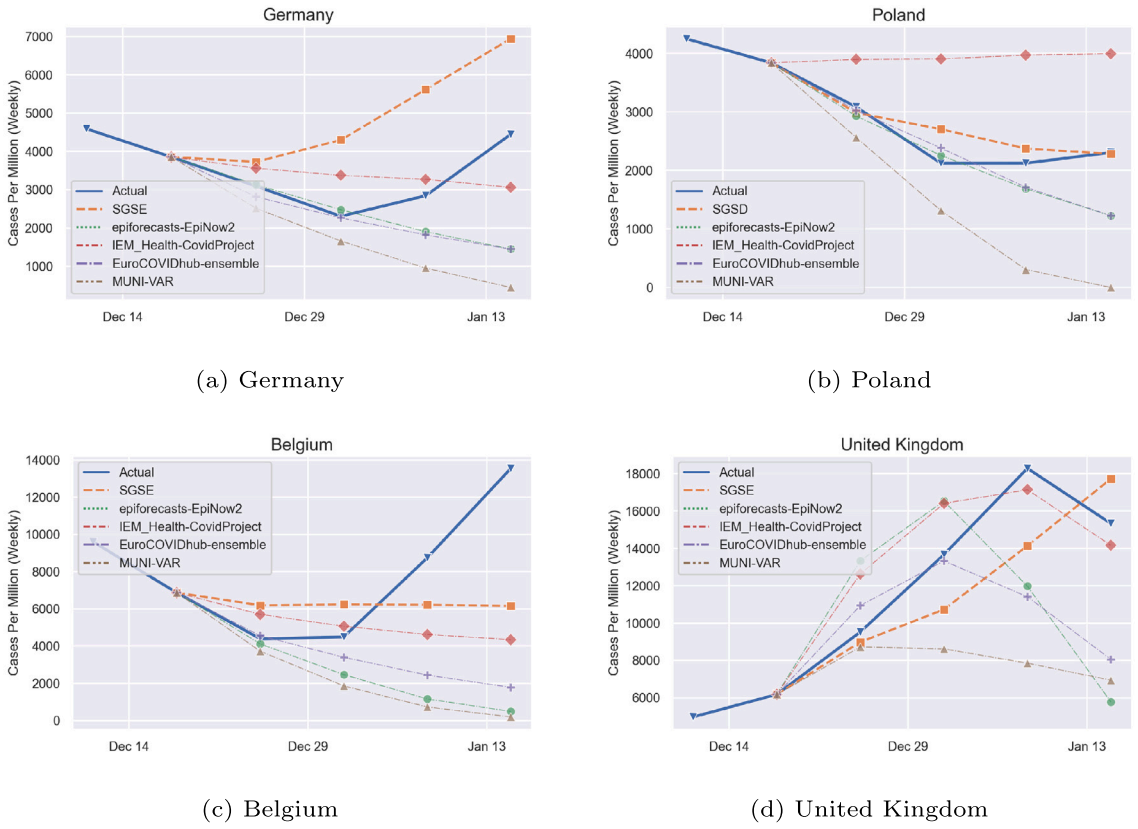
(e) Malaysia (f) Japan

(g) New Zealand (h) Colombia

Fig. 6. Actual/forecast number of cases for eight countries from different continents using Euclidean distance to find similar countries.

A decision-maker can thus take more consistent precautions using the SGSE model; there are two main reasons for this:

- The ARIMA model only considers the values observed until the moment and, therefore, fails to forecast with respect to real-life scenarios. It is useful, though, when there are no steep curves, and the trend is likely to be stable. The parameters should be selected wisely for each country. For instance, Japan's ARIMA forecast values keep decreasing since the plot until November 29th and has a negative slope. However, ARIMA failed to forecast Japan's future because there was a new wave around December 19th.
- The SGSE model forecast values are closer to real-life scenarios and more easily explainable since the SGSE model leverages the past pandemic experience of other countries. Even if the trend changes, the SGSE can quickly adapt to a new trend since countries experience the pandemic at different speeds; in other words, the observation time of steep waves is different. If the trend changes, there is at least one country that has faced a similar



(a) Germany (b) Poland

(c) Belgium (d) United Kingdom

Fig. 7. Visual comparison of the SGSE with four different models from The European COVID-19 Forecast Hub on four European countries.

trend before, and that country's experience can give insight into the predicted country's future.

When the SGSE is compared with various models provided by the European COVID-19 Forecast Hub, it can also be observed that the SGSE model forecast values are closer to the actual values compared to the other models. In Fig. 7, this is shown for four European countries: Germany, Poland, Belgium, and the UK. The SGSE provides more consistent forecasts than other models, especially for Germany, Belgium, and Poland. As mentioned before, for Fig. 6, the SGSE easily adapts to trend changes, and the shape of the plot in the future is more similar compared to the other models.

While leveraging other countries' pandemic experience, the SGSE makes magnitude adjustments while forecasting the future of the predicted country. This is an important feature of the SGSE since the plots seem irrelevant between the countries due to population differences. With magnitude adjustments on the plots, the SGSE can therefore find similarities and provide forecast values that overlap with the actual values in the plot.

The experiment results were obtained for the 20 countries In Table 4, the MAE, RMSE, MAPE, TSS, and ACF1 results for all 20 countries are listed. Table 4 contains error results of three distance methods. As seen in the table, an average of 10.75, 6.44, 5.65 MAE, 12.80, 8.16, 7.53 RMSE, and 0.13, 0.17, 0.14 MAPE errors are obtained with Euclidean, Wasserstein, and Jensen–Shannon distances for the dataset. In addition, average TSS values of 0.94, 0.90, and 0.93 for the new distance method were obtained, respectively.

In Fig. 6, forecast values that resulting from the SGSE and ARIMA models are compared visually by looking at the trend changes on the plots. In Table 5, the comparison is also made using MAE and TSS metrics. TSS metric results show that the SGSE is more successful in capturing trend changes in the data than the ARIMA model. Furthermore, MAE error results indicate that the SGSE produces fewer errors while providing forecasts for the future of the predicted country compared to the ARIMA model.

When the results are observed, this situation seems to be the case for all results obtained using Euclidian, Wasserstein, and Jensen–Shannon distance metrics.

In the second experiment, the weekly case numbers of seven European countries were estimated for four weeks starting from December 20th. The predictions were compared with those of four different models from the European COVID-19 Forecast Hub. Table 6 lists the results of these comparisons in relation to MAE and TSS errors.

4.4. Discussion

Considering the incorrect prediction's large divergence, comparing the results using the median of the loss or similarities instead of raw averages is preferable. The medians of the TSS for these distances are 0.947, 0.903, and 0.927. The Jensen–Shannon distance is more successful on MAE and RMSE scores, whereas Euclidean distance provides better results for MAPE and trend similarity. Wasserstein distance lies between MAE and RMSE, with a lower score for MAPE and trend similarity. As reported in Table 5, the proposed method, with any distance metrics, outperforms the ARIMA method, which only possesses a 0.642 TSS and 11.721 MAE cost.

Table 4 shows that the MAPE results of 13 countries are less than 0.1, which indicates that the distribution for these countries was predicted quite accurately (Lewis, 1982). In addition, the MAPE results for 18 of the countries are less than 0.2. Trend predictions for almost all of the test sets ($p < 0.01$) were made. Another score that shows how accurately the SGSE can predict trends is the TSS: for 19 of the countries, it is higher than 0.9.

When the forecast visualizations are examined, it can be observed that the waves in the plateau, uptrend, and downtrend were successfully estimated. On the other hand, erroneous estimations can be made

Table 4
Presentation of 14 days forecast errors of 20 countries from different continents.

Country	Euclidean distance					Wasserstein distance					Jensen–Shannon distance				
	MAE	RMSE	MAPE	TSS	ACF1	MAE	RMSE	MAPE	TSS	ACF1	MAE	RMSE	MAPE	TSS	ACF1
Albania	28.756	30.357	0.262	0.692	0.697	23.275	25.289	0.214	0.441	0.701	**19.327**	**20.444**	**0.176**	**0.963**	**0.297**
Angola	0.127	0.166	0.216	0.907	0.900	**0.095**	**0.127**	**0.159**	**0.946**	**0.880**	0.151	0.189	0.259	0.899	0.900
Argentina	**1.655**	**2.930**	**0.035**	**0.997**	**0.064**	5.331	6.863	0.106	0.934	0.788	2.517	3.657	0.053	1.000	0.164
Belgium	**208.256**	**221.561**	**0.152**	**0.999**	**0.549**	320.492	413.680	0.250	−0.947	0.926	553.759	642.768	0.412	−0.971	0.771
Bosnia and H.	34.380	40.090	0.202	−0.133	0.926	10.527	12.069	0.061	0.904	0.877	**6.127**	**7.207**	**0.035**	**0.987**	**0.587**
Colombia	**3.011**	**4.240**	**0.076**	**0.949**	**0.835**	3.513	5.680	0.095	0.901	0.941	4.130	6.186	0.110	0.872	0.906
Ecuador	**4.378**	8.002	**0.126**	**0.869**	**0.398**	4.445	**7.977**	0.132	0.856	0.435	5.185	9.148	0.150	0.640	0.516
Ethiopia	0.153	0.172	0.138	0.957	0.904	0.218	0.240	0.196	0.928	0.944	**0.033**	**0.045**	**0.029**	**0.998**	**0.530**
France	56.475	76.216	0.083	0.998	**0.613**	129.416	143.594	0.194	0.923	0.935	**39.874**	**50.884**	**0.060**	**0.999**	0.758
Germany	70.757	87.232	0.115	−0.664	0.798	**57.927**	**71.518**	**0.093**	**0.938**	**0.442**	105.832	125.317	0.171	−0.844	0.858
Japan	0.078	0.095	0.089	1.000	0.872	**0.017**	**0.022**	**0.019**	**1.000**	**0.353**	0.149	0.167	0.172	1.000	0.916
Kenya	**0.245**	**0.414**	**0.132**	**0.945**	**0.903**	0.437	0.643	0.246	0.808	0.956	0.518	0.686	0.311	0.811	0.966
Malaysia	14.815	17.293	0.099	0.949	0.812	6.301	7.557	0.042	**0.991**	**0.777**	**4.117**	**5.038**	**0.027**	0.983	0.839
New Zealand	2.454	3.043	0.122	0.993	0.818	6.583	8.340	0.337	−0.601	0.946	**1.925**	**2.474**	**0.096**	**0.998**	**0.683**
Paraguay	6.701	8.430	0.826	0.684	0.941	**2.165**	**2.698**	**0.267**	**0.976**	**0.696**	6.505	7.858	0.791	0.759	0.915
Poland	87.262	98.092	0.143	0.695	0.837	85.607	98.764	0.140	0.702	0.761	**22.342**	**28.268**	**0.036**	**0.971**	**0.507**
Romania	**15.639**	**17.174**	**0.258**	**0.966**	**0.586**	17.576	19.696	0.295	0.895	0.795	20.207	21.994	0.336	0.875	0.843
Sweden	**16.638**	**21.395**	**0.076**	**0.987**	**0.702**	49.060	59.707	0.221	−0.327	0.964	21.093	25.388	0.096	0.956	0.844
United Kingdom	70.063	78.434	0.100	0.700	0.938	**44.330**	**48.448**	**0.063**	**0.963**	**0.924**	94.298	108.542	0.133	−0.212	0.981
Venezuela	4.156	4.582	0.221	0.897	0.621	**3.999**	**4.432**	**0.213**	**0.900**	0.631	4.156	4.582	0.221	0.897	**0.620**
Median	10.758	12.802	**0.129**	**0.947**	0.815	6.442	8.159	0.176	0.903	0.836	**5.656**	**7.532**	0.142	0.927	**0.805**

Table 5
SGSE vs ARIMA in terms of forecast errors of 20 countries from different continents.

Country	SGSE - Euclidean distance		SGSE - Wasserstein distance		SGSE - Jensen–Shannon distance		ARIMA	
	MAE	TSS	MAE	TSS	MAE	TSS	MAE	TSS
Albania	28.756	0.692	23.275	0.441	**19.327**	**0.963**	52.839	0.864
Angola	0.127	0.907	**0.095**	**0.946**	0.151	0.899	0.373	0.612
Argentina	**1.655**	0.997	5.331	0.934	2.517	**1.000**	7.271	0.979
Belgium	**208.256**	**0.999**	320.492	−0.947	553.759	−0.971	465.389	−0.960
Bosnia and H.	34.380	−0.133	10.527	0.904	**6.127**	**0.987**	21.009	0.068
Colombia	**3.011**	**0.949**	3.513	0.901	4.130	0.872	5.950	0.885
Ecuador	**4.378**	**0.869**	4.445	0.856	5.185	0.640	5.855	0.320
Ethiopia	0.153	**0.957**	0.218	0.928	**0.033**	0.998	0.205	0.869
France	56.475	0.998	129.416	0.923	**39.874**	**0.999**	151.523	0.528
Germany	70.757	−0.664	**57.927**	**0.938**	105.832	−0.844	131.936	−0.885
Japan	0.078	1.000	**0.017**	**1.000**	0.149	1.000	0.426	0.996
Kenya	**0.245**	**0.945**	0.437	0.808	0.518	0.811	0.677	0.681
Malaysia	14.815	0.949	6.301	**0.991**	**4.117**	0.983	15.382	0.882
New Zealand	2.454	0.993	6.583	−0.601	**1.925**	0.998	3.678	**1.000**
Paraguay	6.701	0.684	**2.165**	**0.976**	6.505	0.759	8.060	0.672
Poland	87.262	0.695	85.607	0.702	**22.342**	**0.971**	33.940	−0.025
Romania	**15.639**	**0.966**	17.576	0.895	20.207	0.875	24.129	0.928
Sweden	**16.638**	**0.987**	49.060	−0.327	21.093	0.956	33.541	0.382
United Kingdom	70.063	0.700	**44.330**	**0.963**	94.298	−0.212	105.080	−0.460
Venezuela	4.156	0.897	**3.999**	**0.900**	4.156	0.897	6.570	0.074
Median	10.758	**0.947**	6.442	0.903	**5.656**	0.927	11.721	0.642

Table 6
Performance comparisons of the SGSE model with four different models from the European COVID-19 Forecast Hub.

	Belgium		Romania		France		Poland		United Kingdom		Sweden		Germany	
	MAE	TSS	MAE	TSS	MAE	TSS	MAE	TSS	MAE	TSS	MAE	TSS	MAE	TSS
epiforecasts-EpiNow2	5733.4	0.015	783.7	0.942	10446.2	−0.035	451.8	0.987	5640.8	−0.200	3048.2	0.671	1036.7	0.878
IEM_Health-CovidProject	3799.9	0.334	746.7	0.943	**8227.3**	**0.749**	1533.7	0.991	**2042.0**	0.840	**2797.2**	**0.732**	837.2	0.954
EuroCOVIDhub-ensemble	4829.9	0.128	760.0	0.942	10139.3	0.207	452.9	0.984	3976.6	0.247	3180.4	0.647	1084.4	0.901
MUNI-VAR	6165.0	0.032	**601.7**	**0.957**	14003.3	−0.312	1364.1	0.957	6169.7	0.426	4301.6	0.400	1778.1	0.848
SGSE (Wasserstein)	**3362.3**	**0.529**	694.8	0.948	10154.9	0.576	**238.7**	**1.000**	2506.7	**0.990**	3985.1	0.541	1976.2	0.962
SGSE (Jensen–Shannon)	3545.0	0.435	679.8	0.950	9980.9	0.581	660.1	0.991	5687.7	0.839	3139.6	0.672	676.2	**0.990**
SGSE (Euclidean)	4022.4	0.270	685.6	0.949	10154.9	0.576	260.8	**1.000**	5584.9	0.892	3751.9	0.546	**643.8**	0.974

while estimating the rapid uptrend faced in Sweden and Ecuador. When results from Colombia and Malaysia were considered, the MAE and RMSE errors were found to be relatively high. When the time-series plots of these countries are examined, it can bee seen that there is a difference between the actual daily new cases and predictions during the 14 days. However, when we look at the forecast trend for these

two countries, it overlaps exactly with the actual trend. Although the ARIMA model makes accurate predictions in most countries in the first one or two days, its deviation increases gradually afterward. Bosnia and Herzegovina and New Zealand were in a downtrend during the forecasted period. While the ARIMA model correctly predicted the decline in New Zealand, it predicted a plateau in Bosnia and Herzegovina. The SGSE, on the other hand, accurately forecasted that the number of COVID-19 cases would decrease in both countries. While the ARIMA model in Japan predicted a trend in the opposite direction, the SGSE predicted the trends accurately. ARIMA predicted a plateau for France, while the SGSE predicted a successful uptrend.

The SGSE outperformed four different models submitted to the European COVID-19 Forecast Hub. It has the highest TSS and lowest MAE error values for four out of seven European countries. It also showed the 2nd and 3rd most accurate performances in countries for which it did not have the best forecast. In Germany, the number of cases first decreased and then increased during the forecasted period. All compared models forecasted downtrends, while the SGSE forecasted uptrends. While Belgium had an increasing trend during this period, all models predicted a slight decrease. However, the closest prediction to the actual trend came from the SGSE. For Poland and the UK, the SGSE's estimates are similar to the actual trend.

5. Conclusions

This study proposes a novel method that predicts the smoothed daily new cases per million of COVID-19. GMM representations derived from time-series data of the countries were used. A dataset was created with the representations extracted from the cutoff data of the countries according to different cutoff dates. Using different distance metrics with country representations helps in finding the most similar examples for each country. These examples constitute the histories of other countries. As a result, the SGSE model determines which countries' past COVID-19 data most closely resembles a given country's data. The future of the country observed is predicted from the average of similar samples. The model was tested on 20 different countries, and the results are provided with three different similarity metrics. In addition, the forecasts of eight different countries are visualized. It can be observed that countries that start to apply more stringent precautionary measures may experience a downward trend faster than similar samples. The results show that the SGSE model successfully predicts uptrends, downtrends, and plateaus based on trend similarity scores compared to the baseline method. However, the SGSE does not work as well when there is a rapid uptrend in a country, as seen in Sweden's case. To make a fair comparison between COVID-19 forecast models, one must initiate forecasts on the same countries during the same periods. For this reason, the SGSE model is compared with four models from The European COVID-19 Forecast Hub. It outperformed these four models on the data prediction of seven European countries. The SGSE gave better predictions for Belgium, Poland, the UK, and Germany. Furthermore, it had the second and third best prediction scores for Romania, France, and Sweden.

It is also important to note that the SGSE model is a generic approach that can be applied not only for COVID-19 but also to any dataset containing time-series data produced by different classes or clusters with the same context.

CRediT authorship contribution statement

Emre Külah: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing. **Yusuf Mücahit Çetinkaya:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing. **Arif Görkem Özer:** Data curation, Formal analysis, Methodology, Software, Visualization, Writing. **Hande Alemdar:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is shared publicly by OWID (https://ourworldindata.org/).

References

Abbott, S., Hellewell, J., Sherratt, K., Gostic, K., Hickson, J., Badr, H. S., et al. (2020). EpiNow2: Estimate real-time case counts and time-varying epidemiological parameters. http://dx.doi.org/10.5281/zenodo.3957489.

Akima, H. (1974). A method of bivariate interpolation and smooth surface fitting based on local procedures. *Communications of the ACM*, *17*(1), 18–20.

Ayoobi, N., Sharifrazi, D., Alizadehsani, R., Shoeibi, A., Gorriz, J. M., Moosaei, H., et al. (2021). Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. *Results in Physics*, *27*, Article 104495.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*, Article 105340.

Dehesh, T., Mardani-Fard, H., & Dehesh, P. (2020). *Forecasting of covid-19 confirmed cases in different countries with arima models*. Cold Spring Harbor Laboratory Press, MedRxiv.

Eirola, E., & Lendasse, A. (2013). Gaussian mixture models for time series modelling, forecasting, and interpolation. In *International symposium on intelligent data analysis* (pp. 162–173). Springer.

Farzanegan, M. R., Gholipour, H. F., Feizi, M., Nunkoo, R., & Andargoli, A. E. (2021). International tourism and outbreak of coronavirus (COVID-19): A cross-country analysis. *Journal of Travel Research*, *60*(3), 687–692.

Fisayo, T., & Tsukagoshi, S. (2021). Three waves of the COVID-19 pandemic. *Postgraduate Medical Journal*, *97*(1147), 332.

Gautam, Y. (2022). Transfer learning for COVID-19 cases and deaths forecast using LSTM network. *ISA Transactions*, *124*, 41–56.

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (2020). Artificial intelligence forecasting of covid-19 in china. arXiv preprint arXiv:2002.07112.

Imtyaz, A., Haleem, A., & Javaid, M. (2020). Analysing governmental response to the COVID-19 pandemic. *Journal of Oral Biology and Craniofacial Research*, *10*(4), 504–513.

Ketu, S., & Mishra, P. K. (2022). India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability. *Soft Computing*, *26*(2), 645–664.

Kumar, M., Patel, N. R., & Woo, J. (2002). Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 557–563).

Lewis, C. D. (1982). *Industrial and business forecasting methods: a practical guide to exponential smoothing and curve fitting*. Butterworth-Heinemann.

Liao, Z., Lan, P., Fan, X., Kelly, B., Innes, A., & Liao, Z. (2021). SIRVD-DL: A COVID-19 deep learning prediction model based on time-dependent SIRVD. *Computers in Biology and Medicine*, *138*, Article 104868.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Mousavi, S. H., Zahid, S. U., Wardak, K., Azimi, K. A., Hosseini, S. M. R., Wafaee, M., et al. (2020). Mapping the changes on incidence, case fatality rates and recovery proportion of COVID-19 in afghanistan using geographical information systems. *Archives of Medical Research*, *51*(6), 600.

Naeem, M., Mashwani, W. K., ABIAD, M., Shah, H., Khan, Z., & Aamir, M. (2022). Soft computing techniques for forecasting of COVID-19 in Pakistan. *Alexandria Engineering Journal*.

Pavlik, T., Komenda, M., Pribylova, L., Uher, M., Majek, O., Kraus, A., et al. (2020). *MAMES–monitoring, analyza a management epidemickych situaci: Popis algoritmizace a implementace epidemickych model a mapovani kapacit*. Masarykova univerzita.

Pham, Q.-V., Nguyen, D. C., Huynh-The, T., Hwang, W.-J., & Pathirana, P. N. (2021). Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: A survey on the state-of-the-arts. arXiv preprint arXiv:2107.14040.

Povinelli, R. J., Johnson, M. T., Lindgren, A. C., & Ye, J. (2004). Time series classification using Gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering*, *16*(6), 779–783.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of Biometrics*, *741*, 659–663.

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus pandemic (COVID-19). *Our World in Data*.

Saba, A. I., & Elsheikh, A. H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. *Process Safety and Environmental Protection*, *141*, 1–8.

Sahoo, B. K., & Sapra, B. K. (2020). A data driven epidemic model to analyse the lockdown effect and predict the course of COVID-19 progress in India. *Chaos, Solitons & Fractals*, *139*, Article 110034.

Satrio, C. B. A., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, *179*, 524–532.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Seong, H., Hyun, H. J., Yun, J. G., Noh, J. Y., Cheong, H. J., Kim, W. J., et al. (2021). Comparison of the second and third waves of the COVID-19 pandemic in South Korea: Importance of early public health intervention. *International Journal of Infectious Diseases*, *104*, 742–745.

Sherratt, K., Gruson, H., Johnson, H., Niehus, R., Prasse, B., Sandman, F., et al. (2022). *European Covid-19 forecast hub*. Zenodo.

Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanien, A. E. (2020). Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art. *SN Computer Science*, *1*(4), 1–15.

Singhal, A., Singh, P., Lall, B., & Joshi, S. D. (2020). Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. *Chaos, Solitons & Fractals*, *138*, Article 110023.

Suchoski, B., Stage, S., Gurung, H., & Baccam, P. (2022). GPU accelerated parallel processing for large-scale Monte Carlo analysis: COVID-19 parameter estimation and new case forecasting. *Frontiers in Applied Mathematics and Statistics*, 9.

Sun, Z., Zhang, H., Yang, Y., Wan, H., & Wang, Y. (2020). Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China. *Science of the Total Environment*, *746*, Article 141347.

Tandon, H., Ranjan, P., Chakraborty, T., & Suhag, V. (2020). Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. arXiv preprint arXiv:2004.07859.

Vaserstein, L. N. (1969). Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, *5*(3), 64–72.

Xu, L., Magar, R., & Barati Farimani, A. (2022). Forecasting COVID-19 new cases using deep learning methods. *Computers in Biology and Medicine*, *144*, Article 105342.

Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., et al. (2020). Predicting COVID-19 in China using hybrid AI model. *IEEE Transactions on Cybernetics*, *50*(7), 2891–2904.

Zhou, X.-Y., Lim, J. S., Kwon, I., et al. (2014). EM algorithm with GMM and naive Bayesian to implement missing values. *Advanced Science and Technology Letters*, *46*, 1–5.