Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering

Fernando Rojas[a], Olga Valenzuela[b], Ignacio Rojas[a]

a) Dpt. Computer Architecture and Computer Technology, CITIC-UGR, University of Granada, Spain (frojas@ugr.es,irojas@ugr.es)

b) Dpt. Applied Mathematics, University of Granada, Spain (olgavc@ugr.es)

Abstract

Estimation of COVID-19 dynamics and its evolution is a multidisciplinary effort, which requires the unification of heterogeneous disciplines (scientific, mathematics, epidemiological, biological/bio-chemical, virologists and health disciplines to mention the most relevant) to work together in a better understanding of this pandemic. Time series analysis is of great importance to determine both the similarity in the behavior of COVID-19 in certain countries/states and the establishment of models that can analyze and predict the transmission process of this infectious disease. In this contribution, an analysis of the different states of the United States will be carried out to measure the similarity of COVID-19 time series, using dynamic time warping distance (DTW) as a distance metric. A parametric methodology is proposed to jointly analyze infected and deceased persons. This metric allows to compare time series that have a different time length, making it very appropriate for studying the United States, since the virus did not spread simultaneously in all the states/provinces. After a measure of the similarity between the time series of the states of United States was determined, a hierarchical cluster was created, which makes it possible to analyze the behavioral relationships of the pandemic between different states and to discover interesting patterns and correlations in the underlying data of COVID-19 in the United States. With the proposed methodology, nine different clusters were obtained, showing a different behavior in the eastern zone and western zone of the United States. Finally, to make a prediction of the evolution of COVID-19 in the states, Logistic, Gompertz and SIR model was computed. With these mathematical model it is possible to have a more precise knowledge of the evolution and forecast of the pandemic.

Keywords: COVID-19; Pandemic in the United States ; Time Series; DTW distance; Hierarchical Clustering; SIR model

1. Introduction

The COVID-19 epidemic started in Hubei Province, China, around December 2019. Since then, the disease has been spread to all continents and countries of the world, being categorized as pandemic by World Health Organization on March 11th.

In recent months, contributions have been made that analyze the evolution of different countries, implementing mathematical models to predict their evolution. Traditional predictive models for infectious diseases mainly include models for predicting differential equations and models for predicting time series based on statistics and random processes

For example, in [1] a methodology with the aim of estimating the actual number of people infected with COVID-19 in France is presented, since according to the authors, the number of screening tests carried out and the methodology do not directly calculate the actual number of cases and infection mortality rate (IFR). A mechanistic-statistical approach was developed that combines an epidemiological SIR model that describes this unobserved epidemiological dynamics, a probabilistic model that describes the data collection process and a method of statistical inference.

The logistic growth model, the generalized logistic growth model, the generalized growth model and the generalized Richards model were used to model the number of infected cases in the 29 provinces of China (and several countries), performing a a detailed analysis on the heterogeneous situations by four phases of the outbreak in China [2].

In [3] the Kermack-McKendrick SEIR model (Susceptible, Exposed, Infectious and Recovered) is presented to analyze the effects of behavioral changes on the reduction in community transmission in Mexico. A variable contact rate over time is proposed and the consequences of disease spread in an affected population of non-essential activities is analyzed.

The behavior of the virus in Japan has also been analyzed [4]. By February 29, 2020, in addition to the 619 confirmed cases (passengers and crew members) infected with COVID-19 in a cruise ship (near Tokyo), 215 locally transmitted cases had been also confirmed in Japan. To evaluate the effectiveness of reaction strategies based on avoiding large accumulations or crowded areas and to predict the spread of COVID-19 infections in Japan, in [4] a stochastic transmission model by expanding the epidemiological model based on SIR (Susceptible-Infected-Removed) had been presented. The simulation results showed that the number of Infected and Removed patients will increase rapidly if there is no reduction of the time spent in crowded zone.

In [5] using the Maximum-Hasting (MH) parameter estimation method and the SEIR model, the spread of COVID-19 and its prediction in South Africa, Egypt, Nigeria, Senegal, Kenya, and Algeria under three intervention scenarios (suppression, mitigation, mildness) is presented.

In addition to the most relevant epidemiological models used in the literature, models typically based on time series have also been used to analyze the behavior of the pandemic in different countries. The autoregressive integrated moving average (ARIMA) model is a mathematical model widely studied in the context of time series that is successfully applied in the field of health (estimate the incidence and prevalence of influenza mortality, malaria incidence, hepatitis, and other infectious diseases) as well as in different fields in the past due to its simple structure, fast applicability and ability to explain the data set.

In [6] ten Brazilian states are analyzed using the autoregressive integrated moving average (ARIMA), the cubistic regression (CUBIST), the random forest (RF), ridge regression (RIDGE), the support vector regression (SVR) and the stacking-ensemble learning in the task of time series forecasting of the number of patient infected with COVID-19 with one, three, and six-days ahead. A forecasting model based on ARIMA has also been presented in [7] for Pakistan, presenting the high exponential growth in the number of confirmed cases, deaths and recoveries. In [8] ARIMA time series models were applied to forecast the total confirmed cases of COVID-19 for the next ten days using the model ARIMA (0,2,1), ARIMA (1,2,0) and ARIMA (0,2,1) for Italy, Spain, and France, respectively.

Currently, the analysis of the evolution of COVID-19 in America is of great importance due to the impact of this epidemic on this continent. In this contribution we will focus on the United States. The first patient detected in the United States was a travel-associated case from Washington state on January 19th, 2020. The preponderance of initial cased of infected patients with COVID-19 in the United States were correlated with travel to a "high-risk" country or close contacts of previously identified cases corresponding to the testing criteria adopted by the Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/). From March 1–31, 2020, the number of reported COVID-19 cases in the United States rapidly increased from 30 to 188,172, being the number of deaths from 1 to 5531, and detecting the virus all the states. At the end of April the number of infected reached 1069424 and the number of deceased stood at 62996. At the time of writing this contribution (14th june 2020) the number of infected is more than 2e+6 and more than 1e+5 deaths, being one of the countries of the world that is suffering with greater severity the disease of COVID-19.

In a recent paper [9] an attempt is made to estimate the actual number of infected people, even if they have not been counted. It was estimated that the true number of COVID-19 cases in the United States is likely in the tens of thousands, suggesting substantial undetected infections and spread within the country.

In [10] a relevant contribution is presented analyzing the sequence of nine viral genomes from early reported COVID-19 patients in Connecticut. From the phylogenetic analysis, it can be concluded that the majority of these genomes with sequenced viruses are from Washington State. By coupling their genomic data with domestic and international travel patterns, authors showed that early SARS-CoV-2

transmission in Connecticut was likely driven by domestic travel. The authors hypothesized that, with the growing number of COVID-19 cases in the United States and the large volume of domestic travel, new United States outbreaks are now more likely to result from interstate rather than international spread.

This contribution presents a methodology to analyze the evolution patterns of COVID-19 in the states of United States (including Puerto Rico and District of Columbia). A parametric similarity measure is presented, based on robust distance measure between time series, the dynamic time warping distance (DTW), with which the number of infected and dead in each of the states can be compared simultaneously, even though the start of the epidemic originated on different dates in each zone (therefore, the time series that need to be compared have different lengths).

To the best of our knowledge, this contribution is the first study that tries to develop a hierarchical clustering time series algorithm in order to globally compare and classify the behavior of all the states of United State simultaneously in their evolution of infected and deceased patients suffering COVID-19. Carrying out this classification is very useful, since it will allow to establish similarities and patterns in the evolution of the pandemic among the states of the United States. Once the different states have been grouped into a cluster (nine cluster are obtained), a SIR model, for each of the most representative elements of each cluster is analyzed. The simulation results of the nine SIR models evaluated are presented, indicating the most relevant parameters of the mathematical model and its prediction on the evolution of the pandemic in that state.

2. Material and Methods

A time series is a sequence of numerical (temporal) data points in successive order, which is naturally high dimensional and large in data size. There are two main operations that could be performed when working with time-series with its sequential data: a) the analysis of a single time series; b) the analysis of multiple time series simultaneously. This contribution is concentrated in the analysis of multiple time series for all the states of US suffering COVID-19, with the purpose of finding similarities between multiple time series by performing a clustering time-series methodology.

Clustering such complex objects is particularly advantageous because it may lead to the discovery of interesting patterns in time-series datasets, which contributes to a better understanding of the COVID-19 spread in different regions of the United States.

Clustering of time-series sequences has received noteworthy attention [11,12], not only as a formidable exploratory method and powerful tool for discovering patters, but also as a pre-processing step or subroutine for other tasks [13].

In this section, the database used is presented first (Section 2.1). Subsequently, a review of the most popular distance measures for time series is described (Section 2.2) and a new parametric distance is proposed. Then, existing approaches for clustering time-series data are briefly presented (Section 2.3) and the Logistic, Gompertz and SIR model, which will be used for modelling the time series of the most representative states in each of the cluster obtained, is described (Section 2.4).

2.1. Data set

The COVID-19 epidemic data set used in this contribution was collected from the Johns Hopkins University [14]. In this platform, the number of confirmed, deaths and recovered cases until June 21th 2020 for different countries are presented. For the United States, two additional .csv files are provided, in with detail of administration and province/state is reported (including Puerto Rico and District of Columbia). In order to compare countries behaviour, the time-series data are divided by state population.

2.2. Similarity/Distance measure in Time Series

In a simplified way, the similarity of two simple time series having the same number of points (denoted by m), and defined by X = $\{x_1,x_2,\ldots x_m\}$ and Y = $\{y_1,y_2,\ldots y_m\}$, can be achieved by simply calculating the Minkowski (or Euclidean) distance (shortest path between two points) between points on both time series that happen at the same time. This distance is the measure of similarity, denoted as d(X,Y), and it is a function that takes both times series (X,Y) as input and calculates their distance "d", defined as:

$$d(X,Y) = \left(\sum_{i=1}^{m} |x_i - y_i|^k\right)^{1/k} \quad (1)$$

When k=2, the distance between two series is called Euclidean Distance. Using the Minkowski distance is a good metric to analyze the similarity of two time series, if these time series are synchronized (that is, all similar events in both time series occur at exactly the same time) and have the same length.

The evolution of time series in the different states of the United States present a different start date, both for the number of confirmed and death cases, and therefore its length is also different. Suppose by analogy the time series of the sound of a mother's voice when she speaks slowly to her child. If the mother says the same phrase quickly, the child will most likely recognize that she is still his mother. However, if the Euclidean distance between both series were used as a metric, these two time series would have a very low similarity and would not be considered fundamentally equal. This would lead to the conclusion that the two voices did not come from the same person. To solve this problem, the dynamic time warping distance (DTW) method is frequently used in the bibliography [15].

DTW is a technique that can be considered as an extension of the Euclidean Distance between series [16], that calculates an optimal match between two given time series with certain restriction, performing non-linearly in the series (by stretching or shrinking along its time-axis). This distortion (denoted as warping) between two time series is used to find corresponding regions and determine the similarity between them.

The DTW of two series X and Y, defined as X = $\{x_1,x_2,\ldots x_n\}$ and Y = $\{y_1,y_2,\ldots y_m\}$ is computed in the following way. An *n*-by-*m* matrix D is computed with the (i.j)*th* element, defining the local distance of two elements by:

$$\mathrm{d}(x_i, y_j) = (x_i - y_j)^2 \quad (2)$$

The point-to-point alignment between series X and Y can be represented by a time warping path W, defined as:

$$\mathrm{W} = \begin{pmatrix} w_x(k) \\ w_y(k) \end{pmatrix}, k = 1,2,\ldots p \quad (3)$$

where *p* is the length of the warping path W, and w_x (k) and w_y(k) represent the indexes in time series X and Y respectively. The warping path $\begin{pmatrix} w_x(k) \\ w_y(k) \end{pmatrix}$ indicate that the w_x (k) *th* element in time series X is mapping to the w_y(k) *th* element in time series Y. There are some constraints and rules for the construction of the warping path:

- Every index from the first time series must be matched with one or more indices from the other time series (and vice versa)
- The first (the same for the last index) index from the first time series must be matched (not only this match) with the first (last) index from the other time series. That is, the warping path should start at W(1) = (1, 1) and end up at W(p) = (n,m).
- The mapping of the indices from the first time series to indices from the other time serie must be monotonically increasing, and vice versa. The adjacent elements of path W, W(k) and W(k + 1) must be subject to $w_x(k + 1) - w_x(k) \geq 0$ and $w_y(k + 1) - w_y(k) \geq 0$.

- The warping path should be also have the property of continuity, mathematically expressed as adjacent elements of path W, W(k) and W(k + 1) must be subject to $w_x(k + 1) - w_x(k) \leq 1$ and $w_y(k + 1) - w_y(k) \leq 1$.

The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values. The DTW (minimal distance and optimal warping path) could be found using a dynamic programming algorithm:

$$\mathrm{RD}(x_i, y_j) = \mathrm{d}(x_i, y_j) + \min \begin{cases} \mathrm{RD}(x_{i-1}, y_{j-1}) \\ \mathrm{RD}(x_{i-1}, y_j) \\ \mathrm{RD}(x_i, y_{j-1}) \end{cases} \tag{4}$$

$$DTW(X, Y) = \min\{RD(x_n, y_m)\}$$

where $\mathrm{RD}(x_i, y_j)$ is the minimal cumulative distance from (0, 0) to (i, j) in matrix D. In the methodology proposed in this paper, for each of the states analysed, both the time series of the number of infected and the time series of deaths will be simultaneously taken into account.

If each of these time series needs to be weighted differently, the following parametric metric, $DTW_\propto(S_A, S_B)$ is defined:

$$DTW_\propto(S_A, S_B) = \propto DTW(TSD_A, \mathrm{TS}D_B) + (1-\propto)DTW(TSC_A, TSC_B) \tag{5}$$

that measures the similarity in the evolution of the COVID time series for two states of the United States (S_A y S_B), TSC_A and TSC_B represent the time series of the number of infected, TSD_A and TSD_B represent the time series of the number of deaths for the states S_A and S_B respectively. The parameter α (with $0 \leq \alpha \leq 1$) indicates the relative relevance given to the similarity measure, taking into account the time series of infected or deaths.

2.3 Clustering method for time series

Clustering is a data mining technique in which similar data are divided into related or homogeneous groups, in an unsupervised way, that is, without knowing a priori advanced knowledge of the data. For the problem presented in this contribution, working with time series of states of the United States suffering COVID-19, given a set of individual time series data, the objective is to group similar time series into the same cluster.

The problem of grouping time series data is formally defined as, given a dataset of *N* time series data $Q=\{X_1, X_2 \ldots X_N\}$, find in an unsupervised way, a partition of *Q* into *K* cluster, denoted as $C=\{C_1, C_2, \ldots C_k\}$, taking into account that homogeneous or similar series are grouped together based on a certain similarity/distance measure. In this paper, the parametric metric $DTW_\propto(S_A, S_B)$ is used and there is not intersection between clusters, therefore:

$$Q = \bigcup_{i=1}^{K} C_i \,; \text{with } C_i \cap C_j = \emptyset \ (i \neq j) \tag{6}$$

The methods used in the area of time series clustering [11, 17] are usually based in conventional clustering algorithm by substituting standard distance measurements with a more suitable distance to compare time series (raw methods) or converting series into normal data and using directly classical algorithms (Feature-based methods and models).

Among the most popular clustering algorithms, the hierarchical clustering and the k-means algorithm are widely used in time series clustering. In this contribution the hierarchical clustering is used, mainly due to its great visualization power and its simple and intuitive interpretation.

Hierarchical clustering creates a nested hierarchy of similar time series, according to a pair-wise distance matrix of the time series analyzed. The similarity measure $DTW_\propto(S_A, S_B)$ used is therefore essential in this time-series clustering process.

One of the most relevant characteristics of hierarchical clustering is its generality, since the user does not need to provide any parameters such as the number of clusters. As a disadvantage, hierarchical clustering has a high computational complexity when the number of elements to classify increases (the performance of hierarchical clustering is directly proportional to the squared size of the input data set). The methodology used to build the hierarchical clustering is the following:

1.- Calculate the distance between all the states of the United States using $DTW_{\propto}(S_A, S_B)$, for a certain value of the parameter α. This distance matrix, symmetrical and with the null diagonal, will be essential to analyze the similarity between the behavior of the different states.

2.- Search through the distance matrix in order to select the two most similar elements (in our case the time series of two states).

3.- Join (linkage) these two states to produce a new group that now have at least two objects (states).

4.- Update the distance matrix by calculating the distances between the new cluster and all other clusters.

5.- Repeat step 2 until all cases belong to a group.

The most widely used linkage criteria, such as single, average and complete linkage variants [18], were analyzed. Hierarchical clustering can be converted into a partitional clustering, with *k* cluster, by cutting the first *k* links.

2.3. *Time series modeling*

In this subsection, three models currently used in the bibliography for adjusting the evolution of COVID-19 from data will be briefly described.

2.3.1 Logistic model

Mathematical models are formidable tools for understanding and predicting infectious diseases behaviour, being used in numerous viral diseases. A simple and easy-to-understand mathematical model is logistic regression analysis, used in modelling COVID-19 [19,20], expressed mathematically as:

$$Q^L(\mathrm{t}) = \frac{a}{1 + \frac{a}{b-1} e^{-c(t-t_0)}} \tag{7}$$

where $Q^L(t)$ is the cumulative cases of the logistic model at time *t* (can be the confirmed or dead patients), the parameters *a*, *b* and *c* are fitting coefficients of the model (numerical values of these three parameters depend on available data). Logistic models tend to under-estimate the total size of the infected/death population at the early stage, so they should provide lower bounds.

2.3.2 Gompertz model

This model has been frequently used to describe the growth of animals and plants, as well as the number or volume of bacteria, virus and cancer cells [21]. In [22] this model was used to forecast the impact lethal duration of exposure on the mortality rates of COVID in seven countries (Germany, China, France, United Kingdom, Iran, Italy and Spain). It is based on sigmoid models fitted, that starts with an exponential growth and gradually decreases its specific growth rate, being a special case of the four parameter Richards model, and thus belongs to the Richards family of three-parameter sigmoidal growth models, using the following equation:

$$Q^G(\mathrm{t}) = \mathrm{k} e^{-\log(\frac{k}{n})(-q(t-t_0)} \tag{8}$$

where $Q^G(t)$ is the cumulative cases of the Gompertz model at time *t* (can be the confirmed or deaths patients), n, k and q are fitting coefficients of the model, being q the parameter that modulates how the

spreading rate is slowing down. These these three parameters depend on available data and must be adjusted using the available data from the COVID-19 time series, for each of the states of United States selected in the clustering process.

2.3.3 SIR Model

Modelling the spread of infectious diseases usually performed by the categorization of the individuals in the population as belonging to one of several distinct compartments, which represent their health status with respect to the infection. In this paper the SIR model will be used, having three compartments: *S(t)* is the number of susceptible cases at time *t*, *I(t)* the number of infected cases and the function *R(t)* is the number of recovered persons in time *t* [23].

In order to understand and forecast the evolution of COVID-19 in the different states of the United States, the epidemic can then be analysed as the rates of transfer between these compartments [1], mathematically defined by the following non-linear systems of ordinary differential equations:

$$\begin{aligned} \frac{dS(t)}{dt} &= -\frac{\beta}{N}S(t)I(t) \\ \frac{dI(t)}{dt} &= \frac{\beta}{N}S(t)I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t) \end{aligned} \tag{9}$$

The parameter β and γ are the contact rate and average removal frequency respectively.

In order to solve the non-linear systems of ordinary differential equations presented in equation (9), initial conditions should be defined, being:

$$S(t_0) = (N - \hat{I}_{t0}) \geq 0; I(t_0) = \hat{I}_{t0} \geq 0; R(t_0) = R_0 \geq 0; \tag{10}$$

It follows from equation (9) that:

$$S(t) = S_0 exp\left[-\frac{\beta}{N\gamma}(R(t) - R_0)\right] \tag{11}$$

At the limit time $t \to \infty$, assuming that the number of infected people is practically null, the number of susceptible people $S(t_\infty)$ and recovered persons can be obtained as:

$$\begin{aligned} S(t_\infty) &= S_0 exp\left[-\frac{\beta}{N\gamma}(R(t_\infty) - R_0)\right] \\ R(t_\infty) &= N - S_0 exp\left[-\frac{\beta}{N\gamma}(R(t_\infty) - R_0)\right] \end{aligned} \tag{12}$$

To model the behaviour of the epidemic using the equations presented in (7), the estimation of the parameters β and γ and the values of the initial conditions should be obtained from available data.

3. Results and discussion

To evaluate the performance of the proposed method, several experiments are conducted in this section for three values of the parameter α in the distance metric $\boldsymbol{DTW}_\propto(\boldsymbol{S_A}, \boldsymbol{S_B})$. The time series of the states of the United States has been taken from John Hopkins database. For the computation of the distance metric, a threshold I_{min} has been defined, defining the minimum number of infected people to start the time series, being for this study I_{min}=5 (the number of confirmed was greater than 5). Therefore, the length of the time series is different for each state, being on average 114 days (Figure 1). The index of each of the states is presented in Table 1.

The values of the parameter α analyzed will be {0,0.5,1}. This section of results begins with the value of α = 0.5, that is, the information of the confirmed patients time series having the same relevance as the

time series of deaths for the final computation of the distance $DTW_{\propto}(S_A, S_B)$. The distance matrix between the different states is presented in Figure 2 (for a better visual representation, the distance matrix has been multiplied by a constant and the states are ordered according to the cluster to which each one belongs).

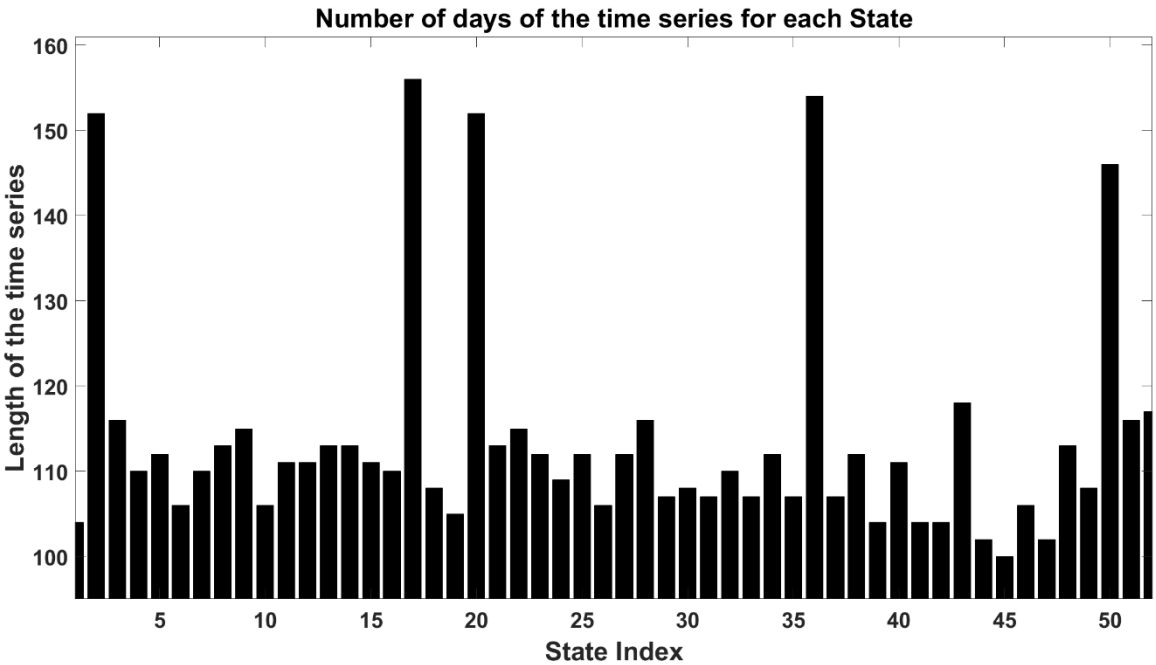


Figure 1 Length of the different time series, corresponding to the states analyzed.

The smaller the $DTW_{\propto}(S_A, S_B)$ distance, the darker its representation. States that have a large distance (therefore have a low similarity in the behavior of their time series), are represented by yellow color (colorbar on the right side of the figure). To carry out the hierarchical cluster tree, average linkage is used. The average distance between all pair of states in any two clusters is defined as:

$$D_{Average}(r,s) = \frac{1}{n_r\, n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} DTW_{\propto}(S_{ri}, S_{sj}) \qquad (13)$$

Where r and s represent clusters, and n_r and n_s is the number of states in cluster r and *s* respectively, being S_{ri} the *i*th state in cluster r and S_{sj} the *j*th state in cluster s. The hierarchical cluster tree obtained is presented in Figure 3.

To analyse the accuracy of the obtained hierarchical cluster, the Cophenetic Correlation Coefficient (CP) is used [24].

The cophenetic correlation coefficient has been widely used in clustering problem, both as a measure of fitting degree of a classification to a set of data and as a criterion for evaluating the efficiency of various clustering techniques. For the problem presented in this contribution CP=0.90 using α=0.5 .

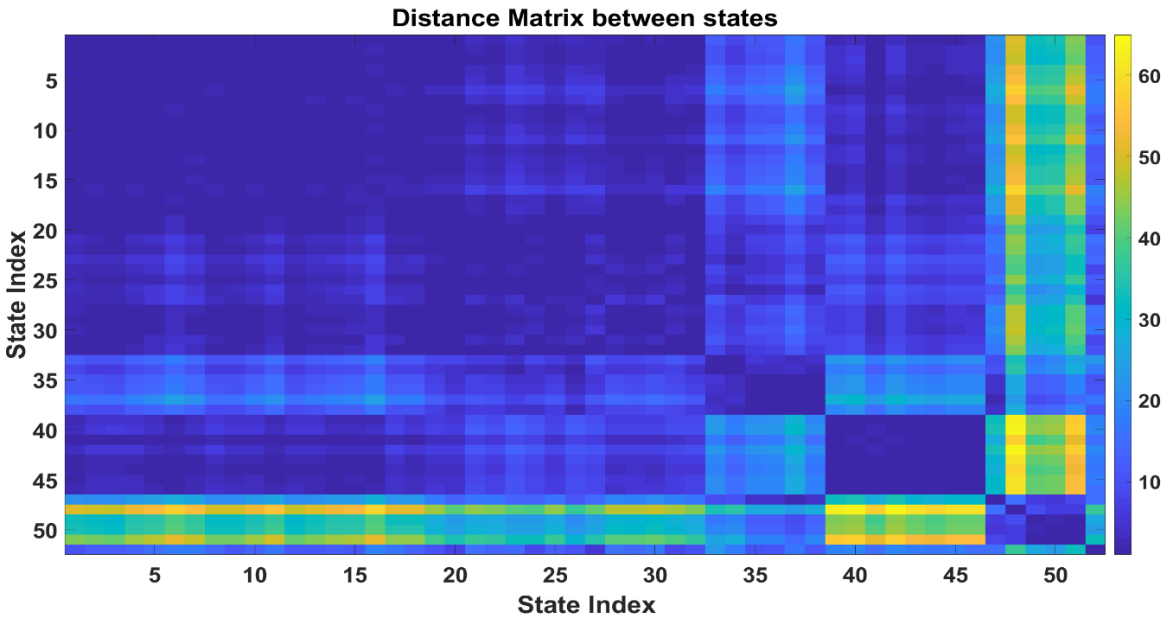


Figure 2 Distance or similarity symmetric matrix to characterize the behavior of the time series for the states of the United States (parameter α=0.5). The greater the similarity, the smaller the distance between the series (being the diagonal of this matrix of zero value).

The Calinski-Harabasz criterion (also denoted as the variance ratio criterion) is used to determine the optimal number, defined as:

$$\mathrm{CHR_K} = \frac{(\mathrm{N}-\mathrm{K})\sum_{\mathrm{i}=1}^{\mathrm{K}} \mathrm{n_i}\|\mathrm{m_i}-\mathrm{m}\|^2}{(\mathrm{K}-1)\sum_{\mathrm{i}=1}^{\mathrm{K}}\sum_{\mathrm{x}\in\mathrm{c_i}}\|\mathrm{x}-\mathrm{m_i}\|^2} \quad (14)$$

where the numerator quantifies the overall between-cluster variance, multiplied by (N-K), where N is the number of observations and K is the number of cluster. The denominator quantifies the overall within-cluster variance. The variable m_i is the centroid of cluster *i*, being *m* the overall mean of the sample data, *x* is a data point, c_i is the *i*th cluster and $\|\mathrm{m_i}-\mathrm{m}\|^2$is the Euclidean distance between two vectors. The larger the $\mathrm{CHR_K}$, the better the data partition (the clustering performed), therefore, the optimal number of clusters is obtained maximizing the Calinski-Harabasz criterion with respect to K. In the problem presented in this paper, the optimal number of clusters was 9 (using the parameter α=0.5) with the distribution shown in Table 1.

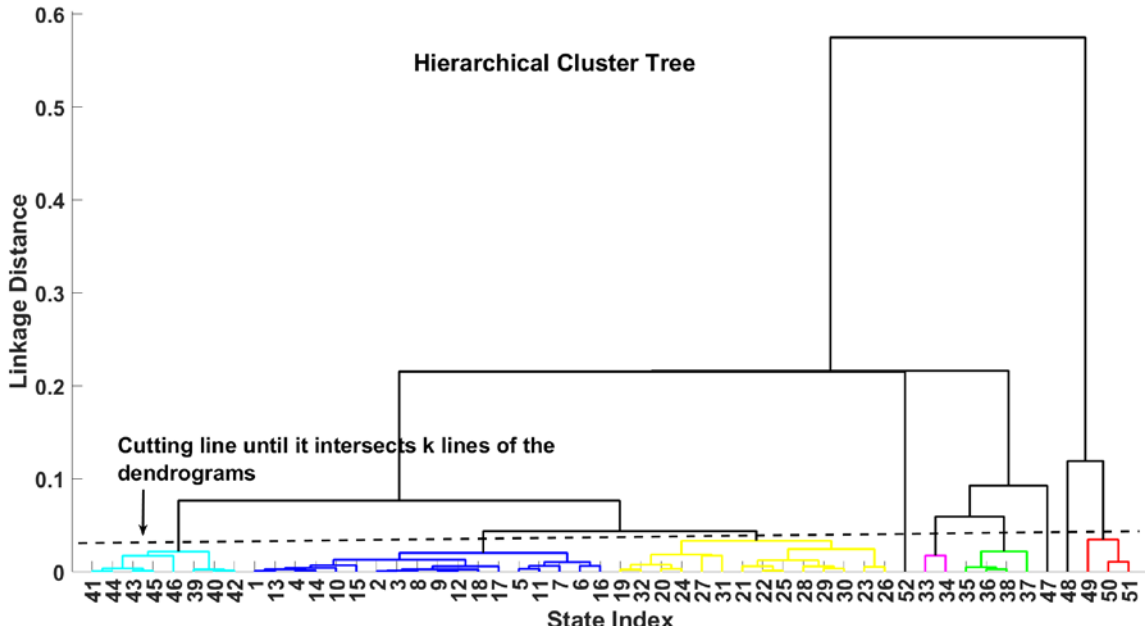


Figure 3 Hierarchical cluster tree obtained using as distance metric the $DTW_{\alpha}(S_A, S_B)$ and α=0.5

Table 1. Distribution of the states obtained by means of hierarchical clustering with 9 clusters (α=0.5 and in bold, the state for which the SIR model is calculated)

Cluster Number (C_N)	**$D_{Cluster}$**	**States (α=0.5)**
1	0.020	(1) Arkansas, **(2) California**, (3) Florida, (4) Kansas, (5) Kentucky, (6) Maine, (7) Missouri, (8) Nevada, (9) North Carolina, (10) North Dakota, (11) Oklahoma, (12) South Carolina, (13) Tennessee, (14) Texas, (15) Utah, (16) Vermont, (17) Washington, (18) Wisconsin
2	0.033	(19) Alabama, (20) Arizona, (21) Colorado, (22) Georgia, (23) Indiana, (24) Iowa, (25) Minnesota, (26) Mississippi, **(27) Nebraska**, (28) New Hampshire, (29) New Mexico, (30) Ohio, (31) South Dakota, (32) Virginia
3	0.017	(33) Michigan, **(34) Pennsylvania**
4	0.022	(35) Delaware, **(36) Illinois,** (37) Louisiana, (38) Maryland
5	0.022	(39) Alaska, (40) Hawaii, (41) Idaho, (42) Montana, **(43) Oregon**, (44) Puerto Rico, (45) West Virginia, (46) Wyoming
6	0	**(47) District of Columbia**
7	0	**(48) New Jersey**
8	0.035	(49) Connecticut, (50) Massachusetts, **(51) New York**
9	0	**(52) Rhode Island**

where $D_{Cluster}$ is the distance between the elements that make up a cluster (its value is zero in the case that there is only one element in a cluster).

It is important to highlight the existence of various clusters with only one state (corresponding with District of Columbia, New Jersey and Rhode Island). Cluster 7 (New Jersey, listed as 48 in Table 1) links directly to cluster 8 ((49) Connecticut, (50) Massachusetts, (51) New York), which denote similar behaviour between these states. For cluster 6, (47)District of Columbia, the linkage is done for both cluster 3 ((33) Michigan, (35) Pennsylvania) and cluster 4 ((35) Delaware,(36) Illinois,(37) Louisiana,(38) Maryland). There are two large clusters (cluster 1 and cluster 2) that contain 18 and 14 states respectively, performing a direct linkage (meaning that these states have analogous performance). Its linkage is done through cluster 5, which contains eight states.

The similarities and distances between the different states and clusters obtained can be analysed

using the results presented in the hierarchical clustering (Figure 3) and distance matrix (Figure 2).

Once the hierarchical clustering of the different states of the United States has been established and performed, it is relevant to model the time series (both infected and dead patients), using the models proposed in Section 2.3 (Logistic, Gompertz and SIR models). Tables 2 and 3 present the parameters of the Logistic and Gompertz models for all the states of the United States, for the time series of confirmed patients and deceased patients of COVID-19, respectively. The variable $Adj_{rsquare}$ is degree-of-freedom adjusted coefficient of determination and the RMSE represents the Root Mean Squared Error (standard error, the difference between the actual data and the data obtained by the model)

Table 2. The prediction epidemic model results of COVID-19 for confirmed cases in all the states of United States, using Logistic and Gompertz model (the similarity metric α=0.5)

Data: Confirmed cases		**Logistic model**					**Gompertz model**				
Cluster	**State**	**a**	**b**	**c**	**$Adj_{rsquare}$**	**RMSE**	**k**	**n**	**q**	**$Adj_{rsquare}$**	**RMSE**
1	(1) Arkansas	3,8E+05	475	0,035	0,995	335	1,5E+08	356,1	0,003	0,995	335
	(2) California	2,7E+05	861	0,043	0,994	4562	6,5E+05	33,12	0,014	0,997	3101
	(3) Florida	2,1E+05	4696	0,032	0,962	5692	4,4E+05	2604	0,010	0,969	5168
	(4) Kansas	1,2E+04	98	0,077	0,993	369	1,3E+04	0,2023	0,043	0,996	280
	(5) Kentucky	1,5E+04	255	0,055	0,992	426	1,9E+04	28,77	0,026	0,997	241
	(6) Maine	3505	140	0,047	0,993	79	4777	69,93	0,021	0,996	61
	(7) Missouri	1,6E+04	646	0,053	0,981	742	1,9E+04	180	0,028	0,991	504
	(8) Nevada	1,7E+04	590	0,039	0,975	661	2,6E+04	267,5	0,017	0,983	536
	(9) North Carolina	1,1E+05	816	0,043	0,997	911	6,6E+05	369,3	0,010	0,999	646
	(10) North Dakota	3462	46	0,068	0,998	59	4133	1,891	0,034	0,999	44
	(11) Oklahoma	1,4E+04	470	0,038	0,970	553	2,3E+04	228	0,015	0,979	468
	(12) South Carolina	4,2E+07	907	0,031	0,982	978	4,9E+08	699,2	0,003	0,981	1005
	(13) Tennessee	4,4E+04	919	0,045	0,990	1102	7,2E+04	299,6	0,018	0,996	745
	(14) Texas	2,6E+05	3377	0,037	0,985	4299	1,0E+06	1767	0,010	0,989	3683
	(15) Utah	2,3E+04	516	0,039	0,988	495	5,0E+04	263,6	0,013	0,993	383
	(16) Vermont	1013	24	0,118	0,965	72	1043	0,9673	0,073	0,977	58
	(17) Washington	2,8E+04	207	0,048	0,983	1286	3,4E+04	0,4778	0,025	0,992	906
	(18) Wisconsin	2,9E+04	522	0,054	0,998	347	4,4E+04	132,9	0,022	0,999	308
2	(19) Alabama	5,6E+04	996	0,041	0,993	779	1,9E+05	585,7	0,011	0,996	605
	(20) Arizona	2,9E+07	121	0,041	0,992	1313	4,5E+08	29,18	0,004	0,988	1547
	(21) Colorado	3,0E+04	633	0,066	0,996	741	3,4E+04	42,47	0,037	1,000	249
	(22) Georgia	6,6E+04	1683	0,050	0,982	2679	8,3E+04	342,5	0,024	0,991	1882
	(23) Indiana	4,3E+04	752	0,064	0,995	1106	5,0E+04	43,27	0,034	0,999	413
	(24) Iowa	2,6E+04	202	0,072	0,995	693	3,1E+04	0,5527	0,038	0,999	361
	(25) Minnesota	3,4E+04	67	0,082	0,999	367	4,0E+04	2,13E-05	0,043	0,999	440
	(26) Mississippi	2,6E+04	630	0,051	0,993	609	3,7E+04	197,1	0,022	0,998	364
	(27) Nebraska	1,7E+04	52	0,083	0,996	410	2,0E+04	2,67E-05	0,047	0,999	199
	(28) New Hampshire	5804	70	0,064	0,998	102	6788	1,814	0,033	1,000	46
	(29) New Mexico	1,0E+04	172	0,064	0,995	238	1,2E+04	12,61	0,033	0,999	94
	(30) Ohio	4,6E+04	1056	0,061	0,992	1368	5,4E+04	129,1	0,032	0,998	693
	(31) South Dakota	6370	113	0,067	0,991	218	7261	5,695	0,037	0,997	127
	(32) Virginia	6,4E+04	579	0,065	0,999	816	8,1E+04	14,89	0,031	0,999	545
3	(33) Michigan	6,1E+04	2378	0,078	0,986	2615	6,4E+04	300	0,050	0,996	1411
	(34) Pennsylvania	8,4E+04	1475	0,076	0,992	2811	9,0E+04	27,66	0,045	0,999	1178
4	(35) Delaware	1,0E+04	140	0,078	0,997	211	1,1E+04	1,44	0,045	0,999	95
	(36) Illinois	1,4E+05	54	0,075	0,999	1402	1,7E+05	6,58E-09	0,034	0,998	2456
	(37) Louisiana	4,7E+04	3258	0,057	0,955	3403	5,1E+04	1178	0,036	0,972	2685
	(38) Maryland	6,7E+04	682	0,068	0,998	1129	7,9E+04	9,668	0,035	1,000	513

5	(39) Alaska	1433	101	0,026	0,913	61	1810	77,31	0,012	0,922	58
	(40) Hawaii	665,4	8	0,144	0,967	46	678,6	0,01096	0,091	0,973	42
	(41) Idaho	4488	404	0,039	0,934	307	5097	245,7	0,023	0,949	270
	(42) Montana	550	46	0,094	0,870	69	579,6	26,46	0,056	0,894	62
	(43) Oregon	9863	288	0,036	0,972	358	1,6E+04	132,1	0,015	0,980	301
	(44) Puerto Rico	1,1E+04	237	0,042	0,993	173	3,5E+04	150,1	0,012	0,995	154
	(45) West Virginia	2752	171	0,050	0,980	114	3255	87,19	0,027	0,989	85
	(46) Wyoming	1252	60	0,051	0,980	55	1497	23,92	0,027	0,989	41
6	(47) District of Columbia	1,0E+04	275	0,072	0,997	184	1,1E+04	33,72	0,040	1,000	84
7	(48) New Jersey	1,6E+05	1926	0,095	0,994	4966	1,7E+05	1,875	0,061	0,999	1689
8	(49) Connecticut	4,4E+04	811	0,084	0,994	1313	4,6E+04	11,39	0,052	0,999	468
	(50) Massachusetts	1,0E+05	50	0,088	0,998	1879	1,2E+05	1E-08	0,041	0,996	2898
	(51) New York	3,7E+05	4777	0,100	0,994	11690	3,8E+05	5,836	0,065	0,999	4212
9	(52) Rhode Island	1,4E+04	58	0,088	0,993	480	1,5E+04	1E-05	0,054	0,997	308

Table 3. The prediction epidemic model results of COVID-19 for death cases in all the states of United States, using Logistic and Gompertz model (the similarity metric α=0.5)

Data: Confirmed cases		**Logistic model**					**Gompertz model**				
Cluster	**State**	**a**	**b**	**c**	**Adj$_{rsquare}$**	**RMSE**	**k**	**n**	**q**	**Adj$_{rsquare}$**	**RMSE**
1	(1) Arkansas	333,7	8,26	0,043	0,979	10,31	596,8	3,702	0,016	0,985	8,53
	(2) California	5833	7,48	0,061	0,995	137,9	7081	1E-07	0,031	0,999	63,76
	(3) Florida	3236	39,27	0,064	0,992	102,2	3764	0,740	0,034	0,998	52,04
	(4) Kansas	244,5	3,47	0,082	0,989	9,856	260	0,019	0,050	0,996	5,70
	(5) Kentucky	531,9	8,38	0,067	0,990	19,01	595,6	0,214	0,038	0,997	10,57
	(6) Maine	103,1	2,63	0,070	0,984	4,758	112	0,174	0,042	0,993	3,23
	(7) Missouri	925,6	10,06	0,080	0,994	27,36	1003	0,038	0,047	0,998	14,76
	(8) Nevada	481,5	7,32	0,071	0,993	14,59	528,7	0,135	0,041	0,999	6,40
	(9) North Carolina	1347	11,19	0,063	0,993	36,67	1662	0,126	0,031	0,998	19,83
	(10) North Dakota	88,54	0,79	0,066	0,994	2,225	117,5	0,036	0,030	0,991	2,63
	(11) Oklahoma	357,4	5,55	0,083	0,992	12,26	378	0,047	0,051	0,998	5,68
	(12) South Carolina	684,9	9,95	0,062	0,992	20,85	825,8	0,496	0,032	0,997	12,72
	(13) Tennessee	601,3	15,78	0,048	0,982	23,1	795	4,147	0,022	0,990	17,01
	(14) Texas	2284	30,13	0,062	0,993	63,3	2739	1,153	0,032	0,998	31,31
	(15) Utah	149,5	1,90	0,060	0,994	3,733	192,6	0,123	0,028	0,998	2,24
	(16) Vermont	54,57	0,12	0,159	0,991	2,136	56	2E-09	0,093	0,989	2,35
	(17) Washington	1216	2,59	0,068	0,994	37,77	1321	6E-09	0,039	0,999	18,90
	(18) Wisconsin	783,5	22,35	0,057	0,988	28,32	921,1	4,154	0,030	0,995	17,71
2	(19) Alabama	886,9	16,27	0,063	0,994	22,81	1083	1,732	0,031	0,998	11,94
	(20) Arizona	1572	1,39	0,059	0,995	33,22	2226	5E-07	0,025	0,998	21,40
	(21) Colorado	1661	17,56	0,075	0,996	41,15	1816	0,067	0,043	0,999	22,06
	(22) Georgia	2700	46,68	0,059	0,989	94,72	3183	2,893	0,031	0,997	54,56
	(23) Indiana	2339	22,94	0,075	0,995	59,65	2581	0,075	0,043	1,000	20,27
	(24) Iowa	731,4	3,00	0,078	0,999	6,828	865,2	0,0007	0,040	0,999	10,04
	(25) Minnesota	1469	6,88	0,073	0,997	27,64	1754	0,003	0,037	1,000	11,26
	(26) Mississippi	1055	14,02	0,064	0,997	18,13	1327	0,94	0,031	0,999	9,70
	(27) Nebraska	345,8	2,20	0,055	0,983	10,7	694,2	0,25	0,019	0,985	10,26
	(28) New Hampshire	368,7	0,18	0,090	0,992	11,47	498,9	5E-09	0,038	0,992	11,47
	(29) New Mexico	485,8	3,39	0,074	0,997	9,983	569,4	0,008	0,039	1,000	3,67
	(30) Ohio	2790	30,37	0,071	0,997	55,6	3209	0,39	0,038	1,000	19,92
	(31) South Dakota	86,66	0,56	0,072	0,993	2,449	104,7	0,001	0,037	0,994	2,18
	(32) Virginia	1631	7,50	0,084	0,994	48,27	1781	7E-05	0,049	0,998	27,42

3	(33) Michigan	5847	80,40	0,097	0,994	174,1	6076	0,219	0,061	0,999	57,16
	(34) Pennsylvania	6381	27,86	0,087	0,997	143,5	6894	0,0001	0,051	0,999	69,05
4	(35) Delaware	460,4	4,04	0,076	0,993	14,4	522,5	0,015	0,041	0,997	9,84
	(36) Illinois	6828	2,87	0,072	0,998	119,4	8880	3E-09	0,031	0,998	96,11
	(37) Louisiana	2880	69,90	0,083	0,991	102,2	3030	2,864	0,052	0,998	44,98
	(38) Maryland	3078	16,57	0,078	0,996	68,37	3438	0,0023	0,044	1,000	18,21
5	(39) Alaska	11,28	0,39	0,083	0,942	1,021	11,85	0,051	0,051	0,949	0,97
	(40) Hawaii	17,12	0,04	0,150	0,990	0,718	17,51	2E-09	0,088	0,989	0,76
	(41) Idaho	84,15	1,90	0,098	0,986	3,909	87,24	0,040	0,063	0,994	2,51
	(42) Montana	18,08	0,54	0,103	0,959	1,37	18,8	0,065	0,062	0,965	1,28
	(43) Oregon	180,5	2,92	0,069	0,988	7,432	197,5	0,058	0,040	0,995	4,68
	(44) Puerto Rico	144,5	4,60	0,084	0,990	5,246	151,7	0,40	0,053	0,997	2,71
	(45) West Virginia	86,5	0,74	0,099	0,982	4,604	90,88	6E-05	0,062	0,989	3,63
	(46) Wyoming	1,004	0,00	0,548	0,968	0,082	1,04	2E-09	0,102	0,885	0,16
6	(47) District of Columbia	525,3	5,73	0,088	0,997	11	564,4	0,0158	0,052	1,000	3,96
7	(48) New Jersey	1E+04	86,32	0,090	0,993	422,3	1E+04	0,0051	0,055	0,998	229,70
8	(49) Connecticut	4175	30,56	0,095	0,996	112,4	4375	0,001	0,059	1,000	35,89
	(50) Massachusetts	7666	1,34	0,092	0,997	178,8	9654	3E-09	0,036	0,993	264,50
	(51) New York	3E+04	179	0,111	0,993	1048	3E+04	0,0002	0,073	0,999	475,90
9	(52) Rhode Island	0,667	1E-5	0,394	0,047	1,107	0,644	2E-14	0,178	0,039	1,11

Finally, the SIR model is calculated for the different representative states of each of the clusters obtained using the parameter α=0.5 (taking into account simultaneously both the number of infected and dead patients), being the parameters presented in Table 4. For the SIR model, the following variables are defined: C_N is the cluster number, N_D is the final number of days take into account for analysing the specific time series, β and γ are average contact and removal frequency used in equation (9). The parameter R is the reproduction number, i.e. number of people infected by a person with COVID-19, defined as:

$$R = \frac{\beta}{\gamma}\left(1 - \frac{C^I_{Total}}{N}\right);\ C^I_{Total} = \sum_{t=t_0}^{t_{end}} I(t) \tag{15}$$

Being C^I_{Total} the total number of people infected at the end of the pandemic (t_{end}) and N is the same constant presented in equation (9) (amount of susceptible population before the outbreak of COVID-19 in a specific state). R_0 is the basic reproduction number, defined as the expected number of secondary cases produced by a single infection in a completely susceptible population and defined by:

$$R_0 = \frac{\beta}{\gamma} \tag{16}$$

The variable T_{EE} is the time estimation of the SIR model in which the number of infected patients is very small, and therefore, it could be affirmed that the pandemic in that state would have ended. NDE denotes the number of days that the infection would have lasted in a given state. The prediction of the SIR model for the different state representative of the cluster obtained, analysing the evolution of total confirmed cases and novel cases per day, is presented in Figure 4.

Table 4. SIR parameters and predictions for representative state in each cluster.

C_N	State	N_D	β	γ	I_0	R_0	R	T_{EE}	N_{DE}
1	California	148	1.615	1.576	36	1.025	0.995	26-nov-2020	342
2	Nebraska	110	0.173	0.083	25	2.075	0.487	08-Oct-2020	245
3	Pennsylvania	111	0.126	0.055	1653	2.297	0.399	17-Nov-2020	297
4	Illinois	148	0.146	0.065	33	2.228	0.447	12-Nov-2020	324

5	Oregon	117	0.172	0.15	110	1.144	1.036	05-Apr-2021	445
6	District of Columbia	101	0.17	0.106	192	1.605	0.619	07-Oct-2020	238
7	New Jersey	112	0.184	0.078	1011	2.354	0.322	10-Oct-2020	247
8	New York	116	0.187	0.075	2305	2.504	0.28	17-Oct-2020	257
9	Rhode Island	117	0.174	0.076	23	2.284	0.372	16-Sep-2020	230

4. Conclusions

A powerful tool for the analysis of time series is the grouping through clustering. Clustering time series is usually an unsupervised process, with the aim of finding behavioral similarities between the different time series that are analyzed. This article has proposed a parametric metric, based on the dynamic time warping distance, in order to measure the distance or similarity between time series corresponding to different states in the United States, taking simultaneously into account the behavior of the number of COVID-19 confirmed cases and deceased persons due to COVID-19 . The proposed parametric metric, named $DTW_{\alpha}(S_A, S_B)$, is robust to the different lengths of data sequences (different beginning of the epidemic in the different states of the United States).

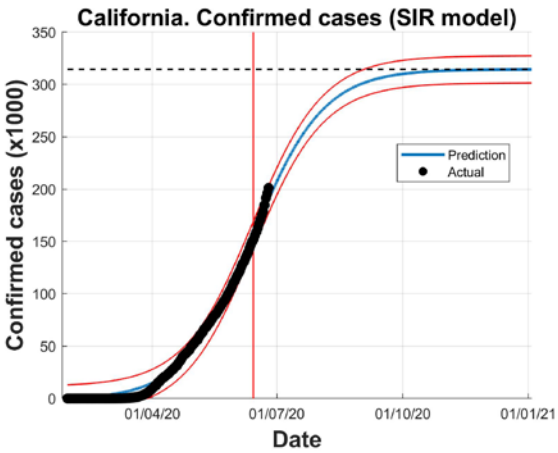
Using the Calinski-Harabasz criterion, the optimal number of clusters in which the different states of United States can be grouped was obtained, taken as value of $\alpha = 0.5$ (same relevance for the time series of confirmed and death patients). A total of 9 heterogeneous clusters were found, in the sense that there are clusters within a large number of states (there are two large clusters, which encompass 18 and 14 states) and other clusters with only one state (indicating that their behavior has been unique, as they do not have excessive similarities with the rest of states).

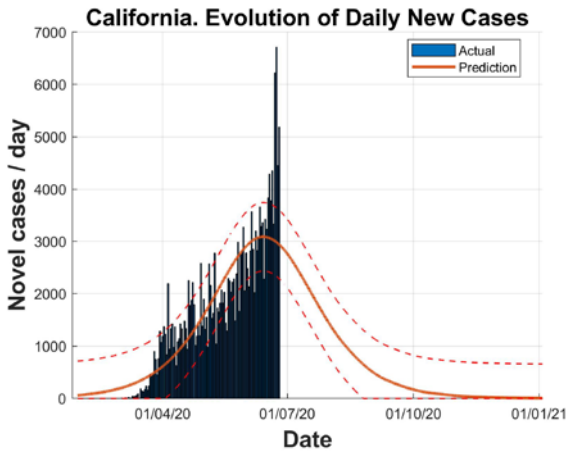
Logistic, Gompertz and SIR mathematical models have been analyzed for the prediction and modeling of the evolution of the epidemic in the different states. For each of the clusters obtained, a representative state is selected and the SIR model was computed. This mathematical model, widely used in the bibliography, allows prediction of the evolution in a given state of the evolution of number of susceptible, infected and recovered patients, being this evolution and the estimate of the final size of the COVID-19 epidemic very relevant for the health authorities.

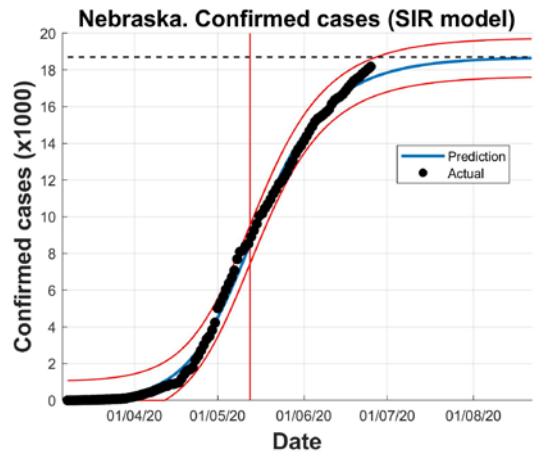
With the proposed hierarchical clustering procedure, it is possible to identify and summarize interesting patterns and correlations in the underlying data of the time series of the states of United States suffering COVID-19 and therefore determine similar behaviors that different states may have.

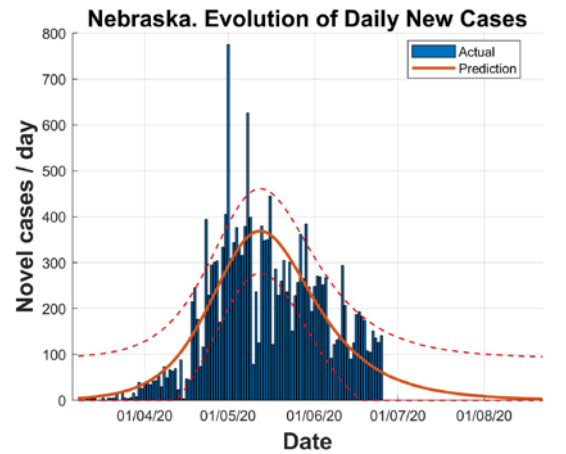
Acknowledgements

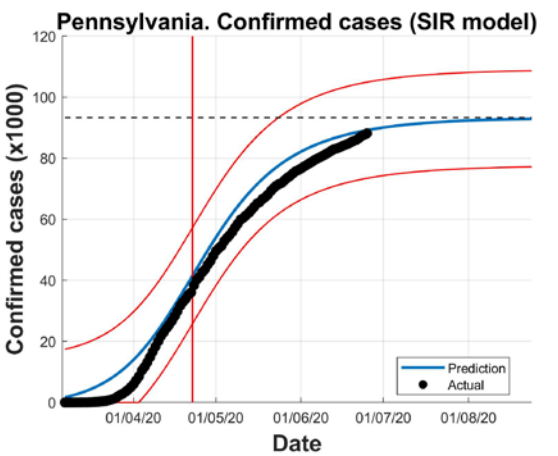
This contribution has been partially supported by the National Spanish project with reference RTI2018-101674-B-I00

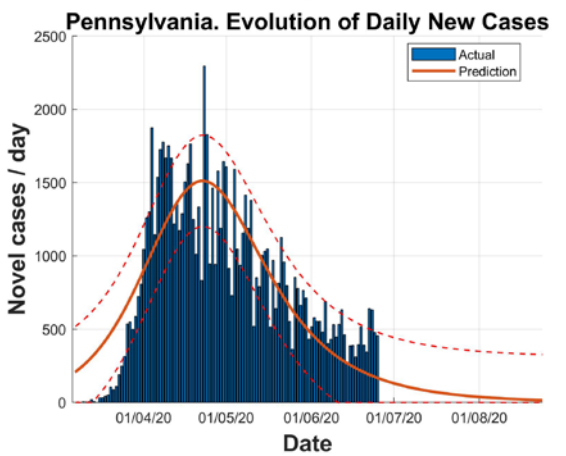


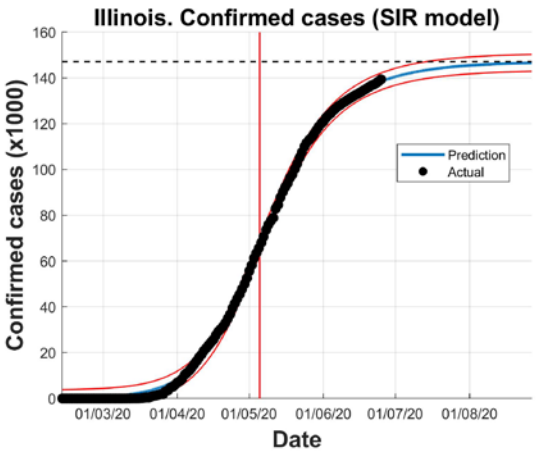


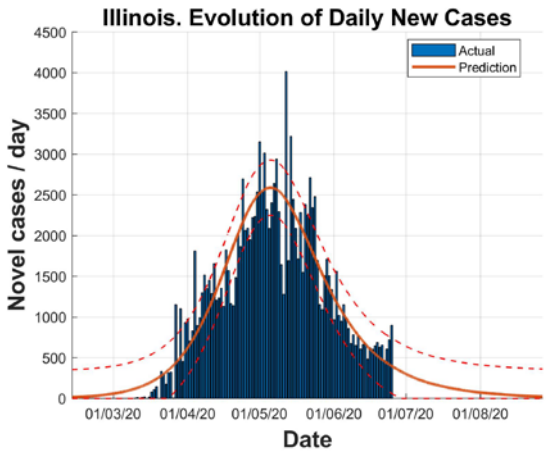


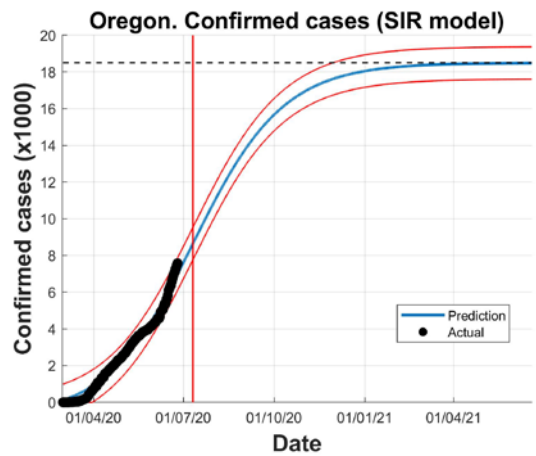


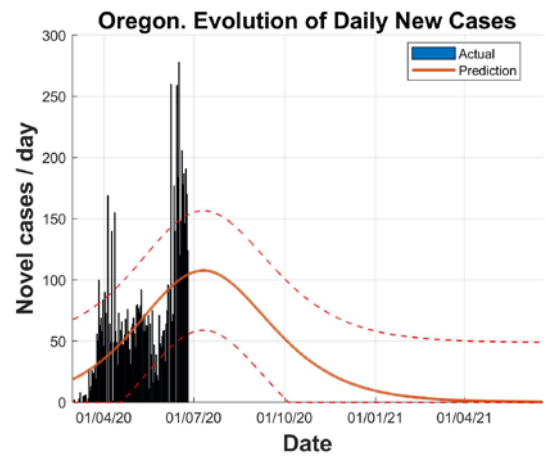


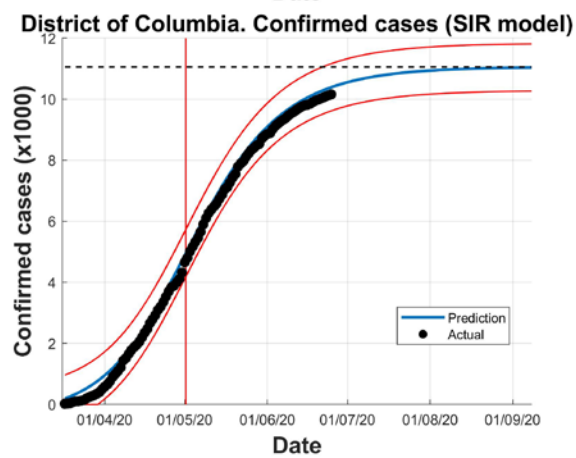


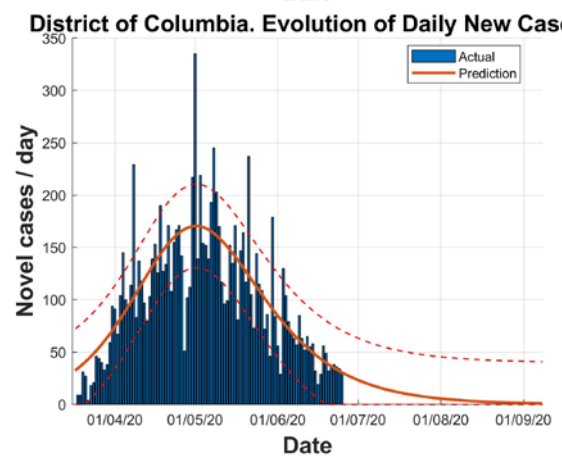












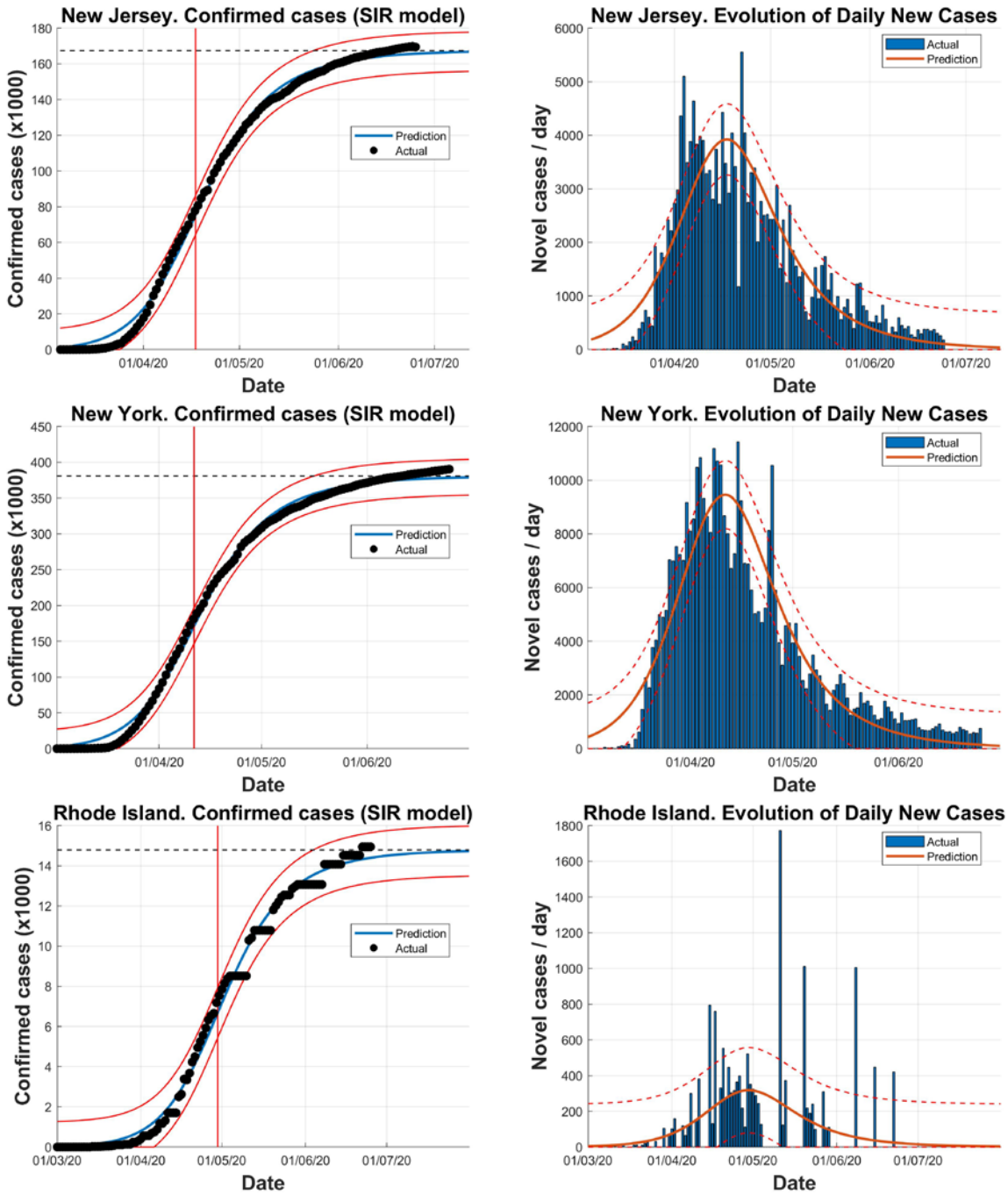


Figure 4 Prediction of the SIR model for the different states of the cluster obtained, analysing the evolution of total confirmed cases and novel infected cases per day.

References

[1] Roques L., Klein E., Papaix J., Sar A. and Soubeyrand S., (2020). Using Early Data to Estimate the Actual Infection Fatality Ratio from COVID-19 in France, Biology, MDPI

[2] Wu, Ke & Darcet, Didier & Wang, Qian & Sornette, Didier. (2020). Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. doi:10.1101/2020.03.11.20034363.

[3] Acuña-Zegarra M., Santana-Cibrian M., Velasco-Hernandez J. (2020). Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance, Mathematical Biosciences.
[4] Karako K., Song P., Chen Y., Tang W., (2020), Analysis of COVID-19 infection spread in Japan based on stochastic transition model, BioScience Trends. 2020; 14(2):134-138.
[5] Zebin Zhao, Xin Li, Feng Liu, Gaofeng Zhu, Chunfeng Ma, LiangxuWang. Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Science of the Total Environment 729 (2020)
[6] Matheus Henrique Dal Molin Ribeiro, Ramon Gomes da Silva, Viviana Cocco Mariani, Leandro dos Santos Coelho. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil, Chaos, Solitons and Fractals 135 (2020)
[7] Muhammad Yousaf, Samiha Zahir, Muhammad Riaz, Sardar Muhammad Hussain, Kamal Shah. Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan, Chaos, Solitons and Fractals 138 (2020).
[8] Ceylan Z. (2020), Estimation of COVID-19 prevalence in Italy, Spain, and France", Science of the Total Environment 729 (2020) 138817.
[9] Perkins, A., Cavany, S.M., Moore, S.M., Oidtman, R.J., Lerch, A., and Poterek,M. (2020). Estimating unobserved SARS-CoV-2 infections in the United States. medRxiv. https://doi.org/10.1101/2020.03.15.20036582.
[10] Fauver et al., (2020), Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States, Cell 181, 990–996. https://doi.org/10.1016/j.cell.2020.04.021
[11] Aghabozorgi, S., Shirkhorshidi, A., Teh Ying W., (2015) Time-series clustering - A decade review. Information Systems, vol.53,
[12] Johnpaul, C., I; Prasad, Munaga V. N. K.; Nickolas, S.; G.R.Gangadharan G.R, (2020), Trendlets: A novel probabilistic representational structures for clustering the time series data, Expert Systems with Applications, vol.145.
[13] Taoying L., Xu W., Zhang, J. (2020), Time Series Clustering Model based on DTW for Classifying Car Parks, Algorithms.
[14] Dong E , Du H , Gardner L . An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 2020;3099(20):19-20
[15] Keogh, E; Ratanamahatana, CA, (2005) Exact indexing of dynamic time warping. Knowledge and Information Systems,Vol.7, n.3, pp.358-386.
[16] Sakoe H., Chiba S., (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Sign. Process. 26, 1 (1978), 43–49.
[17] Bandara K., Bergmeir C., Smyl S., (2020) "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach" Expert Systems with Applications Vol.140, UNSP 112896
[18] Kaufman L., Rousseeuw P., (2009) Finding Groups in Data: An Introduction to Cluster Analysis. Vol. 344. John Wiley & Sons.
[19] Wu, Ke & Darcet, Didier & Wang, Qian & Sornette, Didier. (2020). Generalized logistic growth modeling of the COVID-19 outbreak in 29 provinces in China and in the rest of the world. 10.1101/2020.03.11.20034363.
[20] Lin Jia, Kewen Li, Yu Jiang, Xin Guo, Ting zhao, Prediction and analysis of Coronavirus Disease 2019, 2020, arXiv:2003.05447.
[21] Zwietering, M. H., Jongenburger, I., Rombouts, F. M., & van 't Riet, K. (1990). Modeling of the bacterial growth curve. Applied and environmental microbiology, 56(6), 1875–1881.
[22] Verma V, Vishwakarma RK, Verma A, Nath DC, Khan HTA. Time-to-Death approach in revealing Chronicity and Severity of COVID-19 across the World. PLoS One. 2020;15(5):e0233074. Published 2020 May 12. doi:10.1371/journal.pone.0233074
[23] Harko.T, Lobo. F. S. N and Mak. M. K. (2014). Exact analytical solutions of the Susceptible-Infected- Recovered (SIR) epidemic model and of the SIR model with equal deaths and births, Applied Mathematics and Computation, 236 pp.184–194.
[24] Saracli, S., Dogan, N. & Dogan, I. (2013), Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl. Vol. 203. https://doi.org/10.1186/1029-242X-2013-203