
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.

Applied Soft Computing 129 (2022) 109606

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Analysis of the socioeconomic impact due to COVID-19 using a deep clustering approach

Yullys Quintero [a], Douglas Ardila [a], Jose Aguilar [b,c,d,*], Santiago Cortes [e]

[a] Department of Computer Science, Universidad EAFIT, Medellin, Colombia
[b] CEMISID, Universidad de Los Andes, Merida, Venezuela
[c] GIDITIC, Universidad EAFIT, Medellin, Colombia
[d] Universidad de Alcala, Dpto Automatica, Alcala de Henares, Spain
[e] Factored.ai, Medellin, Colombia

ARTICLE INFO

*Article history:*
Received 18 April 2022
Received in revised form 6 August 2022
Accepted 23 August 2022
Available online 5 September 2022

*Keywords:*
Unsupervised model
Clustering evolution
Socioeconomic model
COVID-19
Time series prediction model

ABSTRACT

One of the main problems that countries are currently having is being able to measure the impact of the pandemic in other areas of society (for example, economic or social). In that sense, being able to combine variables about the behavior of COVID-19 with other variables in the environment, to build models about its impact, which help the decision-making of national authorities, is a current challenge. In this sense, this work proposes an approach that allows monitoring the socioeconomic behavior of the regions/departments of a country (in this case, Colombia) due to the effect of COVID-19. To do this, an approach is proposed in which the behavior of the infected is initially predicted, and together with other context variables (climate, economics and socials) determines the current socioeconomic situation of a region. This classification of a region, with the pattern that characterizes it, is a fundamental input for those who make decisions. Thus, this work presents an approach based on machine learning techniques to identify regions with similar socioeconomic behaviors due to COVID-19, so they should eventually have similar public policies. The proposed hybrid model initially consists of a time series prediction model of infected, to which are added several context variables (climate, socioeconomic, incidence of COVID-19 at the level of deaths, suspects, etc.) in an unsupervised learning model, to determine the socioeconomic impact in the regions. Particularly, the unsupervised model groups similar regions together, and the pattern of each group describes the socioeconomic similarities between them, to help decision-makers in the process of defining policies to be implemented in the regions. The experiments showed the ability of the hybrid model to follow the evolution of the regions after 4 weeks. The quality metrics for the predictive model were around the values of 0.35 for MAPE and 0.68 for $R^2$, and in the case of the clustering model were around the values of 0.3 for the Silhouette index and 0.6 for the Davies–Boulding index. The hybrid model allowed determining things like some regions that initially belonged to a group with a very low incidence of positive cases and very unfavorable socioeconomic conditions, became part of groups with moderately high incidences. Our preliminary results are very satisfactory since they allow studying the evolution of the socioeconomic impact in each region/department.



## 1. Introduction

The pandemic has brought with it, in addition to health problems, an impact on different areas of our society. Some of the areas that have been impacted are social, economic, and educational, among others. Particularly, for the public administration, it has been very difficult to take measures, due to the little experience in similar situations, its rapid expansion, and the multiple effects in such diverse areas of society, among other reasons.

On the other hand, three different sanitary measures can be used to counteract a pandemic: (i) herd immunity development, (ii) vaccination, and (iii) lockdown. The first measure is unreasonable due to a large number of necessary fatalities, and the second measure is in the preliminary stages. Thus, the majority of countries have considered the third option despite its adverse economic impact.

Among the measures taken by governments were lockdown [1], increase Government spending [2], and enforce the use of

* Corresponding author at: CEMISID, Universidad de Los Andes, Merida, Venezuela.
*E-mail addresses:* ymquinterm@eafit.edu.co (Y. Quintero), dardila5@eafit.edu.co (D. Ardila), aguilar@ula.ve, jlaguilarc@eafit.edu.co (J. Aguilar), santiago.cortes@factored.ai (S. Cortes).

https://doi.org/10.1016/j.asoc.2022.109606
1568-4946/

masks [3]. However, justifying draconian measures during the COVID-19 pandemic has not been easy because of individual rights restrictions, but especially, due to its economic and social impacts. Thus, these measures have been considered necessary by the health authorities, but they have not been very popular among citizens for social, economic, legal, among other reasons. Now, public policies during the COVID-19 outbreak should be based on holistic approaches that consider health, economic and social variables in a territory. In this regard, this paper proposes a hybrid machine learning approach to study the evolution of the effects of COVID-19 on the socioeconomic aspects of regions of a country.

### 1.1. Previous works

There are few papers in the literature that focus on this problem based on machine learning techniques, and the majority of the applications are about forecasting models that consider the dynamical behavior of the time-series of COVID-19. These models search to understand the nature of COVID-19 and forecast its spread, considering the relationship among the variables of the SEIRD model (Susceptible, Exposed, Infected, Recovered and Deceased) for COVID-19 infection, but also, weather and transmission variables, among other variables [4–7]. As well, there are some unsupervised approaches to establish COVID-19 impact, like [8,9], but focus mainly on health and not on other types of variables. The authors of [8] analyze the risk factors and clinical outcomes to cluster countries in groups with shared profiles of the COVID-19 pandemic. They use a k-means algorithm to define clusters of countries based on the next variables: disease prevalence estimates, metrics of air pollution, socioeconomic status and health system coverage. With this information, they define clusters in terms of the number of confirmed COVID-19 cases, number of deaths, case fatality rate, and order in which the country reported the first case. In the work of [9], they propose a dynamic clustering for analyzing the lockdown impact due to COVID-19 flare-up. They use healthcare and simulated mobility data to model lockdown as a clustering problem, and design a dynamic clustering algorithm for localized lockdown by taking into account the pandemic, economic and mobility aspects.

In the work of Malki et al. [4], various regressor machine learning techniques are used to determine the relationship between different factors and the spreading rate of COVID-19. The techniques estimate the impact of weather variables (e.g., temperature and humidity) on the transmission of COVID-19 by extracting the relationship between the number of confirmed cases and the weather variables in certain regions. Ze et al. [10] develop a deep learning model to forecast the transmission rate of COVID-19 as a function of features that encompass selected variables of climate conditions, socioeconomic, and government restrictions. In the study of Zhou et al. [11], they analyze the correlation between the land surface temperature (LST) and industrial production, by using the BFAST algorithm and linear regression models on multitemporal MODIS data to derive monthly time-series deviation of LST with a spatial resolution of $1 \times 1$ km. In this way, they explore the spatiotemporal patterns of the COVID-19 control measures impact on industrial production, within Wuhan city. Viezzer and Biondi [12] investigate if COVID-19 variables are higher in cities with higher urbanization, worst socio-economic conditions, and less vegetation cover, in the Atlantic Forest of Brazil. Indexes are created using socio-economic (e.g., absolute built area, average per capita income, population, demographic density), and eco-environmental data (e.g., absolute and relative forest cover, absolute and relative vegetation cover), which are correlated with absolute and relative confirmed deaths, absolute and relative confirmed COVID-19 cases, among other variables. Finally, Guo et al. [13] analyze several confirmed cases of COVID-19 from several countries between January and April 2020, using a nonlinear model, to investigate the associations between COVID19 incidence and daily temperature, relative humidity, and wind speed.

Gupta et al. [14] propose a prediction model for confirmed and dead cases of COVID-19. They use a long short-term memory (LSTM) with two layers to obtain predicted coronavirus cases and deaths for the next 30 days. The authors of [15] define an approach for predictive models for the SEIRD variables, which considers the historical data collected and the context variables. Particularly, the context variables examined include morbidity rates, number of people over 65 years old, and population density, among others. In addition, they carry out an analysis of the dependence of these variables, and also, their relationship with the context variables, to avoid multicollinearity problems and the curse of dimensionality. Finally, several prediction models based on different machine learning techniques and inputs are developed, which include dependence with context variables, temporal interdependence, and temporal intra-dependence. Finally, Camargo et al. [16] define an incremental learning approach to build predictive models of the SEIRD variables for the COVID-19 pandemic. The learning approach is a dynamic ensemble method based on a bagging scheme that allows the addition of new models or the updating of incremental models.

As we can see, there are a large number of studies on predicting the behavior of COVID-19. Also, several studies have analyzed the relationship of certain variables in the socio-economic and environmental spheres with the spread of COVID-19. In addition, some unsupervised learning techniques have been used to analyze the impact of COVID-19. However, there are no studies that analyze the socio-economic impact of COVID-19 based on the predicted behavior of the infection variable, which changes due to the appearance of new strains, new drugs, among others, which affects in real-time to the society.

### 1.2. Contribution

One of the great challenges is to build models that allow describing the behavior of diseases such as COVID-19, and their effects, having little data, and much ignorance of the collateral effects. Another great challenge is to be able to update these models as more knowledge about the disease is obtained. Consequently, the main contribution of this work is to propose a hybrid machine learning model to analyze the socio-economic impact of COVID-19 in a region/department. This model combines variables about the behavior of COVID-19 with other context variables to evaluate its impact in a region. The model uses the variables of the SEIRD model, predicting the infected values, together with other context variables (climate, economics and socials), to determine the current socio-economic situation of a region. Particularly, our approach groups the regions, and the pattern of each group determines the main socio-economic characteristics to be analyzed. The proposed hybrid model consists of a time series prediction model of infected, to which are added several context variables (climate, socio-economics, incidence of COVID-19 at the level of deaths, suspects, etc.) in a clustering model, in order to determine the socio-economic impact in the regions. The clustering model groups similar regions, and the pattern of each group describes its socio-economic characteristics, which can be used by decision-makers to define the public policies to be implemented in the regions. In this work, a dynamic clustering is used, which operates online to update the groups according to new incoming data.

Thus, this work presents a hybrid approach based on machine learning techniques to identify regions with similar socio-economic impacts due to COVID-19, so they should eventually


**Fig. 1.** System's architecture.

have similar public policies. In this work, a predictive model is proposed to foretell the infection rate of COVID-19. Also, several clustering techniques have been analyzed to define the clusters of the regions according to their similitude. Previously, different techniques have been used for the dimensionality reduction of the variables, in order to extract the best clustering variables of the regions. In this way, our approach allows studying the evolution of the socioeconomic impact in each region/department, in order to monitor its socioeconomic behavior due to COVID-19. Specifically, the approach is used to evaluate the evolution of the socioeconomic impact of COVID-19 on Colombia. Our approach could be used in the next waves of corona virus-related diseases, and other viral life-threatening calamities, to significantly reduce the socioeconomic impact of COVID-19.

## 2. Our hybrid machine learning model

Our system consists of two components (see Fig. 1). On the one hand, a forecasting model that predicts the number of infection cases for each of the next 7 days by each department. On the other hand, a clustering process that, with the help of dimensionality reduction techniques and the infection prediction value, generates clusters associated with departments according to their socioeconomic behavior due to the pandemic. In addition, a common task in the two components is the feature engineering process that is carried out in each one, to determine the variables that affect the prediction of infections, or the grouping of departments.

Fig. 1 shows the architecture of our hybrid approach. Initially, the variables to be taken into account in the models are determined, which are measured in each region. Then, the prediction models are built for the variables of interest (previously, the input data are cleaned and analyzed to eliminate variables that do not contribute to the models). The next step creates the clustering model with the result of the predicted variables and the socio-demographic variables.

Thus, our approach is composed of three groups of techniques:

- For the prediction model of COVID-19 infections, we have used a neural network model.
- For the feature engineering process, which in our case consists of dimensionality reduction, we have used three techniques: Principal Component Analysis (PCA), Auto-encoders and Genetic Algorithms (GA).
- For the clustering process, whose groups are used to analyze the evolution of the socio-economic impact of COVID-19, we have used k-means and k-medoids.

Initially, our approach predicts the behavior of the infected with COVID-19. Then, with this information, it makes a vector representation of each political subdivision of interest that contains, in addition, the pandemic's progression, economic, healthcare, geographical and demographic data, to be clustered (after applying a dimensionality reduction technique), and such clusters are the system's final output.

### 2.1. Variable characterization

In this work, different types of variables were considered, which can be grouped into the following groups:

- *Variables linked to the behavior of COVID-19*, which consist of the real-time values of the SEIRD variables. Of this group, only the infected variable is predicted, and the rest are taken from data sources in real-time.
- *Climatic variables*, which are also taken in real-time. Some of the variables considered were precipitation, temperature and relative humidity.
- *Economic variables of the region*, such as poverty level, child labor, economic dependence, informal employment, the area dedicated to the agricultural sector, mining production, etc.
- *Social variables of each region*, such as educational coverage, institutions of higher education, illiteracy, school absence, inhabitants living in houses, inhabitants living in apartments, inadequate flooring, and overcrowding.
- *Geographic variables*, such as height, seismic threat, and forest areas;
- *Demographic variables*, such as rural population, the population under 15 years, the population between 15 and 24 years old, the population over 65, the female population, the population density, the indigenous population, among others;
- *Variables from the health area*, such as mortality from meningitis per 100 thousand inhabitants, mortality from acute respiratory infection per 100 thousand inhabitants, mortality from problems of the cardiovascular system per 100 thousand inhabitants, mortality from chronic respiratory disease, ambulances, number of Intensive Care Unit (ICU), etc.
- Finally, *variables of infrastructure and services in the region*, such as airports, roads, electricity coverage, aqueduct coverage, sewer coverage, internet coverage, etc.

Due to the important number of variables considered (80 variables), one of the fundamental tasks of our approach is the reduction of dimensionality, to determine the appropriate (correlated) variables for the goal of our system.

### 2.2. Prediction model

For the infection forecast, the variables shown in Table 1 have been used for each department. Those variables are static in time, as it is assumed that they do not change significantly during each week of the forecast. The variables are extracted from official sources.

These variables in Table 1 can be grouped in the previous groups of variables: variables linked to the behavior of COVID-19 (e.g., Deaths by chronic diseases, Deaths by a respiratory illness. etc.), climatic variables (e.g., temperature, precipitation, etc.), economic variables (e.g., Informal economy, etc.), social variables (e.g., the population between 15 and 24 years Child labor, Total population with Diabetes, etc.); Demographic variables (e.g., population density, etc.).

The chosen forecaster model is an artificial neural net that only uses the new COVID-19 cases as external variables (SEIRD

**Table 1**
Gathered variables.

| | | | | |
|---|---|---|---|---|
| Altitude | population between 15 and 24 years | Child labor | Total population | population with Diabetes |
| Precipitation | population over 65 years | Dependency ratio | Life expectancy | Deaths by chronic diseases |
| Temperature | population density | Informal economy | Deaths by digestive diseases | Deaths by acute diseases |
| Humidity | women population | illiteracy | Deaths by respiratory illness | Deaths by endocrine disorders |
| Population under 15 years | Multidimensional Poverty Index | school dropout | Deaths by cardiac complications | Death by maligne Neoplasm |

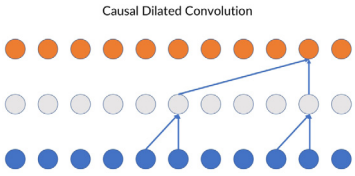



**Fig. 2.** Causal dilated convolutions.

variables). The series of new positive cases by each department can have several peaks, explained maybe by delays in testing. To overcome this problem, three features were included in the model. First, the input data is normalized by taking the logarithm of the raw data. Then, an exponential moving average (Eq. (1)), with $\alpha = 0.1$, is computed for all the series of new infections and used as a feature (as suggested in [17]). Finally, the day of the week was incorporated as an input using an embedding layer.

$$v_t = \alpha x_{t-1} - (1-\alpha)v_t \tag{1}$$

Encoder–decoder architectures have been satisfactorily used for time series forecasting [18]. In particular, 2014 deep Mind's wave net [19] proposed a way to stack convolutional layers to extract features from sequence data, and keep the number of parameters in the model low. This last property reduces the amount of computational power required to train the model. Particularly, in this paper is used an autoencoder neural network. The main parameters in an autoencoder are: the number of nodes in any hidden layer, the number of hidden layers, activation unit (e.g., sigmoid, tanh, softmax, and ReLu activation functions), and regularization parameters on hidden unit weights. Thus, the used architecture consists of an encoder that is made of a series of blocks, each one composed of a dilated convolution (see Fig. 2), with 128 filters of size 2 and causal padding, a Dense layer followed by batch normalization, and finally, a ReLu activation [20] (see Fig. 3.a).

There were 6 of these blocks with 2, 2, 4, 4, 6 and 6 as dilation rates, respectively. For predictions, the decoder takes as inputs the output of the encoder and the one-dimensional embedding for the day of the week in the forecasted window, and then passes these inputs through dense layers with ReLu activations (see Fig. 3). The predictions are made by a dense layer with the size of the forecasted windows as its number of units.

The net is trained using the quantile loss, in order to predict confidence intervals. The 5, 90 and 50 quantiles are predicted, and the last one is taken as the prediction. Before computing the loss, the net output is re-scaled using the last value of the exponential smoothing in the training window:

$$\hat{y}_{t+1,..t+h} = exp(NN(x)) \times I_t \tag{2}$$

$$L(\hat{y}_i, y_i) = max(q(\hat{y}_i, y_i), (1-q)(\hat{y}_i - y_i)) \tag{3}$$

As mentioned before, the smoothing exposed above is necessary in order to overcome the intrinsic heterogeneity of the data. Nevertheless, this approach has its own drawbacks: First, the net is less sensible to instant shocks in the signal. In the context of the COVID-19 outbreak, this can have as a consequence a model that is less efficient to react to sudden downs or spikes created by extraordinary events. The second (and most important drawback) is the necessity of fixing the window size hyper-parameter. The latter implies an added complexity to the model, and creates an extra step to choose this parameter.

### 2.3. Feature engineering process

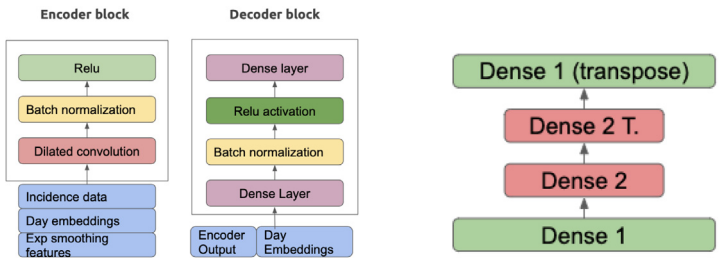
Once the inference is made, the prediction is added to a vector that contains all the other variables gathered. This vector is analyzed in order to avoid the curse of dimensionality [21,22], before applying the k-means or k-medoids algorithm.

For that, a dimensionality reduction is done. Three dimensionality reduction techniques were used: a PCA that captures at least 99 percent of the original variance, GA for feature selection based on the silhouette score [23], and finally, a simple auto-encoder.

- **Principal Component Analysis**
  The principal component analysis is one of the techniques of unsupervised learning, which can be applied for exploratory analysis of the data. One of its applications is for the reduction of dimensionality (number of variables) because in these dimensions the greatest amount of information possible is retained. It is applied to a large number of quantitative variables that are likely to be correlated, achieving the construction of a smaller number of new variables that are not correlated. In our case, it was applied to the set of 80 variables (grouped into the sets of variables already mentioned) and regions (departments) that were taken into account in this study.
- **Auto-Encoders**
  Auto-encoders are a type of neural network architecture that allows building representations of data. For doing so, a Neural Net is trained using the identity as it loses function, and then, some vectors from intermediate layers are extracted. The simpler auto-encoder consists of two parts, an *encoder* and a *decoder*, these can be defined as functions in the following sense:

  $$\phi : \mathcal{X} \longrightarrow \mathbb{R}^n$$
  $$\psi : \mathbb{R}^n \longrightarrow \mathcal{X}$$
  $$\phi, \psi = argmin_{\phi,\psi} \|X - (\psi \circ \phi)X\|^2$$

  Auto-encoders have been successfully applied in information retrieval and dimensionality reduction applications. Our approach uses two autoencoders with the same architectures, but symmetrically such as the weights between layers are shared, thus reducing the number of weights as is shown in Fig. 3. This last is because, with shared weights, it is less prone to overfit in regimes with a small amount of data.
- **Genetic algorithms**
  GA are a metaheuristic based on evolutionary learning that aims to explore the space of solutions in order to find the optimal. In our approach, it was applied to the set of 80 variables as a strategy to reduce the dimensionality (variables) and achieve the best solution. The idea is to find a subset of descriptors (reduction of dimensionality), optimizing a fitness function (Mean Absolute Percentage Error (MAPE)) that corresponds to a quality metric.




(a) Forecasting net architecture

(b) Symmetric autoencoder architecture.

**Fig. 3.** Deep learning architectures used in this work.

The process of selection of descriptors begins with subsets of features/descriptors (population), such that each individual represents a subset of them, which is evaluated with the quality metric (fitness function — MAPE). According to this value, the best individuals are selected to reproduce new individuals using genetic operators, such as mutation and crossover. If the offspring have a better aptitude than their parents, they are more likely to survive. Also, for the generation of the descendants, those individuals that are better reproduce more, and therefore, those that are worse disappear. In the end, the GA find a set of suitable individuals, determined by the quality metrics. In this way, the algorithm returns the subset of descriptors with the best combination of descriptors found (best individual).

### 2.4. Clustering techniques

The Clustering process was carried out taking into account the results obtained in the feature engineering. Two clustering techniques were implemented: k-means and k-medoids.

- **K-means**
  The K-means algorithm is a clustering technique that groups departments into k groups based on information obtained from their descriptors. The aim of this technique is to create clusters that are very different between them, which is done by minimizing the sum of distances between each department and the centroid of its group or cluster. The centroid is the equidistant point of the individuals belonging to that cluster.
- **K-medoids**
  The K-medoids algorithm, like K-means, is an unsupervised learning technique that groups departments into k-groups based on information obtained from their descriptors. It also creates clusters when the departments in a cluster are very similar to each other, but different from departments in other clusters. It is more robust to noise and outliers than K-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances.

## 3. Initial experiments

To evaluate the performance of our approach, we analyze the socio-economic behavior of the Colombian departments generated by COVID-19. In this case, 76 demographic, economic, geographic, and social variables and 4 variables related to COVID-19 according to the SEIRD model, were considered. In addition, the infected variable was predicted.

In our approach, the forecasting model predicts the number of infected for each of the next 7 days in all 33 Colombian departments. On the other hand, the clustering process, with the help of dimensionality reduction techniques, generates clusters associated with each department. It is important to remark that the dimensionality reduction techniques consider in their analysis the predicted infected variable, but to carry out the clustering process, this variable is incorporated into the input variables if it was eliminated by the dimensionality reduction process.

### 3.1. Infected prediction

For the infected forecast, it is used the variables shown in Table 1 for each department. Those variables are static in time because they do not change significantly in a week. The demographic and economic variables were extracted from the official sources of the National Administrative Department of Statistics (DANE) and the National Planning Department (DNP), respectively. The climate and geographical variables were obtained from the official page of Colombia's Institute of Hydrology, Meteorology and Environmental Studies (IDEAM). The health variables are available on the official websites of the National Institute of Health (INS), and of the Ministry of Health and Social Protection.

The prediction of the infection caused by COVID-19 in Colombia was made with the positive cases reported from March to August 2020 by the National Institute of Health (INS). The predictive model predicts a 7-day time window for the positive cases that each department will have. The model was retrained every 7 days. Table 2 shows the average daily predictions for each time interval (of 7 days) in each Colombian department; the intervals are named Week 1, Week 2, Week 3 and Week 4, and the department column presents all the departments of Colombia. The first week of prediction begins on the first week of September.

In a second iteration, weeks 5, 6, 7 and 8 were predicted for the departments. Table 3 shows these predictions.

For each prediction made with the model, the error was measured in percentage terms, particularly, it is calculated the MAPE. Also, $R^2$ was used as a quality metric. The results are presented in Tables 4 and 5. The values presented are the average of the MAPE and $R^2$ values per department in each prediction interval.

In Table 4, it can be seen that the third week had the best predictions for the departments since it has the lowest MAPE value (34%). For the second iteration (see Table 5), week 7 has the best prediction, since the MAPE is 30%. In addition, the $R^2$ values (coefficient of determination) are presented to evaluate the quality of the models. The $R^2$ values indicate that the models explain the variability of the data relatively well.

### 3.2. Feature engineering

In practice, having many variables is considered a problem, given that some of these variables do not provide relevant information for the work being done, making the problem more

**Table 2**
Average daily prediction of infected cases by COVID-19.

(a)

| Department | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 1 | Week 2 | Week 3 | Week 4 |
| Amazonas | 9 | 4 | 3 | 2 |
| Antioquia | 1731 | 1407 | 1369 | 1136 |
| Arauca | 20 | 25 | 25 | 26 |
| Atlántico | 591 | 294 | 224 | 165 |
| Bogotá | 3864 | 3340 | 2890 | 2426 |
| Bolívar | 229 | 251 | 174 | 163 |
| Boyacá | 63 | 98 | 107 | 103 |
| Caldas | 44 | 78 | 77 | 79 |
| Caquetá | 157 | 150 | 134 | 109 |
| Casanare | 20 | 25 | 28 | 36 |
| Cauca | 89 | 127 | 146 | 146 |
| Cesar | 149 | 305 | 332 | 369 |
| Choco | 33 | 20 | 15 | 13 |
| Córdoba | 475 | 404 | 293 | 238 |
| Cundinamarca | 370 | 495 | 489 | 430 |
| Guainía | 0 | 5 | 7 | 14 |
| Guaviare | 5 | 3 | 9 | 13 |

(b)

| Department | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 1 | Week 2 | Week 3 | Week 4 |
| Huila | 85 | 164 | 174 | 201 |
| La Guajira | 80 | 97 | 116 | 107 |
| Magdalena | 240 | 168 | 145 | 128 |
| Meta | 139 | 223 | 267 | 259 |
| Nariño | 228 | 188 | 180 | 161 |
| Norte de Santander | 276 | 326 | 261 | 190 |
| Putumayo | 66 | 60 | 59 | 40 |
| Quindío | 19 | 33 | 47 | 59 |
| Risaralda | 112 | 168 | 165 | 163 |
| San Andrés y Providencia | 2 | 6 | 9 | 18 |
| Santander | 402 | 530 | 530 | 486 |
| Sucre | 162 | 197 | 154 | 127 |
| Tolima | 109 | 172 | 182 | 183 |
| Valle del Cauca | 812 | 635 | 573 | 536 |
| Vaupés | 5 | 8 | 16 | 19 |
| Vichada | 0 | 3 | 6 | 9 |

**Table 3**
Average daily prediction of infected cases by COVID-19.

(a)

| Department | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 5 | Week 6 | Week 7 | Week 8 |
| Amazonas | 2 | 2 | 1 | 1 |
| Antioquia | 1178 | 1181 | 1121 | 1480 |
| Arauca | 26 | 29 | 34 | 38 |
| Atlántico | 139 | 122 | 103 | 125 |
| Bogotá | 2104 | 2009 | 1794 | 1830 |
| Bolívar | 152 | 114 | 119 | 140 |
| Boyacá | 111 | 122 | 130 | 161 |
| Caldas | 97 | 112 | 119 | 152 |
| Caquetá | 99 | 101 | 79 | 99 |
| Casanare | 38 | 49 | 54 | 60 |
| Cauca | 157 | 147 | 134 | 125 |
| Cesar | 321 | 292 | 270 | 242 |
| Choco | 12 | 10 | 8 | 6 |
| Córdoba | 171 | 148 | 122 | 101 |
| Cundinamarca | 389 | 343 | 295 | 309 |
| Guainía | 22 | 23 | 24 | 16 |
| Guaviare | 19 | 20 | 19 | 21 |

(b)

| Department | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 5 | Week 6 | Week 7 | Week 8 |
| Huila | 235 | 240 | 228 | 294 |
| La Guajira | 103 | 102 | 75 | 69 |
| Magdalena | 115 | 106 | 85 | 71 |
| Meta | 253 | 242 | 186 | 232 |
| Nariño | 199 | 146 | 119 | 116 |
| Norte de Santander | 162 | 153 | 145 | 167 |
| Putumayo | 42 | 28 | 25 | 21 |
| Quindío | 68 | 90 | 109 | 125 |
| Risaralda | 172 | 186 | 160 | 186 |
| San Andrés y Providencia | 34 | 36 | 41 | 27 |
| Santander | 451 | 433 | 393 | 393 |
| Sucre | 94 | 78 | 51 | 55 |
| Tolima | 194 | 185 | 135 | 176 |
| Valle del Cauca | 576 | 532 | 573 | 625 |
| Vaupés | 17 | 13 | 13 | 12 |
| Vichada | 17 | 13 | 11 | 14 |

complex. For this reason, it is decided to reduce the dimensionality of the dataset. Three techniques were used to reduce the dimensionality:

- The first technique used was a PCA, with which we kept 20 dimensions. These dimensions retain 99% of the variability present in the original data set. The PCA allowed us to move from a dimension of 80 variables that are correlated among themselves, to a dimension of 20 new variables (principal components) that are not correlated among themselves, retaining 99% of the variability present in Colombia's departments, thus losing the least amount of information possible.
- Another technique used was an auto-encoder with which it was decided to reduce the dimensionality to 3. This dimension was chosen in order to evaluate a much smaller dimension than the previous one.
- Finally, the GA were used to reduce the dimensionality. Each individual represents a subset of descriptors, which is evaluated using as the fitness function the Silhouette index given by the clusters defined by the k-means algorithms. The Silhouette index takes values between $-1$ and 1, a high value close to 1 indicates that the cluster is well-formed, while values close to $-1$ indicates the opposite (clusters poorly formed). In this way, the GA select the best subset of descriptors.

**Table 4**
Quality measures — first iteration.

| | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 1 | Week 2 | Week 3 | Week 4 |
| MAPE | 0.373 | 0.354 | 0.343 | 0.355 |
| $R^2$ | 0.695 | 0.701 | 0.695 | 0.636 |

**Table 5**
Quality measures — second iteration.

| | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 5 | Week 6 | Week 7 | Week 8 |
| MAPE | 0.340 | 0.408 | 0.301 | 0.372 |
| $R^2$ | 0.674 | 0.643 | 0.614 | 0.726 |

### 3.3. Clustering

In our case study, we intend to group the 33 departments of Colombia in k groups, in order to find sets of departments that, given their characteristics, are similar. To group the departments in each time interval, according to their characteristics or similarities, experiments were carried out with 2 different clustering techniques: k-means and k-medoids.

There are several methods to choose the value of k, such as: the elbow method, the Calinsky criterion, the Affinity Propagation, the Gap, Dendrograms, among others. This work takes

**Table 6**
Results of the k-means first and second iteration.

| Methods | Metrics | First Iteration — Forecasting Intervals | | | | | | | | Second Iteration — Forecasting Intervals | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Week 1 | | Week 2 | | Week 3 | | Week 4 | | Week 5 | | Week 6 | | Week 7 | | Week 8 | |
| All variables | Silhouette | k = 6 | 0.248 | k = 3 | 0.246 | k = 5 | 0.249 | k = 3 | 0.246 | k = 5 | 0.250 | k = 3 | 0.247 | k = 3 | 0.252 | k = 3 | 0.252 |
| | Davies–Boulding | | 0.925 | | 1.280 | | 0.898 | | 1.280 | | 0.897 | | 1.279 | | 1.269 | | 1.269 |
| PCA | Silhouette | k = 6 | 0.255 | k = 6 | 0.254 | k = 5 | 0.255 | k = 6 | 0.254 | k = 5 | 0.255 | k = 6 | 0.253 | k = 3 | 0.257 | k = 3 | 0.256 |
| | Davies–Boulding | | 0.914 | | 0.913 | | 0.888 | | 0.913 | | 0.887 | | 0.914 | | 1.257 | | 1.257 |
| Auto-encoder | Silhouette | k = 3 | 0.592 | k = 9 | 0.630 | k = 4 | 0.564 | k = 3 | 0.569 | k = 6 | 0.578 | k = 3 | 0.597 | k = 7 | 0.485 | k = 3 | 0.616 |
| | Davies–Boulding | | 0.466 | | 0.418 | | 0.578 | | 0.524 | | 0.476 | | 0.459 | | 0.603 | | 0.552 |
| GA | Silhouette | k = 3 | 0.584 | k = 4 | 0.386 | k = 3 | 0.438 | k = 6 | 0.383 | k = 5 | 0.375 | k = 3 | 0.401 | k = 4 | 0.385 | k = 3 | 0.398 |
| | Davies–Boulding | | 1.551 | | 1.447 | | 1.346 | | 1.521 | | 1.368 | | 1.192 | | 1.584 | | 1.239 |

**Table 7**
Results of the k-medoids first and second iteration.

| Methods | Metrics | First Iteration — Forecasting Intervals | | | | | | | | Second Iteration — Forecasting Intervals | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Week 1 | | Week 2 | | Week 3 | | Week 4 | | Week 5 | | Week 6 | | Week 7 | | Week 8 | |
| All variables | Silhouette | k = 6 | 0.248 | k = 3 | 0.246 | k = 5 | 0.249 | k = 3 | 0.246 | k = 5 | 0.250 | k = 3 | 0.247 | k = 3 | 0.252 | k = 3 | 0.252 |
| | Davies–Boulding | | 0.925 | | 1.280 | | 0.898 | | 1.280 | | 0.897 | | 1.279 | | 1.269 | | 1.269 |
| PCA | Silhouette | k = 6 | 0.255 | k = 6 | 0.254 | k = 5 | 0.255 | k = 6 | 0.254 | k = 5 | 0.255 | k = 6 | 0.253 | k = 3 | 0.257 | k = 3 | 0.256 |
| | Davies–Boulding | | 0.914 | | 0.913 | | 0.888 | | 0.913 | | 0.887 | | 0.914 | | 1.257 | | 1.257 |
| Auto-encoder | Silhouette | k = 3 | 0.592 | k = 9 | 0.630 | k = 4 | 0.564 | k = 3 | 0.569 | k = 6 | 0.578 | k = 3 | 0.597 | k = 7 | 0.485 | k = 3 | 0.616 |
| | Davies–Boulding | | 0.466 | | 0.418 | | 0.578 | | 0.524 | | 0.476 | | 0.459 | | 0.603 | | 0.552 |
| GA | Silhouette | k = 3 | 0.584 | k = 4 | 0.386 | k = 3 | 0.438 | k = 6 | 0.383 | k = 5 | 0.375 | k = 3 | 0.401 | k = 4 | 0.385 | k = 3 | 0.398 |
| | Davies–Boulding | | 1.551 | | 1.447 | | 1.346 | | 1.521 | | 1.368 | | 1.192 | | 1.584 | | 1.239 |

**Table 8**
Silhouette and Davies–Boulding indexes in the forecast interval, First and second iteration.

| k | Metrics | First Iteration — Forecasting Intervals | | | | Second Iteration — Forecasting Intervals | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Week 1 | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7 | Week 8 |
| 3 | Silhouette | 0.587 | 0.588 | 0.585 | 0.585 | 0.400 | 0.401 | 0.400 | 0.399 |
| | Davies–Boulding | 0.414 | 0.415 | 0.415 | 0.416 | 0.870 | 0.869 | 0.870 | 0.870 |
| 4 | Silhouette | 0.302 | 0.302 | 0.301 | 0.300 | 0.225 | 0.226 | 0.228 | 0.228 |
| | Davies–Boulding | 0.855 | 0.855 | 0.857 | 0.858 | 1.137 | 1.134 | 1.129 | 1.132 |
| 6 | Silhouette | 0.236 | 0.205 | 0.266 | 0.266 | 0.231 | 0.232 | 0.233 | 0.232 |
| | Davies–Boulding | 0.949 | 1.035 | 0.833 | 0.835 | 1.249 | 1.247 | 1.243 | 1.232 |

advantage of the use of the GA for the selection of descriptors, also, as a technique to select the value of k given the fitness function implemented, the Silhouette index. Therefore, the GA was run with these techniques for different k values (k = 1, 2, …, 10), and the k with the highest silhouette value was chosen.

Tables 6 and 7 present the results of the clustering process using the first intervals of prediction of the infected variable like input (week 1 to week 4) and the second intervals of prediction of the infected variable like input (week 5 to week 8). In these Tables, it is possible to observe the methods used for the dimensionality reduction, and the quality metrics (the silhouette and Davies–Boulding indexes). Both the suggested k-value in each interval and the values of the already mentioned metrics are presented.

In the first experiment, it is evident that the dimensionality reduction method with the best results in the different prediction intervals was auto-encoder, since the values for the silhouette are the closest to 1, and the values for Davies are the closest to 0. The order of the techniques in each prediction interval, according to the results, are: auto-encoder, GA and PCA. Also, it is observed that clustering by previously doing a dimensionality reduction is always better than clustering by having all the variables. On the other hand, with the k-means technique, the best results were obtained, which are always above the k-medoids results (see Tables 6 and 7).

Additionally, we carry out an analysis of the k values with the GA. The best results are shown in Table 8. GA suggest the utilization of 3, 4 and 6 groups in the different time intervals, according to the values of the quality metrics.

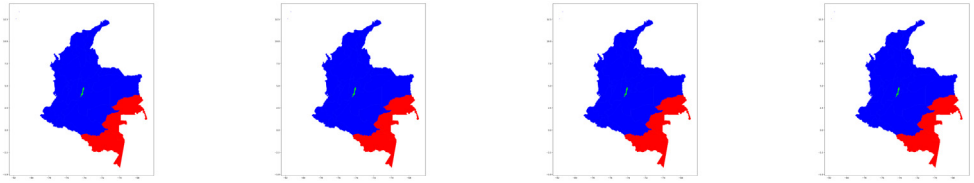
In the second experiment is considered the second interval of prediction of infected cases of COVID-19 (second iteration) during the clustering process. The results for k-means and k-medoids, of the silhouette and Davies–Boulding indices, are shown in Tables 6 and 7.

In this second experiment, it is still shown that using an auto-encoder for the dimensionality reduction and k-means is better since the silhouettes and the Davies–Boulding indexes for each week are the closest to the values 1 and 0, respectively.

Again, we carry out an analysis of the k values for each week with the GA. The best results are shown in Table 8. In this case, GA suggest the utilization of 3, 4 and 5 groups in the different time intervals, according to the values of the quality metrics.

The silhouette indices, for the set of descriptors selected with GA, with k = 3 in each prediction week, are quite close to 1, which indicates that the 3 groups built are very well-formed. In the same way, Davies–Boulding indices close to 0 indicate that there are very small distances between the elements that make up the groups, so it can be said that there is high intra-group similarity and high inter-group differences; that is, the departments that make up each cluster, due to the characteristics that identify them, are very similar to each other.

According to the above results, a good number of groups (value of k) is 3. The values of the silhouette and Davies–Boulding indices support this. In the next section, k = 3 will be used for the analysis. In addition, the best reduction technique was autoencoder, but as is not easy to interpret the centroids of the clusters formed by this method, GA is used as the reduction technique



(a) week 1 (b) week 2 (c) week 3 (d) week 4

**Fig. 4.** Evolution of clusters and Departments — First Iteration.

in the next section since the GA results are very close. Among the most commonly used clustering techniques are k-means and k-medoids, the former takes into account the average behavior of individuals, while the latter is not affected by individuals with atypical behavior, so it was decided to implement and evaluate which one yielded better results. Thus, the best clustering technique was k-means. Therefore, in the next section, the k-means method is used for clustering. Finally, according to the results of the prediction model and its quality in the clustering process, this process should be recalculated (retrained) every 4 weeks to ensure the creation of good clusters.

## 4. Evolution analysis

With the results of the previous experiments, we can set up experiments to analyze the evolution of the behavior of the departments, taking into account the cluster in which it is assigned each week. Taking into account the interest of interpreting the groups formed according to the variables that characterize it, it is used GA as a reduction technique. Thus, to carefully analyze the evolution (week 1, week 2, week 3, week 4, week 5, week 6, week 7, week 8) of the departments using the clusters built by the GA, the centroid of each cluster is studied, and from it, we find the characteristics (pattern) that define them.

### 4.1. Results

Fig. 4 shows the 3 groups formed in the 4 weeks, and the departments that comprise them. It is evident that despite the changes that occurred in the variables in these weeks, these were not significant for the departments to change groups, which means that during these 4 weeks the behavior of each department was very similar.

Now, we will proceed to identify the characteristics of each cluster taking into account its centroids. Cluster 3 is only composed of the capital of the country, Bogotá, which is located in the central part of the country and has an average temperature of 14℃. This cluster is characterized by a high number of infected cases of COVID-19 predicted, a high number of recovered and dead, and a high population density. In addition, it is characterized by medium-low indicators of early childhood care services, economic dependency rates, critical overcrowding, school lags, uninsured health, child labor and informal work. However, a low rate of informal labor does not guarantee that the poverty index will be close to 0. Particularly, in the capital, there is a medium-high poverty index. The percentage of people in the capital between the ages of 15 and 64 is moderately high, and the number of men in this age range is higher, which is another distinctive characteristic of this cluster. On the other hand, Bogotá is characterized by low indicators of illiteracy, low educational achievement, school non-attendance, lack of access to a water source and inadequate excreta disposal.

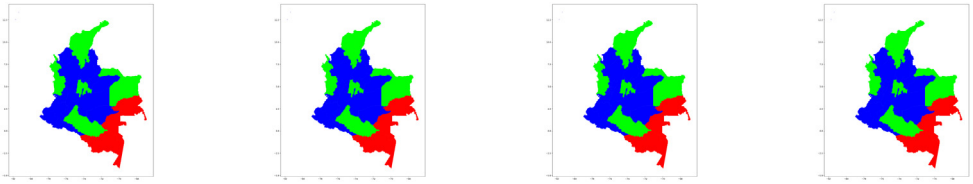
The second cluster is made up of the departments of Amazonas, Guainía and Vaupés. These departments are located in the southeast of the country, and all border Brazil to the southeast. They are located near the equatorial line, and because of this, the temperature during the day is very high, averaging 31.6 °C. This group of departments is characterized by a medium infected number of cases of COVID-19 predicted, recovered and dead, and the population density is low. On the other hand, it presents high rates of poverty, economic dependency and critical overcrowding, inadequate disposal of excreta, non-attendance at school, school backwardness, no early childhood care services, no access to improved water sources and a high percentage of women. Other characteristics present in this cluster are that they have low indicators of literacy, low educational achievement, no health insurance, and a high rate of child labor and informal work. They also have a low percentage of people between 15 and 64 years old, particularly men.

The last cluster, the largest, where the remaining 29 departments of Colombia are located, is a very particular group, given that all of its characteristics are either moderately high or moderately low. This cluster has a positive number of predicted cases of infected by COVID-19, recovered and dead, which are moderately low. In addition, the population density is moderately low, as well as the indices of multidimensional poverty, of early childhood care services, of economic dependency rates, of critical overcrowding, of inadequate excreta disposal, of non-attendance at school, of school backwardness, of no access to an improved water source, of no health insurance, and finally, the percentage of women in the department is moderately low. On the other hand, there are characteristics with moderately high rates, such as illiteracy, low educational achievement, child labor and informal work. In addition, the percentage of people between the ages of 15 and 64 is moderately high; in particular, there are more men than women in this age range.

In the second iteration, the clustering model was retrained, despite the number of groups formed being the same, but there are variants in the formation of the clusters that can be seen in Fig. 5.

### 4.2. Findings

The change in the behavior of some departments in the second iteration is very noticeable, which can be determined by the centroids of each cluster. It can be seen that the cluster that made up the departments of Amazonas, Guainía and Vaupés during the first iteration, in the second iteration continues to be made up of the same departments. These departments are identified by having several infected cases of COVID-19 predicted, which is moderately low, as the indicators of no access to an improved water source, the percentage of people between the ages of 15 and 64, and the number of days where it rains. However, this cluster is also characterized by a high level of multidimensional poverty, critical overcrowding, lack of school attendance, use of inadequate outdoor flooring materials, a high rate of economic dependency and of informal work. These departments also have medium-high rates of illiteracy, of early childhood care services,



(a) week 5 (b) week 6 (c) week 7 (d) week 8

**Fig. 5.** Evolution of clusters and Departments — Second Iteration.

of inadequate elimination of excreta, of non-attendance at school, and of school lags, no health insurance, and a population in miserable conditions, with a medium-high percentage of women in the departments.

The group that initially had the capital city (Bogotá) has grown in size, and now also includes the departments of Antioquia, Atlántico, Boyacá, Caldas, Casanare, Cauca, Córdoba, Huila, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, San Andrés and Providencia, Santander, Tolima, and Valle del Cauca. All of these departments went from having moderately low infected cases of COVID 19 to moderately high, except for Bogotá, which went from a high to a moderately high indicator. In addition, they have medium-high percentages of people between the ages of 15 and 64 and many rainy days. Nevertheless, the indexes of poverty, illiteracy, services for early childhood care, economic dependence rate, critical overcrowding, inadequate elimination of excreta, non-attendance at school, inadequate material for exterior floors, school backwardness, lack of access to an improved water source, lack of health insurance, and population in conditions of misery, are medium-low. The percentage of people and women in these departments is moderately low.

The previous cluster, with almost all the departments, is now made up of only Arauca, Bolivar, Caquetá, Cesar, Chocó, Cundinamarca, Guaviare, la Guajira, Magdalena, Sucre and Vichada. This cluster is identified by its low levels of multidimensional poverty, economic dependence, critical overcrowding, lack of school attendance, informal work, and low percentage of people between the ages of 15 and 64. They also present several infected cases of COVID 19 predicted, and an index of inadequate exterior flooring material, which is moderately low. It is also characterized by some very rainy days and a moderately high indicator of illiteracy. The percentage of people, especially women, is high; there is a high rate of population in miserable conditions. The indexes of services for early childhood care, of critical overcrowding, of inadequate elimination of excrement, of non-attendance at school, of school lags, of no access to improved water sources, and of no health insurance, are high.

### 4.3. Most relevant observations

The observation time period to obtain the above findings in the case study was two months, which made possible to study the behavior of our approach to analyze the evolution of socio-economic variables due to the effect of COVID-19.

A first outstanding aspect is that the ability of the approach to determine the change in the groups over time, and of the members of the group, is clear. So, there is a double dynamic in this process, one in which the groups are redefined by new member regions, and consequently, some that leave; and on the other hand, how their patterns (centroids) are changing due to the effect of COVID-19 on socio-economic variables. The socio-economic variables that are most relevant in the characterization of the groups that are formed over time, due to the effect of COVID-19, are the level of poverty (and everything that involves such as overcrowding, houses in poor condition, etc.), education, public services (water, electricity, etc.) and health.

The groups that are formed have clear socio-economic patterns in these variables, being able to see groups with regions in very poor conditions (critical overcrowding, lack of school attendance, no health services, and in general, miserable conditions), and a medium incidence of COVID-19 (Amazonas, Guainía and Vaupés). This may be because these regions are quite isolated. On the other hand, the groups with a moderate incidence of COVID-19 are regions with higher socio-economic levels, ranging from moderately high to moderately low cases of COVID-19 infection. They are regions with better housing conditions, more opportunities for women, lower levels of poverty, and in general, better public services, economic, educational and health status for their inhabitants. It is the same case of Bogotá, which went from a high to moderately high indicator (improving), because somehow these socio-economic indices were less affected by the special attention given by the national government to these regions. The moderate incidences of COVID-19 in these regions can be understood because they are the most populated areas of the country, so the transmission of the disease is easier.

**Table 9**
Quality measures of the prediction models.

| | Prediction Intervals | | | |
|---|---|---|---|---|
| | Week 1 | Week 2 | Week 3 | Week 4 |
| MAPE | 0.366 | 0.350 | 0.346 | 0.361 |
| $R^2$ | 0.701 | 0.711 | 0.698 | 0.696 |

### 4.4. Evaluation of the generality of our approach

In this section, we carry out the same experiments that were done for the construction of the prediction and clustering models in the previous section, but using the datasets updated for 2021, which are the basis of our hybrid model for the dynamic evaluation of the socioeconomic impact of COVID-19. In this sense, the same previously developed experimental protocols were considered.

The results shown in Tables 9 and 10 reflect the quality of those models with these new data. We see that in general, the quality of the models is maintained, the values of the metrics do not change, they are maintained.

Starting from these partial results, where the individual results of these two models appear, which are integrated into the hybrid model for the socioeconomic analysis, it is natural to suppose that an interpretation of the dynamic behavior of the departments is completely possible with the current dataset.

## 5. Conclusions

Machine learning methods can provide reliable systems to help the development of public policies. We have proposed a hybrid approach to analyze the socioeconomic impact in the

**Table 10**
Quality metrics of the clustering models.

| k | Metrics | Forecasting Intervals | | | |
|---|---|---|---|---|---|
| | | Week 1 | Week 2 | Week 3 | Week 4 |
| 3 | Silhouette | 0.591 | 0.601 | 0.605 | 0.599 |
| | Davies–Boulding | 0.487 | 0.481 | 0.456 | 0.465 |
| 4 | Silhouette | 0.291 | 0.301 | 0.301 | 0.299 |
| | Davies–Boulding | 0.811 | 0.815 | 0.827 | 0.811 |
| 6 | Silhouette | 0.206 | 0.215 | 0.217 | 0.201 |
| | Davies–Boulding | 0.898 | 0.944 | 0.945 | 0.939 |

regions of a country due to COVID-19. Our approach allows monitoring the socioeconomic behavior of the regions/departments of a country. To do this, our approach initially predicts the infected cases, and together with other context variables (climate, economics and socials, among others), determines the current socioeconomic situation of a region. For that, our approach carries out a clustering process of the regions, in order to determine similar ones. With the centroid (pattern) of each group, it is possible to define its socioeconomic characteristics, a fundamental input for those who make decisions.

In this work, we have tested our approach for generating groups of regions in Colombia with similar social, economic and health impacts due to the COVID-19 pandemic. The results obtained were satisfactory, the system naturally suggests three clusters of departments, using social, economic, geographic and demographic variables. Some of these variables describe characteristics like the levels of poverty, economic dependence, overcrowding, school attendance, informal work, among others. Also, our approach considers the behavior of the pandemic, particularly, the infected cases of COVID 19 predicted.

Among the most important limitations of the approach and experiments is that it is a case study based on the behavior of the pandemic in Colombia. So, it is necessary to study its behavior in other countries. This implies adjusting the parameters of the models according to the dynamics of the country of interest, and particularly, adapting them to their data. Another limitation is that according to the behavior of the pandemic, the models must be retrained and parameterized again every certain period. The determination of these retraining periods is not clear, and should be a factor to optimize according to each region, which is being studied for the Colombian case.

Important ideas arise for future works, for example, automatic analysis of the cluster centroids, to gain some autonomous interpretability on the clustering process. Particularly, the idea is to analyze the evolution of knowledge, from the patterns defined by the centroids [24]. Another downstream task is to identify patterns related to the vulnerable population, or specifically, the lockdowns economic impact, adding other variables linked to this decision. On the other hand, our hybrid model is composed of two types of tasks, one for prediction and the other for clustering. A future work should make a more in-depth analysis of the possible techniques to consider in each of them, such as, for example, in the case of prediction, the use of simple regression models, ARIMA, LSTM, among others. Also, the utilization of other feature engineering techniques like the SHAP (https://github.com/slundberg/shap) algorithm will be considered.

Other future works will study the inclusion of concepts like "autonomic cycles of data analytic tasks" [25,26] to generate a self-management environment for monitoring the socio-economic impact in a country due to COVID-19, which could be extended to other diseases or aspects to be evaluated. Also, the addition of contextual ontological information during the analysis process will be defined to introduce contextual reasoning [27]. In addition, our approach will be tested with datasets from other countries on different continents to evaluate its generality and effectiveness. Finally, we hope that this work will encourage more applications of machine learning related to the development of automatic public policies.

## CRediT authorship contribution statement

**Yullys Quintero:** Conceptualization, Methodology, Formal analysis, Simulation, Writing. **Douglas Ardila:** Formal analysis, Simulation. **Jose Aguilar:** Conceptualization, Methodology, Formal analysis, Writing, Funding acquisition. **Santiago Cortes:** Formal analysis, Simulation, Writing.

## 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 will be made available on request.

## Acknowledgments

This research has been carried out in the framework of the project "Plataforma web para la recolección de datos, visualización, análisis, predicción y evaluación de estrategias de control de la enfermedad producida por SARS-CoV-2 mediante herramientas de modelación matemática, simulación e inteligencia artificial" which has been funded by the program MinCiencias (Covid-19 2020) of MinCiencias-Colombia and EAFIT University, Colombia trough the agreement number 1216101576695.

## References

[1] Hien Lau, Veria Khosrawipour, Piotr Kocbach, Agata Mikolajczyk, Justyna Schubert, Jacek Bania, Tanja Khosrawipour, The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China, J. Travel Med. 27 (3) (2020) taaa037.
[2] Giovanni Dell'Ariccia, Paolo Mauro, Antonio Spilimbergo, Jeromin Zettelmeyer, Economic policies for the COVID-19 war, IMF Blog 1 (2020).
[3] Shuo Feng, Chen Shen, Nan Xia, Wei Song, Mengzhen Fan, Benjamin J Cowling, Rational use of face masks in the COVID-19 pandemic, Lancet Respir. Med. 8 (5) (2020) 434–436.
[4] Zohair Malki, El-Sayed Atlam, Aboul Ella Hassanien, Guesh Dagnew, Mostafa A. Elhosseini, Ibrahim Gad, Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches, Chaos, Solitons Fractal 138 (2020) 110137.
[5] Wang Pu, Zheng Xiem, Li Ju, Zhu Ban, Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques, Chaos Solitons Fractals 139 (2020) 110058.
[6] Lamiaa A. Amar, Ashraf A. Taha, Marwa Y. Mohamed, Prediction of the final size for COVID-19 epidemic using machine learning: A case study of Egypt, Infect. Dis. Model. 5 (2020) 622–634.
[7] 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 Fractals 135 (2020) 109853.
[8] Rodrigo M. Carrillo-Larco, Manuel Castillo-Cara, Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach, Wellcome Open Res. 5 (56) (2020) 56.
[9] Md Arafatur Rahman, Nafees Zaman, A Taufiq Asyhari, Fadi Al-Turjman, Md Zakirul Alam Bhuiyan, MF Zolkipli, Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices, Sustainable Cities Soc. 62 (2020) 102372.
[10] Alvin Wei Ze Chew, Ying Wang, Limao Zhang, Correlating dynamic climate conditions and socioeconomic-governmental factors to spatiotemporal spread of COVID-19 via semantic segmentation deep learning analysis, Sustainable Cities Soc. 75 (2021) 103231.
[11] Y. Zhou, L. Feng, X. Zhang, Y. Wang, S. Wang, T. Wu, Spatiotemporal patterns of the COVID-19 control measures impact on industrial production in Wuhan using time-series earth observation data, Sustainable Cities Soc. 75 (2021).

[12] Viezzer J., Biondi D., The influence of urban, socio-economic, and eco-environmental aspects on COVID-19 cases, deaths and mortality: A multi-city case in the Atlantic Forest, Brazil, Sustainable Cities Soc. 69 (2021).
[13] Guo C, Bo Y, Lin C, Li HB, Zeng Y, Zhang Y, Hossain MS, Chan JWM, Yeung DW, Kwok KO, Wong SYS, Lau AKH, Lao XQ, Meteorological factors and COVID-19 incidence in 190 countries: An observational study, Sci. Total Environ. (2021).
[14] Meenu Gupta, Rachna Jain, Soham Taneja, Gopal Chaudhary, Manju Khari, Elena Verdú, Real-time measurement of the uncertain epidemiological appearances of COVID-19 infections, Appl. Softw. Comput. (2021).
[15] Yullis Quintero, Douglas Ardila, Edgar Camargo, Francklin Rivas, Jose Aguilar, Machine learning models for the prediction of the SEIRD variables for the COVID-19 pandemic based on a deep dependence analysis of variables, Comput. Biol. Med. 134 (2021) 104500.
[16] Edgar Camargo, Jose Aguilar, Yullis Quintero, Douglas Ardila, Francklin Rivas, An incremental learning approach to prediction models of SEIRD variables in the context of the COVID-19 pandemic, Health Technol. 12 (2022) 867—877.
[17] Slawek Smyl, A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting, Int. J. Forecast. 36 (1) (2020) 75–85.
[18] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.
[19] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, Wavenet: A generative model for raw audio, 2016, arXiv preprint arXiv:1609.03499.
[20] Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei, Junjun Xiong, Shuicheng Yan, Deep learning with s-shaped rectified linear activation units, 2015, arXiv preprint arXiv:1512.07030.
[21] Jerome Friedman, Trevor Hastie, Robert Tibshirani, The Elements of Statistical Learning. Vol. 1, (10) Springer Series in Statistics, New York, 2001.
[22] F. Pacheco, C. Rangel, J. Aguilar, M. Cerrada, J. Altamiranda, Methodological framework for data processing based on the data science paradigm, in: 2014 XL Latin American Computing Conference, CLEI, 2014, pp. 1–12, http://dx.doi.org/10.1109/CLEI.2014.6965184.
[23] Rosa Lletı, M. Cruz Ortiz, Luis A. Sarabia, M Sagrario Sánchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes, Anal. Chim. Acta 515 (1) (2004) 87–100.
[24] Jose Aguilar, Juan Salazar, Julian Monsalve-Pulido, Edwin Montoya, Henry Velasco., Traceability analysis of patterns using clustering techniques, in: H. Arabnia, K. Ferens, D. de la Fuente, E. Kozerenko, J. Olivas, Fernando G. Tinetti (Eds.), Advances in Artificial Intelligence and Applied Cognitive Computing, Springer International Publishing, Cham, 2021.
[25] J. Aguilar, O. Buendia, A. Pinto, J. Gutiérrez, Social learning analytics for determining learning styles in a smart classroom, Interact. Learn. Environ. 30 (2) (2022) 245–261.
[26] Juan Vizcarrondo, Jose Aguilar, Ernesto Exposito, Audine Subias, MAPE-K as a service-oriented architecture, IEEE Latin Am. Trans. 15 (6) (2017) 1163–1175, http://dx.doi.org/10.1109/TLA.2017.7932705.
[27] Jose Aguilar, Marxjhony Jerez, Ernesto Exposito, Thierry Villemur, CARMi-CLOC: Context awareness middleware in cloud computing, in: 2015 Latin American Computing Conference, CLEI, 2015, pp. 1–10, http://dx.doi.org/10.1109/CLEI.2015.7360013.