



**UNIVERSIDAD POLITÉCNICA SALESIANA**  
**SEDE GUAYAQUIL**

**TRABAJO DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE:**

**INGENIERO DE SISTEMAS**

**CARRERA:**

**INGENIERÍA DE SISTEMAS**

**TEMA:**

**“MODELO DE ANÁLISIS DE LAS SERIES DE TIEMPO DE CONTAGIOS DE  
COVID-19 BASADO EN EL AGRUPAMIENTO JERÁRQUICO”**

**AUTOR:**

**Gisella Belén Montero Castillo**

**TUTOR:**

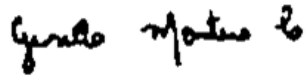
**Msg. Mikel Yelandi Leyva Vazquez**

**Julio 2021**

**GUAYAQUIL-ECUADOR**

## DECLARATORIA DE RESPONSABILIDAD

Yo, **Gisella Belén Montero Castillo**, declaro que los conceptos y análisis desarrollados y las conclusiones del presente trabajo son de exclusiva responsabilidad del/los autor/es.



**Autor: Gisella Montero**



**Tutor: Maykel Leyva**

# COVID-19 time series analysis model based on hierarchical clustering.

Maikel Yelandi Leyva Vazquez<sup>1</sup>, Gisella Belén Montero Castillo<sup>2</sup>

<sup>1</sup> Computer Science Department, Universidad Politécnica Salesiana, Guayaquil, Ecuador  
mleyvaz@ups.edu.ec, gmonteroc@est.ups.edu.ec

**Abstract.** Currently, worldwide the COVID-19 virus has generated impacts on all human development activities. There is a worldwide effort by the research community to explore the impact of the pandemic based on available data. Many different disciplines are trying to find solutions and drive strategies for a wide variety of very different crucial problems, including artificial intelligence. This study aims to cluster the various countries describing the course of the COVID-19 outbreak using hierarchical clustering. This paper presents a proposal of innovative development using visual programming to analyze the data acquired from John Hopkin University and the World Bank. The result shows a correlation between the number of physicians and the number of cases. Hierarchical clustering is performed a group of similar countries is obtained. This study will be of significant importance for showing the similarities and differences among countries in terms of cases.

**Keywords:** COVID-19, artificial intelligence, clustering, hierarchical clustering.

## 1 Introduction

Currently, worldwide the COVID-19 virus has generated impacts on all human development activities[1], [2]. There is a worldwide effort by the research community to explore the impact of the pandemic based on available data. Many different disciplines are trying to find solutions and drive strategies for a wide variety of very different crucial problems, including artificial intelligence. However, little attention has been paid to COVID - 19 infection time series analysis, making public health decision making unstable and inconsistent.

This paper aims to develop a novel analysis resulting in countries' clustering concerning cases based on the John Hopkins dataset[3] and World Bank human development index dataset [4]. It allowed us to handle a new line of cluster analysis adapted to the request to compare the various COVID-19time series of different countries developed in the orange visual programming paradigm for data science[5].

This work contributes to the development of artificial intelligence, the techniques used to analyze the data acquired from John Hopkins University allow the classification of units of analysis that are in groups or clusters not only recognizing quantitative but also qualitative variables, based on the hierarchical grouping of

epidemiological data from different countries using visual programming that contributes to decision making. Additionally, This study will be of great importance for showing the differences among countries in the evolution of the epidemic. Proper use of these data will help decision makers to take precautions regarding COVID-19.

The results presented from the cluster could be useful for a variety of different policymakers, such as physicians, health sector managers, economic/financial experts, politicians and others.

This paper continues as follows. It begins with a preliminary section where concepts of hierarchical clustering and visual programming are discussed to develop the content of this article. The section on material and methods is devoted to methodology and dataset used. Next, the results section is where analysis of the data using the orange data mining tool is shown. Finally, the paper ends with the sections of discussion and conclusions.

## 2 Material and methods

The study was conducted using Orange software with clustering methods included in the tool. The difference of clusters is calculated using Euclidian distance measure[13]

Dataset used were the John Hopkins dataset[3] and World Bank human development index dataset [4]. The initial data used here are obtained from the specific site created by John Hopkins University on Github ([https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse\\_covid\\_19\\_data/csse\\_covid\\_19\\_time\\_series/time\\_series\\_covid19\\_confirmed\\_global.csv](https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv)) through orange file component.



**Figure 1.** Case of Covid reported as 4/04/21.

Data is normalized to get active cases with respect to the total population per country. In the former, the spread of the pandemic is shown for each country throughout the whole period from 1/22/20 through 4/21/21

The Human Development Index (HDI) dataset summarises achievement in key dimensions related to human development: a long and healthy life, being knowledgeable and having a proper living standard. The HDI is the geometric mean of normalized indices for each of three dimensions. The HDI simplifies and captures important part of what human development entails[14]. Comparing this data with the World Bank's data allows us to relate the epidemiological data to data about particular countries.

## 3 Preliminaries

This section contains the main definitions necessary to develop the theory proposed in this paper.

### 3.1 Hierarchical Cluster

The hierarchical clustering method starts by obtaining the initial cluster to form a new one or separating an existing one to give rise to two others to maximize a measure of similarity or minimize some distance. The starting point is as many groups as individuals in the study, and they are grouped until all the cases are in the same group[6].

In this case, the Euclidean distance measure and the hierarchical procedure will be used. This measure makes it possible to measure or evaluate the units used to be worked on[7].

$$d_E(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

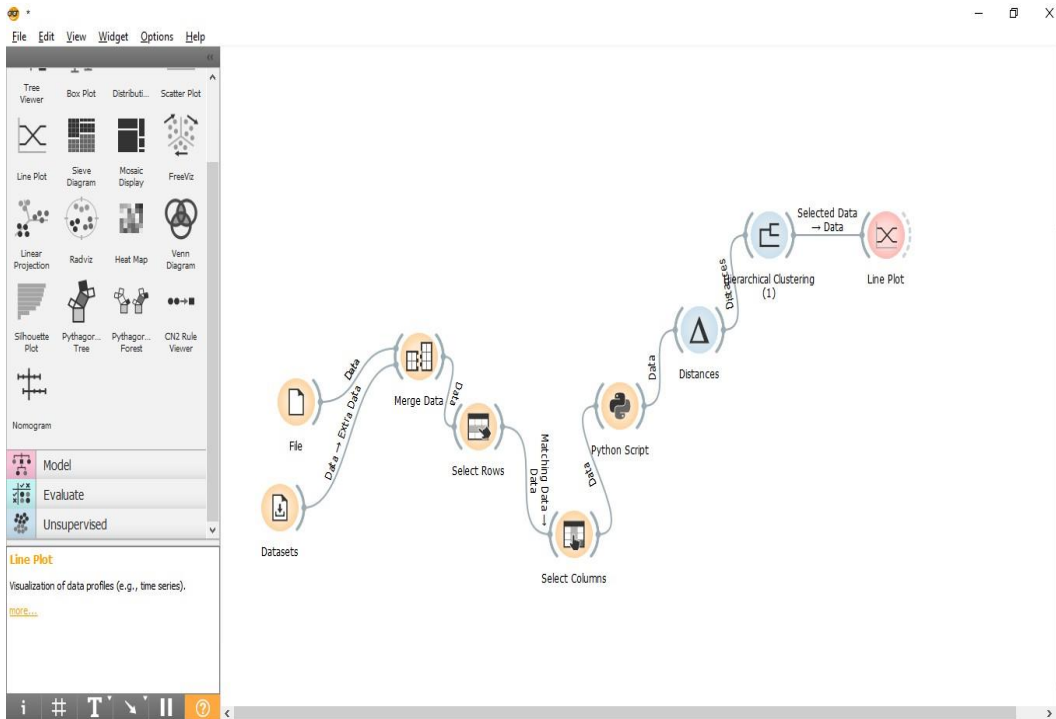
(1)

Agglomerative or ascending and divisive or descending. Agglomerative algorithms, by using some criterion, group units of analysis at each step until a conglomerate that encompasses the totality is reached. Divisive algorithms start from the total set of elements considered a conglomerate. According to some criterion, divide the group into smaller groups, reaching the last stage of the procedure, to consider each element of the initial group as the simplest conglomerate with maximum homogeneity.

Cluster Analysis is often referred to as Cluster Analysis; it is a multivariate statistical technique that seeks to group elements (or variables) to achieve the maximum homogeneity in each group and the most significant difference between groups. It is a multivariate statistical method of automatic data classification. Starting from a table of cases-variables, it tries to place the cases (individuals) in homogeneous groups, conglomerates or clusters, not known beforehand, but suggested by the data, so that individuals that can be considered similar are assigned to same clusters. In contrast, dissimilar individuals are located in different clusters. Time series hierarchical clustering of Covid-19 Data[8] has received close litter attention; some work has included data analysis but lacks integration with external data sources [9], [10].

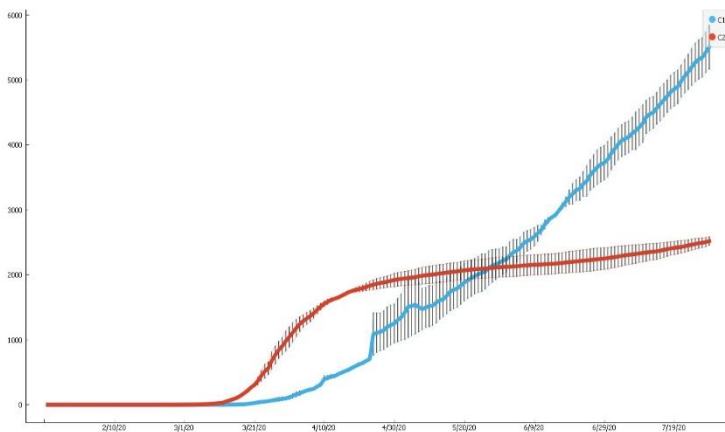
### 3.2 Visual Programming

Orange is a component-based visual programming software package for data visualization, machine learning, data mining and data analytics [5]. Orange components are called widgets and range from data visualization, subset selection and preprocessing to empirical evaluation of learning algorithms and predictive modelling [11], [12].



**Figure 2.** Visual Programming

Visual programming is implemented through an interface in which workflows (Figure 1) are created by linking predefined or user-designed widgets, while advanced users can use Orange as a Python library for data manipulation and widget alteration. (software) -



**Figure 3.** Line Plot component. Both data sources were merged to discard unnecessary information.

As an example of the flexibility and capacity of the tool, Figure 2 shows the Line Plot component for analyzing the evolution of the country curve for different clusters. Orange scripting library is also a part of its visual programming platform with graphical user interface components for interactive data visualization giving more flexibility to the tool[5].

## 4 Results

Pipeline development is shown in figure 3. At the start of the sequence, the two data sources are merged and remove from countries with unknown or zero population. The Feature Constructor widget is used to input a

formula to compute new data columns. Cases per million as the new column's name is incorporated into the data.

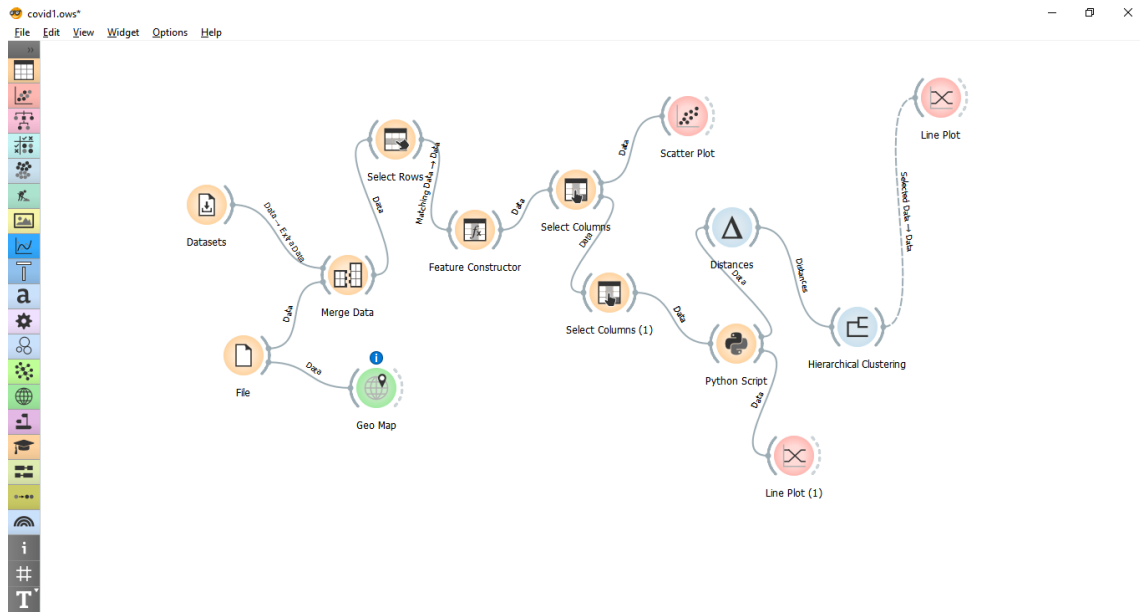


Figure 4. Pipeline

We related the epidemiological data to data about particular countries, for example, the number of physicians and cases per million.

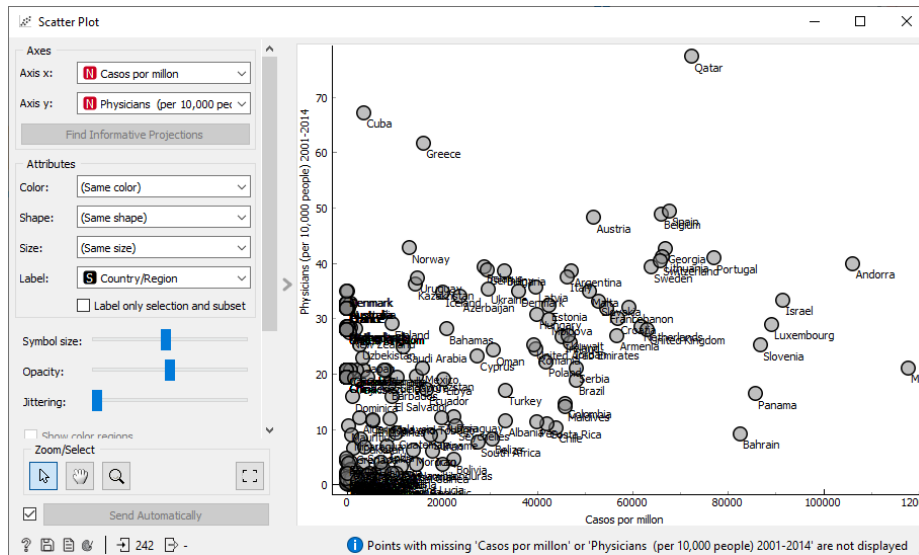
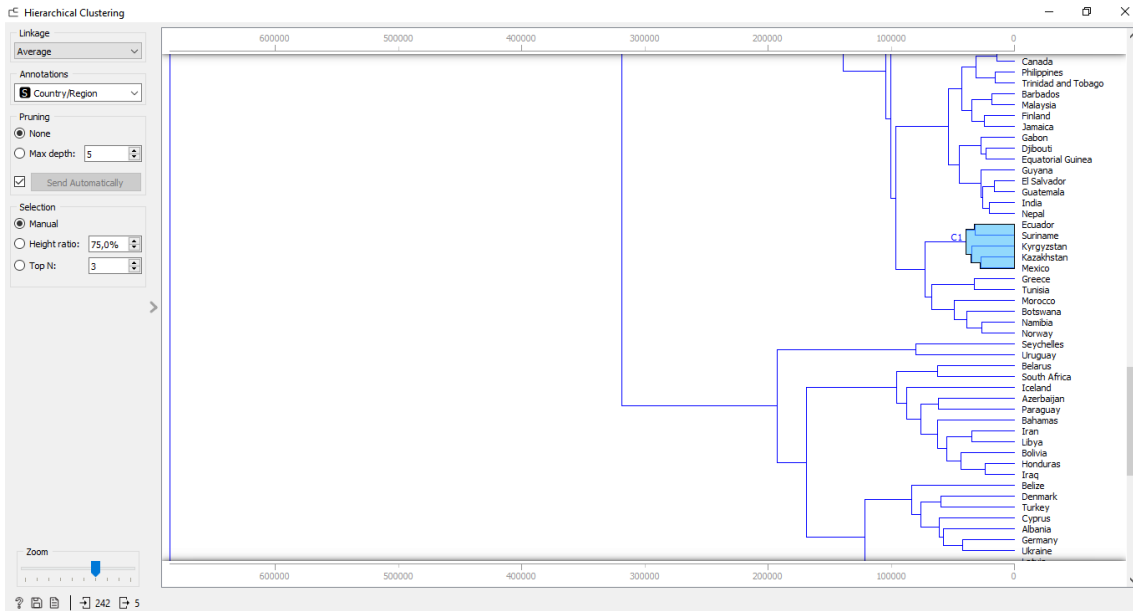


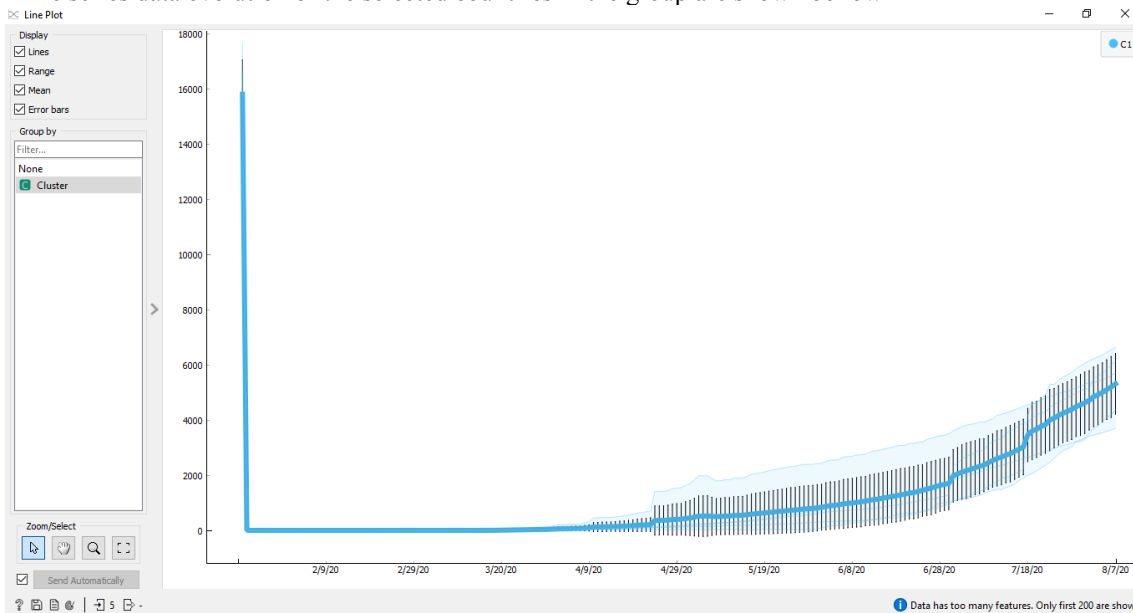
Figure 5. Number of physicians and cases per million.

Visually a correlation arises on the number of cases by millions with the number of physicians. Clustering concerning active cases means that the elements of these clusters have similar time evolution of the active cases, which means that they have faced comparable stresses to the health system. For example, Ecuador, Suriname, Kyrgistan, Kasagastan and Mexico form a group (Figure 6).



**Figure 6.** Dendrogram

Time series data evolution of the selected countries in the group are shown below



**Figure 7.** Time sere data of selected countries

A cluster analysis was used for grouping countries in terms of cases numbers result of utility analyzing the evolution of pandemic; additionally, integration with HDI data allows for enhanced analysis.

## 5 Conclusions

Currently, worldwide the COVID-19 virus has generated impacts on all human development activities. There is a worldwide effort by the research community to explore the impact of the pandemic based on available data. In the present paper, a pipeline using visual programming is developed, resulting in consistent and reasonable clustering. The code was implemented in Orange Data Mining Tool. Python scripting library is used, giving



more flexibility to the pipeline. The overall algorithm follows the concept of hierarchical clustering. Additionally, We related the epidemiological data to data about particular countries. This study will be of significant importance when showing the differences among countries in terms of epidemy evolution in time. Appropriate use of these data will help states take precautions regarding COVID-19. As future work new method for time series prediction based on deep learning will be introduced.

## References

- [1] M. Sigala, "Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research," *J. Bus. Res.*, 2020, doi: 10.1016/j.jbusres.2020.06.015.
- [2] UNESCO, "COVID-19 Impact on Education," *UNESCO Inst. Stat. data*, 2020.
- [3] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons and Fractals*, 2020, doi: 10.1016/j.chaos.2020.109864.
- [4] B. Kabadayı, "Human Development and Trade Openness: A Case Study on Developing Countries," *Adv. Manag. Appl. Econ.*, 2013.
- [5] J. Demšar *et al.*, "Orange: Data mining toolbox in python," *J. Mach. Learn. Res.*, 2013.
- [6] S. Chakraborty, D. Paul, and S. Das, "Hierarchical clustering with optimal transport," *Stat. Probab. Lett.*, 2020, doi: 10.1016/j.spl.2020.108781.
- [7] R. Loochach and K. Garg, "Effect of Distance Functions on Simple K-means Clustering Algorithm," *Int. J. Comput. Appl.*, 2012, doi: 10.5120/7629-0698.
- [8] V. Papastefanopoulos, P. Linardatos, and S. Kotsiantis, "Covid-19: A comparison of time series methods to forecast percentage of active cases per population," *Appl. Sci.*, vol. 10, no. 11, p. 3880, 2020.
- [9] V. Zariakas, S. G. Pouloupoulos, Z. Gareiou, and E. Zervas, "Clustering analysis of countries using the COVID-19 cases dataset," *Data Br.*, vol. 31, p. 105787, 2020.
- [10] O. PASIN and T. PASIN, "Clustering of countries in terms of deaths and cases of COVID-19," *J. Heal. Soc. Sci.*, vol. 5, no. 4, pp. 587–594, 2020.
- [11] B. D. McCullough, T. Mokfi, and M. Almaenejad, "On the accuracy of linear regression routines in some data mining packages," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2019, doi: 10.1002/widm.1279.
- [12] D. J. Shastri, "Machine learning for non-programmers," 2020, doi: 10.1145/3334480.3375051.
- [13] M. R. Berthold and F. Höppner, "On clustering time series using euclidean distance and pearson correlation," *arXiv Prepr. arXiv1601.02213*, 2016.
- [14] D. Dumuid *et al.*, "Human development index, children's health-related quality of life and movement behaviors: a compositional data analysis," *Qual. Life Res.*, 2018, doi:

10.1007/s11136-018-1791-x.