Revista Colombiana de Estadística - Applied Statistics
January 2024, volume 47, issue 1, pp. 25–35
http://doi.org/10.15446/rce.v47n1.108448

Classification of Territorial Entities of Colombia According to the Epidemiological Curve of Sars-Cov2 between 03-06-2020 and 02-04-2021

Clasificación de entidades territoriales de Colombia de acuerdo con la curva epidemiológica de Sars-Cov2 entre el 06-03-2020 y 04-02-2021

Lina Angélica Buitrago Reyes[a],
Sergio Alejandro Calderón Villanueva[b], Isabella Castillo Soria[c]

Department of Statistics, Faculty of Science, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Classify the departments of Colombia and their capitals, according to the behavior of the incidence of Sars-Cov2, between March 6, 2020 and February 4, 2021. The information on daily cases was obtained from the website of the National Institute of Health (INS), the population estimate for each geographic unit was obtained from the population estimated by the National Administrative Department of Statistics (DANE) for 2020. The clusters obtained for both capitals and departments were obtained by non-hierarchical classification methods. Four groups were obtained for both, capitals and departments: the coast zone, the central zone, the eastern zone and the Amazon. In most cases the classification of the capitals coincided with that of the department.

The classification obtained by the k-medoid method, using the Euclidean distance, proposed groups that coincide with different epidemiological behaviors between groups and similar ones within groups, therefore it becomes a useful statistical tool for public health.

Key words**:** Classification; Capitals; Departments; Sars-Cov2 time series.

Resumen

Clasificar los departamentos de Colombia y sus capitales, de acuerdo con el comportamiento de la incidencia de Sars-Cov2, entre el 6 de marzo de 2020 y el 4 de febrero de 2021. La información de los casos diarios se

[a] Ph.D(c). E-mail: labuitragor@unal.edu.co
[b] Ph.D. E-mail: sacalderonv@unal.edu.co
[c] B.Sc. E-mail: icastillos@unal.edu.co

obtuvo de la página del Instituto Nacional de Salud (INS), la estimación de la población para cada unidad geográfica se obtuvo de la población estimada por el Departamento Administrativo Nacional de Estadísticas (DANE) para el año 2020. Los conglomerados obtenidos tanto para las capitales como para los departamentos se obtuvieron por medio de métodos de clasificación no jerárquicos. Se obtuvieron cuatro grupos tanto para las capitales como para los departamentos: zona costera, zona central, zona oriental y amazonia. En la mayoría de los casos la clasificación de las capitales coincidió con la del departamento.

La clasificación obtenida por el método k-medoides, utilizando la distancia euclidiana, propuso grupos que coinciden con comportamientos epidemiológicos diferentes entre grupos y similares dentro de los grupos, por lo que se convierte en una herramienta estadística útil para la salud pública.

***Palabras clave*:** Clasificación; Capitales; Departamentos; Series de tiempo.

1. Introduction

At the end of 2019, alarms were raised in the city of Wuhan (China) due to an outbreak of unknown pneumonia cases (Li et al., 2020), which we now know was caused by the virus called Sars-Cov2. Given its high reproductive number, the outbreak spread fastly around the world, to the extent that on March 11, 2020, the World Health Organization declared the pandemic caused by this pathogen (World Health Organization, 2021*a*). Colombia was not an exception, the first case was reported on March 6, 2020, and by the end of March 2021, more than 2.4 million cases and more than 63 500 deaths due to the disease caused by this virus (COVID-19) had been confirmed (Instituto Nacional de Salud, 2021*a*).

So far, and as generally happens with these types of outbreaks, the presentation of cases worldwide has shown several peaks at different times: the intensity of these peaks and the distance between them has varied in different populations, depending on the governmental measures taken, the appearance of new variants, and the adherence to biosecurity measures by the population (World Health Organization, 2021*b*). In Colombia, such heterogeneity has also been evidenced, for example, the Amazon was hit hard and earlier than other regions (Instituto Nacional de Salud, 2021*b*), likewise, cities on the Atlantic coast such as Barranquilla, Santa Marta, and Cartagena had their first peak before the capital Bogotá (Instituto Nacional de Salud, 2021*a*).

Like predictive models, analyzing the behavior of virus incidence rates allows, in addition to characterizing populations with respect to the pandemic's progress, making decisions for containment or mitigation purposes. Therefore, determining groupings of geographic units with respect to the time series of the incidence of cases in the country is relevant to understanding how the virus's spread has occurred. Additionally, it could be of great help when implementing differential strategies for controlling the pandemic for each of the groups obtained.

Several of the time series clustering methods are adaptations of traditional clustering methods for `static` data, which are based on calculating distance or similarity/dissimilarity measures, and then using non-hierarchical clustering algorithms such as k-Means, k-Medoids, or even hierarchical methods such as the nearest neighbor method or the unweighted pair group method average, see Maharaj et al. (2019) and Peña & Tsay (2021) for details on the methodologies. Additionally, in Aghabozorgi et al. (2015) and Alqahtani et al. (2021), an extensive review of time series clustering can be found, including clustering from the perspective of deep learning called “deep time series clustering”, as well as methodologies based on feature extraction from the time series or model-based methodologies. Many authors have focused on analyzing epidemiological curves, most focused on fitting models that allow predicting or quantifying in some way the effect of different mitigation measures. Likewise, clustering methods have also been applied to these, such as Gohari et al. (2022), who classify countries according to the observed incidence and mortality, including cases presented until August 2021 (Gohari et al., 2022), or Spassiani et al., who classify the number of confirmed cases in the Veneto region, Italy (Spassiani et al., 2021). However, such analysis has not yet been carried out in Colombia.

The objective of this work is to classify the departments of Colombia and their capitals, according to the behavior of Sars-Cov2 incidence between March 6, 2020, and February 4, 2021.

2. Methods

The information on daily confirmed incident cases was obtained from the website of the National Institute of Health (INS) (Instituto Nacional de Salud, 2021*a*), and the population estimate for each geographical unit was obtained from the estimated population by the National Administrative Department of Statistics (DANE) for the year 2020 (Departamento Administrativo Nacional de Estadística (DANE), 2021). For the analysis of time series, the date of case report was taken into account, and the variable analyzed was the number of incident cases per 100 000 inhabitants for each geographical unit. Data was taken from the beginning of the pandemic until before the start of vaccination in Colombia.

2.1. Time Series Classification

The objective of this article is to group geographic units according to their incidence at similar moments, that is, to cluster time series of the number of incident cases per 100 000 inhabitants by capital cities or departments. Thus, similar and homogeneous units form groups with respect to the observed variables, in such a way that the groups themselves are separated from each other, being different from one another. This is one of the methods that currently belongs to unsupervised learning.

Clustering methods are generally classified into hierarchical and non-hierarchical clustering methods. Hierarchical clustering methods group the data through a succession of nested partitions, either starting from a unitary set and arriving at a set or group that includes all individuals for that group, or vice versa. The former is known as agglomerative clustering and the latter as divisive clustering. Both agglomerative and divisive methods use proximity metrics, whether distance measures, dissimilarity measures, or similarity indices, i.e., metrics based on observations, and are recommended for identifying similar geometric profiles. However, there is an approach based on metrics that are based on features, which allows for differentiation between data generating processes. On the other hand, non-hierarchical clustering directly divides individuals without relying on a hierarchical structure.

Our work focuses on non-hierarchical clustering (K-Means and K-Medoids) with Euclidean metrics based on observations, since the objective is to cluster time series without any processing according to incidence at similar moments. This methodology is suggested in Aghabozorgi et al. (2015) as a shape-based clustering. It is important to note that in this case, we will not use the deep time series clustering presented in Alqahtani et al. (2021), as this would require a large number of time series. Next, we will see how Euclidean distance is defined and the steps to carry out non-hierarchical classification. We will follow the notation and ideas given in Maharaj et al. (2019) with the necessary adjustments for time series. Let:

$$X = \{x_{it} : 1, \dots, I; t = 1, \dots T\} = \{x_i = (xi1, \dots, xit, \dots, xiT)' : i = 1, \dots, I\}$$

be the data matrix where x_{it} represents the variable of interest observed at time t on the individual or object i, x_i represents the vector of the i-th observation or the time series corresponding to the i-th individual. For our case, x_{it} will represent the number of incident cases of Sars-Cov2 per 100,000 inhabitants for each capital city or total department i, at time or day t. The Minkowski distance family between seriesi, x_i and series serie i, x_i is defined as:

$$_rd_{il} = \left[\sum_{t=1}^{T} |x_{it} - x_{lt}|^r\right]^{1/r}$$

As a particular case, we have the Manhattan distance when $r = 1$,

$$_1d_{il} = \sum_{t=1}^{T} |x_{it} - x_{lt}|$$

And the Euclidean distance when $r = 2$

$$_2d_{il} = \left[\sum_{t=1}^{T} (x_{it} - x_{lt})^2\right]^{1/2}$$

On the other hand, the K-Means method aims to find an optimal partition of observations into K groups that minimizes the sum of squared errors:

$$\min \sum_{i=1}^{I} \sum_{k=1}^{K} u_{ik} d_{ik}^2 = \sum_{i=1}^{I} \sum_{k=1}^{K} u_{ik} ||x_i - h_k||^2$$

such that

$$\sum_{k=1}^{K} u_{ik},\, u_{ik} \geq 0,\, u_{ik} = \{0, 1\}$$

where u_{ik} indicates the degree of membership of the i-th unit or time series to the k-th cluster or group; $d_{ik}^2 = ||x_i - h_k||^2$ is the squared Euclidean distance between the i-th object and the centroid of the k-th group h_k.

The above minimization can be carried out iteratively by repeating the following steps:

1. Initialize a K-partition randomly or based on prior knowledge. Calculate the centroids or means of each group k, with $k = 1, \dots, K$, that is, calculate the mean time series based only on the observations that belong to each group. The mean time series is calculated time by time using the time series in each group.
2. Assign each unit in the dataset to the nearest group using an appropriate measure between each unit and the centroids.
3. Recalculate the centroids or means based on the current partition.
4. Repeat steps 2 and 3 until there is no change in each group.

An analogous method to K-means is known as K-medoids or Partitioning Around Medoids (PAM) method, which also requires minimizing the sum of dissimilarity of the units to their nearest representative medoids. Unlike means, medoids are units of the observation set. The method requires a phase of exchanging observations. For more details, see Maharaj et al. (2019).

Finally, regardless of the clustering method, it is necessary to identify or select the number of groups to form. For this, there are different types of criteria called validity criteria, which allow for this identification. Among these, we have the Calinski and Harabasz criterion or the silhouette criterion, which identify the optimal number of groups in such a way that the selected criterion is maximized. In this case, the results were obtained using non-hierarchical methods, and we did not consider the use of any of these criteria, because they always pointed out two groups and this cannot be allowed a reasonable interpretation . The number of groups was chosen in such a way that it would show interesting results for interpretation.

3. Results

During the study period, a total of 2 135 412 confirmed cases of Sars-Cov2 and 55 131 deaths due to COVID-19 were reported in Colombia, and the first two peaks of the pandemic occurred in August 2020 and January 2021. The classification obtained for the departmental capitals using the different methods-distances was consistent, with four groups observed in general: one for the central region, one for those located in the eastern region and San Andres islands, one for the coastal regions, and one for the Amazonia region. Specifically, using the Euclidean distance and the K-medoids method, the following four groups were obtained (Figure 1).

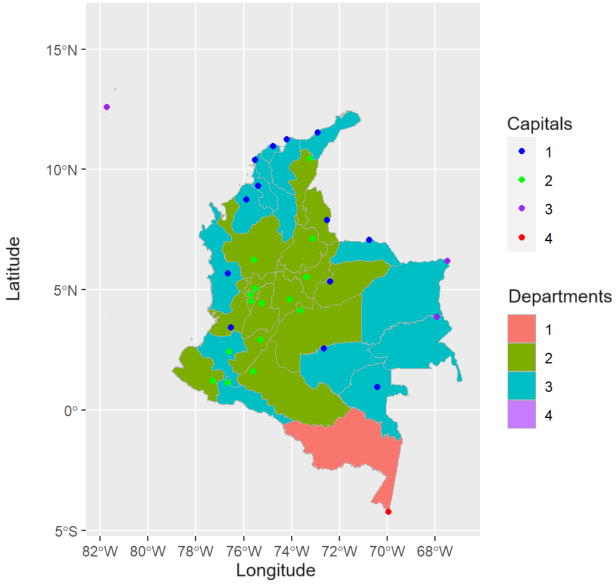

FIGURE 1: Classification of capitals and departments according to the series of the number of cases per 100 000 inhabitants

Group 1 (Coastal area): Corresponds to the capitals of the Atlantic and Pacific coasts, Arauca (Arauca), Cúcuta (Norte de Santander), San José del Guaviare (Guaviare), Yopal (Casanare), and Mitú (Vaupés). These capitals have generally had the lowest daily incidences during a large part of the monitoring period, with maxima in one day between 100 and 150 cases per 100 000 inhabitants, and only one peak of contagion occurred between July and October (Figure 2).

Group 2 (Central area): This group includes the capitals of the departments in the central region of the country and some in the southwestern region, such as Pasto and Popayán. Likewise, this group includes the capitals with the largest populations, such as Bogotá and Medellín, and therefore the behavior of the pandemic has been very similar to that of the country as a whole. These are capitals whose first peak occurred between August and September (close to 100 cases per 100 000), the second in January, and which, in general, was higher than the first (up to 300 cases per 100 000 inhabitants) (Figure 2).

Group 3 (Eastern area): Corresponds to the capitals of the eastern region and the islands of San Andrés. These three municipalities only had one significant peak, between 200 and 300 cases per 100 000 inhabitants, between September and October, except for Puerto Carreño, which had a higher incidence in December (Figure 2).

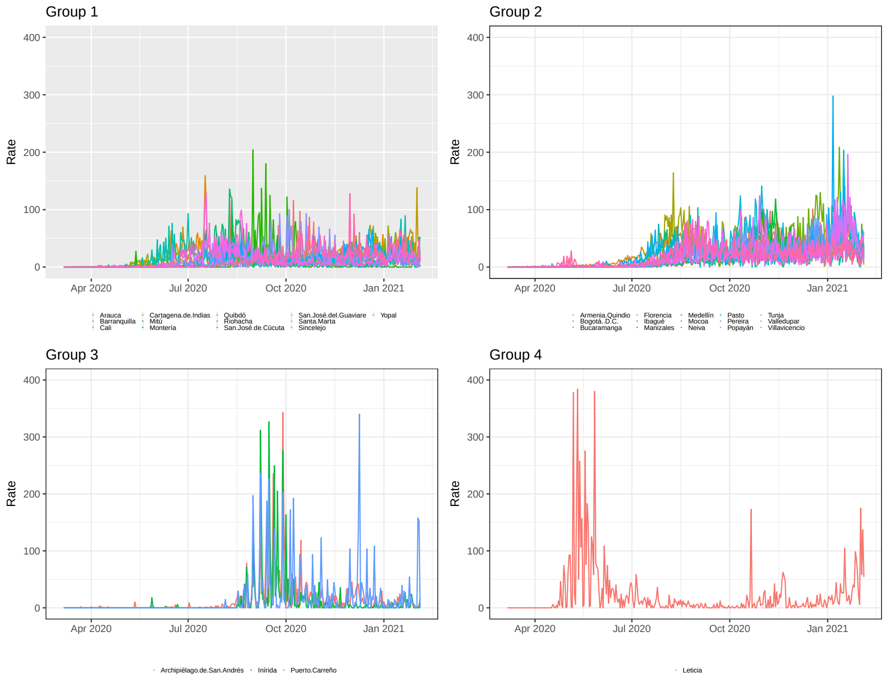

Figure 2: Confirmed cases of Sars-Cov2 per 100 000 inhabitants in Colombia by capital city, according to the obtained classification

Group 4 (Amazon): Only Leticia, the capital of Amazonas, is in this group, which showed a particular behavior: it had the highest daily incidence of all the capitals, close to 400 cases per 100 000 inhabitants, and it also had the earliest peak, between May and June, which remained low until the end of January (Figure 2).

In other words, the groups discriminated the epidemiological behavior of the confirmed case rate. In Groups 1 and 2, there are the capitals that did not have very high peaks, while in Groups 3 and 4, there are the capitals that reached rates above 400 cases per 100 000 inhabitants. These peaks, in the case of Group 3, occurred after the midpoint of the study period, while in Group 4, it occurred at the beginning of the pandemic.

On the other hand, analogously to the department capitals, the classification obtained for the departments using the different methods-distances was also congruent. Classifying the departments into four groups using Euclidean distance and the K-medoids method, two groups covering most departments are obtained, one for San Andrés and Providencia and one for Amazonas (Figure 1).

Coastal and Eastern Zone (Group 2): Corresponds to the departments whose capitals are in Group 1, except Norte de Santander and Casanare, including Vichada, Putumayo, Arauca, Cauca, and Guainía. These departments, in general, have reported less than 100 cases per 100 000 inhabitants and show two peaks towards August and January (Figure 3).

Central Zone: This includes the departments whose capitals were classified in Group 2, including Norte de Santander, Casanare, and excluding Cauca and Putumayo. These departments have reported up to over 200 cases per 100 000 inhabitants in a day, with a predominant peak between September and October and a smaller one in January 2021 (Figure 3).

San Andrés and Providencia: Corresponds to the capitals of the eastern region and the islands of San Andrés. In this department, before August, the incidence was very low, reaching over 300 cases per 100 000 inhabitants between September and October 2020, and after this date, it stabilized below 80 cases per 100 000 inhabitants (Figure 3).

Amazonas: In this department, a significant peak began in May 2020, reaching almost 250 cases per 100 000 inhabitants in June. From July, it stabilized at low incidences, and again towards January, it began to rise towards a new peak, which reaches above 100 cases per 100 000 inhabitants during the period of this study (Figure 3).

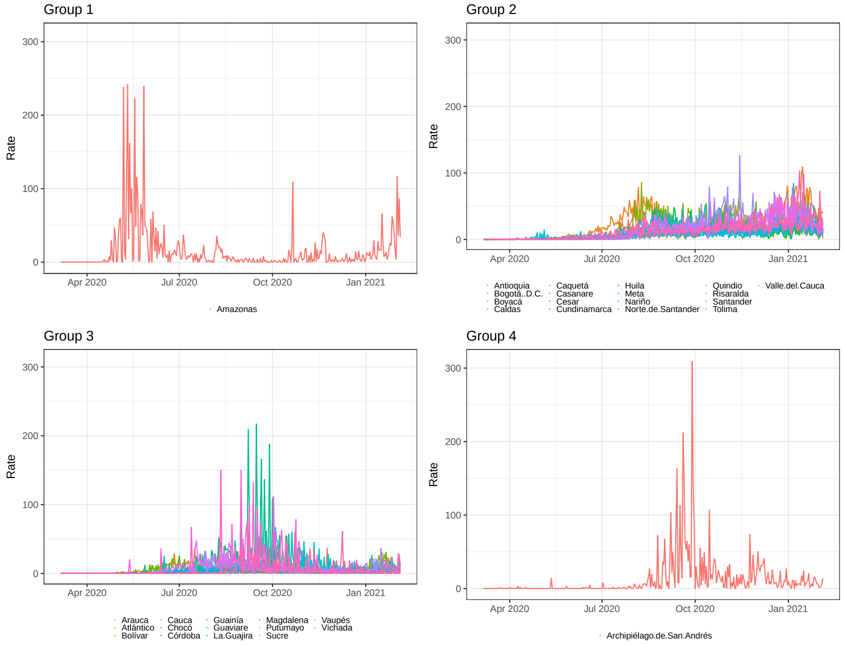

Figure 3: Confirmed cases per 100 000 inhabitants of Sars-Cov2 in Colombia by department, according to the obtained classification

In this case, the clustering of the departments coincided with that of the capitals by creating an exclusive group for Amazonas (Leticia), another group for departments that did not have pronounced peaks (Group 2), and another group for those that had their highest peak after the midpoint of the study period (Group 3). However, there was a difference as the San Andrés and Providencia archipelago was placed in a separate group, as it was the department that experienced the highest peak with a rate exceeding 300 cases per 100 000 inhabitants.

4. Discussion

he behavior of the Sars-Cov2 pandemic has not been homogeneous throughout Colombian territory, to such an extent that, based on incidence in cases per 100,000 inhabitants, four groups of departments and four groups of capital cities can be described, showing different behaviors.

Although the grouping of capitals and departments was generally coincident, in departments such as Casanare, Valle, Putumayo, Cauca, Vichada, Guanía and San Andrés and Providencia, it was not so, indicating that the epidemiological behavior of non-capital municipalities was different from that of the capital, probably due to population density, mobility of its inhabitants or measures taken by each of the municipal governments.

Departments and capitals with larger populations such as Bogotá and Medellín were located in the same cluster, which coincides with having low seroprevalence estimates (below 30%) at the end of 2020, in addition to having the same initial sources of contagion, such as the arrival of travelers from countries with high circulation of the virus and their contacts, how it is evidenced in the "Tipo" column of the historical dataset for the month of April 2020 (Instituto Nacional de Salud, 2021*a*).

On the other hand, it is important to highlight the differential behavior of the Amazonas and its capital (Leticia) compared to the rest of the Colombian territory, since it was the region that showed an epidemiological peak earlier and the highest daily incidence, close to 400 cases per 100 000 inhabitants in a single day (Figures 2-3). This behavior can be explained given the high mobility of the border area with Brazil, a country that has been highly affected by the pandemic since April 2020 (Ministerior de Salud de Brasil, 2021). On the other hand, despite having a high seroprevalence estimate at the end of September, close to 60% (Instituto Nacional de Salud, 2021*b*), the beginning of a second peak is observed in January, again caused by the border area, in this case with the aggravating factor of the widespread circulation of the P1 variant (Brazilian variant) in the region (Faria et al., 2021).

The case of the San Andrés and Providencia islands is also special, since, contrary to the Amazonas, it had a quite high first peak in October. Being a tourist site whose main access is by air, it managed to protect itself until this moment, however, after the opening of some domestic flight routes in September 2020, incidence increased to more than 300 cases per 100 000 inhabitants in one day.

In general, the classification obtained by the k-medoids method, using Euclidean distance, both for capital cities and departments, proposed groups that coincide with different epidemiological behaviors among groups (for instance Leticia had a very particular behavior as it was more affected at the beginning of the pandemic, while Group 2 had its highest peak in January 2021) and similar ones within groups (Figures 2-3), making it a useful statistical tool for public health, as it would simplify the process of adopting measures, proposing specific ones for each of the groups. Likewise, it would be a useful tool for focusing measures within the main cities of the country.

[Received: April 2023 — Accepted: November 2023]

References

Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y. (2015), 'Time-series clustering–a decade review', *Information systems* **53**, 16–38.

Alqahtani, A., Ali, M., Xie, X. & Jones, M. W. (2021), 'Deep time-series clustering: A review', *Electronics* **10**(23), 3001.

Departamento Administrativo Nacional de Estadística (DANE) (2021), 'Demografía y población'. Accessed: 2021 04 05.
https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

Faria, N., Mellan, T., Whittaker, C., Claro, I., da Candido, D., Mishra, S., Crispim, M., Sales, F., Hawryluk, I., McCrone, J. et al. (2021), 'Genomics and epidemiology of a novel sars-cov-2 lineage in manaus, brazil (preprint)'.

Gohari, K., Kazemnejad, A., Sheidaei, A. & Hajari, S. (2022), 'Clustering of countries according to the COVID-19 incidence and mortality rates', *BMC Public Health* **22**(1), 1–12.

Instituto Nacional de Salud (2021*a*), 'COVID-19 en colombia'. Accessed: 2021 04 05. https://www.ins.gov.co/Noticias/Paginas/coronavirus-casos.aspx

Instituto Nacional de Salud (2021*b*), *Estudio nacional de seroprevalencia de Sars-CoV-2*. https://www.ins.gov.co/estudio-nacional-de-seroprevalencia /reporte.html#curso

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M. & Feng, Z. (2020), 'Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia', *New England Journal of Medicine* **382**(13), 1199–1207. PMID: 31995857. https://doi.org/10.1056/NEJMoa2001316

Maharaj, E. A., D'Urso, P. & Caiado, J. (2019), *Time series clustering and classification*, CRC Press.

Ministerior de Salud de Brasil (2021), 'Coronavírus/brasil'.
https://covid.saude.gov.br/

Peña, D. & Tsay, R. S. (2021), *Statistical learning for big dependent data*, John Wiley & Sons.

Spassiani, I., Sebastiani, G. & Palù, G. (2021), ‘Spatiotemporal analysis of COVID-19 incidence data’, *Viruses* **13**(3), 463.

World Health Organization (2021*a*), ‘Coronavirus disease (COVID-19) pandemic’. Accessed: 2021 04 05. https://www.who.int/emergencies/diseases/novel-coronavirus-2019

World Health Organization (2021*b*), ‘WHO coronavirus (COVID-19) dashboard’. Accessed: 2021 04 05. https://covid19.who.int/