



Clustering of cantons in Costa Rica based on interest variables during the beta variant of Covid-19

Agrupamiento de los cantones de Costa Rica con base en variables de interés durante la variante beta del Covid-19

Isaí Ugalde-Araya¹

Ugalde-Araya, I. Clustering of cantons in Costa Rica based on interest variables during the beta variant of Covid-19. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on Bioinspired Processing. Pag. 75-80.

 <https://doi.org/10.18845/tm.v37i7.7302>

¹ Statistician and Researcher. Advanced Computing Laboratory. National High Technology Center. Costa Rica.
 iugalde@cenat.ac.cr
 <https://orcid.org/0009-0003-9517-1653>

Keywords

Cluster analysis; mortality rate; morbidity rate; case fatality rate; third wave pandemic; distance.

Abstract

The study of case behavior and the analysis of bioindicators are relevant and important for decision-making by health authorities worldwide, related to the Covid-19 pandemic. Thus, numerous investigations have been carried out around the world to understand this phenomenon, its variants, and its primary impacts on population health. In this study, a cluster analysis was conducted based on the variables of mortality rate, morbidity rate, and fatality rate, along with cantonal geographical density, for the period of the beta variant in Costa Rica, corresponding to the months from February to June 2021. Therefore, a total of three methods were chosen to obtain groups: k-means, k-medoids, and fuzzy methods; as well as two types of distances: Euclidean and Manhattan. Additionally, the sum of squares within groups and the Dunn index were used to validate the formation of the clusters. It was identified that the method and distance that formed the most compact cantonal clusters with lower intragroup variability were k-medoids and Manhattan, respectively, due to their greater robustness against extreme values. Among the formed groups, cluster 1 has a moderate impact of the pandemic during the specified variant, while groups 2 and 3 have low and high impacts, respectively. Moreover, groups 1 and 2 are predominantly composed of cantons outside the Greater Metropolitan Area, in contrast to the third group. This analysis provides valuable insights for health authorities in understanding the impacts of the Covid-19 pandemic in Costa Rican regions and aids in the development of targeted strategies for effective management.

Palabras clave

Análisis de conglomerados; tasa de mortalidad; tasa de morbilidad; tasa de letalidad; tercera ola pandémica; distancia.

Resumen

El estudio del comportamiento de los casos y el análisis de bioindicadores es relevante para la toma de decisiones por parte de las autoridades sanitarias, relacionado con la pandemia del Covid-19. De este modo, numerosas investigaciones se han llevado a cabo en el mundo para comprender este fenómeno, sus variantes y principales afectaciones en la salud de la población. En el presente estudio, por lo tanto, se realizó un análisis de conglomerados con base en las variables tasa de mortalidad, tasa de morbilidad y tasa de letalidad, así como con la densidad geográfica cantonal, para el periodo de la variante beta en Costa Rica, correspondiente a los meses de febrero a junio de 2021, y se eligieron tres métodos para obtener los grupos: k-medias, k-medoides y métodos difusos; así como dos tipos de distancias: euclídea y de Manhattan. Asimismo, se utilizó la suma de cuadrados dentro de grupos y el índice de Dunn para validar la conformación de los grupos. Se identificó que el método y la distancia que formaban los conglomerados de cantones más compactos, con una menor variabilidad intragrupo, fue k-medoides y Manhattan, respectivamente, debido a que su mayor robustez ante valores extremos. De los grupos formados, el clúster 1 posee un impacto moderado de la pandemia durante la variante en cuestión, y el grupo 2 y 3, un impacto bajo y alto, respectivamente. Asimismo, el grupo 1 y 2 están conformados en mayor medida por cantones no pertenecientes al Gran Área Metropolitana, en contraposición al grupo 3. Este

análisis proporciona ideas valiosas para las autoridades sanitarias al comprender los impactos de la pandemia de Covid-19 en las regiones costarricenses y contribuye al desarrollo de estrategias específicas para una gestión efectiva.

Introduction

Since the beginning of the Covid-19 pandemic, many efforts have been made to identify the main social, economic, and health repercussions of a pandemic. Likewise, numerous investigations have been directed to analyze the behavior of the virus and its respective variants. The question that arises is why it is relevant to continue studying this disease, after three years since its worldwide spread; the answer is based on the fact that the behavior of the virus is not so much related to seasonal effects [1], but rather to social dynamics and interventions or restrictions in health matters, and to the emergence of new variants and mutations [2].

Likewise, in the face of the increase in infections, health authorities have had to take measures aimed at mitigating the growing proliferation of the virus as much as possible. At this point, the importance of data knowledge and analysis as a tool to understand the behavior of the pandemic is highlighted, and based on this, decisions are made in favor of people's health. The disease surveillance constitutes the basis for the response to epidemics [3].

In addition to this, the analysis of the data to determine projections, trends, and indicators has been essential to identify advances and challenges in the management of the pandemic waves, related to the variants of the Covid-19. It is important to highlight that, due to the geospatial, socioeconomic, and access to health care characteristics of the population in different places, it is convenient to analyze the indicators associated with the measurement of an epidemic or pandemic in a more disaggregated way. For this reason, in the present document, the cantonal behavior of Costa Rica is analyzed according to the mortality, morbidity, and lethality rate, during the beta variant wave, which was shown to cause more severe disease and higher transmissibility than the original variant [2].

Therefore, the main objective of this study is to compare clustering methods and their respective distances, and to identify the one that best fits the data, based on the bioindicators already mentioned.

Methodology

The information involved in the analysis came from the epidemiological reports provided by the Costa Rican Ministry of Health (MINSa), which has been compiled by the Centro de Investigación Observatorio del Desarrollo (CIOD) from the University of Costa Rica [4]. Therefore, data related to cantons, specifically daily deaths and cases were used as input for the calculation of relevant indicators such as the mortality, morbidity, and lethality rates. For their calculation, the projected population for 2021 of each canton was obtained from the Instituto Nacional de Estadística y Censos (INEC) of Costa Rica [5]. Moreover, the population density (obtained from INEC), was also considered as a variable for the clustering analysis, as a way of considering the average number of people who live in each canton, and who could eventually be infected.

Due to the behavior of the cases, cantons with average rates corresponding to extreme values are obtained, which, if ignored, generate the presence of new extreme values. In response to this, it is decided to explore three clustering methods to identify which of these forms more compact groups: (a) a method that calculates average points, (b) cantons in the central position (analogous to the median), and (c) a variant of the k-means method, in which each canton has a probability degree of belonging to a group. The three methods are: (1) K-means: it starts with a predetermined number of groups (k) and identifies their average (centroids). It

assigns each observation to the cluster with the closest centroid, (2) K-medoids: it is similar to k-means, but instead of calculating the distance of each point to the group means, it identifies the observation in the central position of the cluster and assigns observations to that cluster, and (3) Fuzzy methods: it considers observations that can be associated with multiple clusters. Each observation has a degree of membership to each group [6].

It is important to highlight that two distances are selected to measure distances between cantons, which corresponds to the Euclidean and Manhattan ones, chosen for their ease of understanding, given that they have similar formula calculation. Also, Dunn index was used to validate the clusters, to quantitatively assess the quality of the data partition into groups. Moreover, to determine the number of clusters, the elbow method was used, that consists in selecting the number of groups in which the within-cluster sum of squared decreases. Also, the study period selected corresponds to the third pandemic wave related to the beta variant, which covered the dates between February 21, 2021, and July 28 of that same year; this period was chosen taken into consideration the characteristics of the beta variant, so the analysis could be used as a reference for future pandemic events under similar conditions.

Furthermore, the analyses were carried out using the R software [7] and the Kabré supercomputer.

Results and discussion

Based on the elbow method, the number of clusters suggested are three. From Table 1, it is possible to identify the values of the sum of squares within cluster for each canton division method and each distance.

Table 1. Sum of squares within cluster by method and distance.

Method	Distance (in millions)	
	Euclidean	Manhattan
k-means	193	193
k-medoids	197	180
Fuzzy methods	193	189

From table 1, it is determined that the minimum sum of square within cluster is 180 million, corresponding to k-medoids method with the Manhattan distance. This means that this method and distance produces clusters with cantons that are more similar to each other, related to the division variables. Moreover, for validating the clusters, the dunn index is shown in Table 2:

Table 2. Dunn index values by method and distance.

Method	Dunn Index	
	Euclidean	Manhattan
k-means	0.066	0.054
k-medoids	0.068	0.072
Fuzzy methods	0.066	0.044

It is possible to identify from Table 2 that the highest Dunn index is 0.072 corresponding to k-medoids with Manhattan distance. In this case, a higher Dunn index means that the groups have a lower intracluster variability. Therefore, both, the sum of squares within clusters and

the Dunn index show that k-medoids with the Manhattan distance is the one that produces the most accurate results; this is because it is more robust to extreme values than the method of k-means. The same behavior is observed in an analogous way with respect to the fuzzy method because this is a variation of the k-means method. However, it is observed that the Dunn index is still small, which indicates that, although despite being the method that generates the greatest similarity between cantons, there is still a lot of dispersion within the groups.

Descriptively, the clustering analysis using the k-medoids method and Manhattan distance reveals distinct patterns among clusters. Cluster 1, comprising mostly non-Greater Metropolitan Area (GAM) cantons, particularly from Alajuela, Guanacaste, and Puntarenas, exhibits a high morbidity rate and moderate mortality and lethality rates, suggesting a moderate impact of the pandemic during the beta variant. Similarly, cluster 2, dominated by non-GAM cantons, especially from Cartago and Limón, shows a moderate morbidity and lethality rate, and a low mortality rate, implying a lower impact of the pandemic. In contrast, cluster 3, primarily composed of GAM cantons, particularly from San José and Heredia, demonstrates high mortality, moderate morbidity, and high lethality rates, indicating an impact with a higher ratio of deaths to infections compared to other clusters.

Conclusions and/or recommendations

It is concluded that the method that performs the best separation of groups is the k-medoids with the Manhattan distance, given that it has the lowest sum of squares and the highest Dunn index. These results coincide with those found by various authors in previous studies [8]. However, it is identified that there is still a high variability between the cantons within the groups, which is due to the presence of extreme values, which occur regularly in a bioinfectious events such as a pandemic. It is suggested to carry out an analysis with other variables, in order to identify the main characteristics of the clusters in terms of health authorities being able to identify the behavior and characterization of the clusters formed.

Acknowledgments

Thanks to the CIODD for compiling COVID-19 data in Costa Rica, and to its researchers for their assistance in a preliminary analysis for this article.

References

- [1] G. L. Vasconcelos, A. A. Brum, F. A. Almeida, A. M. Macêdo, G. C. Duarte-Filho y R. Ospina, «Standard and Anomalous Waves of COVID-19: A Multiple-Wave Growth Model for Epidemics,» *Brazilian Journal of Physics*, pp. 1867-1883, 2022.
- [2] J. L. Jacobs, G. Haidar y J. W. W. Mellors, «COVID-19: Challenges of Viral Variants,» *Annual Review of Medicine*, vol. 74, pp. 31-53, 2023.
- [3] N. Pearce, J. P. Vandenbroucke, T. J. VanderWeele y S. Greenland, «Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions,» *AJPH*, vol. 110, n° 7, pp. 949-951, 2020.
- [4] Universidad de Costa Rica. Centro de Investigación Observatorio del Desarrollo, «Costa Rica,» 30 05 2022. [En línea]. Available: <https://app.powerbi.com/view?r=eyJrljoiMjU3M2NkNjQtMGlyOS00ZjRmLWE3NjY-tNDE2OWNkZjlxZTdjliwidCI6ImFkNjNmZDZmLWE4OTctNDIjZS1hZWU5LTRmYzYxNzY1NjY4YSJ9&pageName=ReportSection>. [Último acceso: 10 09 2023].
- [5] Instituto Nacional de Estadística y Censos, Costa Rica «Estadísticas demográficas. 2011 – 2025. Proyecciones nacionales. Población total proyectada al 30 de junio por grupos de edades, según provincia, cantón, distrito y sexo,» [En línea].
- [6] F. Klawonn, R. Kruse y R. Winkler, «Fuzzy Clustering: More than just fuzzification,» *Fuzzy Sets and Systems*, vol. 281, pp. 272-279, 2015.

- [7] R. C. Team, R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing.
- [8] Suwanda, R., Syahputra, Z., y Zamzami, E. M. «Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid», In Journal of Physics: Conference Series, vol 1566, 2020.