

Equipo de Investigadores:

Julio César Alonso C.
jcalonso@icesi.edu.co

Yoseth Jesualdo Ariza
yjariza@icesi.edu.co

Cristian Camilo Hoyos
cchoyos@icesi.edu.co

Este documento presenta la metodología utilizada para segmentar o agrupar los departamentos de Colombia.

1. Introducción

El aumento de las defunciones en medio de la emergencia ocasionada por la Covid-19 ha generado gran alarma a los gobiernos de diferentes niveles, a los sectores productivos, a los responsables de la salud pública y al público en general. Desde inicios de la pandemia, el Centro de Investigación en Economía y Finanzas (Cienfi) de la Universidad Icesi en colaboración con el Departamento de Salud Pública de la Facultad de Ciencias de la Salud de la misma Universidad, hace un seguimiento al recuento diario de las muertes por COVID-19 que se registran en la base de datos pública administrada por el Instituto Nacional de Salud (INS). Empleando esta información diariamente se generan visualizaciones con un enfoque regional que se encuentran disponibles a una escala departamental en el portal de datos de Pro Pacífico (<https://propacifico.org/covid-19-en-la-region-pacifico/>) y a una escala municipal en el portal de la Alianza Pacífico Task Force (<https://pacificotaskforce.com/datos-covid-19/>)

En general, al analizar las cifras es común realizar comparación de departamentos que pertenecen a una misma región geográfica. Sin embargo, la proximidad geográfica no garantiza que los departamentos tengan las condiciones para realizar una comparación adecuada, y se podría estar “comprando los no comparables”.

Esta metodología tiene como objetivo encontrar grupos de departamentos para los cuáles se puedan realizar comparaciones en términos de recursos de infraestructura para enfrentar a la pandemia, condiciones sociales y

estructura demográfica. En ese orden de ideas, se construye una metodología para contrastar las series departamentales de excesos de muertes (Ver [metodología de excesos de muertes](#)) con un criterio que incluya no solo el carácter económico, si no también de capacidad instalada de los servicios de atención en salud y vulnerabilidad frente a la pandemia del COVID-19. Para lograr el objetivo, se emplean técnicas de clustering para agrupar a los departamentos teniendo en cuenta las características mencionadas.

Este documento presenta la metodología construida por el equipo del Cienfi con el objetivo de generar una alternativa de comparación departamental diferente a las regiones geográficas. La implementación de nuestra metodología se hace en el software R (R Core Team, 2020). A continuación, se presenta en mejor detalle las variables usadas y el proceso de clustering empleado.

2. Datos

Se emplean 3 variables para realizar el agrupamiento de los departamentos de Colombia. La primera es la proporción de personas en Necesidades Básicas Insatisfechas (NBI) obtenido del DANE¹, información calculada a partir del Censo de 2018.

La segunda variable es la proporción de población mayor de 60 años (P_{60}), la cual fue calculada de la siguiente forma:

$$P_{60} = \frac{Pob60_i}{PobTotal_i} \times 100$$

¹ Para más información de esta variable visitar:
[https://www.dane.gov.co/index.php/estadisticas-por-](https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/necesidades-basicas-insatisfechas-nbi)

[tema/pobreza-y-condiciones-de-vida/necesidades-basicas-insatisfechas-nbi](https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/necesidades-basicas-insatisfechas-nbi)

Equipo de Investigadores:

Julio César Alonso C.
jcalonso@icesi.edu.co

Yoseth Jesualdo Ariza
yjariza@icesi.edu.co

Cristian Camilo Hoyos
cchoyos@icesi.edu.co

Este documento presenta la metodología utilizada para segmentar o agrupar los departamentos de Colombia.

donde $Pob60$ es la población con 60 años o más para el departamento i y $PobTotal$ es la población total para el departamento i . La población es obtenida de las proyecciones de población del DANE (Ver: <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>).

La tercera variable empleada fue el número de camas UCI para cada departamento, la cual fue obtenida del Registro Especial de Prestadores de Servicios de Salud (REPS) (Ver: <https://prestadores.minsalud.gov.co/habilitacion/>).

3. Clustering

Luego de contar con la información de cada una de las variables para todos los departamentos, se empleó un algoritmo de clustering para identificar los conglomerados.

Clustering, es una técnica de machine learning que permite segmentar o agrupar individuos u observaciones, de tal manera que los individuos al interior de cada grupo o clúster sean los más similares o parecidos posibles, pero que al mismo tiempo cada grupo sea lo más diferente posible a los otros grupos. En otras palabras, el algoritmo debe encontrar agrupaciones que cumplan con dos características: cada grupo debe tener una mínima variabilidad interindividual y cada grupo debe ser lo más diferente posible de los demás grupos. Para realizar esto es necesario contar con un algoritmo de clasificación y una medida de que tan cerca se encuentran los datos (distancia)².

² Para una explicación intuitiva de esta técnica puede consultarse le siguiente video: <https://youtu.be/z0LX3sBSuXg>.

Existen diferentes algoritmos para desarrollar un proceso de clustering, tales como particionado y jerárquico. En este caso se empleó el clustering jerárquico y el método de Ward (1963)³, que implica minimizar la varianza al interior de los grupos.

Por otro lado, para la estimación de la cercanía entre los grupos se usó la distancia euclidiana, usada de la siguiente forma:

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

donde, p y q representan dos observaciones, en este caso dos departamentos y n el número de variables (en este caso tres).

Por último, de los “ n ” clusters posibles debe seleccionarse el mejor número, para lo cual se emplearon las métricas incluidas en el paquete **Nbclust** (Charrad, Ghazzali, Boiteau & Niknafs, 2015) de R.

4. Resultados

Tras realizar el ejercicio de clustering, se encontraron 6 grupos de departamentos, los cuales se presentan a continuación:

Clúster 1:

- Atlántico
- Antioquia
- Bogotá D.C.
- Valle del Cauca

Clúster 2:

³ Para más información sobre clustering particionado y jerárquico consultar: James, Witten y Tibshirani (2013).

Equipo de Investigadores:

Julio César Alonso C.
jcalonso@icesi.edu.co

Yoseth Jesualdo Ariza
yjariza@icesi.edu.co

Cristian Camilo Hoyos
cchoyos@icesi.edu.co

Este documento presenta la metodología utilizada para segmentar o agrupar los departamentos de Colombia.

- Bolívar
- Cesar
- Córdoba
- Magdalena
- Sucre

Clúster 3:

- Boyacá
- Caldas
- Quindío
- Risaralda
- Tolima

Clúster 4:

- Amazonas
- Arauca
- Caquetá
- Casanare
- Guainía
- Guaviare
- La Guajira
- Putumayo
- Vaupés
- Vichada

Clúster 5:

- Cauca
- Cundinamarca
- Huila
- Meta
- Nariño
- Norte de Santander
- Santander

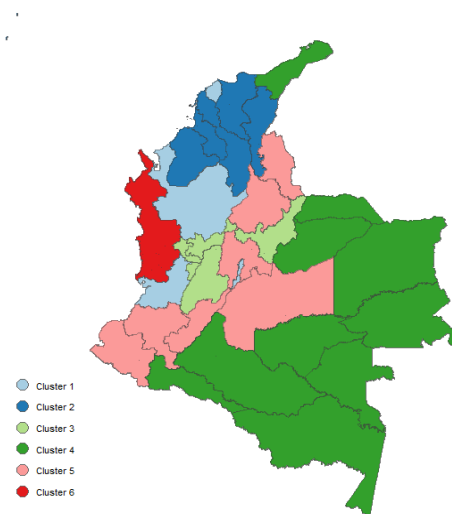
Clúster 6:

- Chocó

5. Comentarios finales

En la Figura 1 podemos ver en el mapa de Colombia los departamentos agrupados por color según los clústeres identificados. Esta agrupación será la base para generar comparaciones del comportamiento de los excesos de muertes durante el 2020. De esta manera se tendrán grupos de departamentos más similares. Por ejemplo, los cuatro dptos. de la región geográfica del Pacífico Colombiano quedaron en clusters diferentes: Valle del Cauca en el número 1, Cauca y Nariño en el número 5 y Chocó es el único dpto. del grupo 6. La distribución descrita permite comparar a cada ente territorial con otros similares en cuanto a condiciones poblacionales, de necesidades básicas insatisfechas y de disponibilidad de infraestructura.

Figura 1 Mapa de los clústeres de departamentos.



Fuente: Cálculos Cienfi.

Equipo de Investigadores:

Julio César Alonso C.
jcalonso@icesi.edu.co

Yoseth Jesualdo Ariza
yjariza@icesi.edu.co

Cristian Camilo Hoyos
cchoyos@icesi.edu.co

Este documento presenta la metodología utilizada para segmentar o agrupar los departamentos de Colombia.

Referencias

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2015). Determining the Best Number of Clusters in a Data Set. . Recuperado de: <https://cran.rproject.org/web/packages/NbClust/NbClust.pdf>.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236–244.