Geomarketing
Ingeniería Civil en Geografía
Departamento de Ingeniería Geográfica
Facultad de Ingeniería
Universidad de Santiago de Chile

DEPARTAMENTO DE
INGENIERÍA GEOGRÁFICA
UNIVERSIDAD DE SANTIAGO DE CHILE

Trabajo 3: Clusterización k-means para el Gran Santiago

Profesor: Ricardo Crespo

Alumno: Pedro Pablo Silva

Fecha: 24 de Septiembre de 2020

Introducción

Objetivo general

Desarrollar un análisis de clusterización para el Gran Santiago (GS), tomando en cuenta variables obtenidas desde fuentes de información pública.

Objetivos específicos

- Aplicar y analizar clusterización con el método k-means para el GS en función de la edad y escolaridad de las personas encuestadas en el Censo de Población y Vivienda 2017.
- Aplicar y analizar clusterización considerando la variable microsimulada previamente “ayuda de pensiones solidarias” desde la encuesta CASEN.

Metodología

Para el desarrollo de este trabajo se utilizaron como base el Censo de Población y Vivienda 2017 y la Encuesta de Caracterización Socioeconómica Nacional 2017 (CASEN). El procesamiento de estas bases de datos se realizaron mediante el software estadístico Rstudio ®.

Marco Teórico

K-Means es probablemente uno de los algoritmos de agrupamiento más estudiados, por su fácil entendimiento y aplicación. Principalmente, separa un conjunto de datos en k grupos o clases, para esto, define centroides de grupos y va iterando esta definición de centroides hasta que forme grupos con distancias similares desde el centro del grupo hacia el límite. Para esto, el algoritmo sigue los siguientes pasos:

1. Se selecciona el número de grupos o clases.
2. Se define un centroide para cada uno de los grupos.
3. Se calcula la distancia desde cada registro hacia todos los centroides y se asigna cada registro al centroide más cercano.
4. Se vuelve a calcular el centroide de cada grupo nuevo.
5. Iterar paso 3 y 4 hasta la convergencia del algoritmo, es decir, los centroides se mantienen sin modificaciones y los datos no cambian de grupo.

Resultados

En primera instancia se trabajó con el Censo de Población y Vivienda 2017 (Censo desde ahora en adelante), donde se utilizaron las variables de escolaridad y edad de cada una de las personas encuestadas a agrupar.

Para encontrar el número óptimo de clusters, se utilizó el método Elbow o el método del codo, el cual grafica el error en función del número de grupos. El resultado se muestra en la Figura 1. A partir de esto se define como óptimo un total de 5 grupos a clusterizar, ya que es en este momento donde el aumento de grupos no influye de manera significativa en el error obtenido. El aumento de grupos podría dificultar su interpretación, por lo que se decide no aumentar el número de clusters.

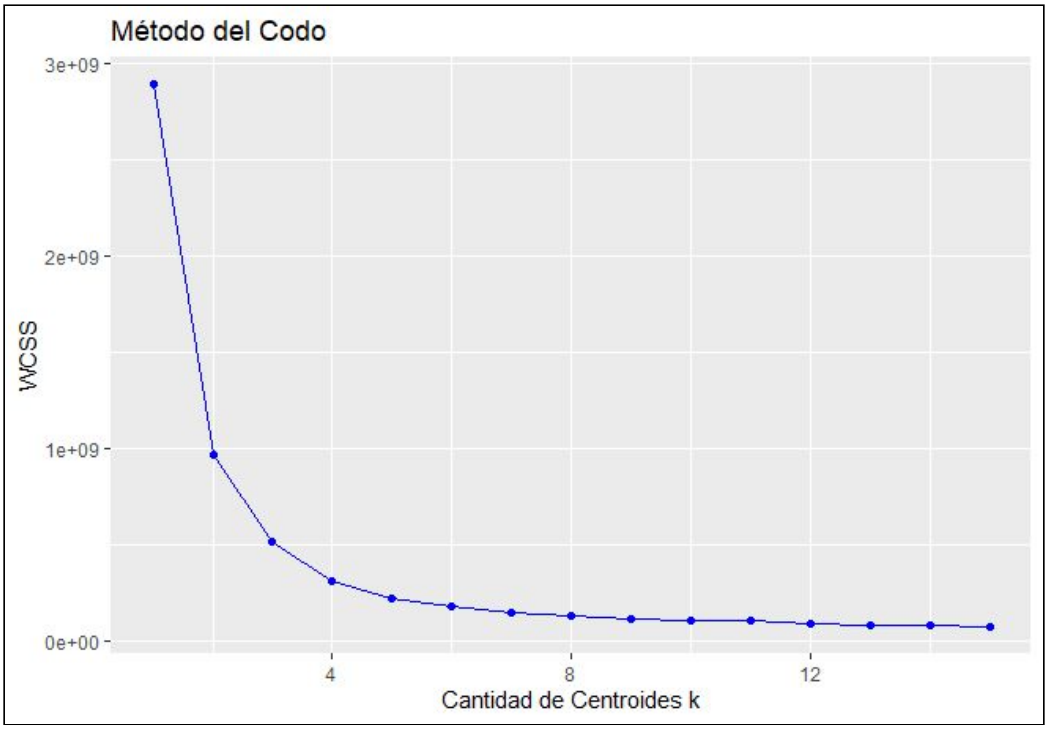

Figura 1. Método Elbow. Fuente: Elaboración propia.

La clusterización generada con 5 centroides, tiene límites cercanos a los valores 20, 30, 45 y 70 para los años de cada personas (Ver Figura 2). Según el histograma de las personas pertenecientes a cada grupo (Figura 3), se observa que el cluster 1, 2 y 5 tienen una frecuencia similar, mientras que el grupo 3 tiene una frecuencia mayor al resto y el grupo 4 una frecuencia menor en comparación de los otros grupos.

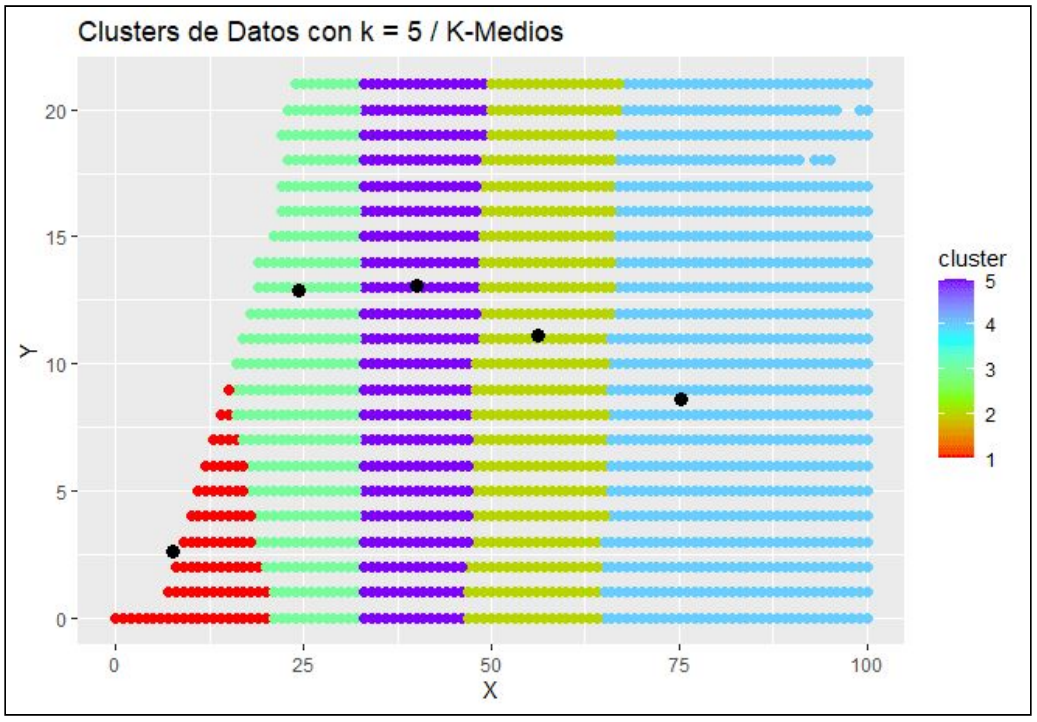

Figura 2. Clusterización k-means según edad (X) y años de escolaridad (Y). Fuente: Elaboración propia.

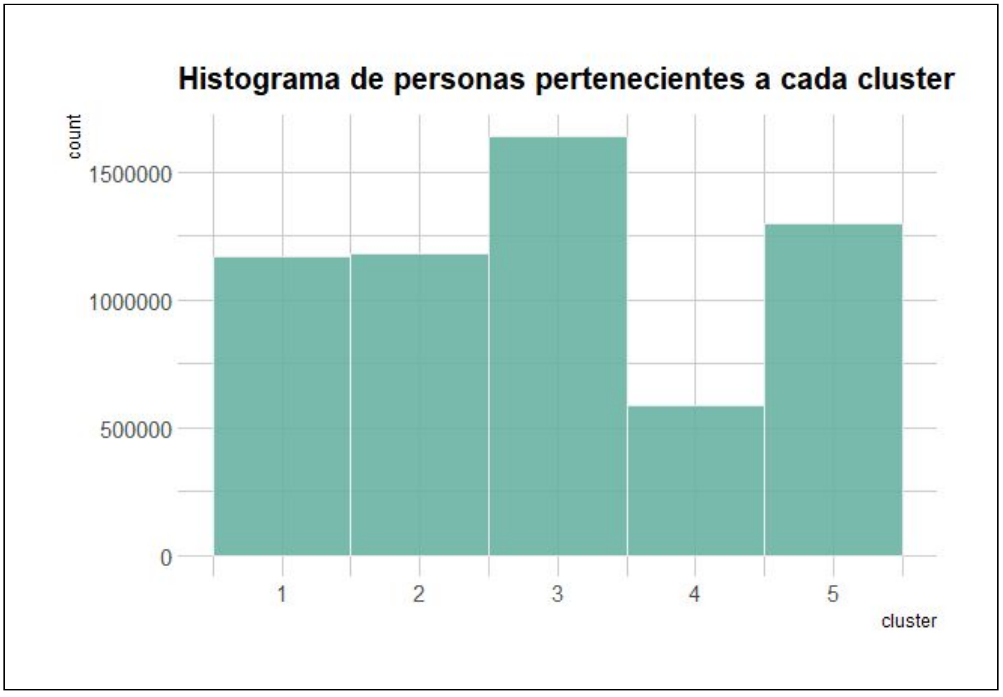

Figura 3. Histograma de personas pertenecientes a cada cluster. Fuente: Elaboración propia.

Luego de generar los cluster, se realiza un conteo de personas que pertenecen a cada uno de los grupos identificados por zona censal y se calcula la proporción respecto al total de personas que habita en cada una de las zonas censales. A partir de este resultado se

mapean las proporciones del cluster número 1 (Figura 4) y del cluster número 2 (Figura 5).

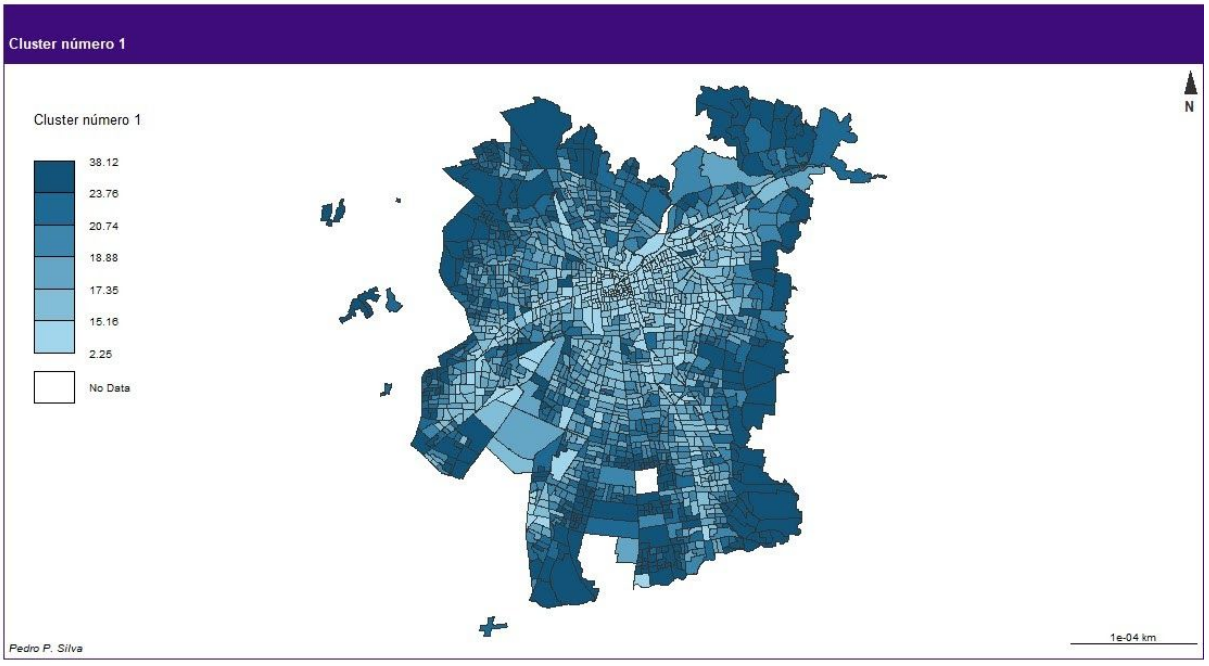

Figura 4. Cartografía del cluster 1. Fuente: Elaboración propia.

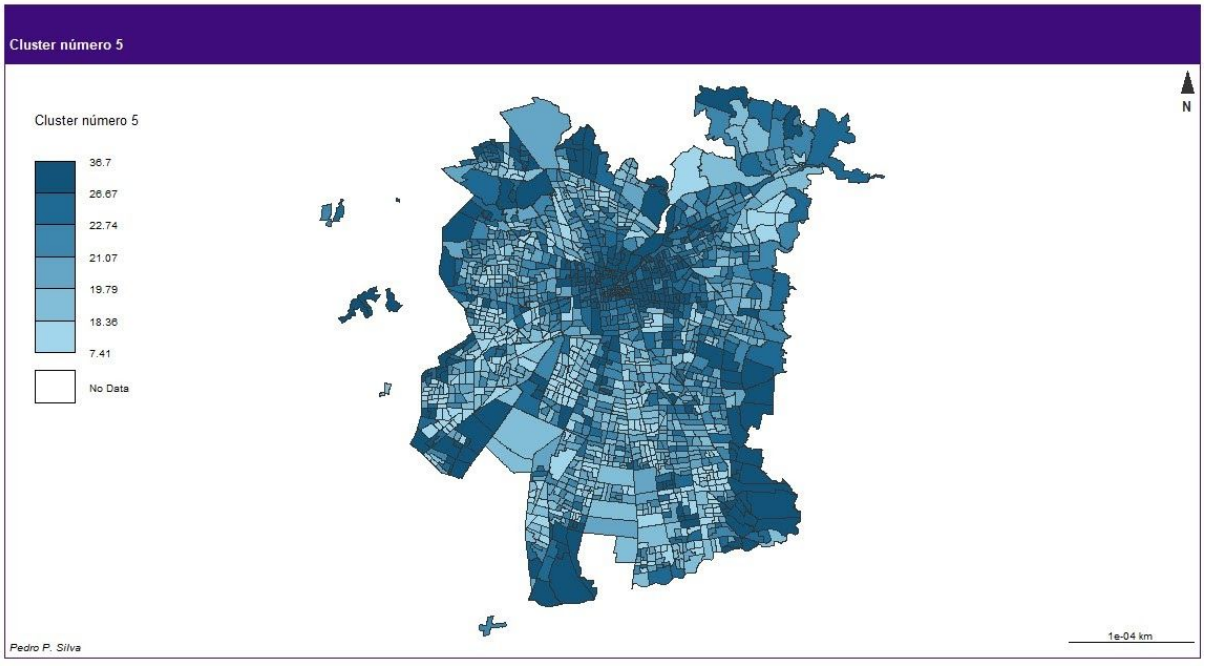

Figura 5. Cartografía del cluster 5. Fuente: Elaboración propia.

Luego, se generó una tabla con la cantidad de personas perteneciente a cada cluster por zona censal. Esto se utilizó como input para calcular el índice de Shannon para medir la diversidad de cada una de las zonas censales respecto a la cantidad de personas pertenecientes a cada cluster. El resultado fue mapeado y es mostrado en la Figura 6.Para la segunda parte del trabajo, se realiza una clusterización, pero ahora considerando como tercera variable la variable microsimulada en la parte 2 del trabajo 1, la cual corresponde a personas que reciben pensión solidaria (total o parcial) por parte del estado. Esta variable

fue extraída desde la Encuesta de Caracterización Socioeconómica Nacional 2017 (CASEN en adelante).

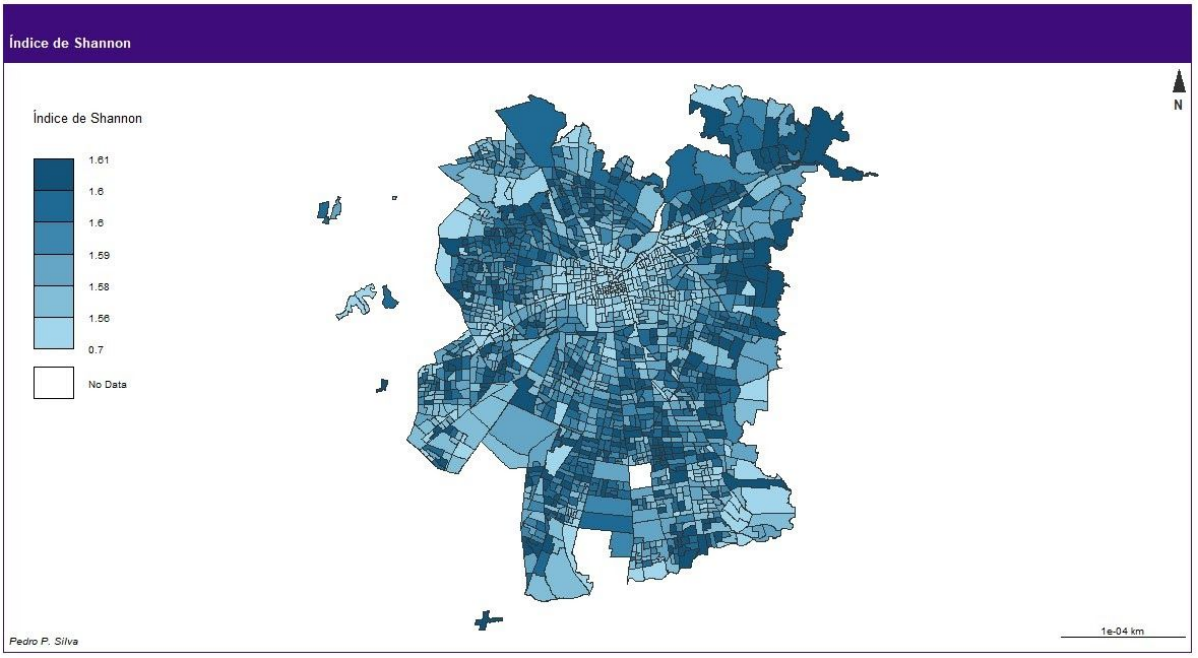

Figura 6. Cartografía del índice de Shannon. Fuente: Elaboración propia.

En la segunda parte de este trabajo se realizó una clusterización de personas caracterizadas en cada una de la zonas censales. Las personas seleccionadas fueron mayores a 60 años, con 12 o menos años de escolaridad y que hayan recibido algún tipo de ayuda estatal de pensión solidaria, esta última fue el resultado de una microsimulación desarrollada en la parte 2 del trabajo 1.

Para encontrar el valor óptimo de cluster se aplicó nuevamente el método Elbow, considerando que 4 clusters era suficiente para obtener grupos que no dificulte su interpretación con un error relativamente bajo (en comparación a clusterizaciones con menos grupos).

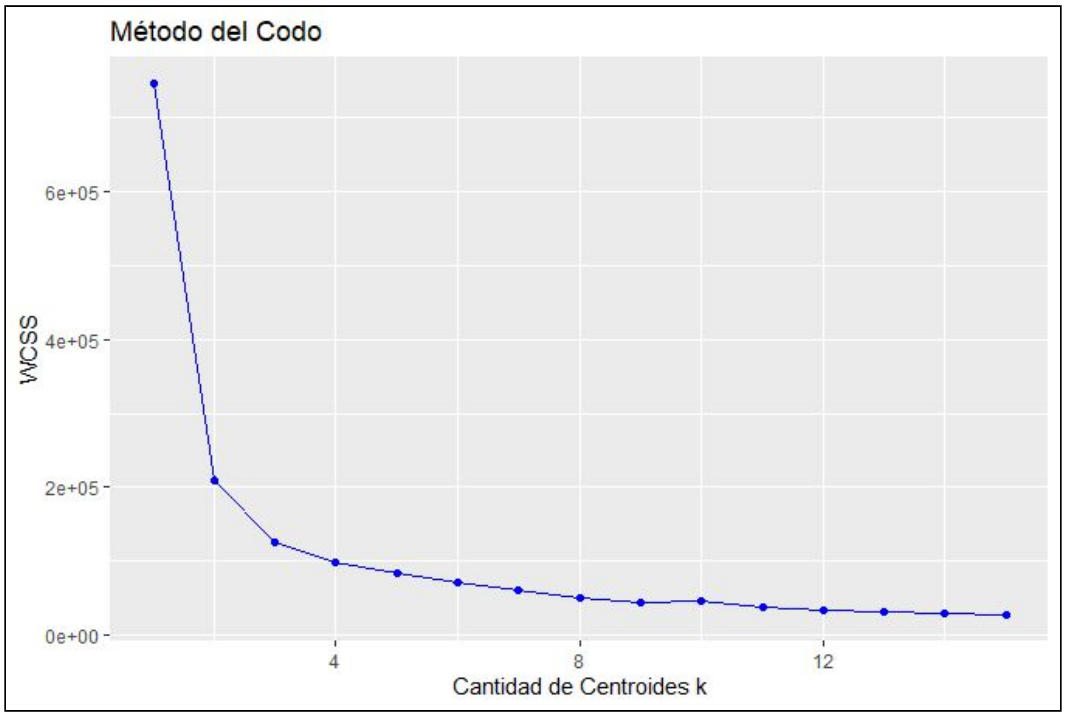

Figura 7. Método Elbow. Fuente: Elaboración propia.

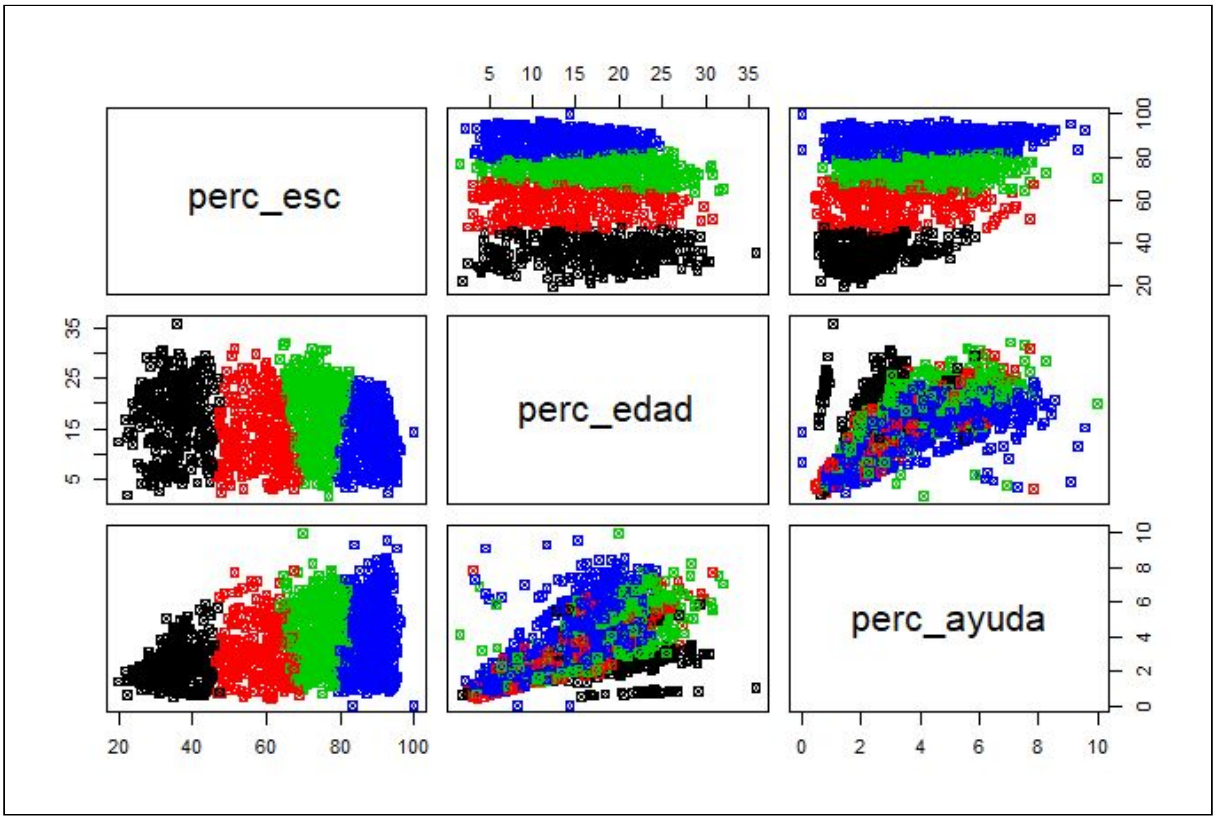

Figura 8. Clusters encontrados según porcentajes de adultos mayores o personas con menos de 12 años de escolaridad. Fuente: Elaboración propia.

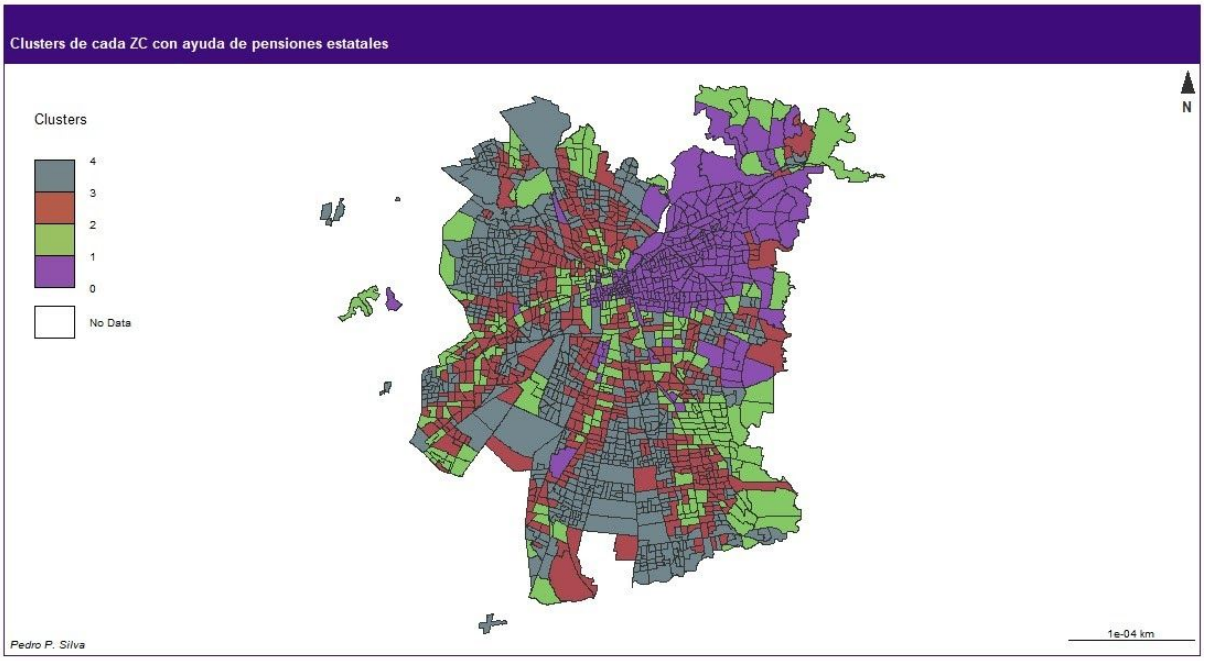

Figura 9. Método Elbow. Fuente: Elaboración propia.

Análisis

Podemos observar que los clusters son limitados principalmente por la edad de las personas, esto se debe a que en general, las personas estudian en ciertas edades, por ejemplo la educación básica se estudia entre los 6 y 14 años, la educación secundaria entre los 14 y 18 años y así. Por lo mismo el primer cluster, el cual contempla personas con edades entre 0 y 20 años tienen estudios que no superan los 12 años de escolaridad, porque en ese rango de edades, aún se estudia en el ciclo educacional obligatorio. El resto de los grupos se distribuyen de una manera más similar, ya que la edad no está muy relacionada con los años de estudio, sino más bien a otras variables socioeconómicas como el ingreso per cápita, entre otros.

Según el histograma realizado, se aprecia que los cluster 1, 2 y 5 presentan frecuencias similares, cercanas a 120.000 personas. El cluster 3 fue el que obtuvo una mayor frecuencia, superando los 150.000 personas. Finalmente el cluster 4 fue el que presenta una menor cantidad de personas con una frecuencia cercana a los 600.000 habitantes.

Al analizar la proporción de personas pertenecientes al cluster 1 y su distribución espacial, podemos observar que se sitúan principalmente en las periferias del GS, el patrón espacial es bastante claro, recordemos que el cluster 1 corresponden a personas jóvenes (hasta los 20 años aproximadamente), por lo que se podría inferir que en estas zonas es donde las familias son más jóvenes que en el resto de la capital o que en estos lugares nacen más personas que en otros sectores.

Para la distribución del cluster 5, el cual corresponde a la población de mayor edad se sitúan en el sector céntrico de la capita, en las comunas de Santiago Centro, Ñuñoa, Providencia, pero también se observa una alta proporción en sectores periféricos como Puente Alto oriente, San Bernardo, Maipú Sur, el Sector de La Pincoya y Lo Barnechea.

Al analizar el índice de Shannon, nos damos cuenta que los sectores intermedios entre el centro y la periferia de la ciudad tienden a ser más diversos respecto a la cantidad de personas pertenecientes a cada uno de los clusters analizados. Esto concordaría con los resultados analizados anteriormente (cluster 1 y cluster 5). Los sectores con mayor diversidad son Lo Barnechea, La Reina, Quilicura, Pudahuel y Cerro Navia.

En la clusterización respecto a los porcentajes de personas caracterizadas podemos observar una clara distribución del cluster número 4 sobre los sectores más vulnerables del GS como las comunas de La Pintana, Santa Rosa, San Ramón, Pudahuel, Cerro Navia y algunas zonas censales en Maipú. Este cluster corresponde a zonas censales con una mayor proporción de recepción de pensiones solidarias. Es importante recordar que estos clusters fueron construidos en base a una proporción, por lo que no necesariamente representa la situación específica a una menor nivel espacial. El Cluster número 1, se concentra principalmente en el sector nororiente de la capital, en las comunas de Vitacura, Providencia, Las Condes, La Reina y algunos sectores de La Florida. Este cluster corresponde a zonas censales con un bajo porcentaje de recepción de aportes a la pensión de los adultos mayores por parte del estado. Los otros dos clusters intermedios se distribuyen de manera heterogénea en los zonas intermedia a los sectores céntricos y periféricos. Esto se puede explicar ya que la diversidad de personas en estas zonas censales son mayores que en los otros sectores, como se vio en el cálculo del índice de Shannon.