Capretto, Tomás
Marí, Gonzalo
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística

MÉTODO DE AGRUPAMIENTO GEOESPACIAL PARA LA SEGMENTACIÓN DE UNA POBLACIÓN DE VIVIENDAS[1]

Resumen:

Generalmente, las encuestas por muestreo realizadas por los organismos oficiales de estadística se valen de un marco muestral que lista y agrupa a viviendas particulares según su ubicación geográfica, de acuerdo con los niveles de desagregación requeridos por el diseño. Los mismos son elaborados a partir de información recogida en los Censos de Población y mediante la asistencia de software cartográfico. La dificultad de este procedimiento no solo se debe al exhaustivo trabajo manual, sino también a imprecisiones respecto a calles y numeración de las viviendas.

Por motivos relacionados a la precisión de los estimadores y de distribución de carga de trabajo es preferible que la cantidad de viviendas por área de muestreo sea uniforme. Este trabajo presenta un algoritmo de agrupamiento espacial que permite especificar tamaño de clusters de antemano y que solo requiere posicionamiento dado por latitud y longitud. Se muestra su desempeño y se lo compara con el método de clustering k-medias en una aplicación a una población sintética de viviendas.

Palabras claves: cluster k-medias, segmentación, conglomerados de igual tamaño

Abstract:

Generally, sampling surveys carried out by official statistical agencies use a sampling frame that lists and groups households according to their geographical location, satisfying the disaggregation levels required by the design. They are elaborated from information collected in Population Censuses and through the assistance of cartographic software. The difficulty of this procedure is not only due to the exhaustive manual work, but also to inaccuracies regarding streets and numbering of homes.

For reasons related to the accuracy of the estimators and the distribution of workload, it is preferable that the number of dwellings per sampling area be uniform. This paper presents a spatial clustering algorithm that allows us to specify size of clusters in advance and only requires positioning given by latitude and longitude. Its performance is shown and compared with the k-means clustering method in an application to a synthetic household population.

Keywords: k-means clustering, segmentation, clusters of equal size

[1] Este trabajo se elaboró en el marco del Proyecto 1ECO199 titulado “Métodos Estadísticos en el Ámbito Oficial”, dirigido por Gonzalo Marí

1. Introducción

Los organismos oficiales de estadística utilizan diseños muestrales en varias etapas para obtener muestras probabilísticas y estimar características de interés de la población. Estos esquemas muestrales generalmente requieren seleccionar unidades espaciales en alguna de sus etapas. Un posible diseño está dado por la selección de regiones geográficas conformadas por localidades, luego un conjunto de áreas compuestas por viviendas, dentro de estas localidades y finalmente, se seleccionan viviendas particulares dentro de estas áreas.

Para obtener una muestra en las condiciones mencionadas, es indispensable contar con un marco muestral que liste y agrupe a viviendas particulares según su ubicación geográfica, de acuerdo con los niveles de desagregación requeridos por el diseño. En general, los Censos de Población constituyen el punto de partida para la construcción de los mismos, dado que cuentan con un listado exhaustivo de viviendas y con información referida a las características de las mismas, y de los hogares y personas que forman parte de ellas. La dificultad generalmente ocurre con respecto a la ubicación geográfica de las viviendas que en muchos casos resulta ser poco precisa por ausencia de nombres de calles y numeración.

En general, resulta deseable que el tamaño de las áreas seleccionadas para realizar el muestreo final de viviendas sea uniforme. Entre los motivos de este requerimiento, se puede mencionar en primer lugar que los diseños que consideran conglomerados de igual tamaño permiten tener una mejor precisión si las medias de las variables bajo estudio resultan ser similares. Por otra parte, y debido a que los marcos muestrales son utilizados para la selección de muestras para Encuestas a Hogares en los períodos intercensales, resulta conveniente por cuestiones operativas que los mismos sean de tamaños similares, para tener una coordinación del trabajo de campo similar en cada una de las áreas.

En la actualidad, gran parte de la elaboración de áreas o segmentos de muestreo se realiza utilizando softwares cartográficos. Por un lado, esto permite que las áreas construidas tengan características deseadas en cuanto al agrupamiento, pero demandan mayor esfuerzo y tiempo.

En el caso de contar con información referida a la ubicación geográfica tal como latitud y longitud, es posible utilizar métodos de agrupamiento de viviendas con ubicaciones cercanas. Los algoritmos de clustering de datos más conocidos, como k-medias, agrupan observaciones basados en un conjunto de variables y su similaridad reflejada en alguna medida de distancia propuesta. Si bien este método permite especificar de antemano la cantidad de clusters a obtener, no permite determinar los tamaños de cada uno de ellos. Esto hace que el método de k-medias no sea satisfactorio para crear una cantidad específica de áreas de muestreo de igual tamaño.

En este trabajo se considera un algoritmo de agrupación de unidades basados en su ubicación espacial. El mismo no solo permite especificar la cantidad de conglomerados a obtener, sino que también permite que estos respeten tamaños especificados de antemano.

El método expuesto en el presente trabajo podría ser utilizado tanto por organismos estadísticos para la obtención de áreas de muestreo con tamaños uniformes, como también por privados que tengan que resolver un problema de distribución de carga de trabajo en una ciudad, provincia, etc., o cualquier interesado en formar segmentos de unidades distribuidas en el espacio que deban respetar tamaños específicos.

2. Metodología

A continuación se presentan dos metodologías orientadas a la construcción de conglomerados de unidades basados en la similitud de un conjunto de variables que caracterizan a las mismas.

En primer lugar, el procedimiento cluster basado en el algoritmo de las k-medias, y luego un procedimiento iterativo que permite la construcción de clusters de tamaños aproximadamente iguales.

2.1. Procedimiento cluster k-medias

El algoritmo de las k-medias es uno de los métodos de cluster más populares. Desarrollado para situaciones en las cuales todas las variables son del tipo cuantitativo, y la distancia cuadrática Euclídea es elegida, generalmente, como medida de diferencia.

El criterio es asignar las n observaciones a los k clusters de modo que dentro de cada cluster el promedio de las diferencias de cada observación a la media del cluster, definido por los puntos del cluster, sea mínima.

En otras palabras, dado un conjunto de observaciones $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$ donde cada una es un vector real p-dimensional, el algoritmo de cluster k-medias busca particionar las n observaciones en k conjuntos $\boldsymbol{S} = \{S_1, S_2, \dots, S_k\}$ de forma tal que la suma de cuadrados dentro de cada grupo sea mínima.

$$\arg\min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{\boldsymbol{x_j} \in S_i} \|\boldsymbol{x_j} - \boldsymbol{c_i}\|^2$$

donde $\boldsymbol{c_i}$ es el centroide de los puntos en S_i. Generalmente, $\boldsymbol{c_i} = \boldsymbol{\mu_i}$ media de los puntos en S_i.

Existen diferentes algoritmos para implementar clustering por k-medias. Algunos de ellos son los propuestos por Forgy & Lloyd (Lloyd, 1957; Forgy, 1965), MacQueen (1967), y Hartigan & Wong (1979). Estos dos últimos se diferencian del primero en que no solo recalculan los centroides cada vez que se termina una iteración, sino que también lo hacen cada vez que se mueve una observación de un cluster a otro.

Si bien es de esperar que k-medias forme segmentos de unidades homogéneas en el espacio al menos cuando el algoritmo converge, no se puede controlar el tamaño de los grupos conformados. Esta característica del método de k-medias no concuerda con el propósito de lograr una cantidad determinada de segmentos, con tamaños establecidos de antemano.

2.2. Cluster de igual tamaño

Se propone utilizar un algoritmo cuyo objetivo es crear agrupamientos espaciales con tamaños específicos, iguales o no.[2]

Este método requiere contar con variables que, para el caso particular considerado en este trabajo, determinen la posición en el espacio de cada unidad a agrupar. El algoritmo calcula la distancia entre unidades en base a la latitud y longitud de las mismas.

La propiedad del método que asegura clusters de igual tamaño está sujeta a que la razón entre la cantidad de unidades a agrupar y la cantidad de grupos sea un número entero. En caso contrario, al menos uno de los agrupamientos tiene un tamaño distinto al resto. En esta

[2] Spatial clustering with equal sizes. Recuperado de https://statistical-research.com/index.php/2013/11/04/spatial-clustering-with-equal-sizes/

situación, el resultado esta dado por agrupamientos espaciales cuyos tamaños difieren a lo sumo en una unidad, y cuyas diferencias con el tamaño deseado esta acotada por el resultado de dividir al resto de la razón anteriormente mencionada por el número de grupos especificados. Si se quiere obtener clusters de tamaños distintos, estos deben ser explícitamente especificados de antemano.

Esta implementación particular tiene en cuenta el hecho que el planeta tierra no tiene un radio constante. La tierra se encuentra girando sobre su propio eje a una velocidad tan rápida, que se achata levemente. La función de distancia propuesta contempla este asunto y calcula las distancias de una manera acorde.

Además, dado que la tierra es mayormente redonda, la medida de distancia que se proponga debe tener en cuenta que las unidades no se encuentran sobre una superficie plana, sino curva.

Por simplicidad de notación, sean

$$x_i = Latitud\ del\ punto\ i-ésimo$$
$$y_i = Longitud\ del\ punto\ i-ésimo$$

Se definen dos puntos cualesquiera en el espacio, ubicados mediante latitud y longitud, a partir de

$$P_1 = (x_1, y_1)\ ;\ P_2 = (x_2, y_2)$$

El factor que contempla achatamiento y curvatura de la tierra está dado por el radio geocéntrico, función de la latitud , el radio ecuatorial y el radio polar[3]

$$a = 3963{,}191\ km\ \ (radio\ ecuatorial)$$
$$b = 3949{,}903\ km\ (radio\ polar)$$

$$r = \sqrt{\frac{\left(a^2 \cos\left(\frac{x_2\pi}{180}\right)\right)^2 + \left(b^2\ \mathrm{sen}\left(\frac{x_2\pi}{180}\right)\right)^2}{\left(a \cos\left(\frac{x_2\pi}{180}\right)\right)^2 + \left(b\ \mathrm{sen}\left(\frac{x_2\pi}{180}\right)\right)^2}}$$

Siguiendo la ley esférica de los cosenos, la distancia en kilómetros entre dos puntos en la tierra resulta

$$d = r \times \mathrm{acos}\left(sen(x_1)sen(x_2) + \cos(x_1)\cos(x_2)\cos(y_2 - y_1)\right)$$

Los pasos del algoritmo se presentan a continuación:

1- Definir la cantidad de grupos a construir y asignarle su tamaño correspondiente. Por ejemplo $m_k = \frac{n}{k}\ \ \forall k$ cuando se desean grupos de igual tamaño y el resto de la división es nulo.

2- Realizar asignación inicial de las unidades a los clusters. En esta propuesta a cada unidad se le asigna un cluster al azar.

3- Calcular el centro de cada grupo, en base a latitud y longitud.

4- Tomar la primera unidad del listado y asignarla al cluster cuyo centro esté más cercano.

[3] Equatorial Radius of the Earth.
Recuperado de http://maia.usno.navy.mil/NSFA/NSFA_cbe.html#EarthRadius2009

Si el centro más próximo es el del cluster al que la unidad ya pertenece, no se produce intercambio.

5- En el caso de existir intercambio, el cluster receptor tiene $m_k + 1$ unidades mientras que el cluster al que la unidad pertenecía tiene $m_{k^*} - 1$. Luego, al cluster cuyo tamaño se vio reducido, se le asigna una unidad de otro cluster que se encuentre más cercana a su centro.

6- Repetir el proceso desde 4 a lo largo de cada una de las unidades. Cada paso por cada una de las unidades a agrupar conforma una iteración.

7- Calcular la suma de distancias de cada unidad al centroide del grupo asignado.

8- Repetir desde 3 hasta que la disminución en la suma de distancias sea menor a un límite asignado o se alcance el número máximo de iteraciones.

2.2.1. Características del método cuando se desean obtener áreas de igual tamaño

Supongamos que se cuenta con un conjunto de n unidades georreferenciadas, y se desea distribuirlos en grupos de igual tamaño, m. Luego, la cantidad de grupos está dada por

$$k = max\{x \in \mathbb{N} \mid x \times m \leq n\}$$

Esto significa que se obtienen grupos de igual tamaño solo cuando el resto de la división $\frac{n}{m}$ es nulo.

En el caso que el resto no sea nulo, es decir $r = n - k \times m \neq 0$, esta cantidad es distribuida lo más uniformemente posible entre los k grupos. En otras palabras, cada grupo está compuesto por $m + \frac{r}{k}$ unidades cuando $\frac{r}{k}$ es entero.

En el caso que la cantidad $d = \frac{r}{k}$ no sea entero es imposible que los tamaños de los grupos sean $m + d$. En esta situación, hay k_1 grupos cuyo tamaño es $m + d_1$, con $d_1 = [d]$; y hay k_2 grupos cuyo tamaño es $m + d_2$ con $d_1 = [d] + 1$.

Los valores de k_1 y k_2 son aquellos que satisfacen la siguiente ecuación:

$$\frac{k_1}{k} d_1 + \frac{k_2}{k} d_2 = d$$

Notando que $k_2 = k - k_1$

$$\frac{k_1}{k} d_1 + \frac{k - k_1}{k} d_2 = d$$

$$\frac{k_1}{k} d_1 + \frac{-k_1}{k} d_2 + d_2 = d$$

$$\frac{k_1}{k} (d_1 - d_2) + d_2 = d$$

$$\frac{k_1}{k} ([d] - [d] - 1) = d - d_2$$

$$-\frac{k_1}{k} = d - d_2 \Rightarrow k_1 = (d_2 - d)k$$

Luego, para obtener k_2 simplemente hacer

$$k_2 = k - k_1$$

3. Aplicación

El algoritmo de agrupamiento presentado en este trabajo fue utilizado para elaborar áreas de muestreo dentro de cada ciudad del Estado de Texas. El objetivo de la aplicación fue crear áreas cuyos tamaños sean homogéneos y en lo posible igual a 200.

En esta aplicación particular, se contaba con un listado de 8.921.047 viviendas con su posición en latitud y longitud, y una variable que indica su pertenencia a uno de los 254 condados de Texas. Esta base de datos se corresponde con la población sintética elaborada por RTI Internacional para el Instituto Nacional de Salud de Estados Unidos (NIH, por sus siglas en inglés) (Wheaton et al, 2009)

Dado que el objetivo es crear áreas dentro de cada ciudad, y puede haber más de una ciudad por condado, fue necesario asignarle a cada vivienda una ciudad de pertenencia dentro de cada condado. Para ello, se utilizó la base de ciudades de Estados Unidos que viene incluida en el software SAS 9.4, que contenía a todas las ciudades del Estado de Texas en conjunto con un centro de georreferenciación en términos de latitud y longitud. Dentro de cada condado, se calculó la distancia de cada una de las viviendas a cada uno de los centros de todas las ciudades de ese condado y se le asignó a cada vivienda la ciudad a la que su centro estaba más próximo.

Esta forma de asignar viviendas a ciudades funciona bien cuando las ciudades están claramente separadas, pero puede ser imprecisa cuando éstas sean contiguas. Según la forma que tome cada ciudad y la cercanía con sus ciudades vecinas, puede ser que la distancia al centro de la ciudad vecina sea menor que la distancia al centro de la ciudad a la que realmente pertenece esta vivienda. En este caso, se incurre en un error de asignación porque se le asigna la ciudad a la que su distancia al centro es menor, es decir, la ciudad vecina.

En esta aplicación, sean

$$n_i = cantidad\ de\ viviendas\ en\ la\ ciudad\ i - ésima$$

$$k_i = cantidad\ de\ grupos\ en\ la\ ciudad\ i - ésima$$

Idealmente,

$$m_{ij} = \frac{n_i}{k_i} = cantidad\ de\ viviendas\ en\ el\ grupo\ j - ésimo\ de\ la\ ciudad\ i - ésima = 200$$

y

$$k_i = max\{x \in \mathbb{N} \mid x200 \leq n_i\} \wedge r = n_i - k_i 200 = 0$$

Sin embargo, cuando $r = n_i - k_i 200 \neq 0 \Rightarrow m_{ij} > 200\ para\ al\ menos\ un\ i$.

Por ejemplo, en una ciudad con 1300 viviendas $k_i = 6$, $r = 1300 - 6{\times}200 = 100$. Si dividimos al resto por la cantidad de grupos, resulta $\frac{r}{k_i} - = \frac{50}{3} = d \cong 16.67$. Como el cociente del resto con la cantidad de grupos no es nulo ni entero, la diferencia entre el tamaño deseado y el tamaño logrado para cada grupo estará dada en algunos casos por $d_1 = [d] = 16$, y en otros por

$d_2 = [d] + 1 = 17$. Luego, los tamaños que se obtienen serán 216 y 217. La cantidad de grupos que se corresponde con cada tamaño, k_1 y k_2 es aquella que satisface:

$$\frac{k_{1i}}{k_i} d_1 + \frac{k_{2i}}{k_i} d_2 = d$$

De allí que $k_{2i} = k_i - k_{1i} = 6 - k_{1i}$.

Siguiendo lo hallado en la sección anterior,

$$k_{1i} = (d_{2i} - d_i)k_i = \left(17 - \frac{50}{3}\right)6 = \left(\frac{1}{3}\right)6 = 2$$

$$k_{2i} = 6 - k_{1i} = 6 - 2 = 4$$

Finalmente, en este ejemplo se obtienen 2 grupos con 216 hogares, y 4 grupos con 217.

Se puede corroborar que $216 \times 2 + 217 \times 4 = 1300$

Tomando como base al conjunto de viviendas georreferenciadas y ubicadas dentro de las diferentes ciudades del estado de Texas, se aplicaron los dos algoritmos introducidos en el presente trabajo para construir conglomerados que constituyen las unidades de muestreo, y comparar su capacidad de producir conjuntos de tamaños uniformes y similares a los preestablecidos.

Dado que en el caso de k-medias no se puede especificar la cantidad de unidades que corresponden a cada cluster, solo se pudo utilizar el algoritmo especificando al valor de k que resulta de la división entera de la cantidad de viviendas en la ciudad por el tamaño deseado, que es 200. Luego, la cantidad de unidades por segmento es determinada por el algoritmo.

Es importante notar que la cantidad de unidades de muestreo a formar por ciudad es independiente del algoritmo aplicado. Consecuentemente, la comparación de la bondad de los algoritmos se basa en los tamaños resultantes de las áreas y no la cantidad de áreas en sí.

Además de utilizar rango, máximos, mínimos y percentiles, una forma posible de medir y comparar el beneficio que resulta de aplicar algoritmos automáticos de agrupamiento de unidades espaciales es mediante el cálculo de la cantidad de áreas con tamaños cuya diferencia con el tamaño deseado se encuentra dentro de límites aceptables.

Es válido argumentar que aquellos segmentos cuyos tamaños sean aceptables son definitivos, mientras que los segmentos cuyos tamaños estén por debajo o por encima de estos limites deberán ser revisados manualmente para elaborar las áreas de forma que los tamaños resultantes estén en concordancia con los propuestos en el plan de muestreo. A mayor porcentaje de áreas aceptables, menor trabajo manual y mayor ahorro en tiempo y esfuerzo.

En la aplicación de k-medias y el algoritmo propuesto se obtienen un total de 43796 áreas de muestreo en todo el estado de Texas. Se compara su capacidad para lograr segmentos uniformes en 4 escenarios distintos.

Tabla 1: Medidas resúmenes de la distribución del tamaño de las áreas para ambos métodos

Algoritmo	Mínimo	Q1	Q2	Q3	Máximo
k-medias	1	62	160	266	743
Nueva propuesta	9	200	201	202	623

Tabla 2: Escenario 1 - Se aceptan áreas con tamaños en el intervalo [120, 280]

Algoritmo	Áreas con tamaños insuficientes.	Áreas con tamaños aceptables.	Áreas con tamaños excesivos.
k-medias	17577 (40,14%)	17524 (40,01%)	8695 (19,85%)
Nueva propuesta	41 (0,09%)	43506 (99,33%)	249 (0,58%)

Tabla 3: Escenario 2 - Se aceptan áreas con tamaños en el intervalo [150, 250]

Algoritmo	Áreas con tamaños insuficientes.	Áreas con tamaños aceptables.	Áreas con tamaños excesivos.
k-medias	20091 (45,87%)	11430 (26,10%)	12275 (28,03%)
Nueva propuesta	54 (0,12%)	43379 (99,05%)	363 (0,83%)

Tabla 4: Escenario 3 - Se aceptan áreas con tamaños en el intervalo [180, 220]

Algoritmo	Áreas con tamaños insuficientes.	Áreas con tamaños aceptables.	Áreas con tamaños excesivos.
k-medias	25587 (58,42%)	3401 (7,77%)	14808 (33,81%)
Nueva propuesta	69 (0,16%)	42332 (96,66%)	1395 (3,18%)

Tabla 5: Escenario 4 - Se aceptan áreas con tamaños en el intervalo [190, 210]

Algoritmo	Áreas con tamaños insuficientes.	Áreas con tamaños aceptables.	Áreas con tamaños excesivos.
k-medias	26939 (61,51%)	1732 (3,95%)	15125 (34,54%)
Nueva propuesta	75 (0,17%)	40540 (92,57%)	3181 (7,26%)

En un primer lugar, se observa de la Tabla 1 que los tamaños de las áreas tienen valores extremos alejados del valor objetivo. Sin embargo, los valores de Q1, Q2 y Q3 evidencian que al menos el 50% central de las áreas formadas tiene un tamaño entre 200 y 202 para el método propuesto, mientras que el algoritmo de k-medias presenta valores entre 62 y 266, indicando gran variabilidad en el tamaño de los segmentos incluso para el 50% central de los datos.

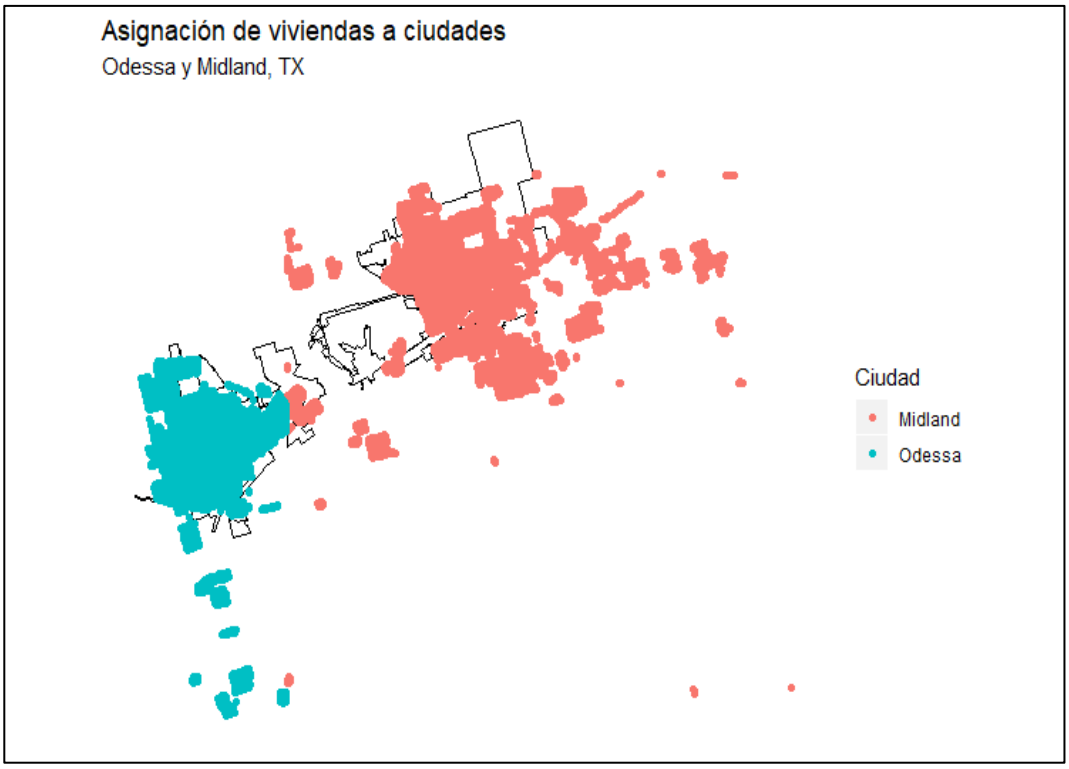
Es relevante notar de las Tablas 2 a 5 que incluso en el escenario mas exigente respecto de los tamaños, el algoritmo propuesto construye áreas de forma tal que el porcentaje de áreas que deben ser revisadas manualmente es menor al 10%. Por otro lado, el algoritmo de k-medias implica revisar manualmente el 60% de los grupos formados incluso en el escenario mas permisivo respecto de los tamaños aceptables.

Si el trabajo propuesto requiere que el tamaño de las áreas sea muy cercano al tamaño especificado, como es el escenario 4, aplicar el algoritmo de k-medias para elaborar áreas no implica ninguna reducción importante en el trabajo manual. Por ejemplo, en esta aplicación,

solo el 4% de las áreas formadas por k-medias resultan aceptables bajo este escenario mas exigente. Contrariamente, el algoritmo propuesto resulta en un 92% de áreas con tamaños aceptables en este mismo contexto.

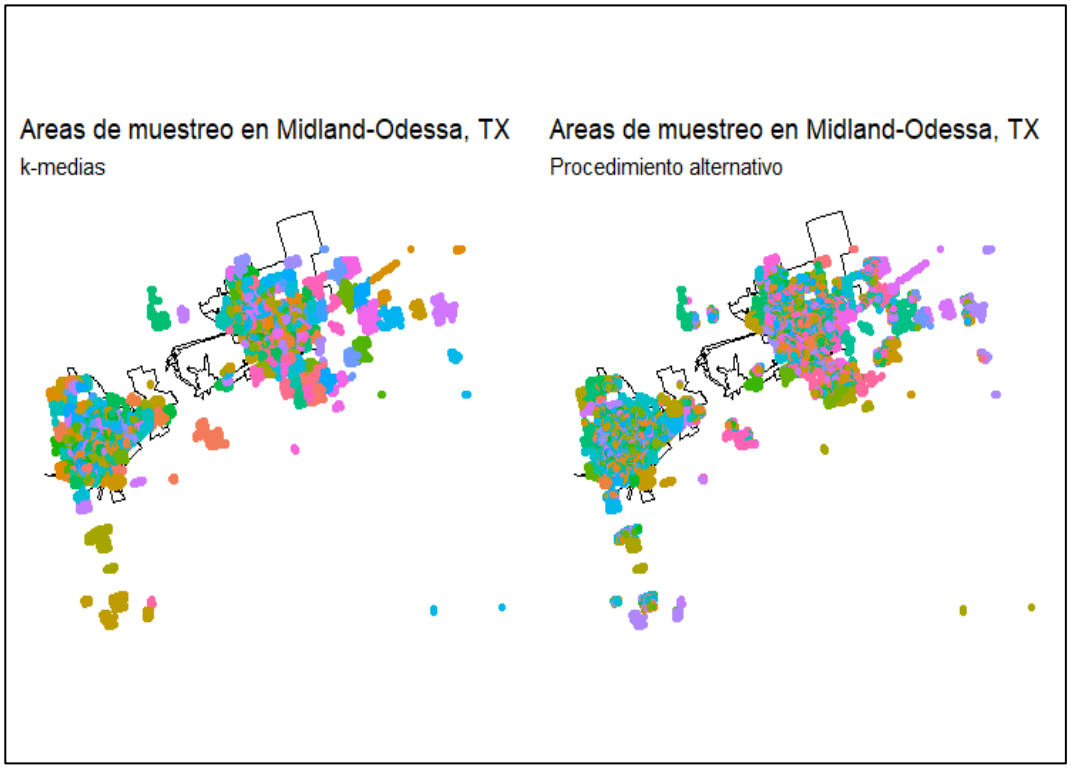
A continuación se presentan, a modo de ejemplo, la asignación de viviendas a ciudades y a la conformación de los conglomerados a través de los dos métodos. Cabe destacar que estos ejemplos corresponden a ciudades pequeñas donde es posible visualizar el trabajo desarrollado.

Gráfico 1: Asignación de viviendas a dos ciudades vecinas.



En el Gráfico 1 se puede observar que la asignación de viviendas a sus respectivas ciudades no fue perfecta porque hay algunas viviendas asignadas a Midland, cuando en realidad se encuentran dentro de los límites de Odessa.

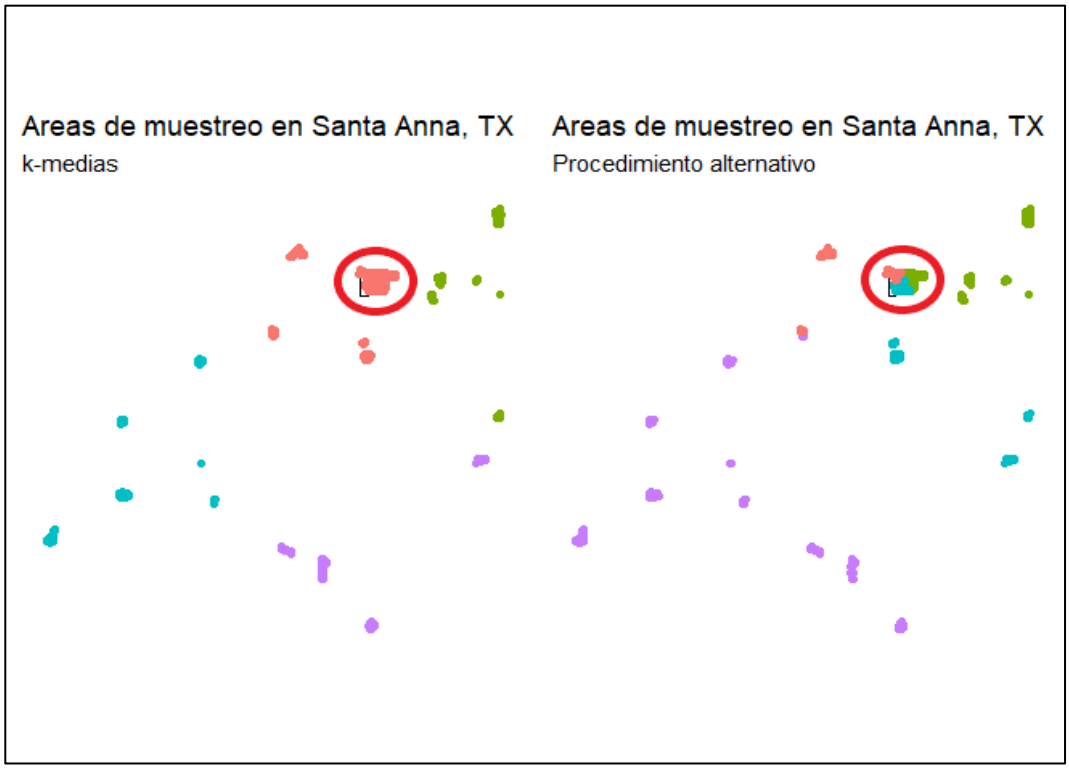
Gráfico 2: Elaboración de áreas de muestreo en dos ciudades vecinas mediante ambos algoritmos.



De acuerdo con el Gráfico 2, ambos algoritmos de agrupamiento elaboran áreas con unidades que se encuentran próximas. Sin embargo, se observa que las áreas elaboradas por el algoritmo de k-medias presentan tamaños variables. Dado que este método busca minimizar suma de cuadrados intragrupo sin ninguna restricción, el mismo tiende a agrupar en un mismo conglomerado a aquellas unidades alejadas del resto, sin importar el tamaño resultante. Si bien esta característica permite identificar mediante el agrupamiento que tales unidades están alejadas del resto, no se corresponde con el objetivo de elaborar áreas de tamaños similares.

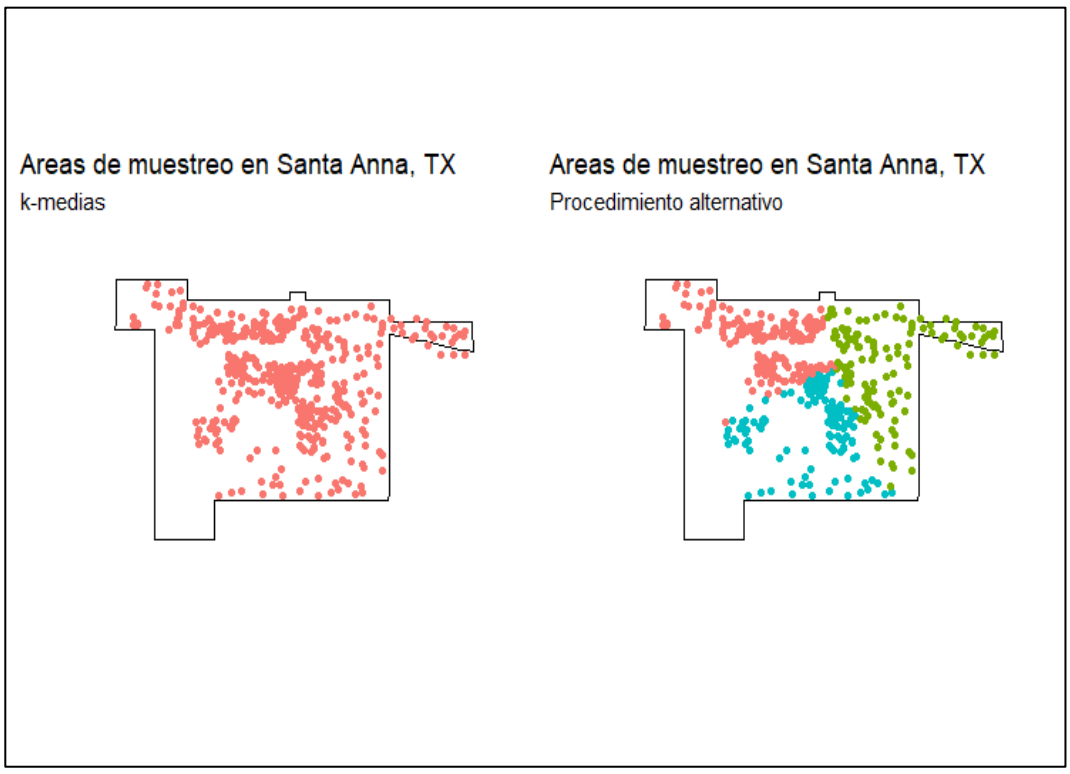
El Gráfico 3 muestra a la ciudad de Santa Anna junto a las viviendas fuera de sus límites geográficos, pero asignadas a esta ciudad por estar en sus cercanías. Ambos algoritmos producen 4 áreas de muestro, ya que no depende del método aplicado.

Gráfico 3: Elaboración de áreas de muestreo Santa Anna y región aledaña.



Se puede observar que k-medias tiende a agrupar a las viviendas dentro de los límites de la ciudad en un único segmento, mientras que el algoritmo propuesto crea 4 grupos de tamaño homogéneo, y las viviendas dentro de los limites de la ciudad se dividen en 3 de estos segmentos.

Gráfico 4: Elaboración de áreas de muestreo en Santa Anna.



4 – Conclusiones y estudios futuros

En primer lugar, se debe notar que en esta aplicación particular se convive con el posible error cometido al asignar ciudades a las viviendas, ya que la base de datos utilizada carecía de dicha clasificación. En el caso que el interés esté no solo en evaluar la calidad de los agrupamientos, sino también en la bondad de la asignación de unidades a regiones geoespaciales, se deberá probar con otra técnica para realizar esta asignación inicial o bien contar con un listado previamente clasificado.

Por otro lado, el algoritmo de agrupamiento propuesto asigna cada unidad a un conglomerado inicial de manera completamente aleatoria. Esta alternativa, posibilita que los grupos configurados como consecuencia de la asignación inicial contengan unidades totalmente disimiles en términos de posicionamiento espacial. Consecuentemente, esto requiere un mayor número de iteraciones para lograr un agrupamiento satisfactorio.

Resulta interesante investigar métodos de asignación inicial que consideren de alguna forma las coordenadas de localización. Por ejemplo, se podría ordenar el listado de las unidades de acuerdo a su latitud y/o longitud y asignar a cada bloque de n unidades ordenadas al mismo grupo. Luego, se debería comparar tanto la similitud de los agrupamientos construidos con esta asignación inicial y los agrupamientos construidos con la asignación al azar, como sus tiempos de cómputo.

Otro aspecto para considerar es el cálculo de la distancia entre dos puntos en la tierra. La

fórmula de cálculo de distancia está dada por la ley esférica de los cosenos. Sería relevante comparar los tiempos de cómputo al utilizar distancias que ignoren la esfericidad de la tierra pero que recurran a fórmulas de cálculo que requieren menos tiempo. Existen propuestas que sugieren utilizar la distancia Euclídea argumentando que la perdida en precisión es depreciable respecto a la ganancia en simplicidad de cómputo, cuando los puntos están cercanos en la tierra.[4]

Dado que el criterio de convergencia del algoritmo depende de la cercanía de las unidades de un mismo grupo a su centro, es de esperar que dos unidades de un mismo grupo se encuentren a una distancia menor que dos unidades de dos grupos cualesquiera. Sin embargo, cuando el número de iteraciones necesarias para lograr esa homogeneidad interna es mayor al límite especificado, no se puede asegurar que el agrupamiento espacial sea satisfactorio. A mayor número de unidades agrupar, peor asignación inicial y mayor cantidad de grupos, hay un mayor tiempo de cómputo por iteración. Es importante establecer un número de iteraciones límite que permita lograr agrupamientos satisfactorios en la mayor cantidad posible de ciudades, pero que no extienda el tiempo de cómputo para una misma ciudad de forma tal que este sea impracticable. En este último caso, el algoritmo no da garantías de la bondad del agrupamiento.

El algoritmo utilizado solo se vale de la latitud y longitud de cada una de las unidades para lograr el agrupamiento. Sin embargo, puede ser que dos unidades que estén cerca espacialmente deban corresponderse a diferentes áreas por pertenecer a diferentes regiones geográficas cuya categorización no esté contemplada. Este algoritmo solo permite establecer restricciones respecto a tamaño, pero no límites geoespaciales.

Idealmente sería bueno un algoritmo pueda hacer variar levemente el tamaño del área especificado si esto implica una gran reducción en la variabilidad de las localizaciones de las unidades de un mismo grupo. Esto evitaría que unidades muy alejadas pertenezcan al mismo grupo, solo por la necesidad de cumplir con una restricción de tamaño.

REFERENCIAS BIBLIOGRÁFICAS

Forgy, E.W. (1965). Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics*, 21, 768-780.

Hartigan, J.A., Wong, M.A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*. Series C (Applied Statistics), Vol. 28, No. 1, pp. 100-108

Lloyd, S.P. (1957). Least squares quantization in PCM. *Bell Telephone Laboratories Paper*.Publicado más tarde en la revista: Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* IT-28: 129–137.

MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bsmsp/1200512992

Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., Roberts, D.J., Allpress, J.L. (2009). *Synthesized Population Databases: A US Geospatial*

[4] Geographic distance can be simple and fast. http://jonisalonen.com/2014/computing-distance-between-coordinates-can-be-simple-and-fast/

Database for Agent-Based Models. RTI Press. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875687/

FUENTES

Población sintética correspondiente al estado de Texas. Wheaton, W.D., Cajka, J.C., Chasteen, B.M., Wagener, D.K., Cooley, P.C., Ganapathi, L., Roberts, D.J., Allpress, J.L. (2009). *Synthesized Population Databases: A US Geospatial Database for Agent-Based Models.* RTI Press. Recuperado de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875687/