ASOCIACION ARGENTINA
DE ECONOMIA POLITICA

LVI REUNIÓN ANUAL| NOVIEMBRE DE 2021

# Métodos de clustering espacialmente restringidos: Un análisis al agrupamiento por nivel de estudio en la provincia de Mendoza

Quintana, Pablo Aníbal

ISSN 1852-0022

# Métodos de *clustering* espacialmente restringidos: Un análisis al agrupamiento por nivel de estudio en la provincia de Mendoza

Pablo Quintana*

**Resumen**

Los avances computacionales en el análisis de datos cada vez ofrecen más herramientas para el estudio de fenómenos sociales. En este trabajo se exploran los avances de aprendizaje automático, particularmente, el aprendizaje no supervisado que no requiere una variable dependiente o de salida para arrojar resultados. Analizando el fenómeno de segregación residencial, se presentan metodologías de *clustering* espacialmente restringidas para determinar regiones de agrupamiento de personas con estudios superiores y personas que no lograron completar estudios básicos. Adicionalmente, se trata de determinar la significatividad de las regiones para saber si realmente se pueden considerar la existencia de segregación regional en dichas zonas, utilizando de indicador el coeficiente de localización.

**Clasificación JEL:** C18, C21, R12.
**Palabras Clave:** Clustering espacial, Coeficiente de Localización, Segregación urbana.

*Facultad de Ciencias Económicas, UNCUYO, Doctorado en Economía; Centro Universitario (M5502JMA), Ciudad de Mendoza, Argentina; email: pabanib@hotmail.com

# 1. Introducción

En economía urbana, el análisis de segregación residencial ha sido un área de creciente preocupación debido a que su dinámica puede finalizar en la formación de guetos o dar lugar a disturbios sociales y enfrentamientos entre grupos antagónicos. Una particularidad sobresaliente de este fenómeno es su naturaleza espacial o geográfica ya que puede concentrarse localmente en áreas urbanas específicas y delimitadas. Tradicionalmente, la medición de la segregación incluye un conjunto de indicadores que pueden dividirse en cinco dimensiones relevantes: similitud, exposición, concentración, centralización y agrupamiento (Massey y Denton 1988). De estas dimensiones, las últimas tres poseen características espaciales o geográficas tal que pueden desagregarse a nivel local, sin embargo, no se desarrollaron hasta recientemente.

Desde inicios del siglo XXI, los estudios de Reardon y O'Sullivan (2004), Wong (2005) y, más cercano temporalmente, Oka y Wong (2014), profundizaron la investigación sobre índices de segregación locales, considerando que las unidades espaciales con las que se calculan los índices globales no son unidades aisladas, si no que las personas se relacionan con sus vecinos. De esta manera logran avances en indicadores tales como disimilitud, entropía, aislamiento (Oka y Wong 2014) y algunos indicadores multigrupos como son el índice de información teórica espacial y el índice de diversidad relativa espacial (Reardon y O'Sullivan 2004). Estos índices se calculan considerando una matriz binaria $C$, de dimensión $n \times n$ donde $n$ son las unidades espaciales, que relaciona cada unidad espacial con todas las unidades que son vecinas tal que $c_{ij} = 1$ si $i$ y $j$ son vecinas, y $c_{ij} = 0$ si $i$ y $j$ no lo son. Estos índices parten de la definición que brinda Anselin (1995) de indicadores locales denominados LISA. Como mencionan (Garrocho y Campos-Alanís 2013) los indicadores locales dan solución a tres problemas clásicos de los indicadores globales de segregación, como son: el problema del tablero de ajedrez, el problema de la unidad espacial modificable y la falta de confiabilidad estadística.

Respecto al avance de las metodologías de *clustering*, Sabatini (2006) sostiene que la realidad se presenta agrupada en aglomerados urbanos y los estudios de segregación tratan de identificar características de dichos aglomerados. Por lo tanto, aplicar metodologías de *clustering* y sobre todo las que consideran la restricción espacial, parecen ser una buena herramienta para detectar patrones de segregación no considerados previamente. En esta dirección avanzan estudios como Openshaw y Rao 1995 que trabajan con datos censales y Aguado-Moralejo, Echebarria y Barrutia (2019) que tratan de identificar tipos de barrios y describir características de los vecindarios previamente construidos con herramientas de regionalización.

Las metodologías de *clustering* se encuentran dentro de las herramientas de aprendizaje automático, más precisamente del aprendizaje no supervisado. Este tipo de aprendizaje pretende reconocer patrones entre variables para poder agrupar las unidades de análisis en una cantidad desconocida de grupos presentes en la población. Es decir, estas metodologías no cuentan con información sobre la cantidad de grupos, como si sucede bajo el aprendizaje supervisado, y se considera que este campo representa la mejor aproximación sobre cómo aprendemos los seres humanos (LeCun, Bengio e Hinton 2015). Como inconveniente, este tipo de problemas es considerado *NP-Hard* (coloquialmente, un sistema cuya complejidad algorítmica es intratable y su solución no es necesariamente óptima, véase Cook 2006), por lo cual no solo dificulta encontrar una solución, sino que también se vuelve complejo saber si la solución hallada es óptima.

El objetivo de este trabajo es realizar un recorrido por diferentes métodos de regionalización actuales que consideran la restricción de contigüidad geográfica y mostrar que son herramientas válidas para tratar problemáticas como la segregación. Estos métodos se aplican empíricamente al agrupamiento de personas, con y sin estudios superiores, en la región norte de la provincia de

Mendoza, en base a datos del Censo Nacional 2010. En una primera etapa se busca conseguir clústeres con la metodología que haya logrado el mejor rendimiento, y posteriormente, evaluar la segregación existente en los mismos con un índice local como es el coeficiente de localización, bien conocido en la literatura.

La estructura del trabajo es la siguiente. La segunda sección revisa las metodologías vigentes en el área de regionalización con restricción espacial. La tercera sección evalúa los clústeres obtenidos de la sección previa utilizando el coeficiente de localización y aplica una estrategia estadística para conocer la significancia de cada agrupamiento. La cuarta sección aplica estas metodologías a los datos censales de Mendoza considerando agrupaciones de personas, con y sin estudios superiores. Finalmente, la quinta sección recoge los comentarios finales.

# 2. Metodologías empleadas

Antes de presentar las diferentes metodologías propuestas, es necesario establecer una notación tal que permita armonizar los resultados de los diferentes métodos. Partiremos de la noción que las áreas u objetos espaciales son las unidades básicas de análisis y que contienen diferentes atributos a considerar.

Sea $A = \{A_1, A_2, ..., A_n\}$ el conjunto de todas las áreas con $n = |A|$.

Los atributos del área $i$-ésima se representan como $A_{iy}$, en donde $y \in Y = \{1, 2, ..., m\}$, siendo que cada área tiene un vector de atributos $x_i = \{a_{i1}, ..., a_{im}\}$ en dónde $a_{iy}$ es un posible valor del atributo $A_{iy}$; y sea $l_i$ un atributo espacial que se extiende a las áreas vecinas del área $A_i$.

Para establecer relaciones entre áreas definimos una función de disimilitud $d : A \times A \to \mathbb{R}^+ \bigcup \{0\}$ que se basa en el conjunto de atributos $Y$. Dicha función $d_{ij} = d(A_i, A_j)$ debe satisfacer que $d_{ij} \geq 0$, $d_{ij} = d_{ji}$, y $d_{ii} = 0$, $\forall i, j = 1, 2, ..., n$. Alternativamente, pueden definirse funciones de distancias.

Sea $G = (V, E)$ el grafo o red asociado con $A$ en dónde el vértice $v_i \in V$ corresponde al área $A_i \in A$ y el eje $\{v_i, v_j\} \in E$ si y solo sí las áreas $A_i$ y $A_j$ comparten un borde en común. Nótese que el grafo $G$, en este caso, es similar a la matriz $C$, definida en la sección previa, que relaciona con un 1 si $i$ y $j$ son áreas vecinas y con un 0 en caso que no lo sean.

Sea $P_p = \{R_1, R_2, ..., R_p\}$ una partición de las áreas $A$ dentro de $p$ regiones con $1 \leq p \leq n$. Denominamos a $\Pi$ como el conjunto de todas las posibles particiones de $A$.

## 2.1. *Clustering* no espacial

Los algoritmos de *clustering* tienen como objetivo reducir el número de observaciones. Básicamente tratan de agrupar $n$ observaciones en $p$ clústeres. Los grupos que así se forman tienen que mantener la mayor semejanza posible entre los elementos que se hayan en el interior y diferenciarse de los otros grupos. Estos objetivos se transmiten a los algoritmos a través de una minimización o maximización, depende el caso, de lo que se denomina función de disimilitud.

Existe una gran diversidad de métodos de *clustering* en la literatura y es difícil clasificarlos de alguna manera ya que comparten estrategias comunes. Si se desea realizar una clasificación genérica, puede decirse que existen métodos de partición, aglomeración y de densidad (una excelente revisión de esta clasificación puede verse en Hastie, Tibshirani y Friedman 2009). Nuestra intención es centrarnos en las metodologías de *clustering* para datos georreferenciados o con información espacial.

Existen alternativas simples para combinar métodos de *clustering* tradicionales y técnicas espaciales. Una de ellas es la implementada por (Anselin, Syabri y Kho 2010) mediante el programa GeoDa en donde se agregan las coordenadas geográficas como dos atributos adicionales a las

unidades espaciales, tal que la función objetivo considera esta información como cualquier otro atributo. Otra alternativa similar es la propuesta por Webster y Burrough 1972, Wise, Haining y Ma 1997, Haining, Wise y Ma 2000 y más recientemente Yuan y col. 2015; Cheruvelil y col. 2017; Chavent y col. 2018, en donde permiten diferenciar la información espacial mediante una ponderación diferente. En este caso, la información espacial está representada por una matriz de contigüidad que contiene pesos espaciales, $W$, marcando una diferencia respecto a los atributos regulares que se están estudiando. Sin embargo, en ambas estrategias, la inclusión de la información espacial no asegura que los grupos se encuentren espacialmente contiguos, algo que intentan garantizar los algoritmos de *clustering* espacial propiamente dichos y que veremos a continuación.

## 2.2. SKATER

El método $SKATER$ (*Spatial Kluster Analysis by Tree Edge Removal*) es un procedimiento espacial que busca clústeres locales de las variables estudiadas en el espacio. El método es relativamente eficiente y opera a través de árboles de decisión eliminando las relaciones de árboles que no son óptimas (ver Assunção y col. 2006).

Para el desarrollo de la técnica SKATER es necesario construir una matriz de pesos espaciales, por lo que es un algoritmo que considera la restricción espacial. En nuestro caso de aplicación, esto tiene importancia porque nos interesa saber si hay relación entre los radios vecinos para determinar los agrupamientos de personas. El método usa el árbol de expansión mínimo (*minimum spanning tree*, MST en siglas) y luego a través de métodos heurísticos se lo "poda" de acuerdo a la cantidad de regiones que se quieren construir, estableciendo así la relación entre unidades geográficas.

Más formalmente, dado un objeto espacial $A_i$ con un conjunto de atributos $\{A_{i1}, ..., A_{im}\}$, en donde todos los objetos tienen un vector de atributos $x_i = (a_{i1}, ..., a_{im})$, tal que $a_{i1}$ es un posible valor del atributo $A_{i1}$. La topología asociada es el grafo $G = (V, E)$ con un conjunto de vértices $V$ y un conjunto de ejes $E$. Dos vértices $v_i$y $v_j$ se conectan entre sí a través de un eje $(v_i, v_j)$ si las áreas $i$ y $j$ son adyacentes. A cada eje le vamos asociar una función de costo $d(i, j)$ que mide la disimilitud que existe entre el área $i$ y el área $j$ considerando sus respectivos vectores de atributos. En caso de que los atributos tengan escalas comparables podemos utilizar como función de disimilitud a la distancia euclídea. El MST es el grafo $G^*$ que contiene a todos los nodos $n$ de $G^*$ conectados por un solo eje por nodo y, además, minimiza el costo considerado a este como la suma de las disimilitudes de cada eje. Cada vez que se corte un eje, el árbol se dividirá en dos grafos.

El algoritmo cortará los ejes que logren optimizar su función objetivo. En este caso lo que se debe lograr es minimizar la suma de las desviaciones cuadradas dentro del clúster, esto es:

$$\text{mín } Q(P_p) = \sum_{i=0}^{p} SSD_i,$$

donde $P_p$ es una partición de los objetos en $p$ árboles y $SSD_i$ es la suma de desviaciones al cuadrado en la región $i$.

La desviación cuadrática dentro del clúster, $SSD_i$, es una medida de la dispersión de los valores de los atributos del objeto en la región. Si la región es homogénea, entonces presentará un pequeño valor del $SSD$. La fórmula sería así:

$$SSD_k = \sum_{j=1}^{m} \sum_{i=1}^{n_k} (x_{ij} - \overline{x}_j)^2,$$

donde $n_k$ es el número de objetos espaciales en el árbol $k$, $x_{ij}$ es el $j$-ésimo atributo del objeto espacial $i$, $m$ es el número de atributos considerados y $\overline{x}_j$ es el promedio de los valores del $j$-ésimo atributo en el árbol $k$.

En cada iteración se va removiendo un eje del grafo $G^*$. La decisión de qué eje remover surge de la siguiente función objetivo:

$$\oint_1 \left(S_l^T\right) = SSD_T - \left(SSD_{Ta} + SSD_{Tb}\right),$$

donde $S_l^T$ es el arreglo que se produce de dividir el árbol $T$ al cortar el eje $l$, $T_a$ y $T_b$ son los árboles que se producen de subdividir el árbol $T$.

## 2.3. REDCAP

Guo 2008 propone la utilización de lo que denomina REDCAP (*regionalization with dynamically constrained agglomerative clustering and partitioning*) que consiste en complementar metodologías comúnmente usadas en los métodos de *clustering* jerárquicos y un tratamiento diferencial de la contigüidad espacial. La base del método consiste en dos etapas: (1) agrupar los datos con árboles con restricción de contigüidad espacial y (2) cortar el árbol generado en regiones mientras se optimiza una función objetivo.

Los métodos jerárquicos implementados pueden ser diferentes y varían según la función de disimilitud que utilizan. Así tendremos la que utiliza el agrupamiento por enlace simple (SLK, por sus siglas en inglés):

$$d_{SLK}\left(R_k, R_{k\prime}\right) = min_{A_i \in R_k, A_j \in R_{k\prime}}\left(d\left(A_i, A_j\right)\right),$$

donde $R_k$ *y* $R_{k\prime}$ son dos regiones, $A_i$ y $A_j$ son dos áreas que perteneces a las respectivas regiones y $d\left(A_i, A_j\right)$ es la función de disimilitud entre ambas áreas.

Por otra parte, tenemos el agrupamiento por enlace promedio (ALK):

$$d_{ALK}\left(R_k, R_{k\prime}\right) = \frac{1}{\left|R_k\right|\left|R_{k\prime}\right|} \sum_{A_i \in R_k} \sum_{A_j \in R_{k\prime}} d\left(A_i, A_j\right),$$

donde $\left|R_k\right|$ *y* $\left|R_{k\prime}\right|$ son la cantidad de áreas que pertenecen a cada región.

Luego tenemos el agrupamiento por enlace completo (CLK):

$$d_{SLK}\left(R_k, R_{k\prime}\right) = max_{A_i \in R_k, A_j \in R_{k\prime}}\left(d\left(A_i, A_j\right)\right).$$

Para mantener la contigüidad espacial se utilizan dos estrategias: la de primer orden y la de orden completo. La estrategia de primer orden solo compara aquellas áreas pertenecientes a las regiones que compartan un borde en común. En el caso de la estrategia de orden completo se comparan todas las áreas pertenecientes a ambas regiones. Es decir, si tenemos dos regiones $R_1 = \{A, B, C\}$ y $R_2 = \{D, E\}$, y suponemos que sólo $A$, $B$ y $D$ comparten borde, entonces la estrategia de primer orden aplicará la función de disimilitud en estas tres áreas solamente. En cambio, en caso del orden completo, se compararán las cinco áreas pertenecientes a los dos grupos. La fusión, en ambos casos, se produce a nivel de región, por lo tanto, ambas estrategias mantienen la contigüidad espacial.

El método de enlace elegido y la estrategia de restricción espacial va formando un árbol que al finalizar el proceso tiene que quedar completamente conectado con diferentes jerarquías. Este árbol se puede graficar de una manera muy práctica a través de un dendograma. Luego de la

obtención del árbol, se procede a cortarlo basándose en la cantidad de grupos elegidos y utilizando una estrategia similar a la que se utiliza en el método SKATER. Podría decirse que la diferencia entre REDCAP y SKATER está en que este último utiliza el MST y REDCAP construye un árbol con las estrategias de los métodos jerárquicos.

## 2.4. AZP

El problema de la zonificación automatizada (*automatic zoning problem*) fue planteado por primera vez por Openshaw 1977. El mismo consiste en agregar los atributos de las $n$ unidades básicas de análisis o áreas de $A$, en $p$ zonas, considerando obtener el menor valor para una función objetivo determinada y, por supuesto, manteniendo la contigüidad espacial en base a la matriz $C$. Por lo tanto, el problema no deja de ser similar a todos los planteados en esta sección, la diferencia como en todos pasa por la forma de la solución. Este tipo de problemas es considerado *NP-hard* por lo que practicar una solución analítica no es posible y por lo tanto hay que recurrir a estrategias heurísticas para su solución.

Las heurísticas utilizadas parten de encontrar una posible solución inicial al problema, lo que convierte al método en sensible ante esta partición inicial $P_p^0$ y, por lo tanto, dificulta la tarea de encontrar alguna solución óptima global. Luego de obtenida $P_p^0$ se procederá a seleccionar alguna región $R_j \in P_p^0$ y se considerarán las áreas vecinas de $R_j$. El algoritmo ahora irá agregando las áreas vecinas a la región $R_j$ y comprobando si dicho intercambio provoca una disminución de la función objetivo, siempre considerando que no se rompa la regla de contigüidad como restricción fuerte. Este proceso se irá iterando entre las diferentes regiones que componen la partición. Cómo en la mayoría de las metodologías planteadas el número de regiones $p$ es elegido a priori de manera arbitraria. El algoritmo planteado puede caer fácilmente en una solución local, lo que es un impedimento para llegar a una solución global óptima. En Openshaw y Rao 1995 se plantean mecanismos de búsqueda local para evitar este tipo de problemas.

## 2.5. Max-p

Otro método propuesto es el denominado Max-p Duque, Anselin y Rey (2012). A diferencia de los métodos vistos anteriormente, en este método la cantidad de regiones es determinada de manera endógena, tratando siempre de minimizar la heterogeneidad intra-cluster. Esto permite tener una ventaja sobre otras metodologías ya que ahora la cantidad de regiones no son definidas arbitrariamente si no que se obtienen a través del propio mecanismo de cálculo.

Una posible partición de $A$ en la metodología Max-p debe cumplir con los siguientes requisitos:

$$|R_k| > 0 \;\; para\, k = 1, 2, ..., p,$$

$$R_k \bigcap R_{k'} = \emptyset \;\; para\, k, k' = 1, 2, ..., p\, y\, k \neq k',$$

$$\bigcup_{k=1}^{p} R_k = A,$$

$$\sum_{A_i \in R_k} l_i \geq umbral \begin{cases} para\, k = 1, 2, ..., p; \\ el\, umbral \in \mathbb{R}^+ \bigcup \{0\} | 0 \leq umbral \leq \sum_{A_i \in A} l_i, \end{cases}$$

$G(R_k)$ es un grafo conectado para $k = 1, 2, ..., p$.

A las posibles particiones $P_p \in \Pi$ las vamos a evaluar con los siguientes criterios:

$$h\left(R_k\right) = \sum_{ij:A_i,A_j \in R_k, i \leq j} d_{ij}.$$

Heterogeneidad de la región $k$ con $R_k \in P_p$:

$$H\left(P_p\right) = \sum_{k=1}^{p} h\left(R_k\right).$$

Heterogeneidad total de la partición $P_p \in \Pi$.

Entonces el problema max-p-regiones puede ser formulado de la siguiente forma:

Determinar $P_p^* \in \Pi$ tal que $\left|P_p^*\right| = max\left(|P_p| : P_p \in \Pi\right)$, y $\nexists P_p \in \Pi : |P_p| = \left|P_p^*\right| y\, H\left(P_p\right) < H\left(P_p^*\right)$.

El algoritmo consiste en buscar una posible solución inicial e iterar diferentes soluciones hasta minimizar la heterogeneidad intra-clúster. El objetivo de Max-p es encontrar la partición óptima $P_p^*$ perteneciente a un conjunto de todas las particiones posibles $\Pi$, tal que no exista otra partición $P_p \in \Pi$, con una heterogeneidad menor que la heterogeneidad de $P_p^*$ (Sáenz Vela 2016) . Además, todo esto lo tiene que hacer cumpliendo por cada región con un umbral mínimo del atributo espacial extensivo $l_i$, y sin perder la contigüidad espacial, está última está garantizada con la utilización del grafo $G$ en la elaboración del algoritmo.

Como se puede ver, considerar todas las posibles particiones no es una tarea viable. Para aproximarse entonces a la partición ideal, Duque, Anselin y Rey 2012 proponen un algoritmo con una heurística particular que lo hace eficiente.

# 3. Evaluación del rendimiento de los clústeres

Hasta ahora se ha conseguido identificar grupos o regiones con contigüidad espacial y características particulares de cada región, pero no se ha hablado si eso realmente da indicios de segregación regional. Por lo tanto, ahora se necesita algún indicador de segregación y así evaluar si los grupos conformados representan realmente a un grupo de población en particular. En nuestro caso, se utilizará el coeficiente de localización, ampliamente utilizado en ciencia regional.

## 3.1. Medidas de rendimiento

Una de las dificultades que presenta el aprendizaje no supervisado es que no existe algún método que funcione mejor que otro, además, la modificación de los denominados hiperparámetros afecta sensiblemente el rendimiento del algoritmo empleado. Por lo tanto, se busca comparar varias metodologías que se usan habitualmente para la detección de clústeres, e incluso comparar varias veces la misma metodología, pero variando los hiperparámetros. Para evaluar cuál de los métodos funciona mejor, habitualmente se considera la suma de cuadrados intra clústeres y entre clústeres, que es la forma de representar la homogeneidad dentro del grupo y la heterogeneidad por fuera de él. Estos indicadores serían:

- *The total sum of square (TSS)*
- *The total within-cluster sum of squares (WSS)*
- *The between-cluster sum of squares (BSS)*
- *The ratio of between to total sum of squares (RBTSS).*

Considerando que $TSS = WSS + BSS$ y que $RBTSS = BSS/TSS$, se busca como objetivo maximizar el $RBTSS$.

El inconveniente que tiene la utilización de este ratio, es que tiende a mejorar cuando se aumenta la cantidad de grupos por más que estos sean similares entre sí. Aunque es cierto que este inconveniente se soluciona con los gráficos de codo (*elbow plot*) que permite obtener la cantidad de grupos cuando el ratio toca el «codo» de la curva.

## 3.2. Coeficiente de localización

El coeficiente de localización (*location quotient*), planteado por Isard (1960), identifica regiones que tienen alguna característica más destacada en comparación con las demás unidades geográficas. De esta forma, como menciona Domínguez Aguilar (2017), el coeficiente de localización es un índice que permite conocer el grado de especialización de una unidad espacial con respecto a otra más amplia. Además, posee la ventaja de ser muy sencillo de calcular como puede verse en la siguiente fórmula:

$$LQ_i = \frac{x_i/t_i}{X/T}, \tag{1}$$

siendo $x_i$ la cantidad de personas del grupo $x$ en la unidad $i$, $t_i$ es la cantidad de personas total en la unidad $i$, $X$ es la cantidad de personas del grupo $x$ en toda la muestra y $T$ es la cantidad total de personas de la muestra.

El $LQ$ ha sido ampliamente utilizado en la literatura de ciencias regionales para medir el grado de aglomeración que existe en una región. Esto quiere decir que el indicador $LQ$ marca por sí solo regiones en dónde existe aglomeración y, por lo tanto, justifica la existencia de un clúster de acuerdo a la característica en particular. Si consideramos como característica distintiva de una región dichas aglomeraciones, esto nos lleva a justificar el uso de este indicador para determinar la «bondad» de los clústeres formados por las diferentes metodologías.

Dada una unidad espacial, decimos que $A_{iy}$ es un atributo que cuenta la cantidad de personas que se encuentran en un área $A_i$ con la característica $y$. Si Suponemos que existe una partición ideal del territorio $P^* = \{R_1, \ldots, R_p\}$, en dónde hay una región $R_j$ que agrupa a las área $A_i$ que contienen a la mayor cantidad de personas con característica $y$, entonces $R_j$ tiene un coeficiente de localización mayor a uno, esto es $LQ_{R_j} > 1$. Además se espera que ese coeficiente de localización sea mayor al del resto de regiones y significativamente distinto. Esto se puede ver en la sección siguiente con la determinación de un intervalo de confianza para $LQ$.

En el presente trabajo, se busca evaluar el funcionamiento de los métodos de regionalización considerando esta idea. Generalmente, en el uso del $LQ$ se parte de unas regiones dadas y se calcula el índice en cada una de ellas para detectar alguna característica. Aquí se busca una partición desconocida a priori y si la misma está bien lograda, entonces el $LQ$ calculado posteriormente la tiene que distinguir.

## 3.3. Intervalo de confianza para $LQ$

Autores como Moineddin, Beyene y Boyle 2003 construyen un intervalo de confianza para $LQ$ de manera teórica como se muestra a continuación.

Sea $R = \begin{pmatrix} r_i \\ r \end{pmatrix}$ un vector aleatorio con media $\mu = \begin{pmatrix} \rho_i \\ \rho \end{pmatrix}$. Luego definiendo a $g(R) = \frac{r_i}{r}$ podemos usar el método delta para aproximar su varianza que queda de la siguiente forma:

$$V(g(R)) = \frac{1}{\rho^2} V(r_i) + \frac{\rho_i^2}{\rho^4} V(r) - \frac{2\rho_i}{\rho^3} Cov(r_i, r).$$

Sea $x_i$ el número de salidas y $t_i$ el tamaño de la población en el área $i$, podemos considerar a $x_i$ como una variable binomial con parámetros $t_i$ y $\rho_i$ siendo este último el ratio de incidencia real de área $i$. Luego es fácil ver que:

$$E\left(x_i/t_i\right) = E\left(r_i\right) = \rho_i \, y \, V\left(r_i\right) = \frac{\rho_i\left(1-\rho_i\right)}{t_i}.$$

$$E\left(X/T\right) = E\left(r\right) = \frac{\sum_{i=1}^{k} t_i\rho_i}{T} \, y \, V\left(r\right) = \frac{1}{T^2}\sum_{i=1}^{k} t_i\rho_i\left(1-\rho_i\right).$$

De esta manera, como muestran Moineddin, Beyene y Boyle 2003 se puede decir que $Cov\left(r_i, r\right) = \frac{\rho_i(1-\rho_i)}{T}$ si asumimos que $Cov\left(x_i, x_j\right) = 0$ para todo $i \neq j$. Luego si aproximamos $V\left(r\right) \cong \frac{\rho(1-\rho)}{T}$ con $E\left(r\right) = \rho$ nos queda la varianza del coeficiente de localización de la siguiente manera:

$$V\left(\frac{r_i}{r}\right) = \frac{\rho_i\left(1-\rho_i\right)}{t_i\rho^2} + \frac{\rho_i^2\left(1-\rho\right)}{T\rho^3} + \frac{2\rho_i^2\left(1-\rho_i\right)}{T\rho^3}. \tag{2}$$

Por lo tanto, podemos definir un intervalo de confianza de la siguiente manera:

$$\frac{r_i}{r} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{r_i\left(1-r_i\right)}{t_i r^2} + \frac{r_i^2\left(1-r\right)}{T r^3} + \frac{2r_i^2\left(1-r_i\right)}{T r^3}}, \tag{3}$$

donde $Z_{\frac{\alpha}{2}}$ es el percentil $100\left(1-\frac{\alpha}{2}\right)$ de una distribución normal estándar.

Al poder construir intervalos de confianzas del indicador, se puede establecer un límite de corte entre regiones y de esta manera determinar que nivel de aglomeración presenta. A los objetivos de este trabajo, este corte no es menor, se puede decir que si dos regiones cuales quiera tienen intervalos de confianza bien separados, que esas regiones están bien distinguidas en cuanto a la característica de aglomeración bajo estudio.

En términos formales podemos decir que *si* $intLQ_i \cap intLQ_j = \emptyset$ entonces $R_i$ y $R_j$ son regiones disímiles entre sí. Siendo $intLQ_*$ los intervalos de confianza del coeficiente de localización de cada región respectivamente.

### Índice $LQ$

Dado el intervalo de confianza podemos validar la disimilitud entre dos regiones. El objetivo buscado cuando aplicamos una metodología de clustering es que todas las regiones sean disímiles entre sí, eso sería una partición ideal. En defecto de esto nos interesaría aquel algoritmo que logre diferenciar más grupos entre sí en comparación con otra metodología. Por eso a los efectos prácticos se puede medir que tan diferenciados son mediante un simple cálculo:

$$LQ_g = \frac{1}{p(p-1)}\sum_{i=1}^{p}\sum_{j\neq i}^{p} I\left(intR_i \cap intR_j \neq \emptyset\right), \tag{4}$$

en dónde $p$ es el número de grupos o regiones, $intR_i$ es el intervalo de confianza de la región $i$. Este indicador va estar entre 0 y 1 considerando al valor 1 como indicador que no existe ninguna región disjunta de las otras.

Este indicador juega en contra de aumentar el número de grupos en la partición. A diferencia del RBTSS mencionado anteriormente. Esto se puede ver estudiando la fórmula en 2 de la varianza del coeficiente de localización. El disminuir $t_i$, por lo general hace aumentar la varianza y por lo tanto la mayor probabilidad de que la intersección entre dos regiones sea no nula.

Por lo general cuando se trata de *clustering*, se busca la agrupación por varias variables, por lo tanto, es conveniente extender el indicador con la siguiente fórmula:

$$LQ_g = \frac{1}{p(p-1)m} \sum_{i=1}^{p} \sum_{j \neq i}^{p} \sum_{y=1}^{m} I\left(intR_{iy} \cap intR_{jy} \neq \emptyset\right) + \lambda \frac{1}{p} \tag{5}$$

En este caso, el incrementar variables hará que el indicador se vaya a cero mostrando que hay más posibilidades de que los grupos sean disjuntos. Si bien se mencionó como ventaja con respecto a RBTSS que este indicador conduce hacia una disminución en la cantidad de grupos, esto puede llevar al error de agrupar con la menor cantidad posible de regiones. Mientras menos regiones haya que comparar es más fácil encontrar que son distintas y puede que conduzca a resultados erróneos. Por lo tanto, se puede agregar un regularizador de la cantidad de grupos, que sería el término $\lambda\frac{1}{p}$ con el parámetro lambda como coeficiente de regularización, que trabaja de manera sencilla y contraria al primer término de la ecuación 5.

# 4. Aplicación empírica

En esta sección se aplicarán las metodologías de *clustering* revisadas a datos empíricos para Mendoza.

## 4.1. Análisis de la base de datos

Los datos fueron tomados de la encuesta de personas del Censo Nacional 2010 que proporciona el INDEC. De dicho censo solo se consideró la información correspondiente a la región norte de la provincia de Mendoza, teniendo en cuenta hasta un radio de 150 km. desde la Ciudad de Mendoza. Se clasificó a las personas en cuatro grupos combinando edad: entre 18 y 25 años y mayores de 25; y en dos grupos de acuerdo al estudio: personas con estudios universitarios y sin estudios básicos. Descartando la información de los individuos que no pertenecen a esta clasificación. Dichas variables se agregaron a nivel radio censal, siendo éste la unidad mínima de información geográfica.

Cuadro 1: Cuadro descriptivo

| Variable | Obs. | Cant./Grupo | Cant./Total | Prom. rad. censal | Sd. rad. censal |
|---|---|---|---|---|---|
| Mayores sin secu | 261448 | 32,21 % | 17,71 % | 170,10 | 102,61 |
| Mayores est. sup. | 114636 | 14,12 % | 7,76 % | 74,58 | 63,88 |
| **Total mayores** | 811592 | 100,00 % | 54,96 % | 528,04 | 201,39 |
| Jóv. sin estudio | 75466 | 36,63 % | 5,11 % | 49,10 | 36,45 |
| Jóv. con estudio | 62874 | 30,52 % | 4,26 % | 40,91 | 24,22 |
| **Total jóvenes** | 206020 | 100,00 % | 13,95 % | 134,04 | 68,17 |

Fuente: Elaboración propia según base de datos.

En el cuadro 1 se muestra la cantidad de personas que hay en toda la muestra en cada una de las categorías propuestas. Algo a destacar es su participación minoritaria en el total de personas y dentro del grupo etario de cada categoría. Esto los clasifica como grupos minoritarios dentro de la población total, con determinar que este grupo de personas se aglomera en una determinada región, estaríamos ante un indicio de segregación regional. Por otro lado, se observa la cantidad promedio

de personas por radio censal clasificadas en cada una de las categorías. Se puede distinguir, además, una alta desviación estándar, lo que indica que hay disparidad entre los radios censales observados.

En un estudio de segregación utilizando el $LQ$ se calcula el índice por alguna división conocida previamente. En el cuadro 2 se puede ver para el nivel de estudio, los valores del índice por departamento. En el mismo se puede ver que existe segregación de acuerdo al coeficiente de localización, en los departamentos de Capital y Godoy Cruz parecen aglomerarse las personas con niveles de estudio más alto. Por el otro lado Lavalle es el departamento que aglomera la mayor cantidad de personas con estudios incompletos.

Cuadro 2: LQ por departamento

| Dpto | lqBE | lqAE | lqJAE | lqJBE |
|---|---|---|---|---|
| Capital | 0,57 | 2,15 | 1,81 | 0,66 |
| Godoy Cruz | 0,79 | 1,31 | 1,26 | 0,90 |
| Guaymallén | 0,96 | 0,99 | 1,05 | 0,96 |
| Junín | 1,18 | 0,65 | 0,99 | 1,00 |
| La Paz | 1,13 | 0,68 | 0,67 | 1,26 |
| Las Heras | 1,05 | 0,72 | 0,79 | 1,12 |
| Lavalle | 1,50 | 0,35 | 0,58 | 1,23 |
| Luján | 0,98 | 1,23 | 1,04 | 0,99 |
| Maipú | 1,15 | 0,64 | 0,80 | 1,08 |
| Rivadavia | 1,16 | 0,80 | 0,92 | 1,03 |
| San Carlos | 1,19 | 0,67 | 0,84 | 1,00 |
| San Martín | 1,17 | 0,74 | 0,82 | 1,09 |
| Santa Rosa | 1,25 | 0,38 | 0,65 | 1,11 |
| Tunuyán | 1,16 | 0,72 | 0,71 | 1,11 |
| Tupungato | 1,31 | 0,52 | 0,61 | 1,07 |

Fuente: Elaboración propia en base al Censo Nacional 2010

## 4.2. Implementación de los algoritmos de regionalización

Si bien el $LQ$ mostrado por departamento ya permite identificar una cierta segregación, ahora se pueden utilizar las metodologías descriptas para encontrar regiones más identificatorias de la segregación. La selección de departamentos es una delimitación política y los fenómenos de segregación suelen traspasar esas fronteras. Para facilitar la descripción posterior se procede a identificar a los grupos de personas propuestos, como atributos de un determinado radio censal $A_i$ de la siguiente forma:

- $A_{i1}$ Cantidad de adultos mayores de 25 años que lograron terminar sus estudios universitarios o terciarios.
- $A_{i2}$ Cantidad de jóvenes entre 18 y 25 años que están estudiando o terminaron sus estudios universitarios o terciarios.
- $A_{i3}$ Cantidad de adultos mayores de 25 años que no lograron terminar el secundario.
- $A_{i4}$ Cantidad de jóvenes entre 18 y 25 años que abandonaron el secundario o antes.

- $A_{i5}$ Cantidad de adultos mayores a 25 años.
- $A_{i6}$ Cantidad de jóvenes entre 18 y 25 años.

Por lo tanto, para el área $i$ nos queda un vector de atributos conformado de la siguiente manera: $x_i = (a_{i1}, ..., a_{i6})$.

Nótese que además de las variables de estudio se han incluido la cantidad de habitantes según los rangos de edad considerados. De esta forma el algoritmo presentará diferencias en áreas con la misma cantidad de personas con estudios, pero con distintas cantidades de población.

Para un mejor resultado de los algoritmos siempre es conveniente trabajar en las mismas escalas, por eso se transformaron los datos de acuerdo a la transformación estándar (z):

$$z = \frac{(x - \overline{x})}{\sigma(x)}$$

o como alternativa la transformación estándar (MAD) que consiste en utilizar desviaciones absolutas respecto a la media en el denominador:

$$mad = \frac{1}{n} \sum_i |x_i - \overline{x}| .$$

Esta última es preferida por varios autores para evitar la influencia de los *outlier* al armar los clústeres. La función de disimilitud usada para todos los algoritmos ha sido la distancia euclidiana. El software utilizado fue GeoDa 1.18.

**Resultados de las diferentes metodologías implementadas**

No existe una solución analítica para determinar que algoritmo utilizar ni que parámetros usar para calcular los grupos. Esta tarea se debe llevar probando distintas alternativas. El número de grupos o regiones en los cuales se quiere dividir el territorio es uno de los hiperparámetros a evaluar en la mayoría de las metodologías de *clustering*. El único que no requiere que se defina arbitrariamente la cantidad de grupos es el algoritmo de Max-p. Por lo tanto, se definió el número de clústeres para el resto de algoritmos partiendo del resultado de este último. El mejor resultado, luego de ir variando los parámetros de cálculo, de Max-p es el que se ve en el cuadro 3 y arrojó un total de cuarenta clústeres. Las restricciones adoptadas para efectuar dicho análisis se trasladaron a los demás algoritmos mientras fue posible. Por un lado, se consideró como matriz espacial a una matriz tipo reina de primer orden. Otro criterio restrictivo fue imponer al algoritmo que al menos tenga treinta radios por región.

Los cuatro primeros métodos mostrados en el cuadro 3 son métodos que no manejan la contigüidad espacial como restricción fuerte. A pesar de poder utilizar la matriz de contactos y darle un mayor peso que a las otras variables, en las pruebas realizadas no se logra la continuidad espacial deseada. Por lo tanto, pese a que K-medias, K-medianas, el método jerárquico y spectral lograron los mejores resultados de RBTSS con un amplio margen frente a los otros, fueron descartados de posteriores análisis por no cumplir con el mínimo requisito de contigüidad para armar las regiones. Un resultado a destacar en estos métodos es que el incluir la contigüidad espacial puede afectar significativamente los resultados de los agrupamientos.

Dentro de los métodos de *clustering* espaciales el que mejor resultados arroja es el SCHC, pero tiene el inconveniente de que ha generado muchas regiones con muy pocos datos, incluso algunos con un único radio censal. Este problema no pudo ser solucionado disminuyendo los números de grupos por lo que tuvo que ser descartado por no cumplir con los requisitos mínimos aquí perseguidos.

Los procesos Redcap y SKATER son, también, de tipo jerárquicos, pero con algunas variaciones tal que se pueden incorporar restricciones sobre la cantidad de unidades por región, lo que hace que este requisito se respete. El problema que tuvieron, como así también lo tuvo AZP, fue con la cantidad de grupos. Contrariamente con el algoritmo Max-p, la cantidad de grupos se tuvo que reducir, si no, de otra forma, las regiones no cumplen con las restricciones establecidas. Esto indica que la cantidad de regiones necesarias se encuentra entre 30 y 35 para el buen funcionamiento de los algoritmos. Todos los algoritmos fueron probados con variaciones en sus parámetros y lo que se alcanza a ver en el cuadro 3 son sus resultados óptimos.

AZP fue el de mejor rendimiento cumpliendo además con las restricciones impuestas. También es un algoritmo que permite muchas parametrizaciones por lo que su implementación incluye muchas opciones. Uno de ellas es definir el algoritmo que se va usar para las búsquedas locales. Hay tres formas que se pueden ver en (Openshaw y Rao 1995), la heurística original que es la que se encarga de intercambiar áreas entre las regiones, luego está la denominada *simulated annealing* y por último la búsqueda tabú que hace búsquedas de forma cíclica. La utilización de uno u otro método de búsqueda local hace variar notablemente el rendimiento del algoritmo. En la figura de la izquierda de la imagen 1 se ve como fue evolucionando el algoritmo al variar los parámetros del método. La búsqueda tabú no proporcionó mejoras, incluso empeoró al principio y luego mejora cuando se incluye el método de inicialización ARiSel. La heurística *simulated annealing* si logra dar un salto en la mejora de la función objetivo, incluso es el que llega al punto máximo. En este mismo gráfico se puede ver que el punto de partida es realmente significativo a la hora de ejecutar el algoritmo, esto se debe a que, el algoritmo intercambia regiones a partir de la partición inicial $P_p^0$. Entonces, una gran mejora ocurre cuando se inicia con la técnica ARiSel, que es un método basado en K-means. Aunque el mejor punto de partida resultó ser incluir el resultado arrojado por el método RedCap que se hizo a partir del octavo intento.

En la figura 1 imagen de la derecha se muestra la búsqueda del número de regiones que optimiza la función a partir de los parámetros seleccionados. Pasados los 34 grupos, se produce una ruptura de la restricción de al menos treinta radios por región, por lo tanto, el punto máximo se logra con 31 regiones como se ve en la imagen. El resultado óptimo de AZP puede visualizarse en el mapa de la figura 2. La mayor diferenciación de las unidades se ve en las zonas más pobladas de Mendoza, teniendo clústeres que abarcan mucho territorio y otros más concentrados.

Si bien no se optimizaron los algoritmos en base al índice $LQ$, se procedió a calcular el mismo luego de que cada uno de ellos finalizara como se muestra en él cuadro 3. La principal diferencia que podemos observar en cuanto al cálculo de del indicador con respecto al utilizado naturalmente, es que el índice $LQ$ castiga a los algoritmos que tienen pocas unidades por grupo o región. Como se muestra en los resultados, los que no cumplieron con la regla de tener al menos treinta unidades son los que presentaron peor índice $LQ$. La justificación de este comportamiento viene explicada por la fórmula de la varianza del coeficiente de localización en 2, pocas unidades generan valores $t_i$ más chicos y por lo tanto varianzas mas grandes.

Para finalizar podemos hacer dos observaciones más. La primera es que el índice $LQ$ condujo a la misma conclusión que el método convencional, colocando a AZP como el algoritmo más óptimo. La segunda, es que, si calculamos en índice $LQ$ de acuerdo al cuadro 2, arroja un resultado de 0.44 lo que lo pone por encima de casi todas las metodologías empleadas en este trabajo, lo que justificaría el uso de ellas para este tipo de estudio

Cuadro 3: Comparación métodos de clústeres

| | TSS | WSS | BSS | RBTSS | Ind LQ | Contig. | > 30 |
|---|---|---|---|---|---|---|---|
| spectral | 15174.9 | 6061.86 | 9113.02 | 0.60 | 0,266 | No | No |
| hierarch | 15174.9 | 4764.67 | 10410.2 | 0.69 | 0,571 | No | No |
| kmeans | 9216.0 | 1593.58 | 7622.42 | 0.83 | 0,334 | No | No |
| kmed | 9216.0 | 2080.83 | 7135.17 | 0.77 | 0,267 | No | No |
| maxp | 15174.9 | 8419.02 | 6755.86 | 0.45 | 0,233 | Si | Si |
| redcap | 9216.0 | 4860.21 | 4355.79 | 0.47 | 0,233 | Si | Si |
| schc | 15174.9 | 6352.16 | 8822.72 | 0.58 | 0,293 | Si | No |
| skater | 9216.0 | 5936.56 | 3279.44 | 0.36 | 0,233 | Si | Si |
| azp | 15174.9 | 6772.33 | 8402.55 | 0.55 | 0,230 | Si | Si |

Figura 1: Evolución del RBTSS variando parámetros de AZP

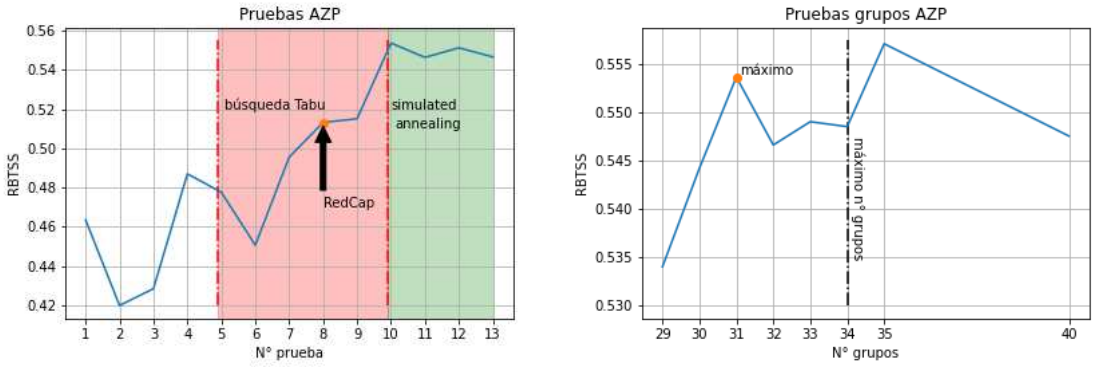
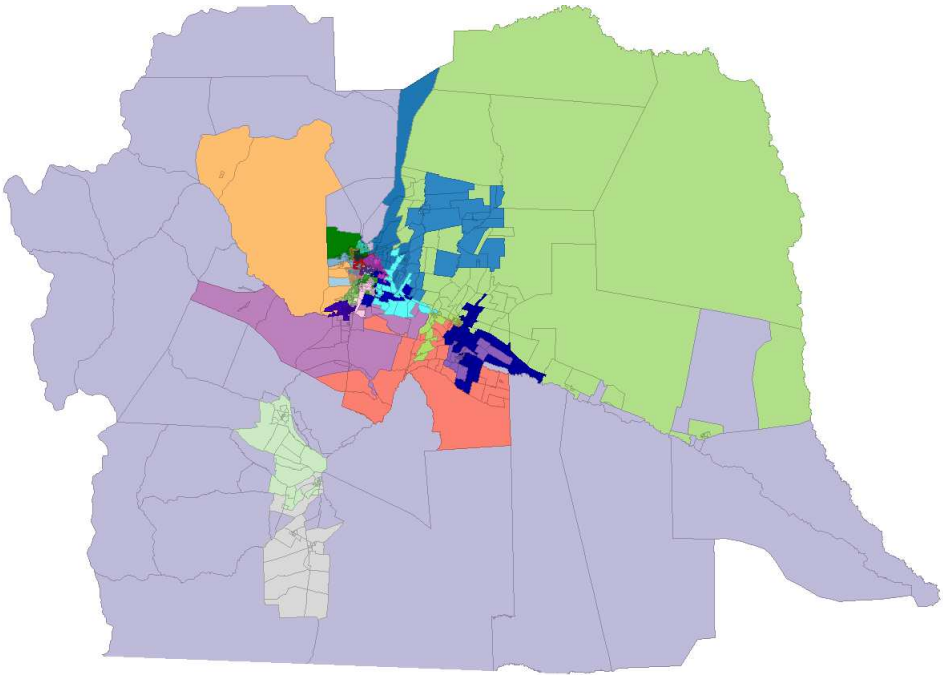



Figura 2: Mapa de clusters AZP



## 4.3. Cálculo del $LQ$ en cada región

El índice $LQ$ nos da un indicio si las metodologías utilizadas han logrado separar correctamente las regiones, pero ahora analizaremos la existencia de regiones destacadas usando los intervalos de confianza mencionados antes. Pero, a los efectos del problema de segregación, basta con encontrar ciertas regiones distinguidas para probarlo, sin necesidad de que todas las regiones se encuentren perfectamente separadas. Entonces, utilizando la regionalización del algoritmo de *clustering* AZP, podemos ahora construir el coeficiente de localización de las variables estudiadas dentro de cada región. El $LQ$ así calculado nos va marcar aquellas regiones en las que se destaque algún grupo poblacional. En particular de acuerdo a las variables vistas en la sección 4.2 el $LQ$ para cada región nos queda:

$$LQ_{ky} = \begin{cases} si\, y \in \{1,3\} & \frac{\sum_{i\in k} A_{iy}}{\sum_{i\in k} A_{i5}} \frac{\sum_{i}^{n} A_{i5}}{\sum_{i}^{n} A_{iy}}, \\ si\, y \in \{2,4\} & \frac{\sum_{i\in k} A_{iy}}{\sum_{i\in k} A_{i6}} \frac{\sum_{i}^{n} A_{i6}}{\sum_{i}^{n} A_{iy}} \end{cases} \tag{6}$$

siendo $LQ_{ky}$ el coeficiente de localización de la variable $A_{.y}$ para la región $k$.

Dado que AZP agrupa los valores semejantes y contiguos, se puede suponer que hay covarianza dentro del clúster $Cov\,(A_{iy}, A_{i'y}) \neq 0$ si $i, i' \in k$ y no hay covarianza en caso contrario $Cov\,(A_{iy}, A_{i'y}) = 0$ si $i \in k, i' \in k'$ con $k \neq k'$, entonces se puede aplicar un intervalo de confianza como el que se muestra en la ecuación 3.

Definidos los intervalos, se puede decir que un clúster es significativamente distinto en al menos uno de los atributos elegidos si los intervalos de confianza para ese atributo de los clústeres no comparten elementos en común. En la figura 3 se pueden ver los respectivos valores del $LQ$ para cada clúster en lo que refiere a jóvenes que van a la universidad y adultos que tienen título universitario. Lo que se puede observar es que se destacan cuatro regiones, identificadas en el mapa de la figura 4. Estas cuatro regiones se muestran significativas entre sí en alguna de las características y muy significativas con respecto a las otras regiones, lo que da indicios de que se produce una verdadera segregación residencial de las personas con estudios universitarios.

Si comparamos los coeficientes de localización calculados sobre las regiones de AZP con los calculados en el cuadro 2 las cuatro regiones con mayor índice, en lo que se refiere a mayores de veinticinco años con estudios universitarios finalizados, superan el 2,5 de $LQ$ y se distribuyen entre los departamentos de Capital, Godoy Cruz, Las Heras y Luján de Cuyo. En cambio al calcularlo por departamento, el que mas tiene es Capital con 2,15 y casi que no podemos decir más nada de los otros departamentos. Con los jóvenes en la Universidad se puede marcar también la diferencia, ya que en estas regiones en las que se encuentran concentrados, las mismas que los adultos, se nota mas la concentración a diferencia con la medición por departamento que no se logra ver una aglomeración tan definida.

Figura 3: Intervalos de confianza

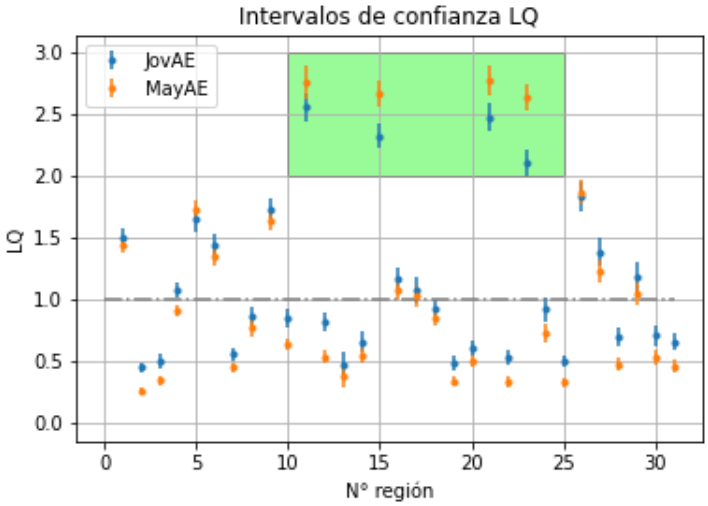
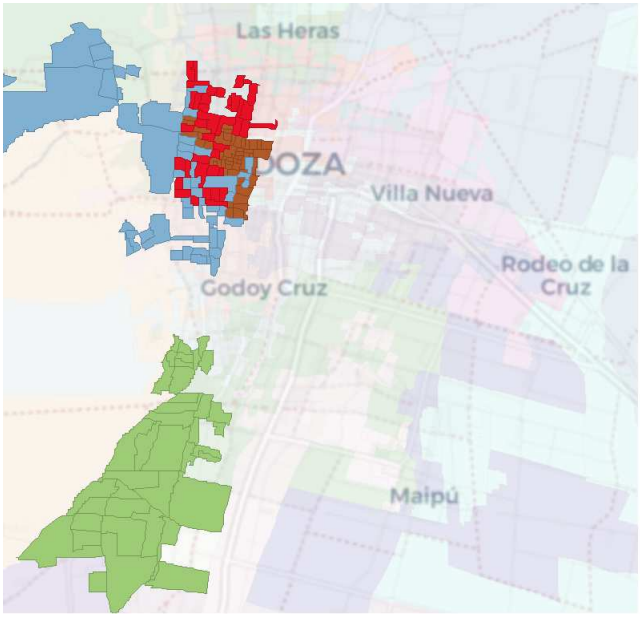



Figura 4: Mapa de regiones significativas jóvenes en la Universidad




## 5. Conclusiones

Las metodologías de *clustering* que forman parte del aprendizaje no supervisado, están mostrando enormes avances en cuanto a resolución de distintas clases de problemas, en muchos casos, llegando a soluciones que no se llegan de manera convencional. Incluso, el uso de estos mecanismos no supervisados permite abordar problemas sin modelos teóricos detrás. Por lo tanto, el estudio de su funcionamiento y de cómo pueden aplicarse a problemas socio-económicos es una tarea que conviene abordar. En el presente trabajo se muestra como las metodologías de regionalización estudiadas contribuyen a estudiar un problema puntual que se da en la realidad, cómo es la segregación urbana. El foco de atención planteado fueron dos características socio-demográficas como el nivel de estudio alcanzado y la franja etaria. La exploración por las diferentes metodologías

de *clustering* permitieron identificar las regiones en dónde las personas con esas características se agrupan, con mayor precisión que si se utilizara otra metodología más convencional.

Por otra parte, la exploración de las diferentes metodologías y la elección de la mejor herramienta no es una tarea fácil ya que al ser el *clustering* un problema *NP-hard* no permite identificar con claridad si realmente estamos ante la mejor partición del territorio. Además, la necesidad de imponer ciertas restricciones a los algoritmos para que la exploración sea acotada, como pueden ser el número de grupos, la cantidad mínima de áreas o la matriz espacial (por lo general definida arbitrariamente por el investigador), conlleva a descartar soluciones que pueden incluir a la solución óptima. Esto, si bien es una dificultad, no es un impedimento para el uso de ellas ya que, como se mostró en este trabajo, se puede ir explorando las diferentes alternativas hasta encontrar un resultado medianamente satisfactorio.

En consecuencia, con la dificultad de entender las soluciones que brindan estas herramientas, en este trabajo se muestra que se puede tener una noción de la veracidad de la solución utilizando herramientas adicionales que tengan que ver con la problemática bajo estudio. En este caso, el coeficiente de localización y la posibilidad de calcular su intervalo de confianza por cada región permite determinar la significatividad de las variables que componen los clústeres. Es decir que se pudo determinar las características que distinguen una región de la otra mediante esta estrategia. Siguiendo este camino, se construye un indicador que cuantifica la cantidad de regiones distinguibles en toda la partición. De esta forma, logramos seleccionar un algoritmo óptimo con las mismas conclusiones que usando la otra métrica de rendimiento usada comúnmente en la literatura.

Para finalizar, se puede decir que, en el ejemplo planteado, se muestra como diferentes regiones del territorio mendocino aglomeran en mayor o menor medida a las personas de acuerdo a su nivel de estudio. Es decir, puede detectarse que existe segregación de las personas con niveles universitarios de estudio y, por otro lado, que las metodologías de *clustering* funcionan satisfactoriamente para detectar este tipo de patrones.

# Referencias


Aguado-Moralejo, I., C. Echebarria y J. Barrutia (2019). "Aplicación de un análisis clúster para el estudio de la segregación social en el municipio de Bilbao". En: *Boletín de la Asociación de Geógrafos Españoles* 81.6, págs. 1-35.

Anselin, L. (1995). "Local indicators of spatial association-LISA". En: *Geographical analysis* 27.2, págs. 93-115.

Anselin, Luc, Ibnu Syabri y Youngihn Kho (2010). "GeoDa: an introduction to spatial data analysis". En: *Handbook of applied spatial analysis*. Springer, págs. 73-89.

Assunção, R. y col. (2006). "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees". En: *International Journal of Geographical Information Science* 20.7, págs. 797-811.

Chavent, Marie y col. (2018). "ClustGeo: an R package for hierarchical clustering with spatial constraints". En: *Computational Statistics* 33.4, págs. 1799-1822.

Cheruvelil, Kendra Spence y col. (2017). "Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method". En: *Ecology and evolution* 7.9, págs. 3046-3058.

Cook, Stephen (2006). "The P versus NP problem". En: *The millennium prize problems*, págs. 87-104.

Domínguez Aguilar, M. (jun. de 2017). "Las dimensiones espaciales de la segregación residencial en la ciudad de Mérida, Yucatán, a principios del siglo XXI". En: *Península* 12, págs. 147 -188.

Duque, Juan, Luc Anselin y Sergio Rey (2012). "The Max-P regions problem". En: *Journal of Regional Science* 52.3, págs. 397-419.

Garrocho, C. y J. Campos-Alanís (2013). "Réquiem por los indicadores no espaciales de segregación residencial". En: *Papeles de Población* 19, págs. 269-300.

Guo, Diansheng (2008). "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)". En: *International Journal of Geographical Information Science* 22.7, págs. 801-823.

Haining, Robert, Stephen Wise y Jingsheng Ma (2000). "Designing and implementing software for spatial statistical analysis in a GIS environment". En: *Journal of Geographical Systems* 2.3, págs. 257-286.

Hastie, Trevor, Robert Tibshirani y Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Isard, Walter (1960). *Methods of regional analysis an introduction to regional science*. The MIT Press, Cambridge.

LeCun, Yann, Yoshua Bengio y Geoffrey Hinton (2015). "Deep learning". En: *nature* 521.7553, págs. 436-444.

Massey, D. y N. Denton (1988). "The Dimensions of Residential Segregation". En: *Social Forces* 67.2, págs. 281-315.

Moineddin, Rahim, Joseph Beyene y Eleanor Boyle (2003). "On the location quotient confidence interval". En: *Geographical analysis* 35.3, págs. 249-256.

Oka, M. y D. Wong (2014). "Capturing the Two Dimensions of Residential Segregation at the Neighborhood Level for Health Research". En: *Frontiers in Public Health* 2, pág. 118. ISSN: 2296-2565.

Openshaw, S. y L. Rao (1995). "Algorithms for Reengineering 1991 Census Geography". En: *Environment and Planning A: Economy and Space* 27.3. PMID: 12346252, págs. 425-446. DOI: 10.1068/a270425. eprint: https://doi.org/10.1068/a270425. URL: https://doi.org/10.1068/a270425.

Openshaw, Stan (1977). "A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling". En: *Transactions of the institute of british geographers*, págs. 459-472.

Reardon, S. y D. O'Sullivan (2004). "Measures of spatial segregation". En: *Sociological methodology* 34.1, págs. 121-162.

Sabatini, F. (2006). "La segregación social del espacio en las ciudades de América Latina". En: *Banco Interamericano de Desarrollo*.

Sáenz Vela, H. (2016). "Revisando los métodos de agregación de unidades espaciales: MAUP, algoritmos y un breve ejemplo". En: *Estudios demográficos y urbanos* 31, págs. 385-411.

Webster, R y P Burrough (1972). "Computer-Based soil mapping of small areas from sample data: I. Multivariate Classification and Ordination". En: *Journal of Soil Science* 23.2, págs. 210-221.

Wise, Steve, Robert Haining y Jingsheng Ma (1997). "Regionalisation tools for the exploratory spatial analysis of health data". En: *Recent developments in spatial analysis*. Springer, págs. 83-100.

Wong, D. (2005). "Formulating a general spatial segregation measure". En: *The Professional Geographer* 57.2, págs. 285-294.

Yuan, Shuai y col. (2015). "Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity". En: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, págs. 1-10.