UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Segmentación geométrica de la ciudad de Quito por medio de un método de aprendizaje no supervisado

Felipe Javier Vaca Ramírez

Luca Guzzardi, Ph.D.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Magister en Matemáticas Aplicadas

Quito, septiembre de 2015

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Segmentación geométrica de la ciudad de Quito por medio de un método de aprendizaje no supervisado

Felipe Javier Vaca Ramírez

Firmas

Luca Guzzardi, Ph.D.,
Director del Trabajo de Titulación

Carlos Jiménez Mosquera, Ph.D.,
Director de la Maestría en Matemáticas Aplicadas y Miembro del Comité de Trabajo de Titulación

Julio Ibarra Fiallo, M.Sc.,
Miembro del Comité de Trabajo de Titulación

César Zambrano, Ph.D.,
Decano del Colegio de Ciencias e Ingeniería

Hugo Burgos, Ph.D.
Decano del Colegio de Posgrados

Quito, septiembre de 2015

© Derechos de Autor

Firma del estudiante: _____

Nombre: Felipe Javier Vaca Ramírez

Código de estudiante: 00118079

C. I.: 1003307418

Lugar, Fecha Quito, septiembre de 2015

DEDICATORIA

A mis amigos les adeudo.

A los espíritus inquietos.

AGRADECIMIENTOS

A quienes han hecho posible este trabajo, la duda y el viaje.

RESUMEN

El presente documento recoge una aplicación de métodos de análisis de datos en urbanismo y Sistemas de Información Geográfica. Se apunta a encontrar patrones de forma y densidad en el mapa de Quito, partiendo del supuesto de que existen similitudes entre las manzanas de la urbe en términos de sus atributos geométricos y su ubicación espacial. Los algoritmos Jerárquico Aglomerativo y DBSCAN han sido considerados como herramientas de análisis. Siendo el resultado una nueva segmentación de la ciudad, en la cual los nuevos barrios o sectores no están definidos de manera arbitraria, se abre la puerta a un replanteamiento de la planificación urbana.

Palabras clave: agrupamiento, polígonos, características geométricas, ciudades, vecindarios, urbanismo, Sistemas de Información Geográfica.

ABSTRACT

This document contains an application of data analysis methods on urbanism and Geographic Information Systems. It aims to find shape and density patterns in the map of Quito, starting from the assumption that there are some similarities between the city blocks in terms of its geometric features and its spatial location. The Hierarchical Agglomerative and DBSCAN algorithms have been considered as analysis tools. Since the result is a new partition of the city, in which the new neighborhoods or sectors are not defined in an arbitrary way, we open the door for rethinking urban planning.

Key words: clustering, polygons, geometric features, cities, neighborhoods, urbanism, Geographic Information Systems.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

INTRODUCCIÓN

La ciudad es un concepto que involucra una gran variedad de aspectos: sociales, económicos, arquitectónicos, geográficos, ambientales, políticos, etc. De ahí la complejidad de estudiarla, pues no sólo basta caracterizar cada uno de dichos aspectos, sino también entender las relaciones que existen entre ellos y su evolución.

La estructura y la forma, en tanto variables espaciales, constituyen características ineludibles en el estudio de la urbe. Según Clifton *et al*. (2008), "el progreso hacia un mejor crecimiento o a un nuevo urbanismo puede únicamente lograrse con un entendimiento más completo de la forma de la ciudad y qué medidas de política se necesitan como respuesta".

A su vez, la estructura urbana está constituida por la red vial y otros elementos físicos como edificios, manzanas y vecindarios. La red vial constituye una parte importante dicha estructura en tanto puede capturar información sobre mecanismos subyacentes de formación y evolución de la urbe (Louf y Barthelemy, 2014), mientras los otros elementos dan cuenta del uso del suelo y de la ubicación de puntos de interés. Estos elementos pueden estudiarse como problemas complementarios: por un lado, se puede caracterizar el grafo de la red vial, y por otro, las manzanas se pueden tratar como polígonos y por ende usar sus atributos (Louf y Barthelemy, 2014).

De hecho, la mayor parte de los trabajos dedicados a estudiar la estructura de las ciudades utiliza como recurso el análisis de la topología de los grafos de las redes viales correspondientes.[1] Son pocos los estudios que explotan el potencial de los algoritmos de clasificación para caracterizar la estructura de una ciudad. Más aún, entre estos, gran parte son de tipo normativo, es decir, parten definiendo tipos de manzanas o calles de acuerdo a

[1] Véase Schirmer y Axhausen (2015) para consultar una revisión al estado del arte sobre este enfoque.

cierto criterio de experto o a un ideal de buena ciudad. Con ello la clasificación se vuelve supervisada.[2] Entre las aproximaciones que utilizan algoritmos de agrupamiento o clasificación no supervisada, cabe citar (en orden cronológico) a:

Joshi *et al*. (2009) proponen una extensión al algoritmo DBSCAN para agrupar polígonos. En particular, adaptan las definiciones para puntos a polígonos y utilizan como métrica la distancia de Hausdorff. Finalmente, usan algunos ejemplos de prueba para probar el algoritmo y comparan los resultados con el DBSCAN, obteniendo grupos más compactos con la propuesta.

Gil *et al*. (2009) clasifican manzanas y calles de dos barrios de Lisboa (Portugal), uno antiguo y otro contemporáneo, considerando datos geométricos y de uso de suelo y utilizando el algoritmo *k-means*. El propósito es identificar tipologías comunes y distintas en dichos barrios.

Amindarbari y Sevtsuk (2013) hacen una revisión de indicadores de forma urbana y uso del suelo (como tamaño, densidad, cobertura, multicentralidad, entre otros) con el fin de describir el cambio y el crecimiento del *built environment* de áreas metropolitanas.

Hecht *et al*. (2013) utilizan diferentes herramientas de clasificación supervisada para clasificar edificios de la ciudad de Dresden (Alemania). Se agregan los resultados a nivel de manzana, se derivan tipos de estructuras urbanas y se evalúa la precisión del procedimiento propuesto.

Louf y Marc Barthelemy (2014) clasifican ciudades de acuerdo a los patrones de sus calles utilizando la distribución de probabilidad condicional de los factores de forma de las

[2] Aunque existe una rama importante que utiliza reconocimiento de imágenes y morfología matemática, una mención a la misma está fuera del alcance del presente trabajo.

manzanas con un área determinada y un método de agrupamiento jerárquico simple. Adicionalmente, los autores llevan a cabo el mismo proceso en barrios de Nueva York.

Finalmente, Schirmer y Axhausen (2015) discuten los atributos y medidas para caracterizar a la morfología urbana. Proponen una metodología para realizar clasificación a diferentes escalas (edificio, manzana, vecindario, distrito y municipio) y lograr una tipología utilizando los algoritmos *k-means* y *k-medoids*. El conjunto de datos de prueba corresponde al cantón de Zurich.

Es necesario notar que, la mayoría de estos trabajos consideran grupos de manzanas cuya manifestación más frecuente son los barrios. Si bien sus resultados son de gran utilidad, estas aproximaciones no están exentas de críticas. Uno de sus posibles sesgos radica en que los barrios o sectores censales suelen ser divisiones administrativas cuyo establecimiento puede carecer de sustento o volverse obsoleto, principalmente para las partes más contemporáneas de la ciudad. Esto provoca la omisión de potenciales relaciones espaciales entre las manzanas y calles, el sometimiento del análisis a cómo se define la unidad de área y la consecuente elaboración de políticas inadecuadas en materia urbana (Louf y Barthelemy, 2014; Serra *et al*., 2013). Por esta razón, se hace necesario extraer estos "nuevos barrios o sectores" y estudiar las relaciones espaciales entre los distintos grupos que puedan surgir de este proceso.

Asumiendo que existen semejanzas entre los elementos de la estructura urbana, el presente trabajo asume este reto y apunta a buscar patrones de tipo geométrico-espacial en la estructura urbana de la ciudad de Quito a partir del estudio de sus manzanas. Se realiza clasificación no supervisada (agrupamiento) en tanto se procura dejar de lado tipos

predefinidos de formas urbanas y explotar la flexibilidad de algunos métodos de agrupamiento.

Puesto que dichos patrones pueden tener relaciones con la accesibilidad, flujo vehicular, uso del suelo, calidad del paisaje urbano, crecimiento, forma, evolución y otros aspectos de la ciudad, el presente trabajo abre una gama de potenciales estudios en cuanto a planificación urbana y brinda insumos para el desarrollo de una nueva política pública.

Para concluir esta sección, cabe mencionar que, el resto del documento consta de: una referencia a los métodos utilizados, el análisis de los resultados obtenidos y algunas conclusiones y recomendaciones.

METODOLOGÍA

Algoritmo de Agrupamiento Jerárquico Aglomerativo

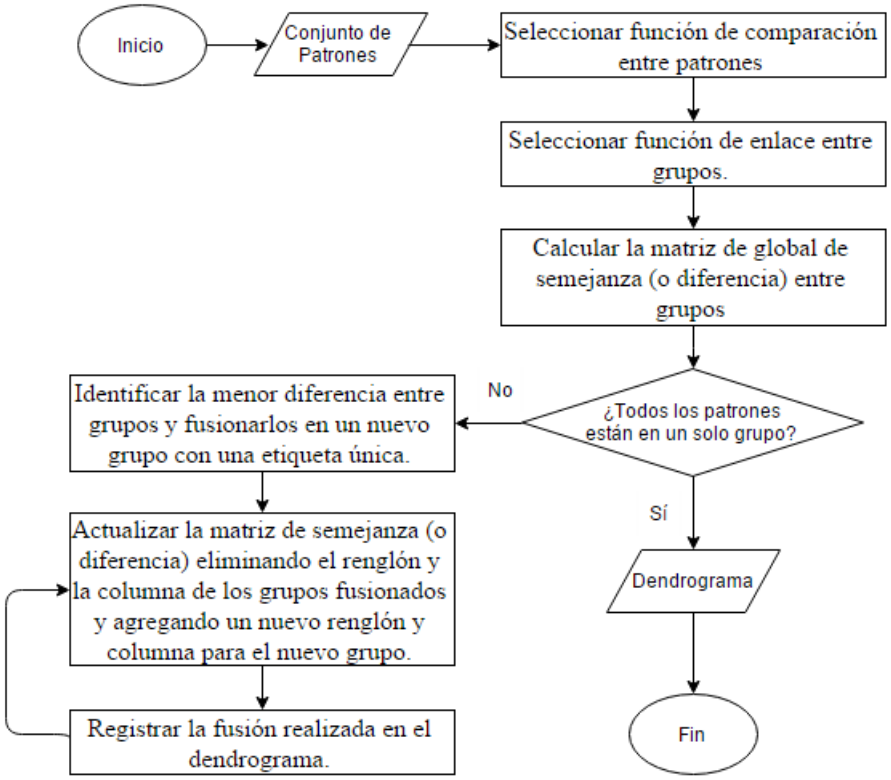
Los algoritmos jerárquicos dividen el conjunto de datos en una secuencia de particiones anidadas. En el caso de los algoritmos jerárquicos aglomerativos el proceso empieza considerando a cada uno de los patrones (u objetos) en un grupo unitario. Luego se fusionan los pares de grupos más cercanos de acuerdo a algún criterio de semejanza o diferencia hasta lograr un solo grupo que contenga todos los patrones. A dichos criterios de comparación se los conoce como funciones de enlace (en inglés, *linkage functions*) o métodos de aglomeración.

Una de los métodos ampliamente utilizado es el método Ward (1963) o método de mínima varianza. El nombre obedece a que, en cierto modo, minimiza la pérdida de información asociada a cada fusión de grupos. Usualmente, la información se cuantifica en términos de una suma de cuadrados de los errores. Así, en cada etapa de este método se fusionan dos grupos cuando dicha fusión resulte en el mínimo incremento en la pérdida de información. Sean C_r y C_s dos grupos con centroides $\overline{C_r}$ y $\overline{C_s}$, respectivamente. La función de enlace de Ward está dada por:

$$d(C_r, C_s) = \frac{|C_r||C_s|}{|C_r| + |C_s|} \|\overline{C_r} - \overline{C_s}\|^2 ,$$

Cabe mencionar que, una forma de representar los agrupamientos posibles es a través de un dendrograma. Así, el número de grupos que se desee formar se ve reflejado como un corte en dicho dendrograma.

Descripción del algoritmo



Medidas de Validación Interna[3]

La determinación del número de grupos puede realizarse mediante un contraste de Medidas de Validación Interna. Estas medidas reciben como argumento al conjunto de datos y la partición del mismo y utilizan información intrínseca de los datos para evaluar la calidad del agrupamiento. Los indicadores considerados hacen referencia a tres criterios:

1) Compacidad. Evalúa la homogeneidad del agrupamiento, observando usualmente la varianza intra-grupos.

[3] Tomado de Brock et al. (2011)

2) Conectividad. Evalúa en qué medida las observaciones están en el mismo grupo en el que se están sus vecinos más cercanos en el espacio de datos. Una forma de cuantificarlo es con el indicador *Connectivity* (Handl *et al*., 2005).

3) Separación entre los grupos obtenidos. Usualmente medido a través de la distancia entre los centroides de los grupos.

De acuerdo a Handl *et al*. (2005), mientras que la homogeneidad intra-grupos mejora cuando el número de grupos se incrementa, la distancia entre grupos tiende a deteriorarse. Por esta razón, algunos indicadores ampliamente utilizados combinan los dos índices en uno. Entre estas medidas de validación interna se encuentran: *Connectivity*, *Silhouette Width* y *Dunn Index*. Sus definiciones y características son recogidas en el Anexo B.

Density-based spatial clustering of applications with noise (DBSCAN)

El algoritmo de agrupamiento por densidad DBSCAN propuesto por Ester *et al*. (1996) encuentra grupos con forma arbitraria en un conjunto de datos espaciales recibiendo como argumento dos parámetros: ε (radio de las vecindades) y μ (umbral de densidad). Algunas definiciones previas se requieren antes de describir el algoritmo.

Definición. Sea D un conjunto de puntos. La ε-vecindad de un punto p, denotada por $N_\varepsilon(p)$, se define como $N_\varepsilon(p) = \{q \in D \mid d(p,q) \leq \varepsilon\}$

Definición. Se llama umbral de densidad, denotado con μ al mínimo número de puntos que una vecindad debe contener para considerarse de "alta densidad".

Definición. Un punto p es directamente denso-alcanzable desde un punto q con respecto a ε y μ si y sólo si:

1) $p \in N_\varepsilon(q)$, y

2) $|N_\varepsilon(q)| \geq \mu$.

Definición. Un punto p es denso-alcanzable desde un punto q con respecto a ε y μ si existe una cadena de puntos $p_1, p_2, \dots, p_n$, con $p_1 = q$ y $p_n = p$, tal que p_{i+1} es directamente denso-alcanzable desde p_i.

Definición. Un punto p es denso-conexo a un punto q con respecto a ε y μ si existe un punto r tal que tanto p y q sean denso-alcanzables desde r con respecto a ε y μ.

Definición. Sea D un conjunto de datos. Un *cluster* C con respecto a ε y μ es un subconjunto no vacío de *D* que satisface las siguientes condiciones:

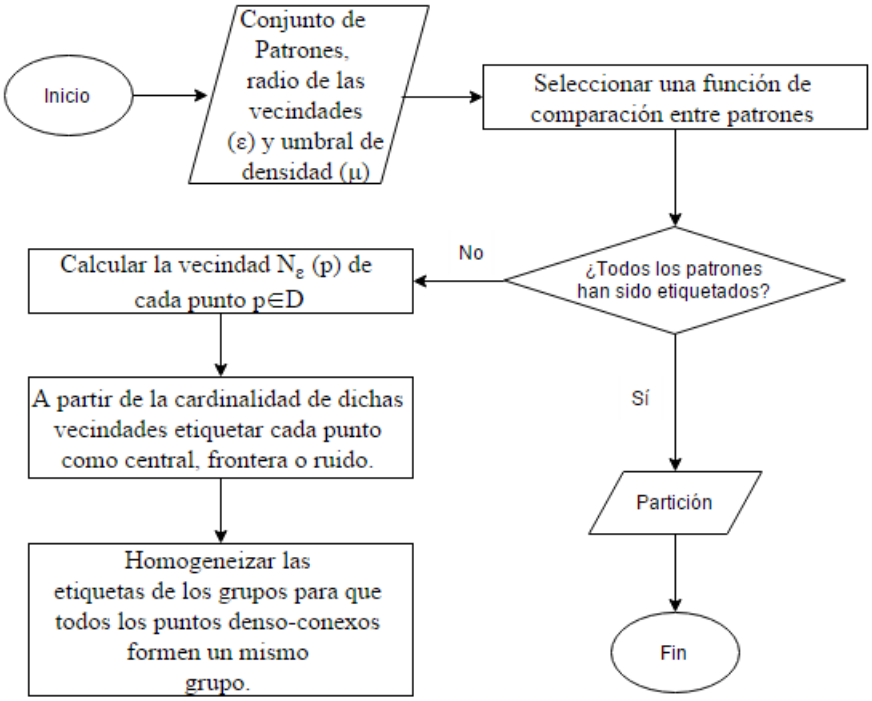
1) $\forall\, p, q$: si $p \in C$ y q es denso-alcanzable desde p con respecto a ε y μ, entonces $q \in C$ (Maximalidad).

2) $\forall\, p, q \in C$: p es denso-conexo a q con respecto a ε y μ (Conectividad).

Definición. Sean $C_1, C_2, \dots, C_k$ *clusters* de un conjunto de datos D respecto a parámetros ε_i y μ_i, $i = 1, \dots, k$. Se define como ruido al conjunto de puntos en D que no pertenecen a ningún grupo C_i.

Definición. Los puntos cuya ε-vecindad es de "alta densidad" se denominan puntos centrales de un grupo (en inglés, *core points* o *seed points*). Los puntos que no tengan una ε-vecindad de "alta densidad" pero pertenezcan a alguna que lo sea se denominan puntos frontera (en inglés, *border points*). Los puntos que no tengan una ε-vecindad de "alta densidad" y tampoco pertenezcan a alguna que lo sea se denominan puntos ruido (en inglés, *noise points*).

Descripción del algoritmo



Formas de escoger ε y μ

Ester *et al*. (1996) fijan $\mu = 4$ y estudian un gráfico en el cual el eje *y* contiene el valor de la distancia al cuarto vecino más cercano de cada observación y el eje *x* un índice asignado a cada observación. [4] Puesto que los valores de dicha distancia están ordenados de forma descendente, el gráfico tiene forma de codo. Así, el codo corresponde al umbral ε sobre el cual están los puntos que pertenecen a los grupos y bajo el cual están los puntos ruido u *outliers*.

La forma en que se escogió ε y μ en el presente trabajo, aunque considera a dichos autores, difiere de manera considerable del mismo:

[4] De acuerdo a dichos autores, $\mu = 4$ es un valor "adecuado" para datos en dos dimensiones.

Se parte determinando el radio de la vecindad ε en base al histograma y al gráfico de codo de las distancias al vecino más cercano de cada observación como en Ester *et al*. (1996). Posteriormente, para establecer el parámetro μ se corre DBSCAN con $\mu = 1, 2, \ldots, k$ (donde *k* es un entero menor que el número de puntos)[5] y se calculan: 1) el porcentaje de puntos ruido respecto al total de puntos, 2) el número de grupos "grandes" o significativos (cuya cardinalidad sea mayor que la media más dos desviaciones estándar de las cardinalidades de los grupos), y 3) el porcentaje de puntos que están en dichos grupos "grandes" respecto al total de puntos. A medida que el número mínimo de puntos requeridos en una ε-vecindad se incrementa se espera que el porcentaje de *outliers* también se incremente, y por ende, el porcentaje de puntos en grupos "grandes" disminuya. De este modo, el parámetro μ más adecuado será aquel que minimice el ruido y maximice los grupos significativos. Nótese la disyuntiva entre ambos objetivos.

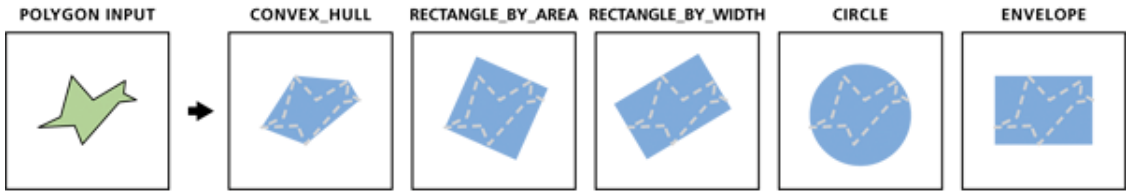
Cabe mencionar que, los indicadores de validación interna podrían utilizarse para escoger los parámetros de DBSCAN. Sin embargo, se los ha descartado puesto que resultan inadecuados para el tipo de datos en cuestión. En efecto, no necesariamente son malos aquellos agrupamientos con alta varianza intra-grupos y tampoco son necesariamente buenos aquellos con alta distancia entre centroides de grupos.

[5] Los experimentos del presente trabajo determinaron que *k* = 20 es adecuado.

DATOS Y RESULTADOS

El punto de partida es un archivo .shp de líneas que corresponde a la red vial de Quito y que es convertido a polígonos; estos representan las manzanas de la ciudad. Para cada polígono se mide su área y el área de algunos tipos de geometrías delimitadoras mínimas (en inglés, *minimum bounding geometry types*) especificados en la figura 1. Siguiendo a Louf y Barthelemy (2014), se calculan las razones entre el área del polígono y el área de dichas geometrías delimitadoras como medidas de forma. Adicionalmente, se añaden el número de lotes de cada manzana como una medida de densidad intra-manzana y la elevación del centroide de cada manzana (obtenido de un modelo de elevación digital) como *proxy* de la altura promedio de la manzana.

Figura 1: Tipos de Geometrías Delimitadoras Mínimas



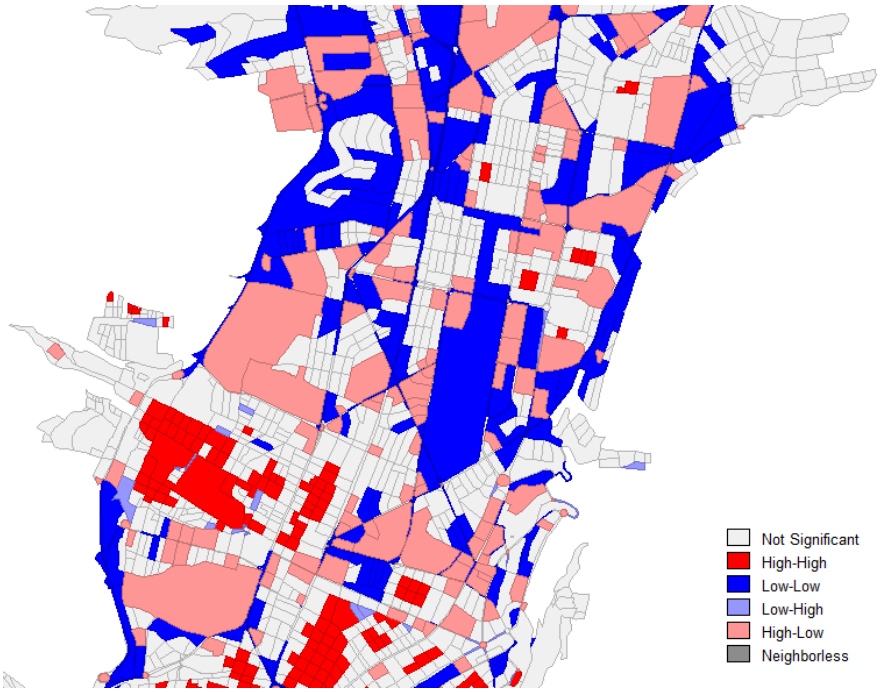
Fuente: http://resources.arcgis.com/en/help/

Puesto que estamos buscando variables que no estén altamente correlacionadas (véanse la tabla 1 y la figura 15), se excluye de este conjunto de variables a la razón correspondiente a RECTANGLE_BY_WIDTH.

En cuanto a los métodos cabe plantearse si indicadores de dependencia espacial como el Índice de Moran son suficientes para capturar y extraer la estructura de la ciudad. Los resultados de este índice se muestran en las figuras 2, 17 y 18 y la tabla 2. Aunque

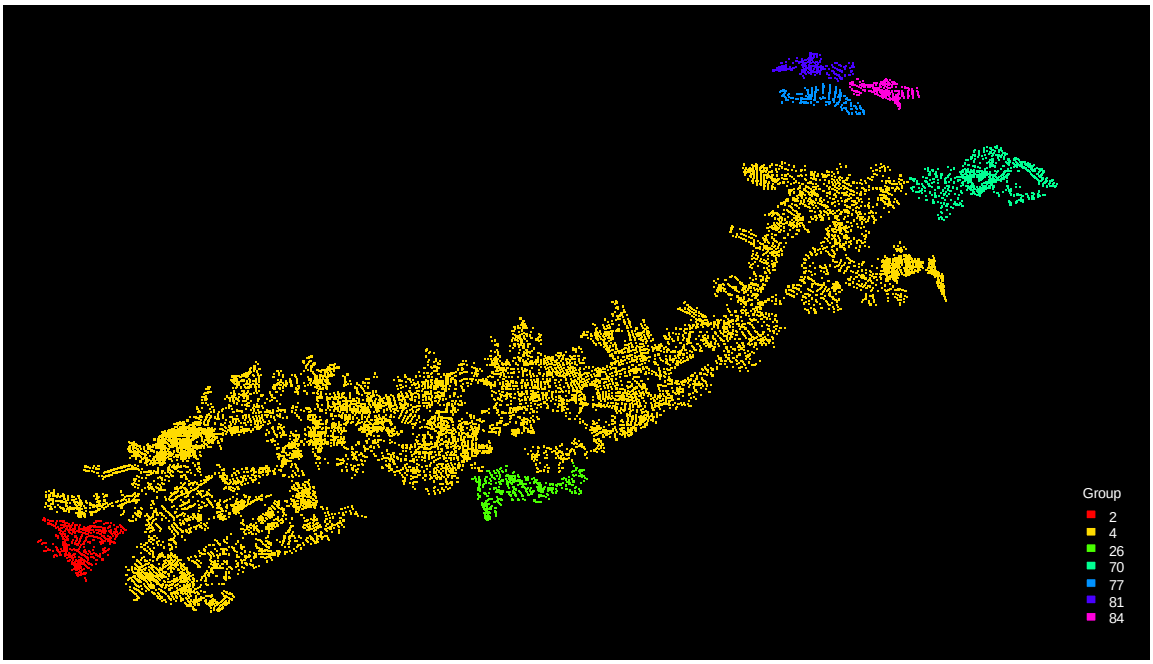
existen indicios de dependencia espacial de las variables, esta herramienta no resulta ser útil para extraer la estructura de la urbe así como se demuestra en la siguiente figura. Nótese que las manzanas de tipo rectangular pertenecen a distintos grupos.

Figura 2: Mapa del Índice de Moran Local de *Ratio Circle*



Tampoco resulta adecuado utilizar DBSCAN directamente sobre todas las observaciones, pues siendo los parámetros adecuados ε = 200 y μ = 4, cerca del 74% de observaciones pertenecen al grupo 4 (en color amarillo en la siguiente figura) y 79 de los 88 grupos tienen participaciones menores al 1%. Los puntos ruido representan el 3.7%.

Figura 3: Partición de DBSCAN con todos los centroides [6]



Por esta razón se utiliza una técnica de agrupamiento en dos etapas. En la primera se trata de extraer una especie de esqueleto de la ciudad utilizando un algoritmo de agrupamiento jerárquico con función de enlace de Ward y distancia euclideana. De los resultados obtenidos nos interesan principalmente aquellos grupos cuya disposición se asemeje a una rejilla. El propósito de este paso es disminuir de manera significativa el número de objetos a analizar en la siguiente fase, tanto para considerar únicamente aquellas manzanas relevantes como también para reducir el costo computacional.

De los resultados del agrupamiento representados en un dendrograma y de acuerdo a índices de validación interna (figuras 4 y 5), las particiones que consideran 3, 8 y 11 grupos son las candidatas. Una exploración visual sobre los resultados, nos lleva a elegir la de 8 grupos. Dentro de esta partición, los grupos 1, 3 y 4 contienen a la mayor parte de las

[6] Se muestran únicamente aquellos grupos cuya cardinalidad es mayor que la media.

manzanas cuya disposición en el mapa es de tipo rejilla. El número de observaciones a ser utilizadas en la siguiente etapa es 9160, es decir, 30% menos objetos que en el conjunto original. Los resultados del agrupamiento están representados en las figuras 6, 7 y 8.

Figura 4: Cluster Dendrogram

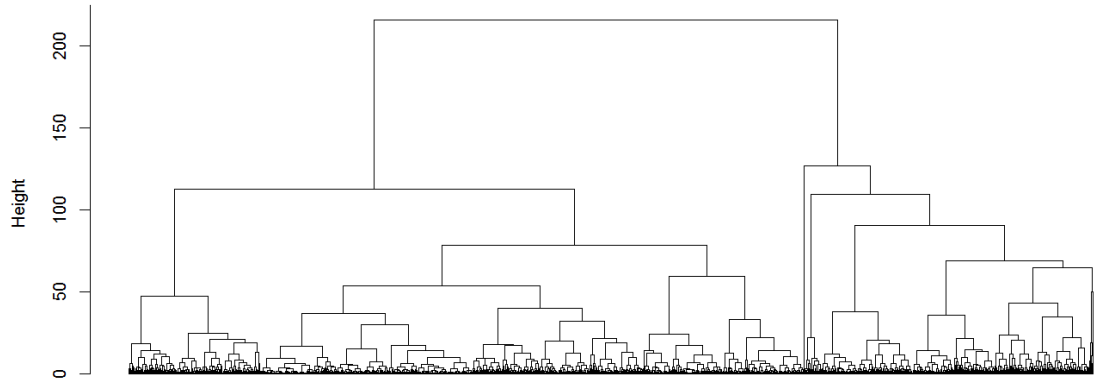


Figura 5: Índices de Validación Interna del Agrupamiento Jerárquico

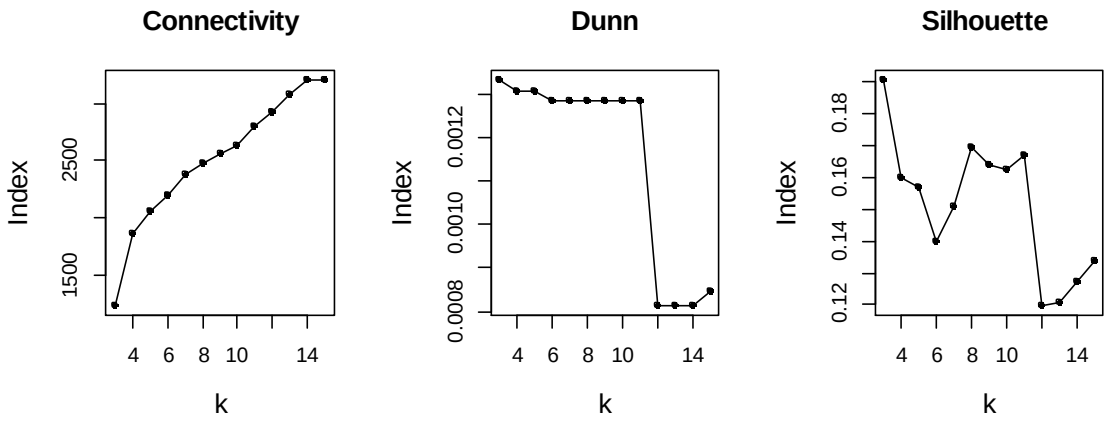


Figura 6: Mapa de agrupamiento jerárquico

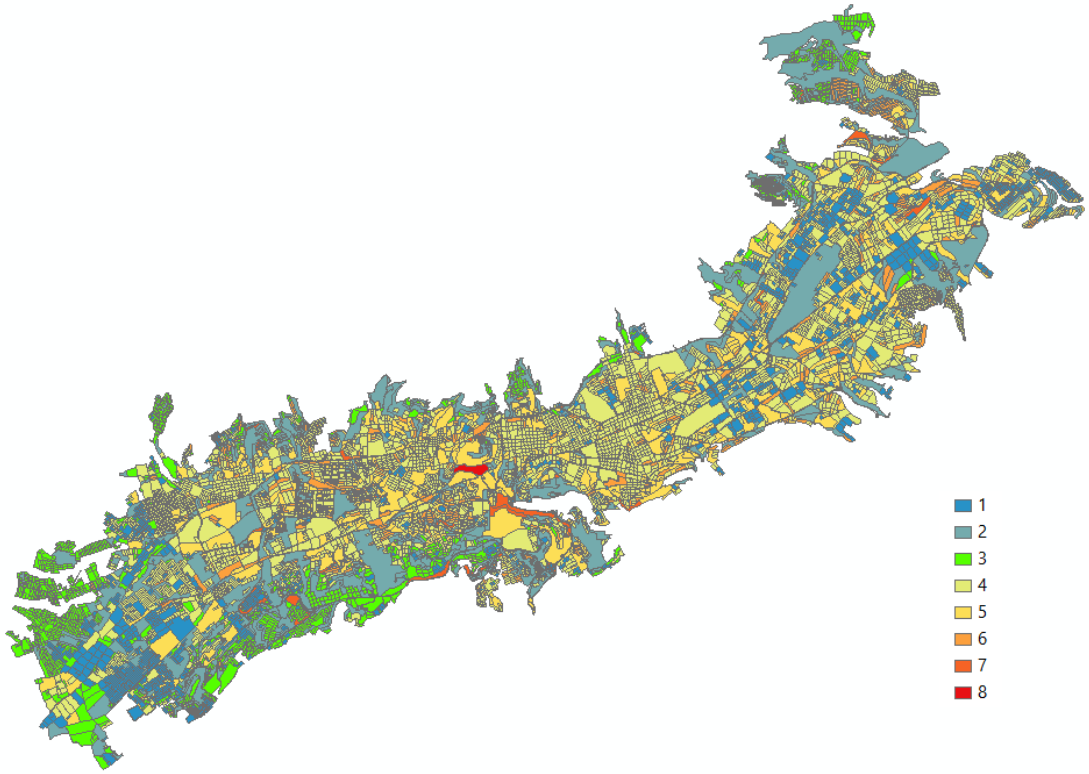
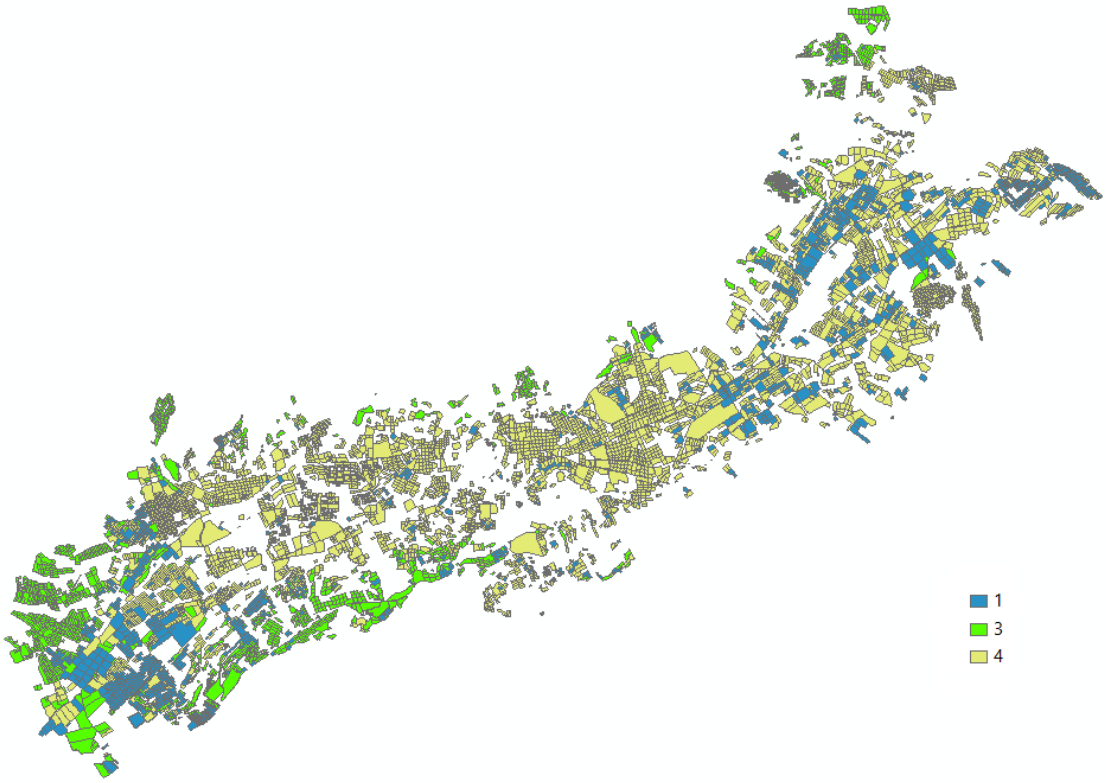


Figura 7: Ejemplos de grupos de manzanas

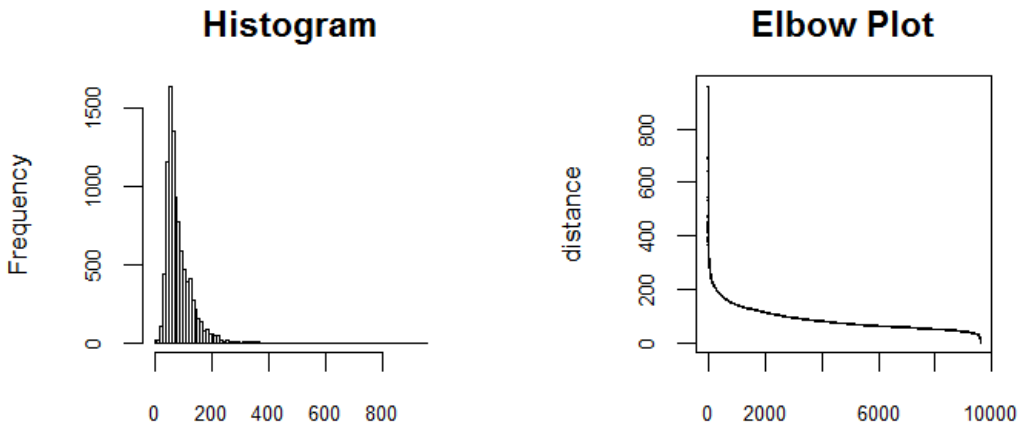
Figura 8: Grupos con disposición tipo rejilla



En la segunda etapa se apunta a segmentar la ciudad por densidad, considerando como observaciones a los centroides de los polígonos pertenecientes a los grupos tipo rejilla y como herramienta al algoritmo DBSCAN. Se utiliza la distancia de Manhattan en tanto ésta tiene mayor concordancia con la forma en que ocurren los desplazamientos de una manzana a otra en una ciudad.

Observando el histograma y el gráfico de codo de las distancias de las observaciones a sus vecinos más cercanos (figura 9), se fija el parámetro ε en 200, con lo cual se considerará al menos al 3% de las observaciones como ruido.

Figura 9: Distancias de los centroides de rejillas a sus vecinos más cercanos



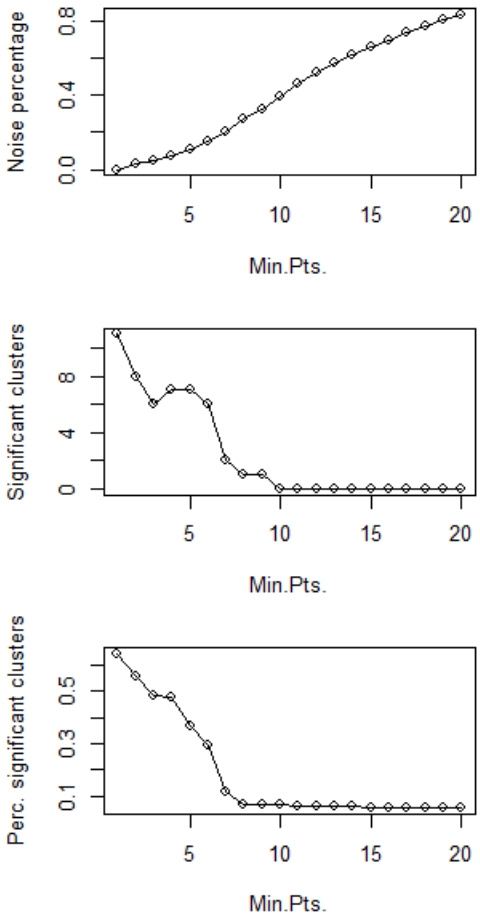
Dado ε, el número mínimo de puntos μ adecuado es a lo sumo 5 (véanse la figura 10 y la tabla 4). Considerando adicionalmente la cantidad de grupos "grandes" y el porcentaje de observaciones en los mismos, se escogió $\mu = 4$.

Con estos parámetros, el resultado de DBSCAN es una partición del mapa en 169 grupos (figura 11) con puntos ruido ubicados principalmente a lo largo de la periferia y sin concentrarse en lo que podría considerarse un barrio o sector. Cabe notar que, aunque existen grupos pequeños que podrían asociarse a un barrio, ciertamente los más grandes encierran a varios de ellos. Estos nuevos sectores, se diferencian principalmente en su altura, año de construcción, diámetro (máxima distancia entre dos puntos) y distancia de sus manzanas al centroide (figuras 20 y 21).

Por otra parte, se observa que aunque el agrupamiento se basó en atributos distintos al año de edificación de la manzana, sus resultados tienen concordancia con dicho año. En efecto, en la figura 12 se observa que la mayoría de grupos formados con $\varepsilon = 200$ y $\mu = 4$

(polígonos de borde azul) contienen manzanas de a lo sumo dos épocas, y en tal caso, dichas épocas son cercanas.

Figura 10: Indicadores de μ

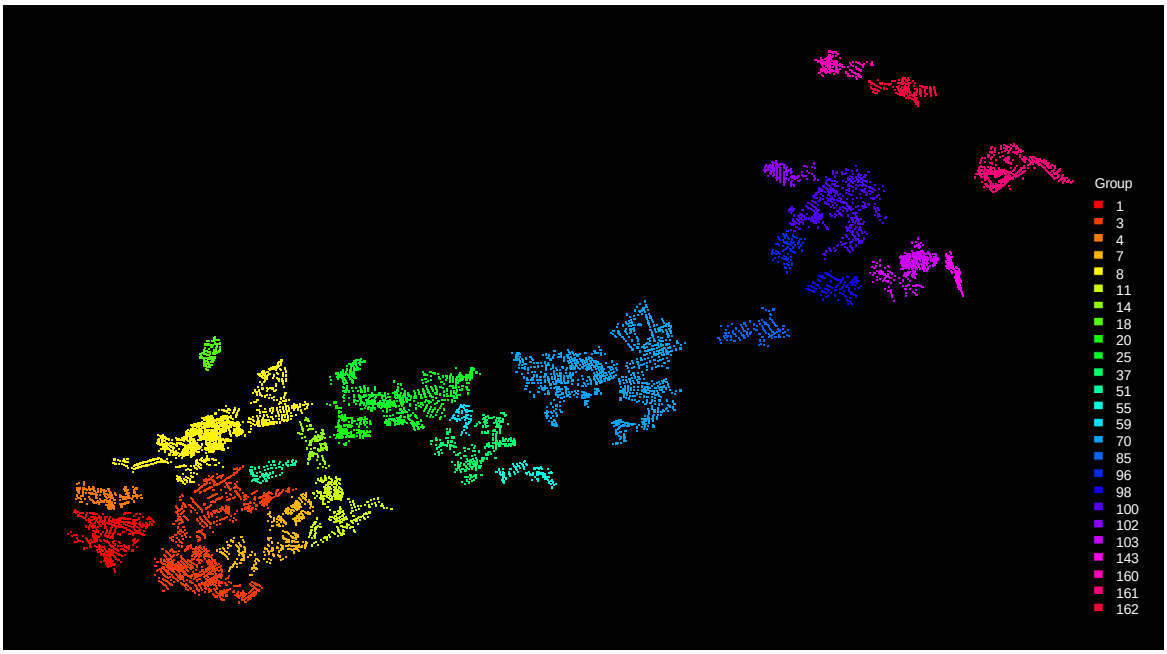


Adicionalmente, si se mantiene fijo μ y se establece ε = 261.83 (el 1% de las observaciones es ruido), los grupos resultantes (polígonos de borde rojo) siguen manteniendo la división entre norte y sur. Nótese que, en este último nivel de agregación algunas manzanas que no fueron consideradas en DBSCAN y que en parte corresponden a espacios verdes "grandes", son ahora capturadas por las envolventes de borde rojo. Cuando dichos espacios verdes se ubican en la periferia, éstos son más propensos a no ser

encerrados por las envolventes mencionadas (un ejemplo de ello es el Parque Metropolitano Guangüiltagua). Se evita su asignación a uno de los sectores obtenidos con DBSCAN, en tanto estos espacios verdes suelen ser compartidos por habitantes de diversas partes de la ciudad.

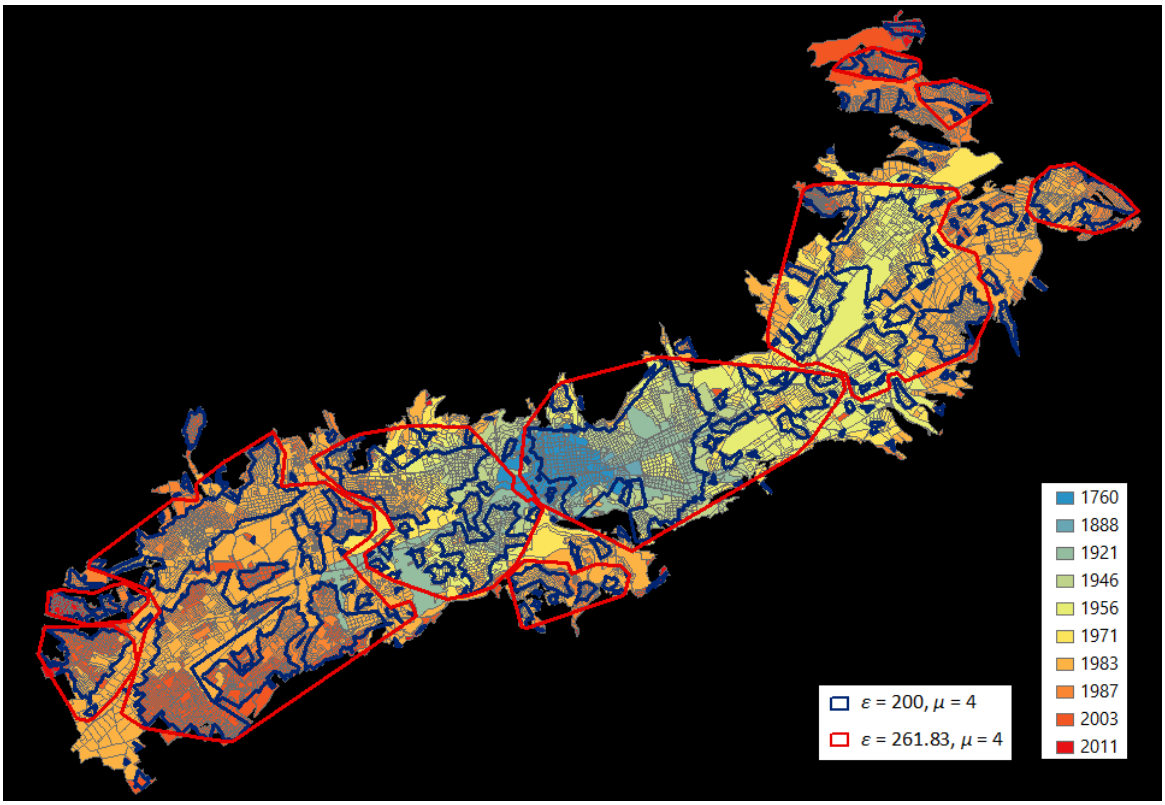
Esto da pie para pensar que, por un lado, las huellas de la planificación urbana pueden estudiarse a partir de los patrones existentes en los atributos geométricos y la densidad de las manzanas de la urbe, y por otro, que existe una especie de jerarquía en la estructura de la ciudad que determina la forma y estructura de la ciudad y que concuerda con la lectura clásica de la urbe.

Figura 11: Partición de DBSCAN con centroides de rejillas [7]



[7] Se muestran únicamente aquellos grupos cuya cardinalidad es mayor que la media.

Figura 12: DBSCAN vs. Año de edificación



Finalmente, cabe mencionar que, en el presente trabajo la diferenciación entre norte y sur se da principalmente por la altura y las distancias entre centroides de manzanas (figuras 13 y 14). Ciertamente, la diferenciación entre estas dos zonas va más allá del aspecto geométrico o espacial, sin embargo, el entendimiento de dicha diferenciación escapa al alcance del presente trabajo.

Figura 13: Diagramas de caja de Norte, Centro y Sur

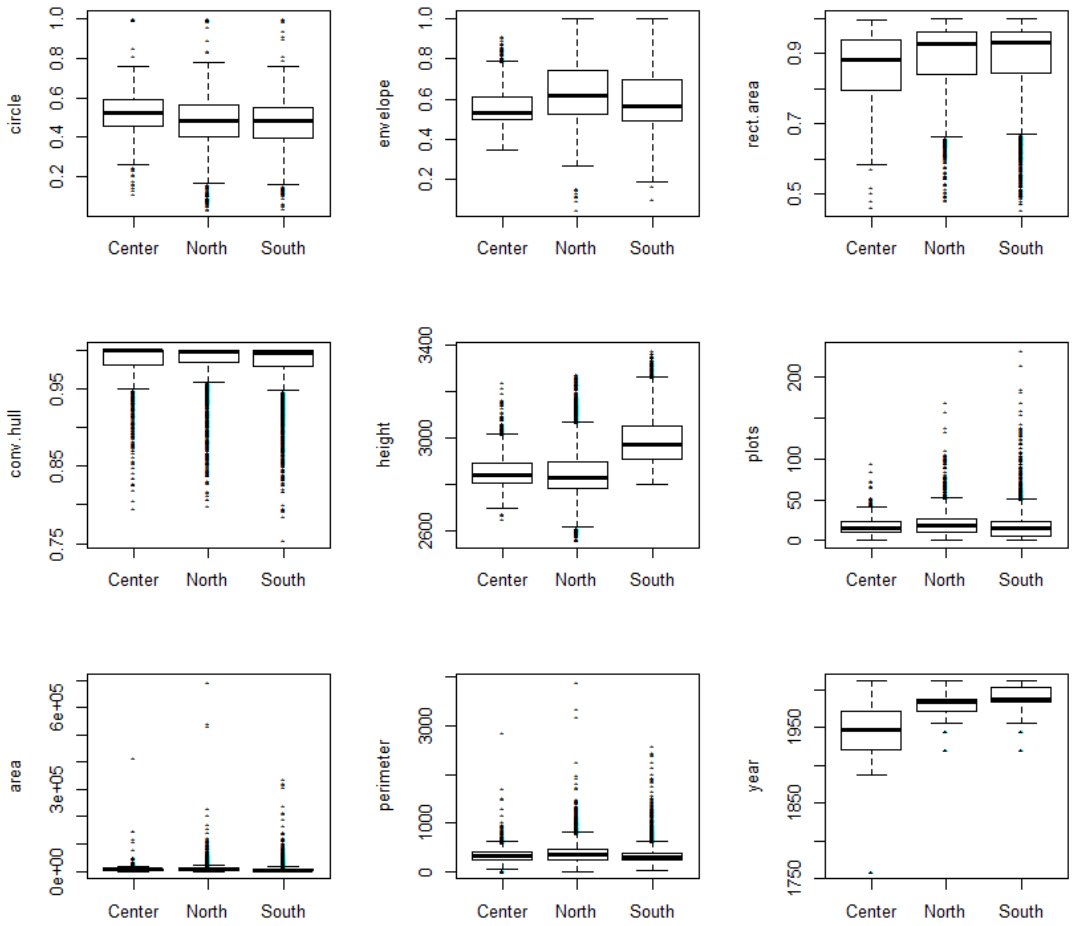
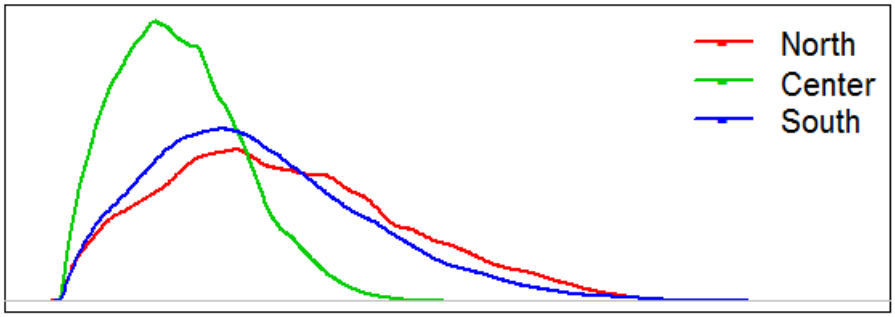


Figura 14: Densidad de distancias entre centroides



RECONOCIMIENTOS

La obtención de los datos que sirven de insumo para el agrupamiento jerárquico y las coordenadas de los centroides de cada manzana fue realizada en ArcGis 10.2. A excepción de los Índices de Moran para cuyo cálculo se utilizó GeoDa 1.6.1., el análisis de datos fue realizado en su totalidad en RStudio 0.98.1103 con los paquetes *maptools, spdep, fpc y clValid*.

CONCLUSIONES

El presente documento recoge una propuesta metodológica para segmentar una ciudad a partir de los atributos geométricos de sus manzanas y la densidad de las mismas en el mapa. Se consideró a la ciudad de Quito como caso de estudio y se observó que los resultados del agrupamiento tienen concordancia con la huella de la planificación urbana en dicha ciudad. Estos nuevos grupos pueden servir como insumo de política pública en tanto permitirían tratar problemas locales y estudiar sus interrelaciones en lugar de abordar directamente el problema global.

Puesto que existen limitaciones de información, los trabajos futuros apuntan en primera instancia a incorporar datos a nivel de edificación, uso de suelo, puntos de interés, sistemas de transporte público, flujo vehicular, semáforos, intersecciones, accesibilidad, etc. En segundo lugar, es necesario encaminar esfuerzos a caracterizar y segmentar la ciudad a partir del grafo de la red vial y contrastar los resultados con los obtenidos en el presente trabajo. Finalmente, la consideración de factores socioeconómicos (como la producción, empleo, migración, formación de barrios marginales y/o clandestinos y relaciones centro-periferia) y ambientales (como emisiones de CO_2 y generación de ruido) son vitales para avanzar hacia una mejor comprensión de la ciudad, y por ende, hacia una adecuada planificación urbana..

REFERENCIAS

Amindarbari, R., & Sevtsuk, A. Measuring Growth and Change in Metropolitan Form. Sciences, 104(17), 7301-7306.

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: an introduction to spatial data analysis. *Geographical analysis*, *38*(1), 5-22.

ArcMap, A. (2013). 10.2. *Environmental Systems Research Institute, Redlands, CA*.

Batty, M. (2012).Building a science of cities. *Cities*, *29*, S9-S16.

Batty, M., & Longley, P. A. (1994). *Fractal cities: a geometry of form and function*. AcademicPress.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., ... & Carvalho, M. (2015). Package 'spdep'. https://cran.r-project.org/web/packages/spdep

Bivand, R., & Lewin-Koh, N. (2014). maptools: Tools for reading and handling spatial objects. R package version 0.8-29. http://CRAN.R-project.org/package=maptools

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software (Brock et al., March 2008)*.

Clifton, K., Ewing, R., Knaap, G. J., & Song, Y. (2008). Quantitative analysis of urban form: a multidisciplinary review. *Journal of Urbanism*, *1*(1), 17-45.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

Gan, G., Ma, C., & Wu, J. (2007).Data clustering: theory, algorithms, and applications (Vol. 20). Siam.

Gil, J., Montenegro, N. C., Beirão, J. N., & Duarte, J. P. (2009). On the Discovery of Urban Typologies: Data Mining the Multi-dimensional Character of Neighbourhoods.

Guo, J. Y., & Bhat, C. R. (2007).Operationalizing the concept of neighborhood: Application to residential location choice analysis. *Journal of Transport Geography*, *15*(1), 31-45.

Handl, J., Knowles, J., & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. Bioinformatics, 21(15), 3201-3212.

Hecht, R., Herold, H., Meinel, G., & Buchroithner, M. (2013). Automatic derivation of urban structure types from topographic maps by means of image analysis and machine learning. In *26th International cartographic conference–Proceedings: International cartographic association.< http://icaci.*

org/files/documents/ICC_proceedings/ICC2013/_extendedAbstract/362_proceeding.pdf> Accessed (Vol. 17, p. 14).

Hennig, C. (2013). fpc: Flexible procedures for clustering. R package version 2.1-5. http://CRAN.R-project.org/package=fpc

Joshi, D., Samal, A. K., & Soh, L. K. (2009, March). Density-based clustering of polygons. In Computational Intelligence and Data Mining, 2009.CIDM'09. IEEE Symposium on (pp. 171-178). IEEE.

Louf, R., & Barthelemy, M. (2014).A typology of street patterns. Journal of The Royal Society Interface, 11(101), 20140924.

Marshall, S. (2004). *Streets and patterns*. Routledge.

Murtagh, F., & Legendre, P. (2011). Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. arXiv preprint arXiv:1111.6285.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schirmer, P. M., & Axhausen, K. W. (2015).A multiscale classification of urban morphology. Journal of Transport and Land Use.

Serra, M., Lopes Gil, J. A., &Pinho, P. (2013, October). Unsupervised classification of evolving metropolitan street patterns. In *Proceedings of Ninth International Space Syntax Symposium, October 31-November 3, 2013, Seoul, Korea. Eds.: Kim, YO, Park, HT, Seo, KW Paper 46*. Sejong University Press.

Sevtsuk, A. (2010). *Path and place: a study of urban geometry and retail activity in Cambridge and Somerville, MA* (Doctoral dissertation, Massachusetts Institute of Technology).

Southworth, M., & Ben-Joseph, E. (2003).*Streets and the Shaping of Towns and Cities*. Island Press.

Studio, R. (2012). R Studio: integrated development environment for R.

Wang, X., Gu, W., Ziebelin, D., & Hamilton, H. (2010). An ontology-based framework for geospatial clustering. *International Journal of Geographical Information Science*, *24*(11), 1601-1630.

Wilson, A. (2012). *The Science of Cities and Regions: Lectures on Mathematical Model Design*. Springer Science & Business Media.

Zhang, J., Samal, A., &Soh, L. K. (2005).Polygon-Based Spatial Clustering. In *Proceedings of the 8th International Conference on GeoComputation*.

ANEXO A: TABLAS Y GRÁFICOS ADICIONALES

Tabla 1: Matriz de Correlaciones de Pearson

	circle	envelope	rect.area	rect.width	conv.hull	height	plots
circle	1	0.637	0.542	0.519	0.503	0.11	0.058
envelope	0.637	1	0.493	0.481	0.485	0.075	0.108
rect.area	0.542	0.493	1	0.996	0.777	0.094	0.008
rect.width	0.519	0.481	0.996	1	0.768	0.093	0.007
conv.hull	0.503	0.485	0.777	0.768	1	0.084	-0.076
height	0.11	0.075	0.094	0.093	0.084	1	-0.054
plots	0.058	0.108	0.008	0.007	-0.076	-0.05	1

Figura 15: Diagramas de dispersión de los datos

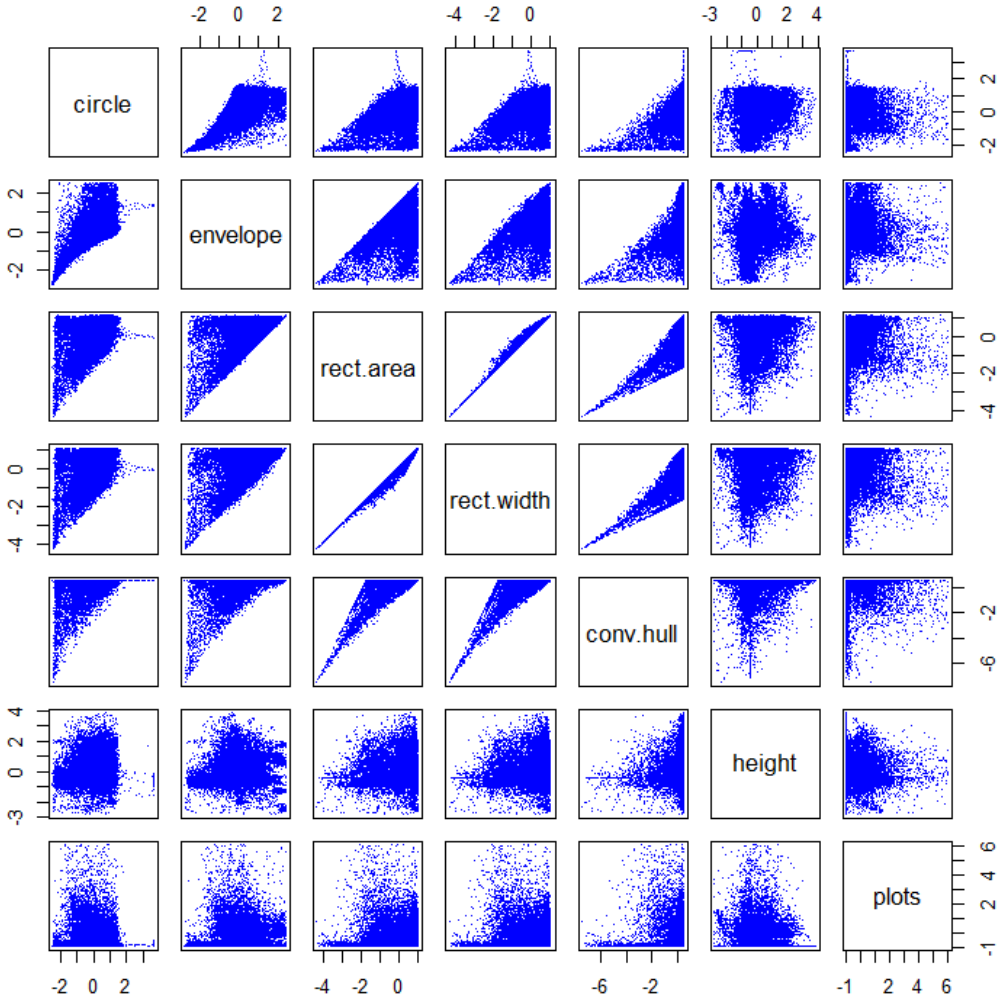


Figura 16: Histogramas de los datos

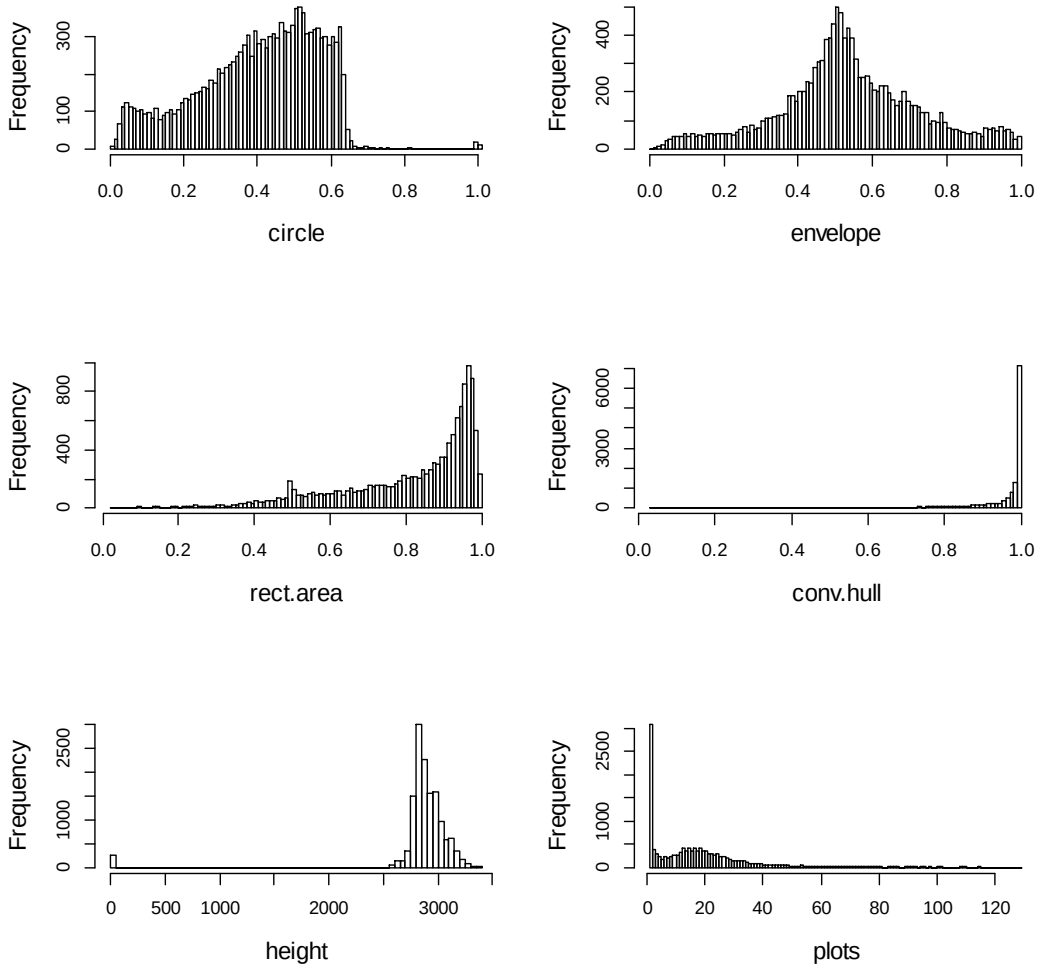


Tabla 2: Índice de Moran

	circle	envelope	rect.area	conv.hull	height	plots
I	0.318	0.521	0.394	0.248	0.971	0.051
p-valor*	0.001	0.001	0.001	0.001	0.001	0.002

* pseudo p- valor basado en 1000 permutaciones. Ho: aleatoriedad espacial.

Figura 17: Diagramas de dispersión de Moran

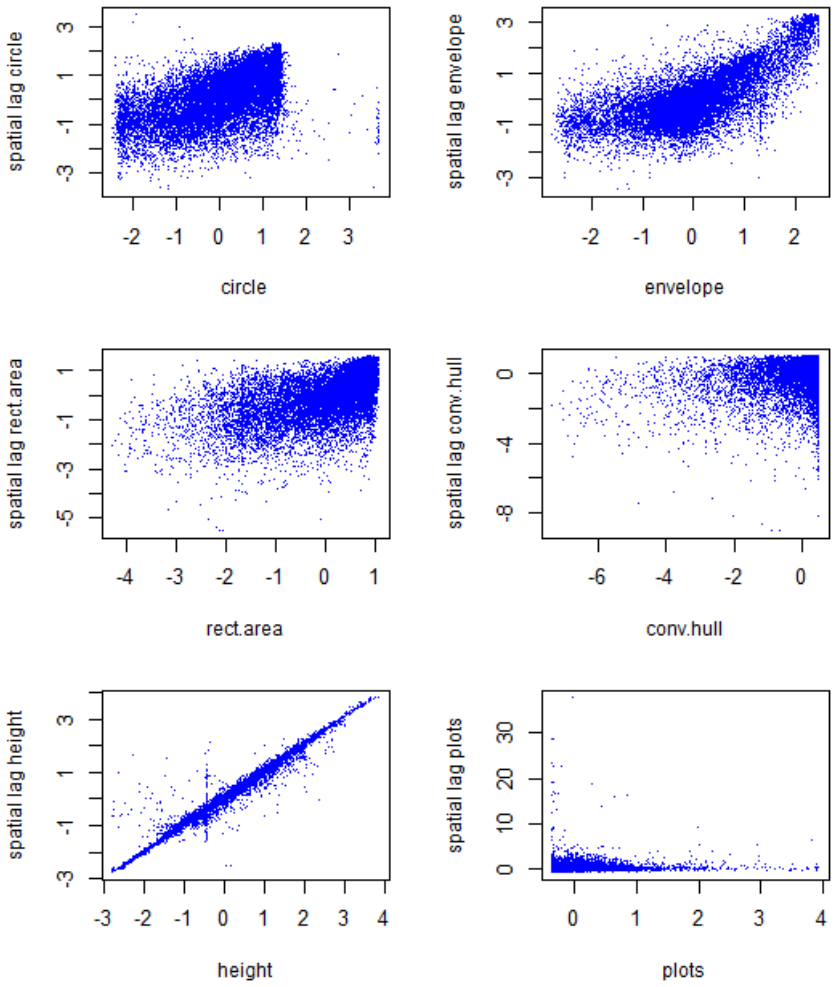
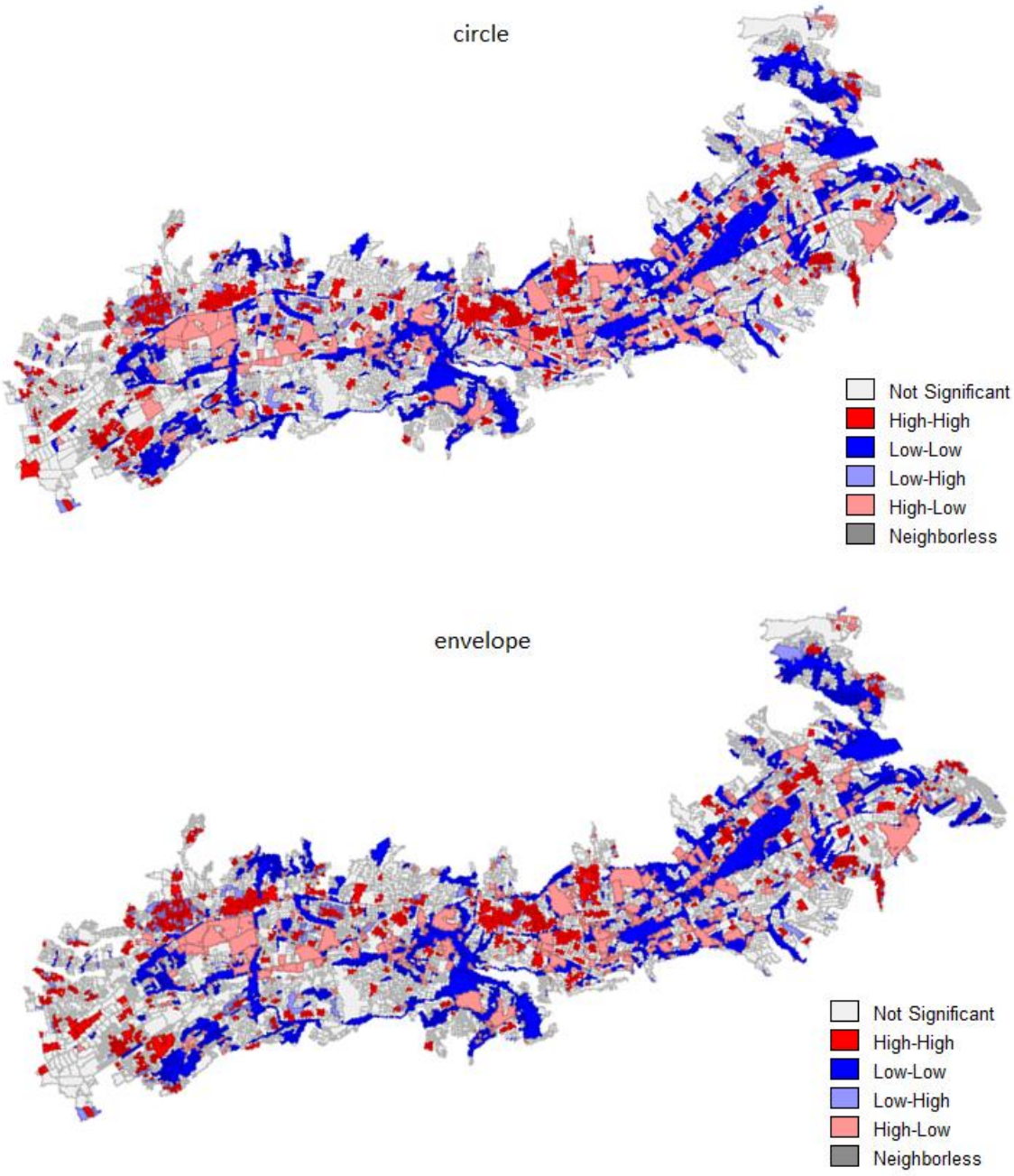
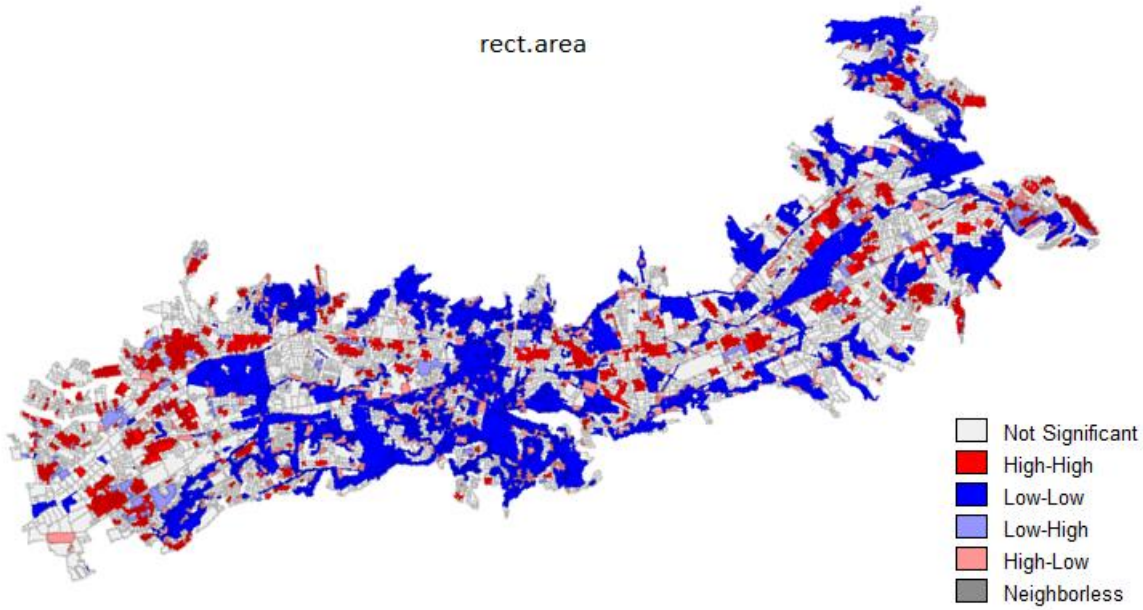
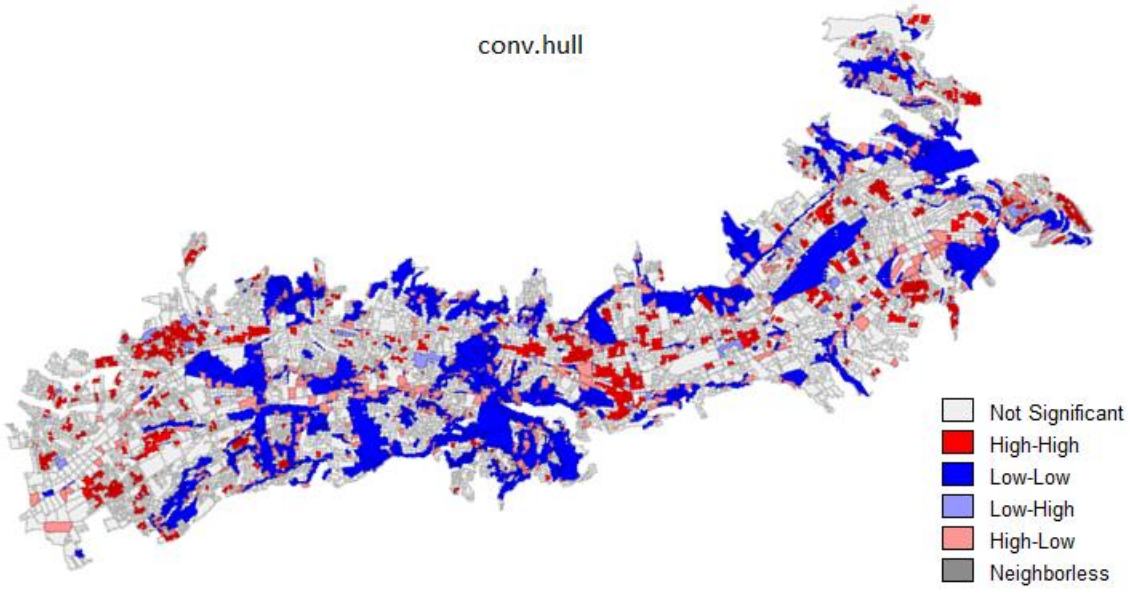
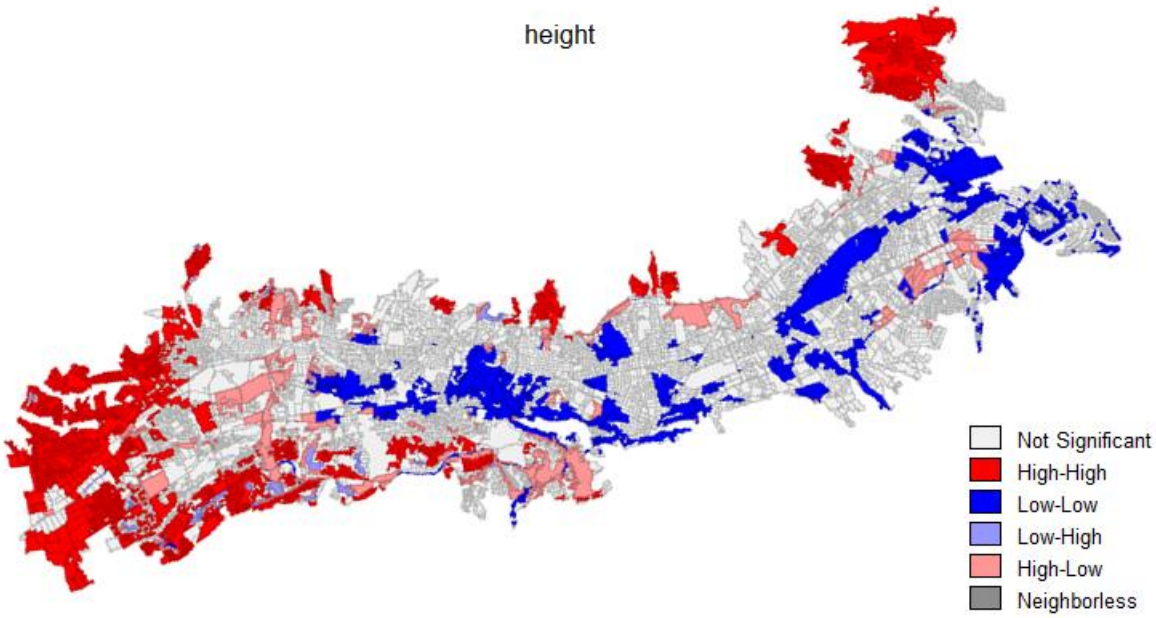


Figura 18: Mapas del Índice de Moran Univariante Local




rect.area
Not Significant
High-High
Low-Low
Low-High
High-Low
Neighborless


conv.hull
Not Significant
High-High
Low-Low
Low-High
High-Low
Neighborless


height
Not Significant
High-High
Low-Low
Low-High
High-Low
Neighborless

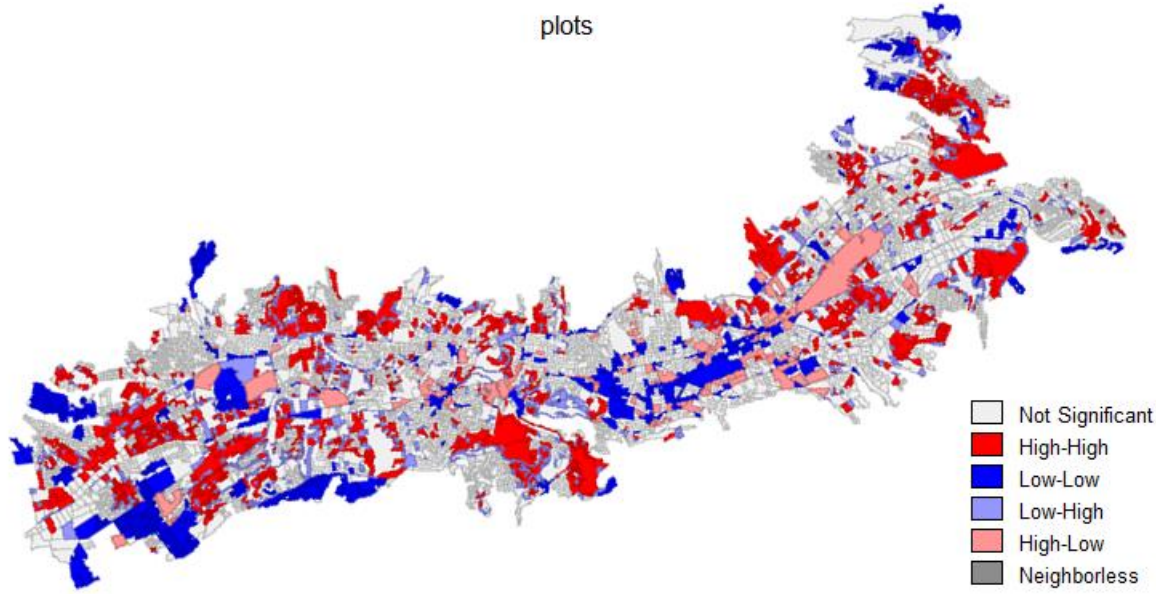

plots
Not Significant
High-High
Low-Low
Low-High
High-Low
Neighborless

Tabla 3: Índices de Validación Interna del Agrupamiento Jerárquico

N. grupos	Índice		
	Connectivity	Dunn	Silhouette
3	1226.629	0.00133	0.19065
4	1859.541	0.00131	0.15987
5	2055.294	0.00131	0.15675
6	2199.174	0.00128	0.13984
7	2374.060	0.00128	0.15092
8	2474.865	0.00128	0.16938
9	2558.349	0.00128	0.16395
10	2633.680	0.00128	0.16252
11	2796.524	0.00128	0.16696
12	2922.676	0.00081	0.11967
13	3071.795	0.00081	0.12046
14	3201.800	0.00081	0.12734
15	3202.466	0.00085	0.13370

Figura 19: Distancias de todos los centroides a sus vecinos más cercanos

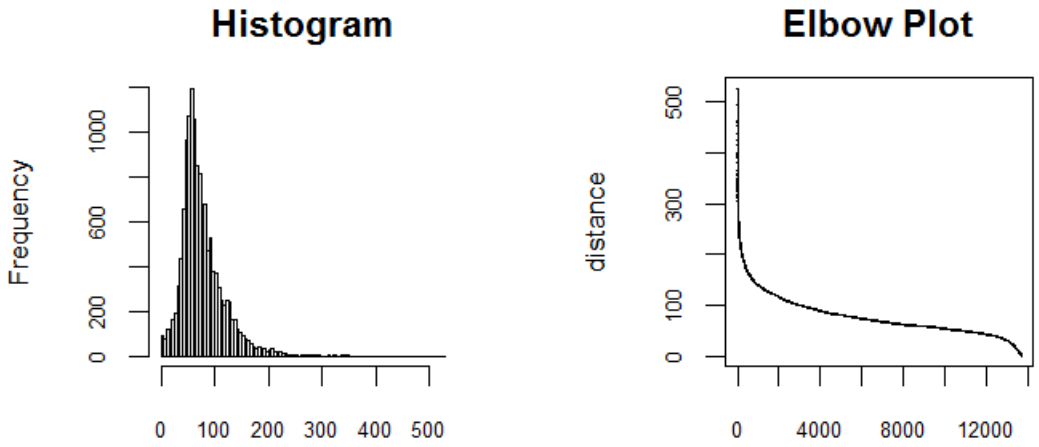
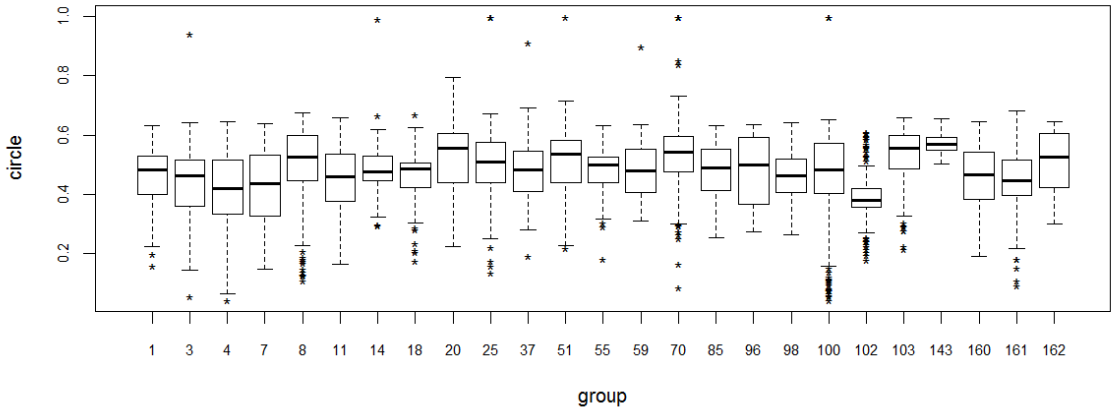


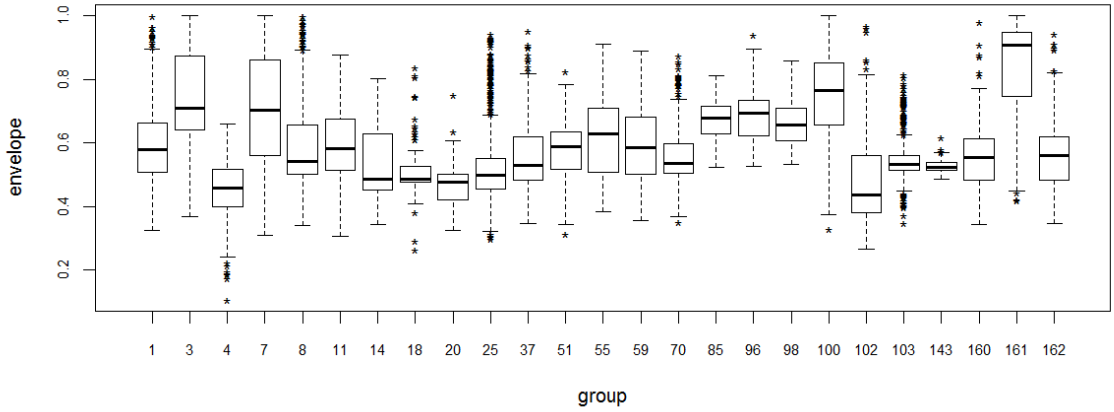
Tabla 4: Indicadores DBSCAN con centroides de rejillas

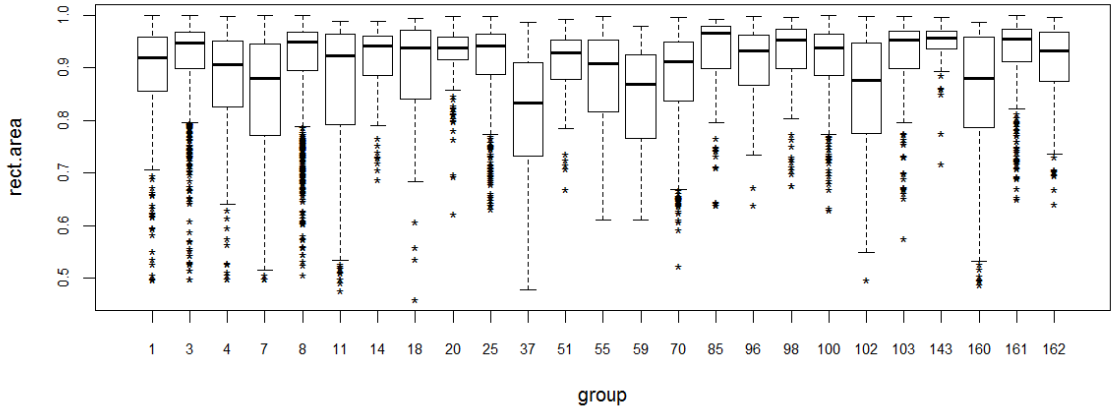
mu	% noise	N. sig. clust.	% Sig. clust.
1	0	11	0.645
2	0.032	8	0.556
3	0.047	6	0.485
4	0.072	7	0.476
5	0.108	7	0.366
6	0.152	6	0.294
7	0.207	2	0.116
8	0.272	1	0.065
9	0.324	1	0.065
10	0.392	0	0.065
11	0.463	0	0.064
12	0.524	0	0.064
13	0.574	0	0.063
14	0.620	0	0.063
15	0.657	0	0.056
16	0.699	0	0.056
17	0.735	0	0.056
18	0.774	0	0.055
19	0.807	0	0.054
20	0.833	0	0.054

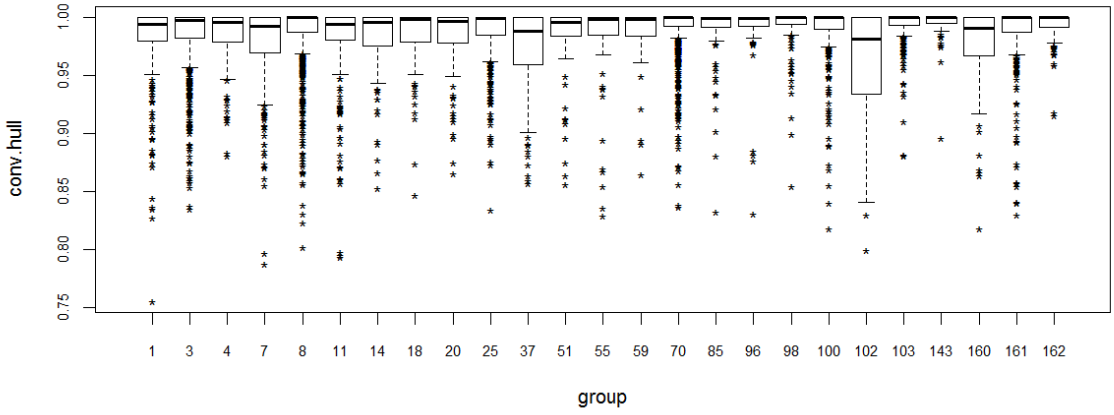
Figura 20: Diagramas de caja de grupos finales [8]

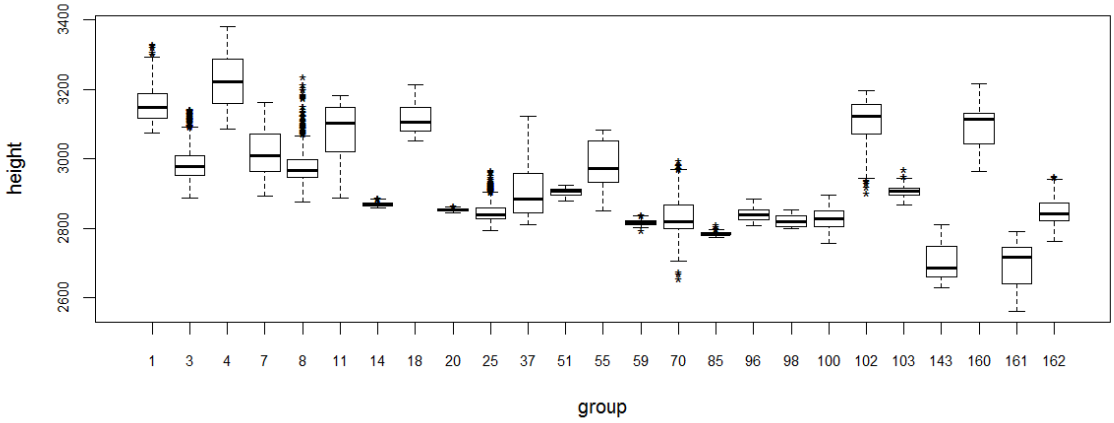


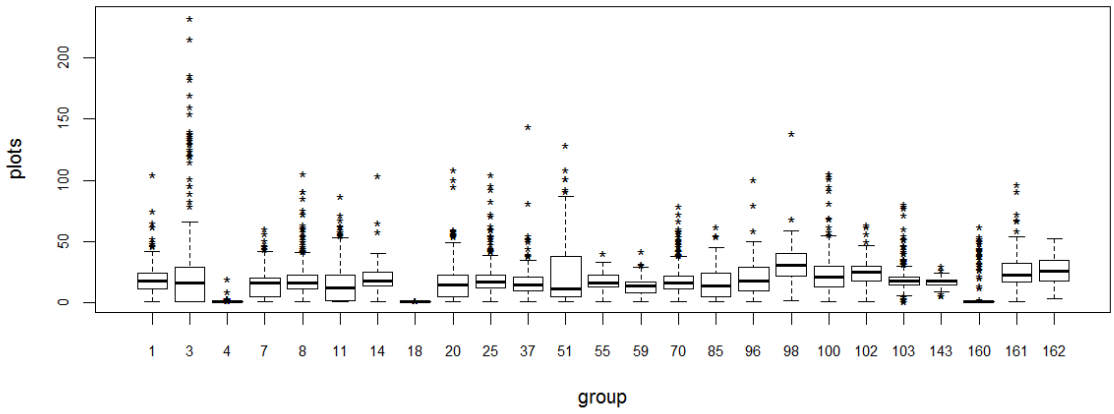
[8] Se muestran únicamente aquellos grupos cuya cardinalidad es mayor que la media.

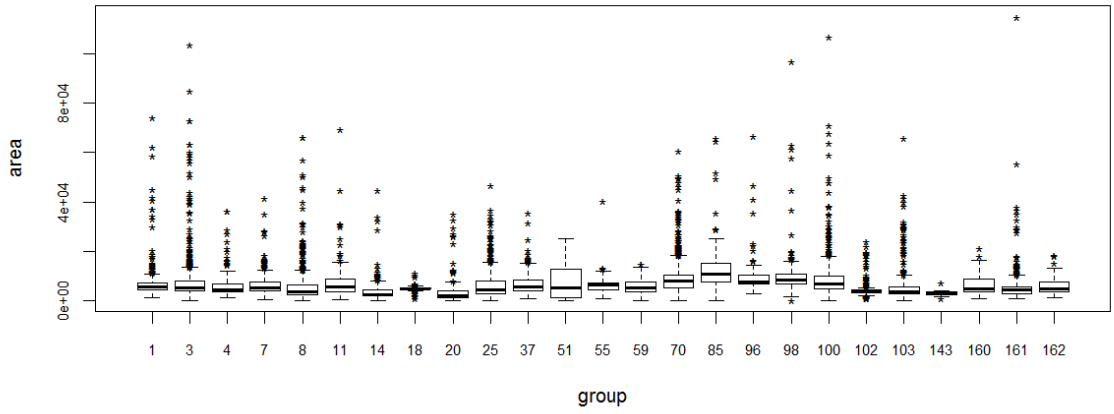

envelope
0.2
0.4
0.6
0.8
1.0
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group

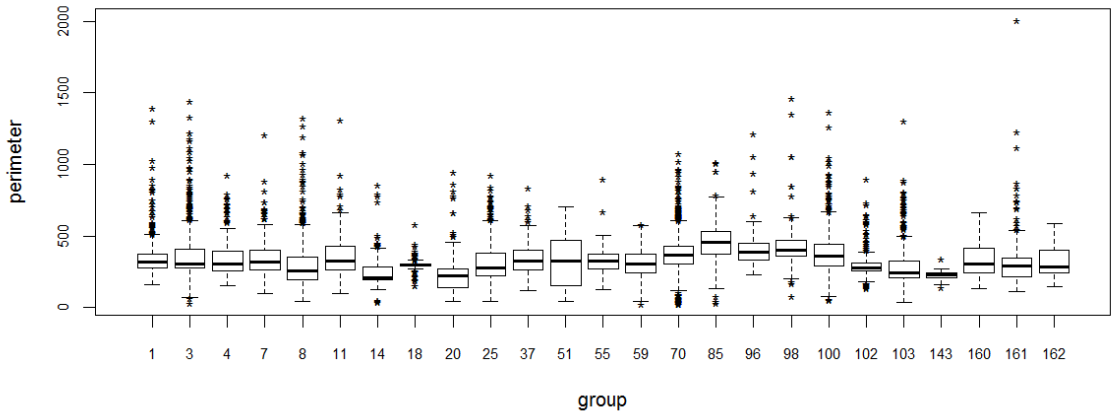

rect.area
0.5
0.6
0.7
0.8
0.9
1.0
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group

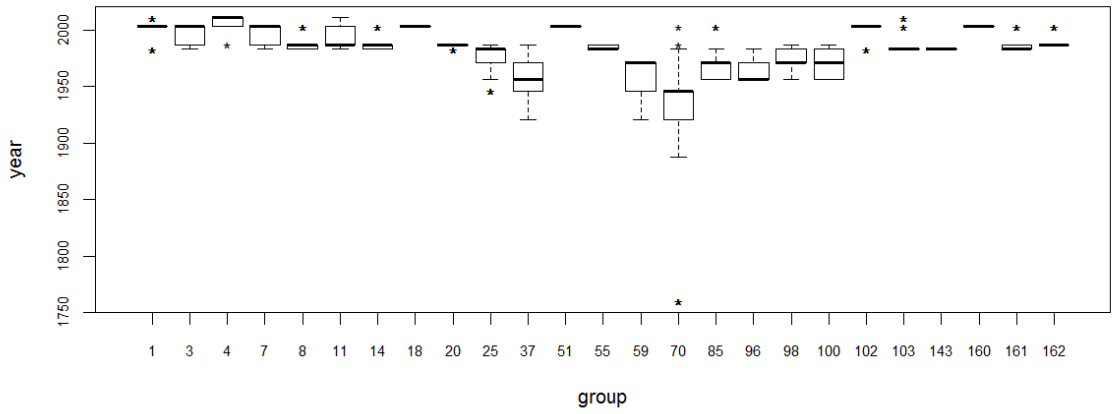

conv.hull
0.75
0.80
0.85
0.90
0.95
1.00
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group


height
2600
2800
3000
3200
3400
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group


plots
0
50
100
150
200
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group


area
0e+00
4e+04
8e+04
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group


perimeter
0
500
1000
1500
2000
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group


year
1750
1800
1850
1900
1950
2000
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group

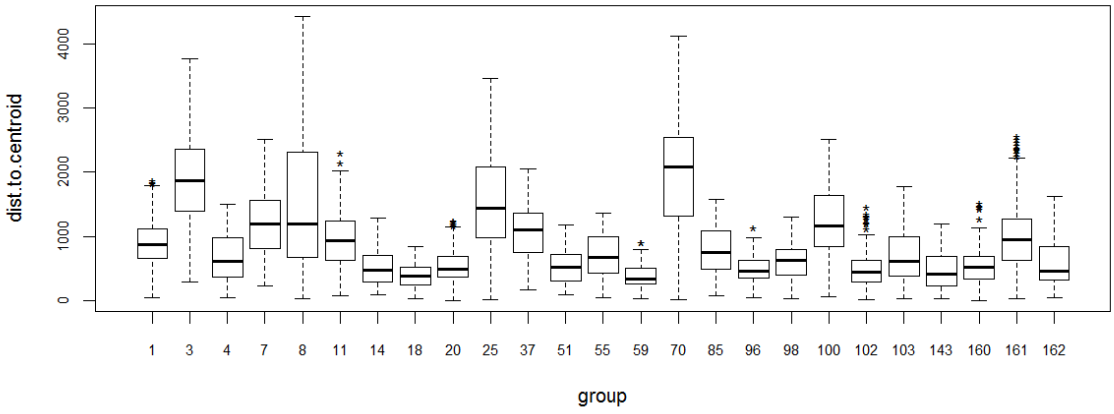
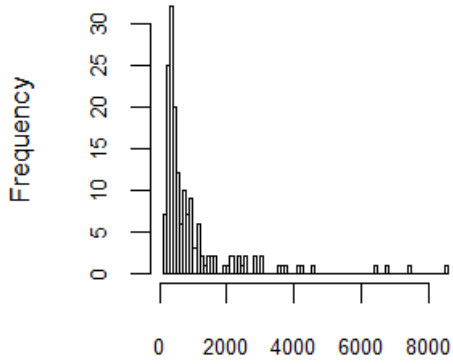

dist.to.centroid
0
1000
2000
3000
4000
1 3 4 7 8 11 14 18 20 25 37 51 55 59 70 85 96 98 100 102 103 143 160 161 162
group

Figura 21: Histograma del Diámetro de Grupos



ANEXO B: MEDIDAS DE VALIDACIÓN INTERNA[9]

Connectivity. Sea un conjunto de datos de *N* observaciones y *M* variables. Sea $nn_{i(j)}$ el *j*-ésimo vecino más cercano de la observación *i* y defínase

$$x_{i,nn_{i(j)}} = \begin{cases} 0, & si\ i\ y\ j\ están\ en\ el\ mismo\ grupo \\ \frac{1}{j}, & caso\ contratrio \end{cases}$$

Para una partición de los *N* datos $C = \{C_1, C_2, \dots, C_k\}$ en *k* grupos disjuntos, la conectividad se define como:

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}},$$

donde *L* corresponde al número de vecinos más cercanos a considerar. La conectividad es un número mayor o igual que cero y debe ser minimizado.

Silhouette Width. Es una combinación no lineal de compacidad y separación. En particular, es el promedio del valor de la *Silhouette* de cada observación. Ésta última mide el grado de confianza en la asignación de una observación a un grupo, tomando valores cercanos a -1 cuando la observación está "mal asignada" a un grupo y cercanos a 1 cuando está "bien asignada". Para una observación *i*, este indicador se define como:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)},$$

donde a_i es la distancia promedio entre la observación *i* el resto de observaciones del mismo grupo, y b_i es la distancia promedio entre *i* y las observaciones del "grupo más cercano", es decir

[9] Tomado de Brock et al. (2011)

$$b_i = \min_{C_k \in C \setminus C(i)} \sum_{j \in C_k} \frac{d(i,j)}{|C_k|},$$

donde $C(i)$ es el grupo que contiene a la observación *i* y $d(i,j)$ la distancia entre las observaciones *i* y *j*. El indicador *Silhouette Width* está en el intervalo $[-1,1]$ y deber ser maximizado.

Dunn Index. Es una combinación no lineal de compacidad y separación. En particular, es la razón entre la menor distancia entre observaciones que no se encuentran en el mismo grupo y la mayor distancia intra-grupo, es decir:

$$D(C) = \frac{\min_{C_k, C_l \in C, C_k \neq C_l} \left(\min_{i \in C_k, j \in C_l} d(i,j) \right)}{\max_{C_m \in C} diam(C_m)},$$

donde $diam(C_m)$ es la máxima distancia entre observaciones del grupo C_m. *Dunn Index* toma valores mayores o iguales que cero y debe ser maximizado.