Pontificia Universidad
JAVERIANA
Cali

Aplicación de modelos de clusterización para analizar patrones comerciales en la Calle 5 de Cali: Impacto de equipamientos y estructura vial

Juan Fernando Gutiérrez
ID. 8992507
Juan Camilo López
ID. 0224993

Proyecto Aplicado para optar al título de Magister en Ciencia de Datos

Director
Gustavo Adolfo Arteaga

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS SANTIAGO
DE CALI, 2025

Contenido

Índice de Ilustraciones

Índice de Tablas

Índice de Gráficos

INTRODUCCIÓN

La configuración urbana obedece a procesos complejos y multivariados, los estudios realizados sobre procesos de consolidación urbana son tan diversos como las ciudades que existen. Lo anterior no es casual, pues la manera en la que se disponen los espacios y se planifican los territorios puede obedecer a mecanismos económicos [1], sociales ([2], [3]), culturales [4] y arquitectónicos [5]. Entender cómo se dan estas transformaciones se vuelve cada vez más relevante para la planificación urbana dada la velocidad con la que se dan los cambios en la composición de una ciudad. Por eso, este proyecto apunta a examinar un fragmento de este fenómeno en la ciudad de Cali, analizando patrones en los establecimientos comerciales y su relación con los equipamientos urbanos y la estructura vial del corredor de la calle 5.

Teniendo en cuenta la complejidad en la composición de una ciudad, la relevancia que han tomado las ciencias de la computación en años recientes y el crecimiento exponencial en las capacidades computacionales para el manejo de datos, no es casual que las herramientas del análisis de datos estén siendo implementadas en los análisis urbanísticos. Algunos ejemplos de las ventajas que esto ha generado son entendimiento de patrones en volúmenes altos de datos [6], la predicción del avance de la gentrificación en los barrios [7], y la estimación de lugares óptimos para instalar paradas de buses [8]. Siguiendo los aprendizajes de otras investigaciones que emplean Machine Learning para estudiar las configuraciones urbanas, este proyecto pretende generar un modelo de agrupamiento que permita identificar patrones en la consolidación de usos comerciales del corredor de la Calle 5 entre la carrera 1 y carrera 50.

Para esto, se recogieron datos sobre la composición de las manzanas en el sector seleccionado, entre los que están la presencia de equipamientos urbanos, estaciones de transporte, espacios públicos como plazas y parques, además de establecimientos comerciales de diferente naturaleza. Adicionalmente, se calcularon indicadores de movilidad urbana para el sector priorizado, que permitieran complejizar el análisis e involucrar la manera en que las personas recorren los espacios urbanos.

El resultado de este ejercicio aporta un análisis cuantitativo que impulse la generación de discusiones y soluciones frente a los procesos de transformación urbana a través de dos resultados concretos: (1) un modelo que permita identificar patrones consolidación de usos comerciales en el corredor de la calle 5 y (2) un conjunto de conclusiones sobre las variables que inciden en la configuración urbana en el corredor de la calle 5.

1. Contextualización del proyecto

1.1 Definición del problema

1.1.1 Planteamiento del problema

El corredor de la calle 5 comprende varios lugares de interés para el desarrollo urbano de la ciudad de Cali, entre ellos están el barrio de San Antonio y buena parte de la comuna 3 que son áreas de Interés Patrimonial y preservación urbanística. Debido al atractivo que genera la ubicación de las manzanas de este corredor, en las últimas décadas se ha convertido en un epicentro de importantes actividades culturales y comerciales que han derivado en marcados cambios estructurales, funcionales, territoriales y poblacionales que exigen nuevos retos en su tratamiento e intervención. La centralidad de la zona más las dinámicas comerciales y culturales que tienen lugar en ella han sido detonadoras de procesos de transformación urbana y social que han impulsado la reestructuración del sector.

Actualmente, siendo una zona consolidada de la ciudad, pero aun sujeta a cambios por su atracción de distintos usos comerciales y residenciales, entender las realidades del corredor constituye un elemento esencial en la gestión del territorio al ser uno de los ejes que mayor impacto tienen en la estructura y función de la ciudad. De esta forma, pensar en una herramienta que sirva como un insumo que sea susceptible de replicación, no solo en otras zonas de la ciudad, sino en función de otras variables que puedan complementar el análisis, plantea un paso para comprender de manera más integral las realidades del territorio en función de los datos que arroja y, así, generar respuestas normativas y de política pública más aterrizadas y flexibles.

Por esta razón, se requiere comprender las dinámicas que han llevado a esta estructuración socioespacial dinámica y fluctuante y, así, plantear herramientas analíticas que permitan entender el funcionamiento del territorio, caracterizar los procesos de transformación urbana que pueden estarse dando y sus implicaciones en términos de retos y oportunidades que deben ser abordadas desde la gestión pública.

Ahora, usando las herramientas de la Ciencia de Datos, se aborda este proyecto con la prueba y validación de modelos de agrupamiento de Machine Learning que permitan identificar patrones en la consolidación de usos comerciales del corredor, una dimensión importante para la compresión de la configuración urbana del sector. En ese sentido, el problema fundamental en el aspecto técnico de la Ciencia de Datos gira alrededor de la definición de variables relevantes para la generación de modelos de agrupación, la escogencia y depuración de datos

con valor analítico que permitan el correcto funcionamiento del modelo y la optimización de los modelos para el caso particular del proyecto.

1.1.2 Formulación del problema

La pregunta de investigación que guía el proyecto es: ¿Qué modelo de agrupación de Machine Learning se ajusta mejor a las particularidades del caso del corredor de la calle 5 en Cali para la identificación de patrones en la consolidación de usos comerciales del corredor?

1.1.3 Preguntas de sistematización

Las preguntas de sistematización que alimentan el proceso son: ¿cuáles son las variables socioeconómicas y territoriales y los modelos de Machine Learning que se han priorizado en estudios previos que analizan las transformaciones urbanas a escalas barriales?, ¿qué técnicas se deben emplear para la armonización de fuentes de datos oficiales y no oficiales que hacen seguimiento al desarrollo socioeconómico y urbano de Santiago de Cali?, ¿cuál es el modelo de agrupación más apto para caracterizar, a través del agrupamiento, la transformación urbana la consolidación de usos comerciales en el corredor de la calle 5 en Cali? Y ¿qué dinámicas aportan a la consolidación de los usos comerciales en el corredor de la calle 5 en Cali?

1.2 Objetivos del Proyecto

1.2.1 Objetivo General

Generar un modelo de agrupamiento basado en Machine Learning que permita identificar patrones en la consolidación de usos comerciales del corredor de la Calle 5 entre la carrera 1 y carrera 50.

1.2.3 Objetivos Específicos

1) Revisar bibliografía pertinente sobre la identificación de variables relevantes para la consolidación y perfilamiento de usos de suelo en las ciudades mediante técnicas de análisis espacial y Machine Learning.
2) Construir una base de datos con valor analítico para la generación de un modelo de agrupación de Machine Learning.
3) Validar modelos de agrupamiento según su capacidad para la identificación de patrones en la consolidación de usos comerciales en el área de estudio priorizada.
4) Reconocer el impacto que tienen las dinámicas urbanas en la consolidación de usos y características concretas de determinadas zonas de la ciudad.

1.3 Marco de Referencia

1.3.1 Marco Teórico

El análisis urbano y las técnicas de Ciencia de Datos

Desde el análisis urbano se han adelantado aplicaciones de técnicas que han llevado a la generación de conocimiento y soluciones en la materia, reconociendo que, por ejemplo, enfoques tradicionales de la planificación basados en la zonificación y el transporte se han quedado cortos para atender las exigencias de las ciudades modernas [9]. Esta reconceptualización de la planificación urbana implica un enfoque integral frente a las problemáticas de las ciudades, reconociendo el papel fundamental que desempeñan, de manera interrelacionada, las infraestructuras físicas y digitales en la transformación de los espacios urbanos en sistemas más adaptables y resilientes. Bajo este contexto, la Ciencia de Datos es un instrumento para soportar el futuro desarrollo urbano a través de modelos analíticos y predictivos frente a problemáticas urbanas específicas.

Entonces, se requiere aprovechar la producción acelerada y en masa de información que se genera mundialmente en la actualidad y ponerla al servicio de la planificación urbana e, incluso, la arquitectura, de manera que los planificadores urbanos también aborden la aplicación de métodos y tecnologías para la recolección y procesamiento de información en tiempo real. De hecho, estos cambios en la disciplina han derivado en la conceptualización de un nuevo campo de conocimiento que, desde la academia y la práctica, deberá fomentarse y robustecerse: Ciencias Urbanas (*urban science*). Esto implica también la necesidad de lograr balances en la forma de interpretar los datos y no caer en lo que se denomina "tiranía de los datos" [10], en la que se pasa por alto, de manera reduccionista, la dimensión humana e impredecible de los comportamientos individuales y sociales. Así mismo, la arquitectura puede hacer uso de la Ciencia de Datos para la identificación de la idoneidad de materiales, consumo de energía, interacción con aspectos climáticos y comportamientos del usuario para diseñar mejores infraestructuras.

De esta forma, haciendo uso de datos para abordar comportamientos ciudadanos y dinámicas urbanas, se han hecho aportes a la planificación urbana a partir de las siguientes aplicaciones de la Ciencia de Datos, las cuales no sólo potencian herramientas y técnicas tradicionales de planeación, sino que introducen nuevos paradigmas [11]:

1. Modelamiento predictivo: con el fin de llevar a cabo, anticipadamente, medidas que den respuesta al crecimiento poblacional a través de la generación de infraestructura, vivienda y transporte.
2. Optimización de servicios por medio de la recolección de datos en tiempo real.
3. Optimización de la movilidad urbana a través de la identificación de patrones de tráfico.
4. Ordenamiento Territorial, a partir de imágenes satelitales, se identifican patrones para definir las localizaciones y usos del suelo óptimos.

5. Análisis de equidad social: observando patrones demográficos y socioeconómicos se pueden plantear intervenciones urbanas que mejoren el acceso a servicios sociales y urbanos de manera equitativa.
6. Gestión del impacto ambiental: proyección y reducción de consecuencias ambientales de distintos proyectos y dinámicas urbanas (ruido, agua y aire).
7. Respuesta a emergencias: se analizan datos en tiempo real frente a catástrofes con el fin de priorizar estrategias de mitigación y rescate.
8. Análisis de opinión pública: identificación de sentimientos y opiniones ciudadanas en torno a la estructura, funcionamiento y servicios de la ciudad.
9. Mitigación de islas de calor urbanas a través de su identificación y aplicación de estrategias de urbanismo verde y ambiental.
10. Planificación del turismo: se observan patrones de acceso a servicios y se toman decisiones informadas para incentivar el turismo y mejorar las experiencias de los visitantes y habitantes.

De esta forma, diferentes ciudades han avanzado en esta reconceptualización y aplicación de nuevos paradigmas [12], resaltando casos como los de Barcelona, donde la gestión de los recursos hídricos y energéticos se soporta enteramente en el análisis de bases de datos integradas, o Singapur, que adelanta una planificación territorial basada en datos para anticipar retos urbanos. Así mismo, se resalta lo adelantado por Corea del Sur en el manejo de datos espaciales para la gestión de la propagación del virus Covid-19.

La técnica del Space Syntax

Para el caso de este proyecto, se utilizó la Sintaxis Espacial (Space Syntax en inglés) como herramienta para incorporar el análisis del espacio entre las variables de con las que trabaja el algoritmo. Según la plataforma pedagógica de Space Syntax creada por University College of London, la sintaxis espacial se define como

> "un set de técnicas para analizar distribuciones espaciales y patrones de actividad humana en edificios y áreas urbanas. También es un conjunto de teorías que vinculan el espacio y la sociedad. La sintaxis espacial aborda dónde están las personas, cómo se mueven, cómo se adaptan, cómo evolucionan y cómo hablan sobre ello". [13]

De esta manera, bajo principios similares de la sintaxis lingüística, en la que ningún elemento (palabra o espacio) tiene significado pleno por sí solo, sino que se define por su posición y relación con los demás, aborda una lógica relacional para abordar los patrones de disposición espacial de los elementos y no se limita a atributos individuales de los mismos. [14]

Su construcción y desarrollo se soporta en dos enfoques surgidos desde la antropología estructural y los estudios del ambiente construido, siendo este último el que funciona como

base conceptual, técnica y tecnológica del Space Syntax como se conoce y práctica actualmente y, por ende, del presente proyecto. Este enfoque nace a partir del trabajo de Bill Hiller y colegas en University College London, dando la estructura teórica y técnicas analíticas del Space Syntax a través de su trabajo "The Social Logic of Space" [14], del cual, precisamente, el centro de investigación de la Facultad Bartlett del Entorno Construido de dicha universidad ha sido precursor en su desarrollo.

La sintaxis espacial aplica un conjunto de conceptos y herramientas analíticas para cuantificar y mediar la configuración espacial y sus implicaciones sociales, de manera que hace uso de representaciones gráficas para comprender el espacio y estimarlo de manera cuantificable [15]. Dentro de las principales representaciones gráficas se encuentran [13]:

Mapas axiales: se construyen trazando las líneas rectas más largas y mínimas necesarias para cubrir todos los espacios abiertos y transitables de un entorno construido. Estas líneas representan los caminos principales por los que las personas pueden desplazarse, y sus intersecciones marcan posibles puntos de encuentro. Cuando varias líneas axiales convergen en un mismo lugar, suele indicar zonas importantes de interacción social, como plazas, corredores principales o cruces concurridos. Estos mapas ayudan a entender cómo la forma del espacio facilita o dificulta el movimiento y el encuentro entre personas.

Ilustración 1. Mapa Axial *[13]*

Mapas de segmentos: representa el entorno construido dividiendo los caminos en pequeños tramos rectos entre intersecciones. A diferencia del mapa axial, que usa las líneas más largas posibles, en el mapa de segmentos cada vez que dos caminos se cruzan, la línea se corta, creando segmentos más cortos y detallados. Esto permite analizar con más precisión cómo las personas realmente se mueven, considerando giros, desvíos y cambios de dirección. Los mapas de segmentos son muy útiles para estudiar el flujo de movimiento en calles, corredores o rutas peatonales de una forma más realista y granular.

Ilustración 2. Mapa de Segmentos *[13]*

Grafos de visibilidad: se basan en el concepto de *isovista*, que es el área visible desde un punto específico en un espacio. A partir de varios puntos generadores, se crea una red de conexiones visuales que muestra qué tan visible es cada parte del entorno desde cada uno de esos puntos. Estos grafos permiten analizar patrones de integración visual, control visual y hasta aspectos como la privacidad o la vigilancia natural en edificios y ciudades.

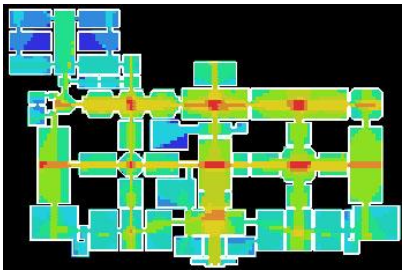

Ilustración 3. Grafo de Visibilidad *[13]*

Grafo justificado: es un diagrama que representa cómo se conectan los espacios en un edificio o ciudad en términos de facilidad de acceso (permeabilidad). Parte de un punto inicial, como una entrada, y muestra cuántos pasos o transiciones se necesitan para llegar a otros espacios. Cuanto más "profundo" esté un espacio en el grafo, más aislado o segregado está del punto de inicio. Este tipo de análisis ayuda a entender la jerarquía de accesos y qué tan accesibles o privados son diferentes espacios.

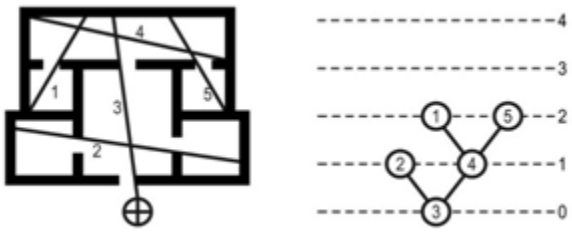

Ilustración 4. Grafo justificado (derecha) elaborado a partir de un mapa axial (izquierda) *[16]*

Estos gráficos son la base para desarrollar análisis matemáticos y computacionales que asignan al espacio analizado métricas para estudiar la configuración espacial y relación física de cada elemento que compone lo compone. Entre las métricas implementadas para este proyecto están: Integración global, Integración local, Profundidad, Valor de control, Conectividad, Elección (Choice) y Asimetría Relativa. Si bien en el apartado 2, objetivo específico 1, se realiza una explicación de cada métrica, es importante resaltar algunos de los usos relevantes:

Para la planificación y diseño urbano, su aplicación se centra en entender cómo la disposición de calles, plazas y corredores influye en el flujo peatonal, el uso del suelo y la vitalidad de los espacios públicos. Al analizar índices de integración y conectividad en la red vial, se pueden diseñar entornos más accesibles, navegables y socialmente dinámicos. De este modo se optimiza la implantación de infraestructuras y se fomenta la cohesión comunitaria [17].

En los Estudios Rurales y Análisis Territorial, Space Syntax examina la conectividad y accesibilidad entre núcleos de población, recursos naturales y vías de comunicación. Mediante grafos configuracionales se detectan corredores rurales, puntos de convergencia y áreas relativamente aisladas que afectan la movilidad y el desarrollo local. Este tipo de análisis contribuye a la planificación sostenible, orientando la gestión de infraestructuras y servicios en territorios dispersos [18].

Por otro lado, en los estudios sociales y antropología se investiga cómo la configuración del espacio habilita o restringe interacciones sociales y refleja jerarquías culturales. A través de métricas como integración, control y profundidad, se cuantifica la probabilidad de encuentros y se interpretan diferencias en la coexistencia de grupos en contextos contemporáneos e históricos [14].

Por su parte, la investigación cognitiva conecta las medidas de sintaxis espacial con la percepción, la formación de mapas mentales y la navegación interna [19].

Finalmente, la principal herramienta para su aplicación es el software Depthmap, desarrollado por la UCL, predominando sobre otras como AGRAPH, Ajanachara, Syntax 2D, Mindwalk o AJAX Light. Además, también se puede aplicar a través de librerías de Python como sDNA+py o NetworkX [15].

Las técnicas de Aprendizaje Automático

Como afirman Boelaert y Ollion [20] las distintas técnicas de aprendizaje automático o machine learning han permeado los procesos de investigación en ciencias sociales, ofreciendo algoritmos y modelos que han desencadenado la transformación de distintas disciplinas tanto en sus procedimientos como alcances. En este sentido, uno de los factores que ha impulsado dicha transformación de las ciencias sociales responde a la creciente disponibilidad y volumen

de datos, lo cual ha conllevado a nuevas concepciones y tratamientos de la generación de conocimiento a través de la evidencia desde distintas escalas [21].

Si bien los enfoques cuantitativos de las ciencias sociales se han enfocado en la construcción y evaluación de modelos estadísticos, las técnicas de machine learning también han facilitado que se promueva la escalabilidad de los análisis en tanto permiten abordajes que potencian el poder predictivo de los modelos al fomentar su aplicación en contextos que trascienden los datos elegidos para estudios en contextos específicos y, a través de la simplicidad, la explicación de problemas vastos y complejos [22]. Como afirma Hindman, esto representa un salto fundamental en la manera de entender y ejercer las ciencias sociales, ya que irrumpe con tendencias de décadas en limitar los análisis cuantitativos al uso de regresiones lineales con resultados que implican retos considerables en su interpretación y que tienden a ser sobre ajustados, por lo que propone la utilización de modelos de aprendizaje automático para la aplicación de técnicas que permitan la generación de perspectivas y conocimiento con bases de datos medianas y la combinación de distintos algoritmos y técnicas.

No obstante, para que la utilización de las técnicas de machine learning sea funcional a las disciplinas de ciencias sociales se requieren abordar discusiones como la demostración de resultados óptimos sin comprometer la flexibilidad que le da poder predictivo a los modelos [20]. En este sentido, Grimmer *et al* [21] proponen abordajes concretos desde la ciencia de datos que deben contemplarse en el marco de los procesos intrínsecos a estas disciplinas, de manera que para los procesos de descubrimiento se propone de manera general la utilización de técnicas como clusterización y embeddings, mientras que para ejercicios de medición se proponen las regresiones, algoritmos de aprendizaje supervisado y modelos preentrenados. Por otra parte, en cuanto a los ejercicios de inferencia, se identifican dos ramas: la causal, que se basa en la identificación de efectos heterogéneos y análisis multidimensionales y la predictiva, que se soporta en la predicción social y el forecasting o pronosticación.

Lo anterior, implica también una discusión epistemológica de las ciencias sociales, ya que la aplicación de estos métodos significa una revisión del modelo deductivo de las ciencias sociales, de manera que

> En vez de suponer que nuestras teorías están completamente desarrolladas antes de observar los datos, enfatizamos en que el aprendizaje automático ofrece herramientas que ayudan a generar preguntas de investigación, conceptos e hipótesis que pueden ser posteriormente testeadas de manera rigurosa con nuevos datos (…) enfocándose en seleccionar el método que optimiza el desempeño para cada tarea investigativa en vez de la búsqueda de un modelo real de los datos. [21, p. 403](traducción propia).

Esto, según los autores, será más efectivo en tanto haya teorías establecidas que tengan implicaciones medibles, permitiendo desarrollar un enfoque que permite la redefinición de conceptos y el desarrollo de nuevas teorías e hipótesis.

Específicamente, Hindman [22] identifica los algoritmos y técnicas de machine learning más funcionales a los objetivos investigativos de las ciencias sociales, donde se destacan los árboles de clasificación y regresión (CART), la reducción de dimensiones y descomposición de valores singulares, métodos de vecinos más próximos (Nearest neighbor) y máquinas de vectores de soporte. En esencia, afirma el autor, estos algoritmos son diametralmente distintos a los modelos de regresión que predominan en las ciencias sociales, dan poder predictivo frente a los propósitos propios de la disciplina y son la base de modelos de aprendizaje compuestos para contrarrestar incertidumbres generadas por los parámetros de las regresiones y en la generación de datos.

De esta forma, se promueve la predicción y replicabilidad en los métodos cuantitativos de las ciencias sociales, a partir de modelos robustos alejados del sobreajuste que tradicionalmente han promovido estas disciplinas en las técnicas estadísticas predominantes hasta hace unos años [22], de manera que se fomenta la posibilidad de generar hipótesis y conclusiones funcionales a otros estudios e, incluso, otras áreas de conocimiento.

Para los fines de este trabajo se exploran particularmente los algoritmos de aprendizaje automático no supervisado que, como se muestra en la sección de antecedentes, han tenido una amplia implementación en el campo del análisis urbano. Especialmente, se presentan los algoritmos de DBSCAN, clasificación Jerárquica y K-Means, que son los implementados para el desarrollo de los objetivos del proyecto, en tanto muestran mayor robustes y coherencia con el objetivo de este ejercicio, analizar patrones en la configuración de locales comerciales desde la agrupación de las manzanas de la calle quinta en Cali.

El algoritmo de agrupación DBSCAN

El algoritmo de Agrupamiento Espacial Basado en Densidad para Aplicaciones con Ruido (DBSCAN por sus siglas en inglés) es un algoritmo de agrupamiento basado en la detección de áreas de mayor densidad en el conjunto de datos, que están separadas por áreas de menor densidad. La agrupación la realiza basado en dos parámetros fundamentales, el máximo de la distancia entre dos puntos para ser considerados vecinos (épsilon) y el número mínimo de puntos cercanos a un punto central de clúster para ser considerado como tal [23, p. 23].

Este algoritmo se ejecuta en tres etapas fundamentales:

1. Encontrar los puntos en la vecindad. Para cada observación se considera el número de puntos en una vecindad máxima de *épsilon.*
2. Encontrar los puntos centrales. Considerando el número mínimo de puntos cercanos en el radio *épsilon* se asignan los puntos centrales y a estos sus respectivos grupos.
3. Asignar los puntos que no están en una vecindad de los puntos centrales como ruido.

A diferencia de otros algoritmos de agrupación no se requiere especificar el número de clústeres y la forma de estos es arbitraria, además, está diseñado para manejar el ruido y los

valores atípicos. Finalmente, la calidad de los clústeres es determinada por la distancia euclidiana, es decir, el cuadrado de las distancias entre los puntos de un clúster y su centro.

El algoritmo de agrupación Jerárquico

Los algoritmos de clustering aglomerativo de tipo jerárquico funcionan contemplando cada uno de los datos en su propio clúster y fusionando en varias iteraciones los clústeres más cercanos hasta que todos los puntos de datos pertenezcan a uno sólo [24, p. 277]. Existen diferentes tipos de algoritmos jerárquicos, que se diferencian por el criterio empleado para determinar qué clústeres fusionar, entre ellos se destacan 4 [25]:

1. Algoritmo del enlace simple: Se basa en la distancia mínima entre los puntos de dos clústeres.
2. Algoritmo del enlace completo: Utiliza la distancia máxima entre puntos de dos clústeres.
3. Algoritmo del promedio entre grupos: Apunta a un balance entre los métodos anteriores, basándose en la distancia media entre todos los pares de puntos de dos clústeres.
4. Algoritmo empleando el método de Ward: Es uno de los más implementados porque fusiona los grupos determinando cuáles ocasionarían el menor incremento en la suma de los cuadrados de las distancias internas del grupo, reduciendo la varianza total de los clústeres y produciendo grupos bien balanceados [24].

El método Ward puede ser implementado usando el algoritmo de Lance-Williams, que se usa para actualizar las distancias entre los puntos $d_{ij,}d_{ik}$ y $d_{jk,}$ cada vez que se hace una nueva unión entre los clústeres C_i, C_j y $C_k,$ reduciendo cada vez la varianza. Este algoritmo simplifica la búsqueda del par óptimo. La distancia $d_{(ij)k}$ del nuevo clúster puede ser computada por la fórmula:

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|$$

Donde $\alpha_i, \alpha_j, \beta, \gamma$ son parámetros que dependen del tamaño del clúster, junto con la función de distancia $d_{(ij)k}$. El método de varianza mínima de Ward puede ser implementado entonces por la fórmula de Lance–Williams para clúster C_i, C_j y C_k y con tamaños n_i, n_j y n_k respectivamente [24, p. 282]:

$$d(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} d(C_i, C_k) + \frac{n_j + n_k}{n_i + n_j + n_k} d(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} d(C_i, C_j)$$

El algoritmo de agrupación K-Means

El objetivo del algoritmo K-Means es particionar un conjunto de datos en un número predeterminado de grupos (k clústeres) logrando que los datos al interior de cada grupo se

organicen alrededor de un 'centroide' que representa la media de los puntos en el clúster [8, p. 6]. Esencialmente, el algoritmo funciona en 4 etapas.

1. Inicialización: Elige un conjunto aleatorio de puntos como centroides iniciales. El número de centroides debe ser previamente determinado y pueden usarse estrategias para hacer la selección de centroides más alejados ayudando la convergencia del algoritmo.
2. Asignación: Los puntos en el dataset son asignados al clúster con el centroide más cercano, esto se calcula usando la distancia euclidiana. Aquí se crean los "k clústeres".
3. Actualización: Nuevamente se calculan los promedios para cada clúster y se reasignan los centroides según sea el caso.
4. Iteración: La asignación y actualización se repiten hasta que los centroides no cambien significativamente, entendiendo que se logra la convergencia, o hasta que se logra el número máximo de iteraciones que había sido asignado [8, p. 6].

Este algoritmo busca que los clústeres sean internamente lo más similares entre sí y, al tiempo, lo más diferentes con otros clústeres. Esta similitud se mide con la distancia euclidiana a los centroides de cada grupo.

> Los objetos se representan con vectores reales de *d* dimensiones (x1, x2, …, xn) y el algoritmo k-means construye k grupos donde se minimiza la suma de distancias de los objetos, dentro de cada grupo S= {S1, S2 ,…, Sk}, a su centroide. El problema se puede formular de la siguiente forma:
>
> $$\min_{\mathbf{S}} E(\boldsymbol{\mu_i}) = \min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x_j} \in S_i} \|\mathbf{x_j} - \boldsymbol{\mu_i}\|^2$$
>
> donde *S* es el conjunto de datos cuyos elementos son los objetos *xj* representados por vectores, donde cada uno de sus elementos representa una característica o atributo. Tendremos k grupos o clusters con su correspondiente centroide μi [26].

Una característica particular del algoritmo K-Means es que no determina por sí sólo la partición óptima del conjunto de datos, es necesario determinar el número de clústeres antes de correr el algoritmo. Para esto, el método más usado es graficar la 'varianza explicada' de diferentes valores de k, y a partir de ello evaluar dónde está el "codo" o punto en que la varianza explicada deja de disminuir significativamente y se estabiliza [23, p. 22]. En pocas palabras, el algoritmo se ejecuta para un rango dado de k (por ejemplo, desde k=2 hasta k=11) para cada valor se calcula la inercia, o la suma de las distancias entre cada punto y su centroide, finalmente, se realiza un gráfico de líneas de tendencia para la inercia según su respectivo número de clústeres. La curva trazada en el gráfico disminuye a medida que los grupod aumentan, hasta cierto punto donde la inercia no disminuye significativamente, esto sería en el número óptimo para k.

1.3.2 Antecedentes

A continuación, se presentan trabajos donde se investigan y analizan transformaciones urbanas desde la implementación de algoritmos de *Machine Learning* así como otras técnicas derivadas de las ciencias de la computación. En principio se utilizaron los términos "transformación urbana", "Machine Learning", "análisis de datos" y "algoritmos no supervisados" como términos fundamentales para las sentencias de búsqueda. Esta estrategia se empleó en dos buscadores especializados: Google Scholar y IEEE Xplore, teniendo en cuenta la amplitud de posibilidades que brinda el primero y el foco en estudios de ingeniería del segundo buscador. La búsqueda se realizó en inglés y español, apuntando por mayor alcance en los resultados. Se escogieron entonces los textos más representativos sobre el uso de herramientas de la Ciencia de Datos a la hora de medir el fenómeno, predecir su alcance y analizar el impacto que tiene la transformación urbana en la estructura socioeconómica, espacial y funcional en sectores urbanos.

Uno de los fenómenos de transformación urbana estudiados en los artículos encontrados es la gentrificación, lo que no es casual, ya que este fenómeno ha cobrado relevancia en los últimos años gracias a las posibilidades que brinda un mundo más interconectado. Reades, De Souza y Hubbard [27] buscan llevar el estudio de este fenómeno más allá, utilizando patrones y procesos de cambio en los barrios de Londres para identificar áreas que posiblemente se vean gentrificadas en el futuro. Sobre este mismo proceso de transformación urbana existen otros trabajos que abordan ciudades como Ciudad de México ([28]; [29]), Madrid y Barcelona [30], Dublin [31], Seúl [32] y Quito [33], cada una de estas investigaciones aborda la problemática desde ángulos diferentes; algunos desarrollan índices propios ([34], [30]), hacen seguimiento a datos demográficos ([28], [33]) y del mercado inmobiliario ([30], [33]) y utilizan tecnologías como los Sistemas de Información Geográfica (SIG).

Teniendo en cuenta que una de las preocupaciones fundamentales a la hora de formular este proyecto, y en general cualquier proyecto dentro de la Ciencia de Datos, es la obtención de los datos. Entre los documentos revisados se encontraron métodos útiles y fuentes creativas de datos que pueden considerarse para el desarrollo de proyectos de esta índole, como el uso de datos extraídos de la red social Instagram referentes a características demográficas (como edad y género), frecuencia de palabras en las publicaciones usando World2Vec, datos Geoetiquetados y otras técnicas computacionales para analizar las imágenes y textos publicados [32]. Otro ejemplo interesante es la extracción de datos de la plataforma Airbnb, los investigadores utilizaron la plataforma InsideAirbnb para extraer 24 puntos georreferenciados de propiedades listadas desde enero del 2016 a febrero del 2020, se lograron aprovechar datos valiosos como coordenadas de las propiedades, fechas en que se dispusieron para ser arrendadas, tipos de propiedad, capacidad, precio, entre otras [31]. Adicionalmente, una fuente

de datos alternativa entre las investigaciones compiladas fue Google Street View, herramienta que compila imágenes desde 2007 y se actualizan cada 1 a 3 años, generando una línea del tiempo de las calles [35].

Luego de la identificación de fuentes de datos y su recolección, es importante definir los métodos estadísticos y computacionales que se implementarían en la investigación. Martí-Costa, Durán y Marulanda [33], por ejemplo, realizan un análisis del desplazamiento de población urbana en Quito, Ecuador, colectan sus datos de fuentes tradicionales, como censos y registros municipales, y procesaron los datos empleando técnicas estadísticas como el Análisis de Componentes Principales para identificar las variables más influyentes [33]. Por otro lado, Palafox y Ortiz-Monasterio [28] emplean técnicas computacionales más avanzadas en su esfuerzo por identificar los factores que impulsan dinámicas de transformación urbana en diferentes áreas de la Ciudad de México. Los investigadores emplean redes neuronales para la predicción de la gentrificación en la ciudad y utilizan el método de interpretabilidad LIME (Local Interpretable Model-agnostic Explanations) ofreciendo una mejor interpretación de los factores específicos que impulsan este fenómeno en vecindarios individuales [28].

Finalmente, De Nadai y Lepri [36] desarrollan un estudio orientado a estimar los precios de viviendas basándose en datos abiertos y sin la necesidad de transacciones históricas. Su aporte fundamental está en la propuesta metodológica que contiene, pues los investigadores desarrollan un modelo predictivo que integra las características del vecindario con las características de la propiedad para estimar los precios de las viviendas, esta es una variable significativa a la hora de abordar transformaciones urbanas, sobre todo en barrios residenciales. El modelo empleado en este caso involucra técnicas de regresión ponderada, redes neuronales recurrentes y análisis de imágenes para extraer características de las propiedades y sus entornos [36].

Uso de algoritmos de agrupación para el análisis urbano

Ahora bien, los métodos predictivos como las redes neuronales no son los únicos implementados en el análisis urbano, aunque este tipo de algoritmos son útiles para algunos fines específicos, para el caso abordado en este documento, es fundamental conocer los casos de aplicación de algoritmos de clustering o agrupamiento.

Un caso interesante el uso de algoritmos de agrupación para la detección de posibles paradas de autobús [8], en él se compararon cuatro algoritmos de agrupación - K-Means, DBSCAN, HDBSCAN y OPTICS- en su capacidad de agrupar códigos postales para el establecimiento de paradas de bus escolares. Aunque K-Means resultó ser el algoritmo con mejores métricas de evaluación, la cantidad de clústeres resultó poco favorable para la construcción de paradas de bus en Ciudad de México.

En un estudio centrado en Inglaterra y Gales también se usan algoritmos de agrupación, esta vez para el análisis espacial de sistemas urbanos, tomando como unidad de análisis los barrios [6]. En este estudio se buscó descubrir patrones y procesos urbanos a lo largo del tiempo en Inglaterra y Gales utilizando datos a lo largo de 25 años de transacciones inmobiliarias y el censo. Tras el procesamiento y análisis de los clústeres los investigadores caracterizaron los 5 grupos resultantes como: "Urbano rico, Urbano pobre, Suburbano de clase trabajadora, Periurbano rico y Rural" [6, p. 6]. Una conclusión relevante de este artículo es que los algoritmos no supervisados pueden ser una herramienta fuerte para analizar dinámicas de transformación urbana, pues tienen la capacidad de revelar procesos que no son evidentes de otra manera, ni se alinean con los lineamientos oficiales de planeación.

Finalmente, otro de los estudios donde se implementan algoritmos no supervisados tiene que ver con el establecimiento de una diferenciación socioeconómica del espacio urbano en Ciudad de México [37]. Para ello, utilizaron una clasificación geodemográfica basada en estilos de vida provenientes de geomarketing, además de información censal desagregada al nivel de manzanas. Los investigadores priorizaron 34 variables de las 170 disponibles inicialmente, para la selección se apoyaron del análisis de correlaciones, análisis de componentes principales (PCA) y un árbol de decisiones. Mediante el algoritmo K-Means se particionó el dataset en 6 clústeres, definidos por la literatura en estratificación, de forma similar al estudio en Inglaterra los clústeres en México se denominaron: Periferia urbano-rural marginal, Empleados de oficina en unidades habitacionales, Proletariado periférico, Élites urbanas, Zonas mezcladas y Clase media educada [37, p. 22]. Es de resaltar que fue valioso para los investigadores contar con información en el nivel de manzanas, pues les permitió capturar la complejidad de los estilos de vida y diferenciación socioespacial en la ciudad.

Los textos revisados dan luces importantes sobre los avances y vacíos de la investigación de la consolidación y transformación urbana, los métodos y herramientas para estudiarla y cómo se ha abordado el tema en diversas ciudades del mundo. Ahora bien, se reconoce que las preocupaciones de los autores están sobre la medición de las transformaciones urbanas, en su magnitud e incluso en predecir la posibilidad de predecir dichos fenómenos, además se emplean métodos de extracción de datos en fuentes oficiales y no oficiales y se procesa la información a través de métodos de analítica de datos como las redes neuronales, sin embargo, no se identifican proyectos que trabajen utilizando modelos de agrupación de Machine Learning para analizar patrones comerciales como parte de los procesos de transformación urbana. Adicionalmente, no se encontraron estudios de esta naturaleza para el caso de Cali. En ese sentido, este proyecto se ubica como una posibilidad de aportar al conocimiento y la gestión pública en el nivel local.

2. Revisión de literatura e identificación de variables (OE 1)

Este apartado pretende desarrollar el fundamento técnico y teórico para la selección de variables en el entrenamiento del modelo de aprendizaje automático. En ese sentido, se ha realizado una revisión de literatura sobre estudios de urbanismo y planificación urbana, así como, artículos que analizan la evolución de un territorio urbano y, en general, textos que pretenden conocer más sobre los fenómenos de transformación urbana.

Adicionalmente, en el desarrollo de las actividades del proyecto fue necesario además investigar sobre estrategias de análisis espacial que sirvieran para enriquecer el dataset con el que se entrenaría el algoritmo de clustering. En ese sentido, y con la guía del tutor del proyecto, se escogió el Space Syntax como la metodología con mejor respaldo para esta tarea. Por eso, este apartado también aborda el Space Syntax desarrollado en la universidad de Londres (UCL), sus conceptos principales y sustenta las métricas utilizadas en el algoritmo.

El análisis urbano y la cuestión de la ciudad:

Novick, en el 2004 [2] hace un recorrido histórico sobre la concepción del urbanismo y la ciudad, con un foco importante en Argentina. Una idea relevante es la ampliación de la concepción de ciudad, de la restricción de lo técnico o lo político, a la apertura de las diferentes dimensiones que comprenden la ciudad. La ciudad deja de ser sólo un diseño para volverse un proyecto social y espacial, que amerita un abordaje interdisciplinario para entender sus dinámicas. Ahora bien, esta perspectiva presenta una desventaja en la medida que concibe la ciudad como un catálogo de conocimientos, dejándola en cierta medida deshabitada desde su concepción. Como respuesta surge desde los estudios culturales una perspectiva que vincula a las personas y sus representaciones a la forma como se configura y entiende la ciudad "se trata al mismo tiempo de dos hilos estrechamente anudados: actores sociales que construyen formas de ver pero también herramientas cognitivas que les permiten hacer o cambiar las formas de hacer" (Topalov en Novick, [2, p. 16]). La autora muestra entonces una primera visión panorámica del entendimiento de la ciudad desde los estudios urbanos, que no se limita sólo a los aspectos técnicos, sino también a las historias y representaciones que tienen los ciudadanos de la ciudad.

Aunque esta concepción es ciertamente difícil de incluir en un conjunto de datos para entrenar un algoritmo, arroja luces sobre la selección de variables, en tanto sugiere que es importante contemplar la ciudad más allá de su composición física incluyendo también sus interacciones con los habitantes. Sería relevante, por ejemplo, pensar cómo las personas la recorren, que es un aspecto rastreable y cuantificable para ser incluido en un dataset que sea insumo para estudiar aspectos concretos de una ciudad.

De manera similar, Castells [1] habla en La Cuestión Urbana sobre el proceso de urbanización. Desde el materialismo histórico como enfoque teórico, el autor plantea que la formación de áreas metropolitanas es producto de la especialización sectorial e interdependencia funcional propias de las sociedades industriales capitalistas. Las áreas metropolitanas reemplazan entonces la figura del barrio tradicionalmente autónomo, donde confluían funciones diversas que le daban cierta independencia sobre el resto del territorio. La especialización del trabajo en los territorios resalta además las diferencias socioeconómicas desde las que empiezan a configurarse las ciudades modernas, diferencias más significativas marcadas por la distinción entre los territorios rurales y urbanos [1, p. 42]. El autor tiene un acercamiento más centrado en la creación de los espacios urbanos por parte de sus habitantes, sobre todo impulsados desde el campo de la política, cuestionándose incluso si hay una determinación inherente de lo social por lo urbano o si lo urbano es un reflejo de dinámicas sociales preexistentes. En últimas, Castel invita a comprender la ciudad no solo como un espacio físico, sino como un producto social, un escenario de conflictos y transformaciones. Esta cuestión resulta interesante al preguntarnos cómo la ciudad se configura de manera que favorezca los modos de producción del sistema político y brinda un soporte material a los ejercicios de dominación estatal.

Al igual que Novick, Castells apunta a una realidad mucho más compleja que supera las consideraciones técnicas de lo que es una ciudad y cómo esta se configura. En particular la perspectiva de Castells rescata la importancia de una mirada crítica sobre los lineamientos y presupuestos del gobierno sobre la ciudad, pues reconoce que estas ideas de ciudad están pensadas para la delimitación de espacios para la reproducción de las formas de gobierno. En ese sentido, las consideraciones de este autor resultan relevantes para el trabajo de reconocer las dinámicas urbanas en Cali, sobre todo al sugerir un acercamiento crítico a las determinaciones de zonificación de la ciudad y pensar cómo y por qué existen como lo hacen ahora, qué relación tienen con la composición real de los espacios analizados y cómo los lineamientos de planeación urbana han impactado la configuración del espacio.

Otros autores exploran los estudios urbanísticos desde ángulos diversos, aunque todos coinciden en la naturaleza compleja y las múltiples variables que juegan un rol en el desarrollo y composición de las ciudades. Por ejemplo, Smolka y Goytia [5] centran su análisis en los mercados de suelo urbano, y cómo se han transformado desde los trabajos originales sobre el tema. En especial los autores examinan cómo las nuevas formas de estudiar estos mercados brindan herramientas novedosas para comprender el acceso a la vivienda, la segregación residencial, la movilidad intraurbana, el crecimiento urbano desordenado y la informalidad. No es casual que haya surgido la necesidad de actualizar las herramientas y metodologías de análisis, pues la manera como se vienen configurando los costos del suelo no obedecen sólo a las normas de oferta y demanda, exigiendo la intervención de técnicas más sofisticadas de análisis espacial por medio de sistemas de información geográfica (SIG) y nuevas técnicas de

análisis estadístico [5, p. 9].

Un acercamiento adicional al análisis de la aparición de las ciudades se aproxima más a la sociología urbana y los estudios sociales del espacio, con una perspectiva similar a la de Casterls, el Banco de Desarrollo Latinoamericano (CAF) [3] plantea la accesibilidad urbana como concepto fundamental para entender los retos más críticos para el desarrollo de las ciudades, especialmente las latinoamericanas. La accesibilidad urbana comprende la regulación del uso del suelo, la oferta e infraestructura de transporte y el mercado de vivienda como sus determinantes esenciales [38, p. 21]. El término de accesibilidad urbana se refiere a la capacidad de las familias y las empresas (firmas) para acceder a las oportunidades que una ciudad brinda, en ese sentido, se comprende que para el CAF resulta fundamental conocer esta idea de accesibilidad, pues puede ayudar a orientar las políticas públicas de un territorio para mejorar las oportunidades de sus ciudadanos.

Partiendo de este trabajo tan relevante para el contexto latinoamericano, el presente trabajo debe determinar variables que apunten a los determinantes de la accesibilidad urbana, en el contexto de Cali podría pensarse en: 1) los datos sobre la infraestructura del transporte masivo de la ciudad (MIO), así como la información sobre la conectividad y accesibilidad de las vías; 2) la información sobre el mercado de vivienda o en este caso, sobre la localización y naturaleza de los establecimientos de comercio y; 3) el POT (Plan de Ordenamiento Territorial) como la herramienta fundamental para la regulación del suelo. Sin embargo, para el caso abordado en este proyecto, respecto al uso del suelo el análisis se centra en los equipamientos urbanos y su localización dispuesta por la gestión pública de la ciudad.

En lo que se refiere a la infraestructura de transporte, los datos sobre las estaciones, paradas y número de vagones del MIO serán integrados al conjunto de datos para el entrenamiento del algoritmo; en este mismo campo se pensó que el transporte privado de los ciudadanos también resulta relevante, por lo que algunas métricas de la conectividad de las vías y su accesibilidad –generadas a través de Space Syntax– también serán incorporadas a la base de datos. Por último, en tanto este trabajo se enfoca en los establecimientos comerciales, se contemplan los datos sobre la ubicación y tipo de establecimientos que hay en el sector, como reemplazo del mercado de vivienda cuya información es más difícil de acceder.

Los procesos de transformación y consolidación urbana:

Los fenómenos de transformación urbana, en la medida en que han sido estudiados en la actualidad, son procesos complejos que implican cambios significativos en aspectos de la ciudad como su estructura física, uso del suelo, y dinámica social y económica (algunos ejemplos son los procesos de renovación urbana, el crecimiento a grandes dosis y el constante crecimiento informal en las periferias urbanas). La transformación puede ser desencadenada

de manera espontánea o como resultado de intervenciones de los responsables de la planeación urbana ([39], [4]). Para el caso Latinoamericano, Romero [4] realiza un recorrido por el proceso de transformación urbana entre 1950 y el 2020 donde resalta particularmente la incidencia del modelo capitalista en la configuración de las ciudades. El autor coincide con Castells en tanto identifica una especialización de los territorios y habla de una acumulación del capital en ciertos sectores de la ciudad, la dinámica capitalista eventualmente devendrá en la desregularización y apertura económica que ha resultado en lo que Romero llama la desposesión del espacio que obliga a una reorganización de la ciudad. La desposesión de las ciudades, según el autor, resultó en una crisis de vivienda en Latinoamérica, a la que los Estados respondieron con políticas "vivendistas" que apuntan a la financiación de la compra de propiedades, pero no necesariamente a devolver los espacios a la ciudadanía. En ese sentido, el autor propone un rol estatal equilibrado, que no sólo se preocupe por la regulación de los mercados, sino que trabaje desde la organización popular para la producción social de la vivienda ([4, p. 99]).

Romero aporta a este trabajo una reflexión interesante sobre la forma en la que se dieron las transformaciones urbanas en el continente, así como una mirada crítica sobre el papel del Estado en este fenómeno, desde su gestación hasta el intento de controlarlo. Lo anterior significa un llamado para la interpretación de los resultados del proyecto, pues propone que la configuración de la ciudad no sólo compete a sus gobernantes, sus habitantes también tienen un papel en la configuración de los espacios y en la mitigación y atención de las problemáticas que pueda traer un proceso de transformación urbana descontrolado.

Ahora bien, Barrera [40] plantea un acercamiento mucho más particular al fenómeno de la transformación urbana, en su análisis lote a lote de este fenómeno en el que estudia el proceso en la ciudad de Bogotá, precisamente valiéndose de la técnica del análisis espacial. El autor, a través de su análisis de datos catastrales, encuentra un proceso de densificación 'lote a lote' que, aunque parezca pequeño al observar sus apariencias individuales, al acumularse en el tiempo y espacio representa un cambio significativo de la ciudad, incluso en los sectores periféricos. Barrera describe la demolición de estructuras viejas para construir edificios de varios pisos, que poco a poco van densificando la población en Bogotá. Reconoce además que no es un proceso que haya sido previamente estudiado por su carácter gradual, que lo hace difícil de detectar en el corto plazo [40, p. 19]. La investigación del colombiano apunta a la relevancia de confrontar los datos con las realidades de los habitantes al plantear que la información cuantitativa debe ser confrontada con las experiencias de los ciudadanos y sus perspectivas. Además, resulta interesante cómo los métodos de análisis espacial del autor descubren un fenómeno poco estudiado, debido a su gradualidad, y revelan en potencial de estas técnicas de estudio.

En ese sentido, al igual que Barrera, este trabajo se plantea explorar en un área de gran relevancia para la ciudad de Cali posibles estructuras y servicios urbanos, así como patrones comerciales que tengan injerencia en la configuración del corredor de la Calle 5, recurriendo a datos por manzanas que permitan a los algoritmos de agrupamiento una mirada detallada de la composición del territorio. Como se ha señalado, estos datos comprenden información sobre los establecimientos comerciales, la composición de la infraestructura de transporte, presencia de equipamientos y servicios urbanos y la configuración vial del sector. Esta última información es recogida mediante el proceso de Space Syntax, cuyas métricas son explicadas a continuación.

Uso del Space Syntax para el enriquecimiento del dataset:

Para integrar el análisis espacial a los datos del algoritmo se consideraron las siguientes métricas, que se calcularon utilizando el programa DepthMapX como se explica más arriba. Las métricas calculadas son:

- Integración global: mide la accesibilidad tomando en cuenta todo el sistema.
- Integración local: mide la accesibilidad solo en el vecindario cercano de ese espacio [41].
- Profundidad: mide cuántos pasos o cambios de dirección se tienen que hacer para llegar de un lugar a otro. Cuanto menos profundo es un espacio (menos pasos), más integrado está. Por ejemplo, un pasillo que te lleva directo a muchas habitaciones es poco profundo; un cuarto al final de varios pasillos es muy profundo [18].
- Valor de control: mide cuánto poder tiene un espacio para dar acceso a otros espacios cercanos. Un espacio con un alto control value es como una puerta clave o un nodo de paso obligatorio: controla el flujo hacia varios otros lugares. Por ejemplo, un vestíbulo que conecta varias salas tendría un valor de control alto [14].
- Conectividad: representa cuántas conexiones directas tiene un espacio con sus vecinos inmediatos. Cuantos más caminos o puertas lleven a otros espacios desde un punto, mayor es su conectividad. Se utilizan las métricas de conexión Axial y Angular [14].
- Asimetría Relativa: compara la profundidad real de un espacio contra un modelo ideal donde todo estaría conectado directamente, como los rayos de una estrella [15].
- Choice: calcula cuántas veces un segmento o un nodo está incluido en los caminos más cortos entre todos los pares posibles de nodos en la red. Es decir, mide la importancia de un segmento o espacio como ruta intermedia en el movimiento entre diferentes puntos [18].

3.Compilación de base de datos (OE 2)

Una vez reconocidos los conceptos fundamentales para la identificación de fenómenos de transformación urbana, es claro que consolidar una base de datos robusta, con información sobre los predios en el área priorizada del proyecto será fundamental para alimentar el modelo de agrupación con información adecuada. Para eso se realizaron 6 pasos, desde la obtención de datos, preparación, limpieza y disposición:

1. **Obtención de bases de datos iniciales**

En principio, las bases de datos que alimentaron el ejercicio corresponden a los establecimientos comerciales de la ciudad, suministrado por la Cámara de Comercio y, por parte de la Infraestructura de Datos Espaciales de Santiago de Cali, desarrollada por el Departamento Administrativo de Planeación Distrital, se facilitaron shapes y bases de datos en torno a vías y sus jerarquías, estaciones y paradas de transporte público, equipamientos (de salud, educación, bienestar social, cultura, parques y otros), manzanas y espacios públicos. Adicionalmente, para el recorte de los límites espaciales del corredor de la calle 5 se utilizó la herramienta Google Earth, generando un archivo KMZ que abarca todo el corredor desde la carrera primera hasta la 50, comprendiendo al menos 5 cuadras desde la calle principal hacia occidente y oriente. El recorte se pensó como una forma de priorizar sectores relevantes para el corredor y limitar el número de manzanas para tener una cantidad manejable de datos, teniendo en cuenta los recursos de computación disponibles.

2. **Designación de comercios a cada manzana.**

Utilizando el archivo KMZ y el programa ArcGIS, se hizo una intersección espacial con los shapes de manzanas y establecimientos con el fin de designarle a cada manzana el tipo y cantidad de comercios relacionados a su polígono. En este punto la base de datos tenía 484 manzanas, que es la unidad de análisis y aproximadamente 300 variables.

3. **Generación de métricas de Space Syntax.**

A partir del shape de vías y equipamientos, se hizo un procesamiento que derivara en las dinámicas de acceso de los equipamientos con relación al polígono priorizado. Sin embargo, es importante aclarar que estas métricas, dado que tienen un sentido sistémico, no se realizaron de manera reducida al área de interés, en cambio, se generó un buffer de 1000 metros sobre dicho polígono y sobre el área resultante se establecieron las métricas. Nuevamente el tamaño del buffer se definió teniendo en mente el balance de dos aspectos; un espacio que tuviera una influencia significativa en la movilidad del sector y una cantidad de datos manejable para los recursos de computación. En un escenario ideal, contando con mayor poder computacional, el ejercicio podría ser más completo si se aplica sobre la totalidad de la ciudad, metodología que se exploró sin resultados al no contar con dichos recursos. Con este shape de líneas, en DepthmapX se crearon los mapas axiales y de segmentos para proceder al cálculo de métricas

para cada vía y sus distintos tramos, entendiendo que cada vía tiene distintos puntos con distintas métricas en tanto las cantidades de intersecciones o tamaño de manzanas en cada tramo.

4. **Asignación de métricas a las manzanas.**

Con estas métricas de las vías, se creó un shape de puntos con estos valores que siguen el trazado de cada vía, esto con el fin de asignarle dicha métrica de accesibilidad a cada manzana en función de la cercanía a la misma con el punto medido de la vía en cuestión. Para esto, se utilizó la técnica de generación de polígonos Thiessen [42, p. 54], entendidos como una herramienta que divide un espacio en áreas donde cada punto dentro de una zona está más cercano a un punto generador específico que a cualquier otro. Esta técnica se utiliza en análisis espacial para modelar influencias o distancias en geografía y otras disciplinas.

5. **Definición de cantidades de equipamientos más próximos.**

Por otra parte, a través del uso de la librería Shapely, en Python se definió, para cada manzana, la cantidad de equipamientos por tipo dentro de un radio de 500 metros, considerando esta como la distancia caminable real, es decir, la distancia accesible a pie desde cada manzana hasta los equipamientos, en lugar de utilizar un buffer geométrico simple. De esta forma, cada manzana tuvo la asignación de métricas de accesibilidad e integración de Space Syntax (definidas anteriormente), cantidades de comercios (por tipo) contenidas en el polígono de la manzana y, además, cantidades de equipamientos, estaciones de transporte público y espacios públicos (por tipo) a una distancia inferior de 500 metros de la manzana.

6. **Priorización de variables.**

Dada la cantidad de tipos de equipamientos, estaciones, comercios y espacios públicos, la base de datos inicial resultó con 350 variables, un número que, para los análisis planteados, podría ser muy extensa, complejizaría el análisis y afectaría los resultados. Por dicha razón, de manera manual, se eliminaron las variables que resultaron en un valor de 0 en todas o casi todas las manzanas, ya que existen comercios que no están presentes (o lo están de manera moderada) en la zona de estudio, siendo un ejemplo claro de esto los cementerios. Este primer ejercicio permitió reducir el dataset a 75 variables.

Así mismo, se redujo la dimensionalidad al aplicar, en primera medida, un filtro de varianza con un umbral de 0.1 para eliminar las variables que presentan poca variabilidad en los datos, ya que estas no aportan información relevante para el análisis. Esto permitió reducir la cantidad de variables eliminando aquellas con varianza baja. Luego, sobre el conjunto de variables restantes, se calculó la matriz de correlación absoluta para identificar variables altamente correlacionadas (correlación > 0.8). Se eliminaron aquellas variables redundantes para evitar multicolinealidad y reducir la dimensionalidad del conjunto de datos sin perder información relevante. Finalmente, se aplicó un Análisis de Componentes Principales (PCA) para condensar

la información en un número reducido de componentes principales que retienen el 95% de la varianza total de los datos. Esto permitió una reducción adicional de la dimensionalidad, facilitando el análisis posterior y mejorando la eficiencia computacional.

En pocas palabras, este proceso combinó selección basada en varianza, eliminación por alta correlación y reducción dimensional mediante PCA. Usando este proceso en tres frentes el dataset se redujo a 64 variables, lo que permitió tener un conjunto de datos optimizado y representativo. Sin embargo, este número de variables todavía se considera muy alto para realizar los procesos de agrupamiento, además para generar un análisis de los resultados podría ser confuso abordar este número tan elevado de variables. Por eso, de manera complementaria, se entrenó un modelo de regresión basado en Random Forest utilizando como variable dependiente la suma total de las columnas del conjunto de datos que representaban los equipamientos urbanos. Esto permitió capturar la relación global entre las variables y su aporte al resultado agregado.

Así, se configuró un modelo RandomForestRegressor con 100 árboles y una semilla fija para garantizar reproducibilidad. Tras el entrenamiento, se extrajeron las importancias de cada variable predictora, ordenándolas de mayor a menor relevancia. Finalmente, se seleccionaron únicamente aquellas variables cuya contribución acumulada explicara hasta el 99.2% de la varianza total según el modelo, con el fin de conservar la mayoría de la información relevante y reducir una última vez la dimensionalidad del conjunto de datos.

Al finalizar la reducción de variables y limpieza de la base de datos, se obtuvo un dataset con 484 manzana y 38 variables, resumidas a continuación:

- Equipamientos de Salud
- Equipamientos de Educación
- Equipamientos de Administración Publica
- Equipamientos Parques
- Servicio de transporte público. Pretroncal
- Comercio: Expendio a la mesa de comidas preparadas
- Espacios de Culto
- Servicio de transporte público. Alimentadora
- Comercio: Comercio al por menor de productos farmacéuticos y medicinales
- Equipamientos Plazoletas
- Integración Total
- Conectividad Angular
- Servicio de transporte público: Dos (2) Vagones
- Integración
- Equipamientos: Administración de Justicia y Convivencia

- Equipamientos: zonas verdes
- Bienestar Social
- Comercio: Otros tipos de expendio de comidas preparadas
- Comercio: Actividades de la práctica médica sin internación
- Comercio: Expendio de comidas preparadas en cafeterías
- Elección
- Elección Normalizada
- Comercio: Actividades de hospitales y clínicas con internación
- Comercio: Actividades de apoyo terapéutico
- Comercio: Comercio al por menor de otros productos nuevos en establecimientos
- Comercio: Actividades de consultoría de gestión
- Conexión Axial
- Seguridad Ciudadana
- Otras actividades de atención de la salud humana
- Deporte
- Actividades inmobiliarias realizadas a cambio de una retribución
- Actividades inmobiliarias realizadas con bienes propios o arrendados
- Mantenimiento y reparación de vehículos automotores
- Otras actividades de servicio de apoyo a las empresas
- Comercio al por menor en establecimientos no especializados
- Actividades de apoyo diagnóstico
- Expendio de bebidas alcohólicas para el consumo dentro del establecimiento
- Comercio al por menor de prendas de vestir y sus accesorios

4.Aplicación de Modelos de Clustering (OE 3)

Algoritmo Jerárquico:

Una vez definidas las variables a utilizar para correr los algoritmos de agrupamiento, y con el objetivo de garantizar que las escalas de las variables no afectaran el funcionamiento de estos, se estandariaron los datos utilizando la función MinMaxScaler del paquete sklearn.

Para el primer algoritmo se implementó el de análisis de conglomerados jerárquico, utilizando el método de enlace de Ward, el cual minimiza la varianza intragrupo al momento de fusionar clústeres. Para ello, se empleó la función linkage del paquete scipy.cluster.hierarchy sobre la matriz de datos previamente normalizada. Posteriormente, se construyó un dendrograma que permitió observar la estructura jerárquica de los datos y establecer criterios preliminares para la selección del número óptimo de clústeres.

A partir de los resultados obtenidos en el dendrograma, se definió un punto de corte en una distancia de 7 unidades, criterio que permitió segmentar las observaciones en un número específico de clústeres mediante la función fcluster del módulo scipy.cluster.hierarchy. La clasificación fue incorporada como una nueva variable categórica (cluster) en la base de datos

original.

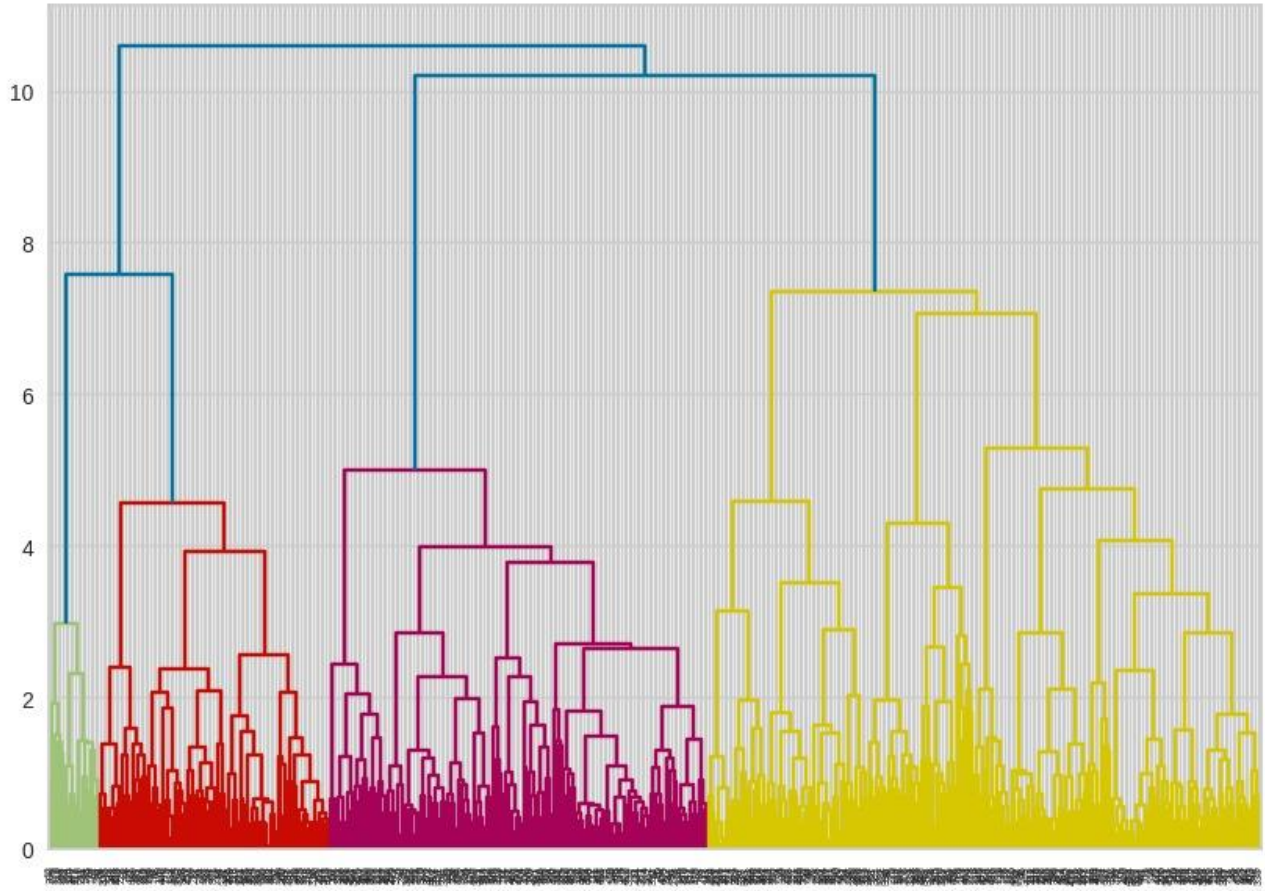

Gráfico 1. Dendrograma de la Clusterización Jerarquica

Para evaluar qué tan compactos y separados están los clústeres obtenidos por los diferentes algoritmos de agrupación, se calcularon los índices de Davies-Bouldin y Calinski-Harabasz. El primero, cuantifica la relación entre la dispersión intra-clúster y la distancia entre clústeres, de modo que valores más bajos indican una mejor estructura de agrupamiento. La evaluación para algoritmo jerárquico, que usó la configuración de 7 clusters y método de unión Ward, se realizó sobre la base de datos sin la columna de asignación de clúster, y utilizando como etiquetas de grupo la clasificación previamente determinada. El resultado obtenido fue un valor de 10.23 para el índice Davies-Bouldin, lo que sugiere una estructura de clústeres con un grado considerable de solapamiento o baja separación entre grupos, teniendo en cuenta que un resultado ideal debe ser cercano a 0.

Para el caso del índice de Calinski-Harabasz en los resultados del agrupamiento jerárquico, hay que tener en cuenta que este mide la proporción de la dispersión entre clústeres y la dispersión dentro de los clústeres, siendo deseable obtener valores más altos, ya que indican una mayor separación entre grupos y clústeres más compactos. El cálculo se realizó sobre la matriz original

de variables y utilizando las etiquetas generadas por el método jerárquico. El valor obtenido fue de 2.70, lo cual sugiere que la estructura de los clústeres formados presenta una baja separación relativa entre grupos en relación con la variabilidad interna de cada clúster.

El resultado de estos dos índices sugiere la necesidad de revisar el número de clústeres o explorar técnicas complementarias de agrupamiento. Por dicha razón, se hizo un proceso de optimización de hiper parámetros utilizando esta vez el índice de silueta, el cual permite medir la calidad del agrupamiento con base en la cohesión interna y la separación entre clústeres. Esta métrica varía entre -1 y 1, donde valores más cercanos a 1 indican una mejor estructura de agrupamiento.

Se exploraron distintas combinaciones de los parámetros método de unión (Ward, completo o promedio) y permitiendo que el algoritmo encuentre el número óptimo de clústeres para la convergencia. Los mejores 5 resultados se presentan en la siguiente tabla:

Configuración	Davies-Bouldin	Calinski-Harabasz	silhouette
complete, k=2	0.3914	6.0304	0.4798
average, k=2	0.3914	6.0304	0.4798
complete, k=3	0.9790	23.5989	0.2959
average, k=3	1.2810	6.3006	0.2953
average, k=4	1.1053	5.1247	0.2470

Tabla 1. Resultados optimización del algoritmo Jerárquico

Para cada combinación, se entrenó un modelo con AgglomerativeClustering y se evaluó su desempeño con silhouette_score. El proceso permitió identificar los mejores hiper parámetros asociados al mayor valor del índice de silueta observado. En este proceso, el resultado derivó en que el número de clústeres óptimo sería 2, por lo que se descartó la utilización de este algoritmo en la medida que clasificar 484 manzanas en solo dos categorías podría resultar en la pérdida de información frente al comportamiento de las variables en el sector analizado.

Algoritmo DBSCAN:

Con el objetivo de definir los mejores parámetros del algoritmo, se corrió un ciclo que ejecutara el algoritmo con valores de épsilon (distancia mínima entre vecinos) entre 0.1 y 1.1 y con el número mínimo de vecinos para configurar el clúster entre 2 y 20. La mejor combinación de hiper parámetros se evaluó mediante el Siluhouette Score y el resultado fue que el algoritmo con épsilon de 1.0 y número mínimo de vecinos de 4 tuvo un Score de 0.24.

Al correr el algoritmo con estos parámetros se obtuvo un solo clúster con 450 datos y un segundo grupo identificado como ruido con 34 datos. Este resultado da indicios de que el algoritmo DBSCAN no es el más apropiado para encontrar las agrupaciones adecuadas en este

dataset en tanto los datos no presentan múltiples regiones densas bien diferenciadas sino un conjunto grande de datos con ruido.

Para asegurar un resultado óptimo en la utilización del algoritmo, se ejecuta un nuevo ciclo en el que se evalúan y registran Silhouette Score y los índices de Davies Bouldin y Calinski Harabasz.

Épsilon	Min samples	# clústeres	Silhouette score	Davies Bouldin	Calinski Harabasz
1.0	2	3	0.114382	2.804616	6.143126
0.9	8	2	0.083807	3.472025	12.329741
0.8	8	2	0.081369	3.460892	14.570123
0.9	6	2	0.080726	3.533366	10.788797
0.8	6	2	0.078524	3.577828	13.153491

Tabla 2. Resultados optimización del algoritmo DBSCAN

En esta nueva evaluación el mejor resultado es la combinación de un Épsilon de 1.0 y Min samples de 2. Esta combinación arroja una agrupación de 3 clústeres, pero al estudiar el detalle de los resultados la distribución de estos clústeres es de 450 manzanas en un grupo, 3 en el segundo y 2 en el tercero, con 29 manzanas identificadas como ruido. Este resultado no tiene un valor analítico para el objetivo de este proyecto, pues sólo de identifica un gran grupo de alta densidad y dos muy pequeños, lo que se en términos prácticos es lo mismo que el primer resultado. Cabe agregar que los resultados de los índices tampoco son favorables, por un lado, el de Davies Bouldin supera el puntaje aceptable de 1, lo que se corresponde con clústeres muy similares entre sí, mientras el Calindki Harabasz tiene resultados muy bajos, sobre todo si los comparamos con las agrupaciones realizadas luego con el algoritmo K-Means. Los resultados de este último índice nos muestran una dispersión en los resultados de los clústeres.

En lo que respecta al uso del algoritmo DBSCAN se encontró que los resultados no son favorables para su implementación pues el algoritmo sólo reconoce un gran clúster y ruido, además arrojó índices con muy malos resultados. Este algoritmo entonces no tiene valor para la interpretación de la agrupación y no es el más óptimo para analizar los fenómenos de transformación urbana y la incidencia que el comercio puede tener en las dinámicas en el corredor de la calle 5 en Cali.

Algoritmo K-Means:

Como alternativa a los resultados anteriores, se realizó una clusterización utilizando el algoritmo k-means. Con el fin de identificar un valor apropiado para el número de clústeres (K) en el modelo, se aplicó el método del codo que, como ya explicó en el marco teórico, consiste en analizar la evolución de la inercia (suma de los errores cuadráticos intra-clúster) al incrementar

el número de grupos.

Se evaluaron valores de K desde 2 hasta 20, y para cada uno se entrenó un modelo KMeans con 10 inicializaciones (n_init=10) y una semilla fija (random_state=42) para garantizar la reproducibilidad. Además, se registró la inercia asociada a cada modelo, entendida como una medida de qué tan compactos son los clústeres. Los resultados se visualizaron en una gráfica donde se representó la inercia en función del número de clústeres, con el fin de encontrar el "codo".

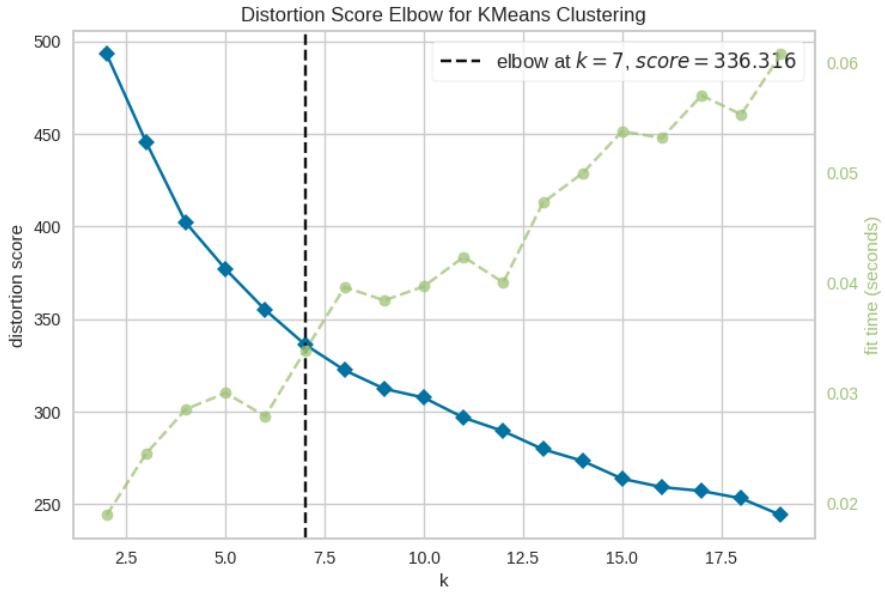

Gráfico 2. Resultados método del "Codo"

Adicionalmente, con el objetivo de complementar el método del codo y fortalecer la decisión sobre el número óptimo de clústeres para el algoritmo K-means, se calculó el índice de silueta para cada modelo generado con valores de K entre 2 y 20. Como ya se explicó, la silueta permite evaluar la calidad del agrupamiento en función del grado de separación entre clústeres y la cohesión dentro de cada uno, oscilando entre -1 y 1, donde valores más cercanos a 1 indican agrupamientos más definidos. Para cada valor de K, se entrenó un modelo con el algoritmo KMeans utilizando 10 inicializaciones (n_init=10) y una semilla fija (random_state=42). Luego se calcularon los respectivos puntajes de silueta usando silhouette_score.

K	Silhouette Score
8	0.165254
7	0.156479
4	0.155633

6	0.154786
9	0.152667
12	0.151145

Tabla 3. Resultados índice de silueta Algoritmo K-Means

De esta manera, el método del codo arrojó que el mejor k corresponde a 7 y el puntaje de silhouette k=8. Por dicha razón, se realizó el ejercicio de clusterización con estos dos valores y los siguientes parámetros:

- n_clusters=7 y 8: Se estableció la creación del número de grupos dentro del conjunto de datos.
- init="k-means++": Se utilizó una estrategia de inicialización avanzada para elegir de forma más eficaz los centroides iniciales, reduciendo la probabilidad de converger a soluciones subóptimas.
- n_init=10: El algoritmo se ejecutó 10 veces con diferentes inicializaciones para asegurar una mejor estabilidad del resultado final.
- max_iter=300: Se fijó el número máximo de iteraciones para alcanzar la convergencia en cada ejecución del algoritmo.
- random_state=42: Se usó una semilla aleatoria fija para garantizar la reproducibilidad de los resultados.

Como resultado, para k=7 se obtuvo un índice Davies-Bouldin de 1.856 y un Calinski-Harabasz de 53.803, mientras que para k=8 arrojó 1.723 y 50.938 respectivamente. De esta manera, se tuvo como resultado la siguiente distribución de clústeres, cuando k fue 7:

Clúster	Cantidad
3	126
5	92
6	91
4	53
0	46
2	43
1	33

Tabla 4. Resultado de asignación de manzanas por clúster del algoritmo k7

Mientras que con 8, resultó en la siguiente:

Clúster	Cantidad
4	129
0	91
7	82
1	52
6	44
5	41
3	31
2	14

Tabla 5. Resultado de asignación de manzanas por clúster del algoritmo k8

Sin embargo, ya definido que el k más eficiente sería igual a 8, y con el objetivo de identificar la combinación de parámetros que maximizara la calidad del agrupamiento en términos de cohesión interna y separación entre grupos sin perder eficiencia computacional, se realizó una búsqueda exhaustiva (grid search) de hiperparámetros relevantes, evaluando diferentes combinaciones sobre el conjunto de datos normalizado.

Concretamente, se implementó una exploración de hiperparámetros mediante el uso de ParameterGrid de sklearn, evaluando un total de 168 combinaciones posibles. Para cada configuración, se entrenó un modelo KMeans sobre los datos normalizados y se calcularon las etiquetas resultantes. Luego, se utilizó el índice de silueta como métrica principal para evaluar la calidad del agrupamiento, registrando el puntaje obtenido para cada combinación. El proceso permitió comparar sistemáticamente el desempeño de cada conjunto de hiperparámetros en términos de cohesión intra-clúster y separación entre clústeres. Finalmente, se seleccionó como configuración óptima aquella que obtuvo el valor más alto de silueta, lo que permitió mejorar ligeramente la definición de los grupos respecto a las configuraciones iniciales.

Los parámetros optimizados y sus rangos fueron:

- Número de clústeres (n_clusters): 8 (ya que resultó ser el número óptimo de clústeres en ejercicios anteriores)
- Método de inicialización de centroides (init): 'k-means++' y 'random'.
- Número de inicializaciones independientes (n_init): 10, 20, 30, 50, 75 y 100.
- Número máximo de iteraciones (max_iter): 200, 300, 500, 800, 1000, 1500, 2000.
- Algoritmo interno para el cálculo (algorithm): 'lloyd' y 'elkan'.

Los valores seleccionados para cada hiperparámetro se definieron con base en buenas prácticas comunes y en la literatura relacionada. Por ejemplo, n_init se varió ampliamente (10 a 100)

para evaluar la estabilidad de la solución frente a diferentes inicializaciones; max_iter se amplió progresivamente (200 a 2000) para observar si el algoritmo requería más iteraciones para converger; y se probaron los algoritmos 'lloyd' y 'elkan' para contrastar sus eficiencias dependiendo de la geometría de los datos. Aunque no se aplicó validación cruzada en un sentido más formal, el uso de una semilla fija y múltiples inicializaciones contribuyó a obtener un resultado reproducible y menos dependiente del azar.

Cada combinación fue evaluada usando el índice de silueta como métrica de calidad para determinar qué configuración ofrecía un mejor balance entre cohesión y separación de los grupos.

El mejor resultado obtenido correspondió a la configuración con:

- n_clusters = 8
- init = 'k-means++'
- n_init = 75
- max_iter = 200
- algorithm = 'lloyd'

Resultó en un puntaje de silueta de aproximadamente 0.166, indicando una mejora sutil en la definición de los grupos en comparación con configuraciones previas, donde el k=7 había dado 0.157 y el k=8, 0.165. Así mismo, en cuanto a los índices, el Davies-Bouldin resultó 1.72 y el Calinski-Harabasz dio 50.977, teniendo en cuenta que, más allá de que esta selección de hiperparámetros fue más eficiente en todas las métricas usadas, estos dos índices se utilizaron como criterios complementarios, mientras que el índice de silueta fue el criterio principal para seleccionar la mejor combinación de hiperparámetros durante la optimización.

En este sentido, esta configuración resultó ser la óptima al ofrecer clústeres bien separados y compactos, con menor solapamiento y una explicación de la varianza total eficiente, de manera que esta combinación resultó con mejores resultados en las tres métricas priorizadas en el proceso:

Configuración	Silhouette	Davies-Bouldin	Calinski-Harabasz
k = 7 (inicial)	0.157	1.856	53.803
k = 8 (inicial)	0.165	1.723	50.938

k = 8 (óptimo)	**0.166**	**1.72**	**50.977**

Tabla 6. Comparación de algoritmos k7, k8 y k8 optimizado

En resumen, el proceso de optimización permitió mejorar ligeramente la definición de los clústeres y seleccionar una configuración de hiperparámetros que, si bien no generó cambios drásticos en los indicadores, sí representó un modelo más robusto, reproducible y con un desempeño levemente superior frente a las configuraciones iniciales.

Resultados algoritmo K-Means (K=8):

En este sentido, la agrupación con el algoritmo optimizado derivó en 8 clústeres con la siguiente distribución de manzanas:

Clúster	Cantidad de Manzanas
0	49
1	81
2	42
3	128
4	14
5	31
6	48
7	91

Tabla 7. Resultado de asignación de manzanas por clúster del algoritmo k8 optimizado

En el siguiente mapa (Gráfico 2) se pueden apreciar las agrupaciones a nivel espacial. Es importante resaltar que, aunque el algoritmo no recibió valores sobre la ubicación espacial de las manzanas, el resultado si tiene una configuración espacial organizada, es decir, las manzanas de cada clúster tienden a tener una relación espacial compacta. Lo anterior con la excepción del clúster 6, aspecto que se estudia con detalle más adelante.

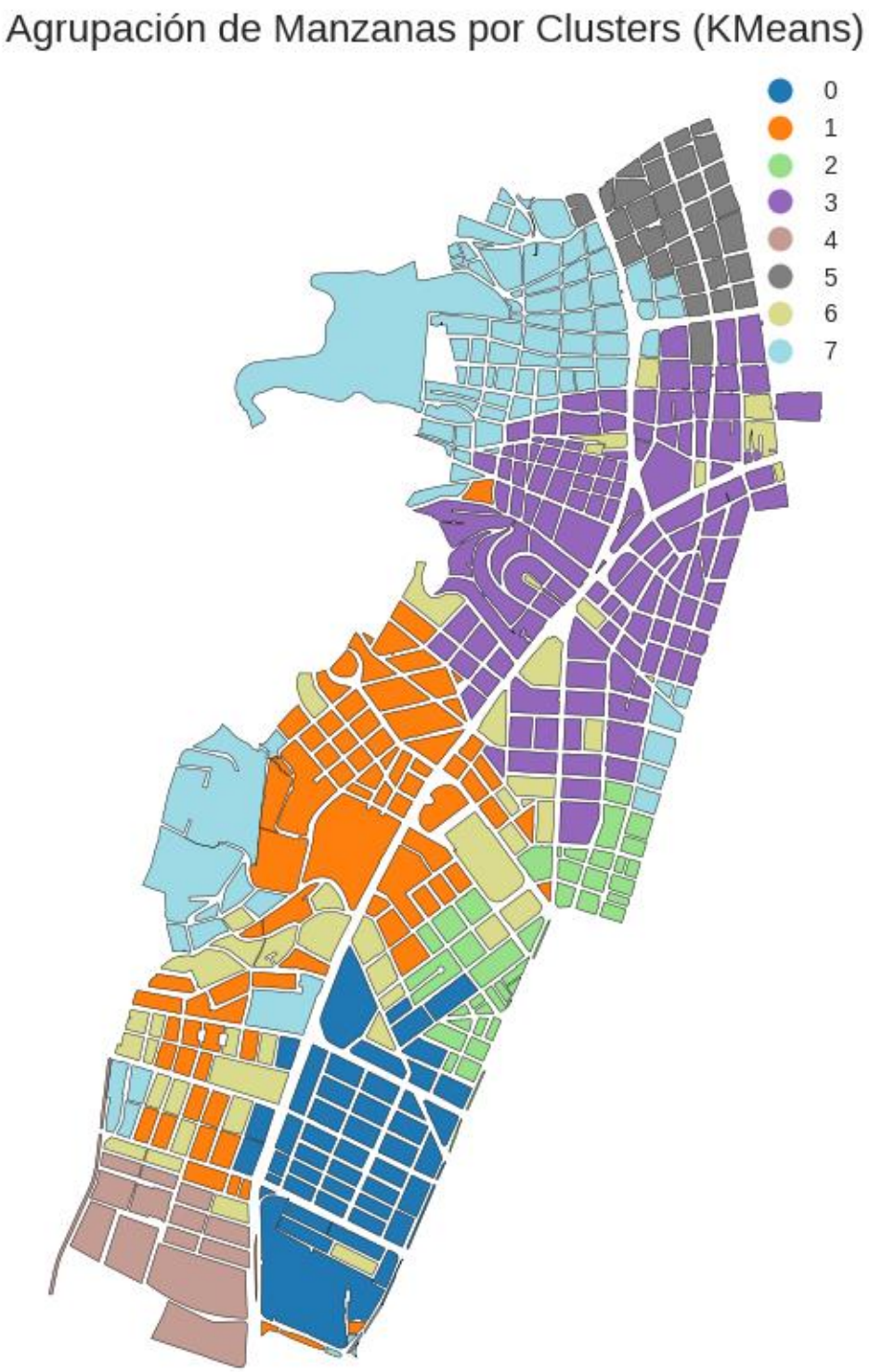

Gráfico 3. Resultados algoritmo de agrupación K-Means (K8 optimizado)

De esta manera, el comportamiento de los clústeres identificados corresponde a:

Clúster 0:
Este clúster corresponde a la zona de Tequendama y Nueva Tequendama, entre la Calle 5 y la Avenida Roosvelt y las carreras 44 y 50. En este sentido, se trata del clúster con la vocación más marcada, mostrando una predominancia de equipamientos, comercios y servicios de salud (Gráfico 2), así como uno de los que más zonas verdes tiene en promedio. Por otro lado, cuenta con buenos índices de conexión axial (Gráfico 3), entendida como la cantidad de caminos que se tocan directamente con un camino dado, lo que implica más opciones de movimiento inmediato desde un punto determinado, lo cual ha sido aprovechado desde la planificación urbana al ser una centralidad fundamental para la ciudad con al ser uno de los clusteres que más cuentan con la presencia y tránsito de rutas pretroncales (Gráfico 8).

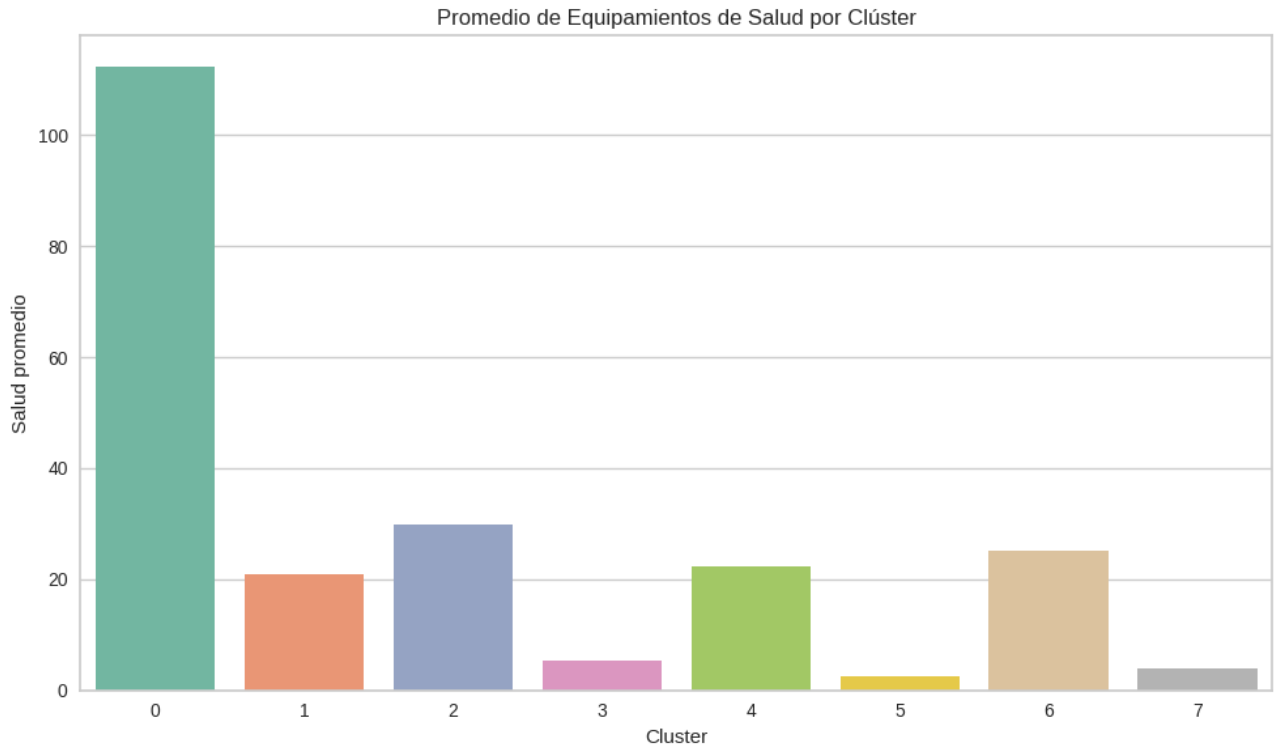

Gráfico 4. Promedio de Equipamientos de Salud por Clúster

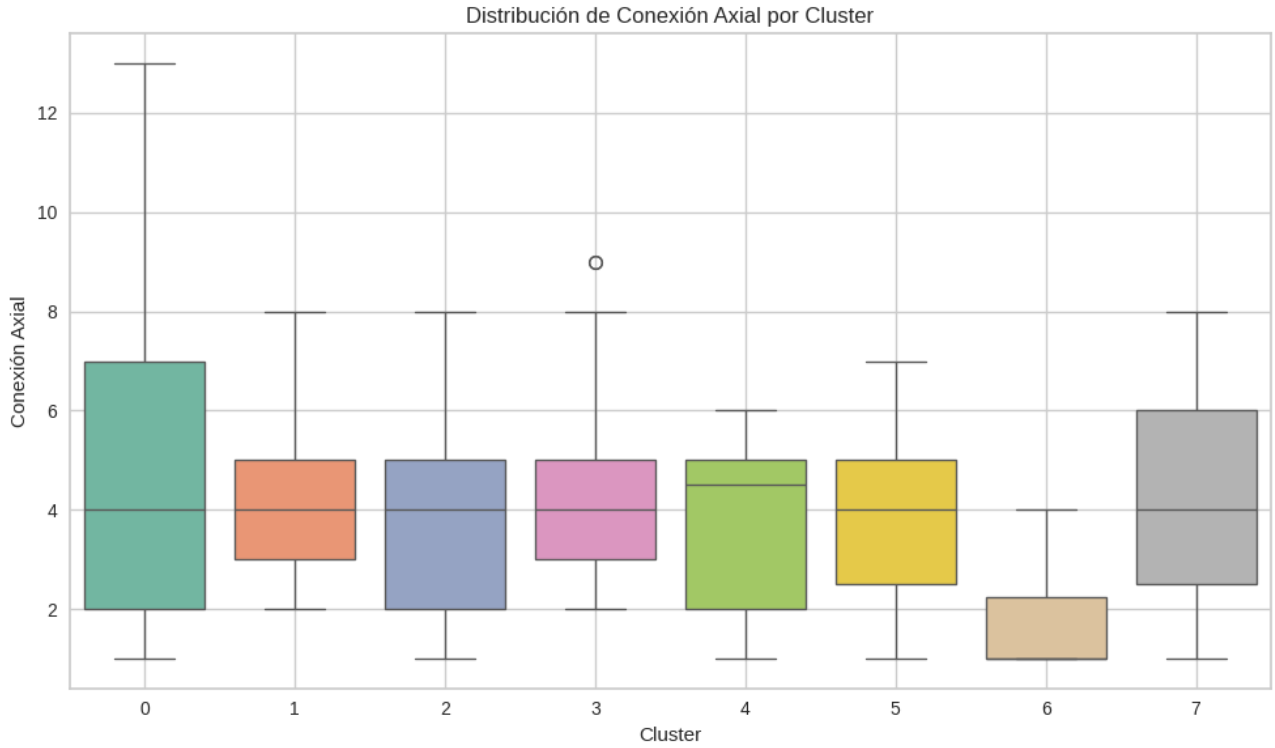

Gráfico 5. Conexión Axial por Clúster

En principio, como lo establece el Documento Técnico de Soporte de formulación de la Unidad de Planificación 10 [43], esta zona se configuró urbanísticamente en los 50 como periferia con asentamiento de trabajadores de clase media y alta, lo que explica la existencia de manzanas con frentes relativamente más largos que los de otras zonas residenciales de la ciudad y la oferta de espacio público más consolidada a lo largo de los años.

En el transcurso de las últimas décadas, aprovechando el cambio de las estructuras familiares a menor cantidad de personas y su relocalización a casas y apartamentos con tamaños más acordes, por lo que la consolidación del barrio consta de la densificación a través de construcción en altura y la adecuación de casas para servicios médicos, dinámica que fue impulsada por la atracción de centros médicos complementarios que generó la construcción del Centro Médico Imbanaco en 1976 [44]. Su cercanía a vías principales y zonas residenciales ha facilitado la configuración de una centralidad con vocación esencialmente de servicios y comercios ligados a la salud.

Clúster 1:
Este clúster corresponde al Barrio San Fernando Viejo y la mayoría de las manzanas de Santa Isabel, Nueva Granda y San Fernando Nuevo. Al tener cercanía relativa con la zona especializada en salud de la ciudad, alcanza a ser un clúster con buena accesibilidad a servicios de salud y,

gracias a su centralidad y cercanía con el centro de la ciudad y otros equipamientos educativos como la Universidad Libre o la Universidad del Valle, también presenta buen acceso a servicios educativos (Gráfico 5). Al igual que el clúster 0, cuenta con buena cantidad de zonas verdes (Gráfico 6), sin embargo, estas zonas verdes se han establecido primordialmente por el sistema de concesiones de espacio público realizadas por las construcciones en altura que se establecieron a partir de los años 90. Además, teniendo en cuenta que el desarrollo urbano del sector fue impulsado por el establecimiento de la Universidad del Valle y el Hospital Universitario, la configuración de la zona fue pensada en términos de manzanas extensas, lo cual facilita la generación de zonas verdes y espacio público.

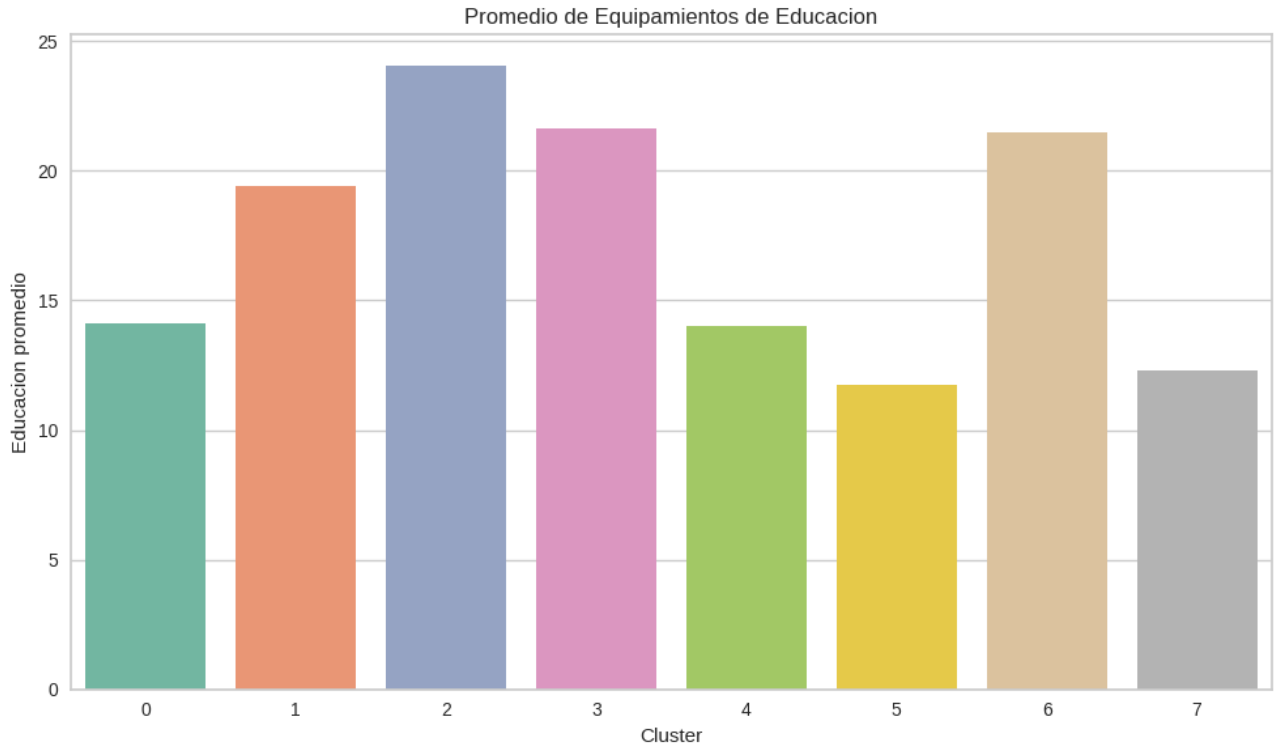

Gráfico 6. Equipamientos de Educación por Clúster

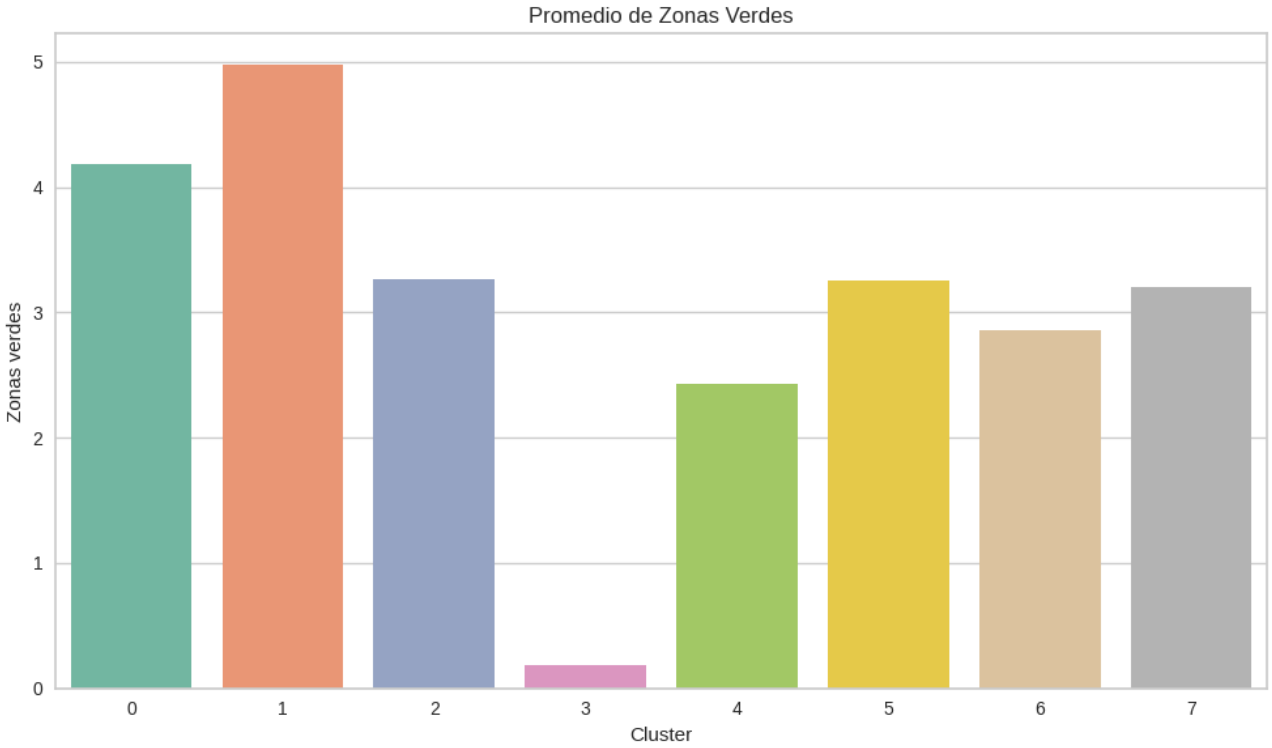

Gráfico 7. Zonas verdes por Clúster

Es relevante observar que el indicador de Elección (Gráfico 7), que indica cuantas veces, en este caso una manzana, se encuentra en la ruta más corta entre todas las posibles combinaciones de puntos en una red espacial, es alto en promedio pero disperso en sus valores, esto podría explicarse a la cercanía de muchas manzanas con vías de acceso importantes, como la Calle 5 o la Carrera 39 y su contraste con las manzanas en las zonas más altas, las cuales, por topografía, tienen una configuración vial más irregular y prevalecen las manzanas más extensas.

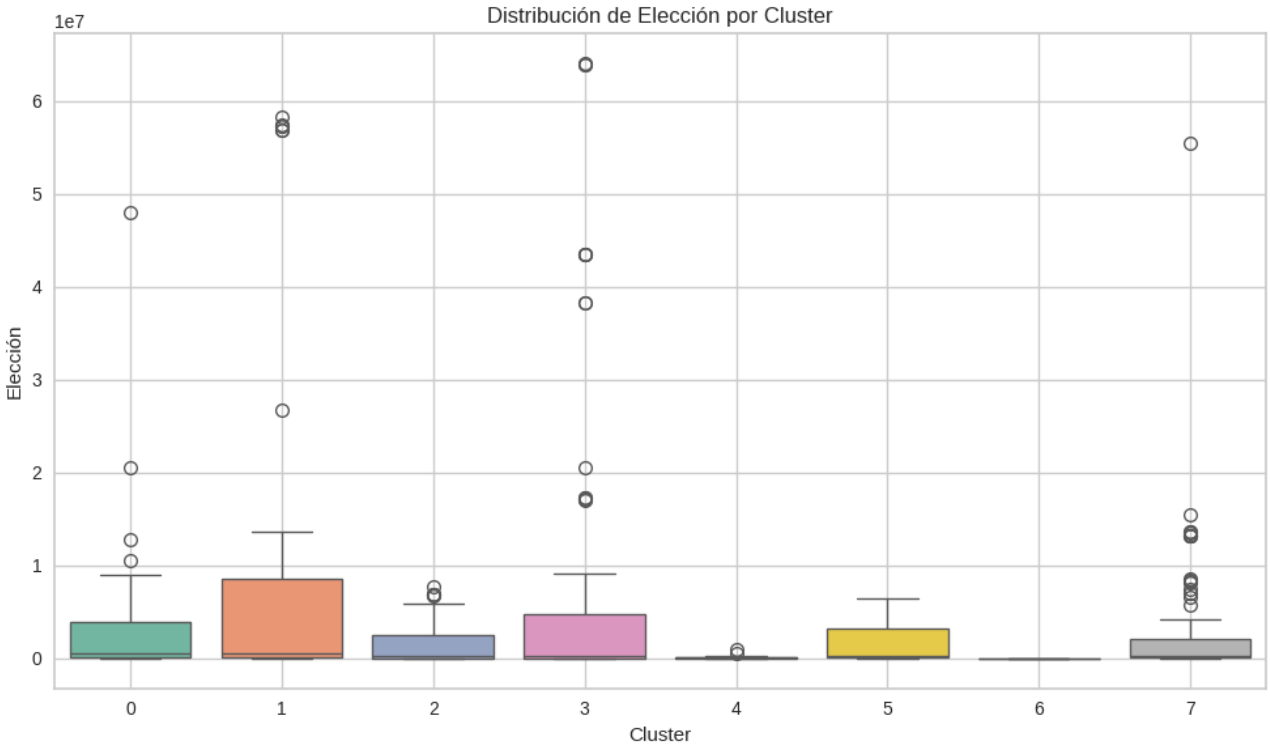

Gráfico 8. Distribución de Elección por Clúster

Clúster 2:

El clúster 2 también revela una vocación clara en oferta de servicios y comercio relacionado al deporte, complementándose con una fuerte presencia de oferta en salud, educación, bienestar social y espacios públicos, lo cual explica que esta centralidad sea la que más rutas pretroncales (Gráfico 8) tiene en su espectro.

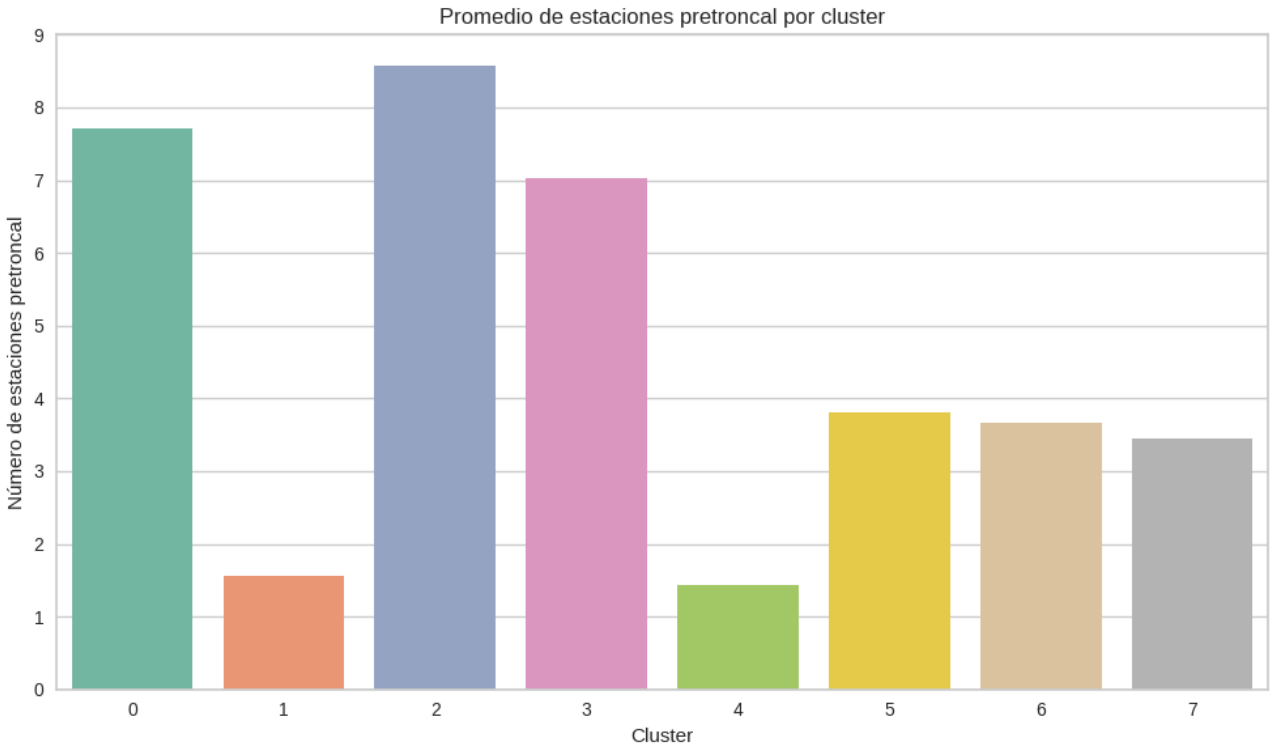

Gráfico 9. Estaciones pretroncales por Clúster

En este sentido, se trata de las manzanas cercanas a las Canchas Panamericanas (antes Hipódromo de San Fernando) y el Estado Pascual Guerrero, los cuales funcionaron como factor atrayente de comercio y servicios relacionados al deporte. Sin embargo, las dinámicas relacionadas con los eventos deportivos en el Estadio y sus implicaciones en la percepción de seguridad de la población que lo habitaba, su naturaleza residencial se fue debilitando, dando entrada a comercios que funcionaron en torno a la vocación deportiva en consolidación y a servicios de salud, que se centraron tanto en temas deportivos como en turismo estético. Así mismo, al haber sido una zona cuya vocación fue residencial de estratos medios y altos, el espacio público fue un elemento central en su configuración inicial y que aún se conserva.

Clúster 3:
Este clúster, por su parte, cuenta con servicios de educación más predominante (Gráfico 5), desde una perspectiva comparativa, pocas zonas verdes (Gráfico 6), muy pocos equipamientos de deporte (Gráfico 9) y salud (Gráfico 2) y una alta presencia de equipamientos de bienestar social (Gráfico 10) y seguridad (Gráfico 11). Espacialmente corresponde, en el costado occidental de la calle 5 con los barrios Miraflores y Libertadores y en el oriental con San Juan Bosco y Alameda.

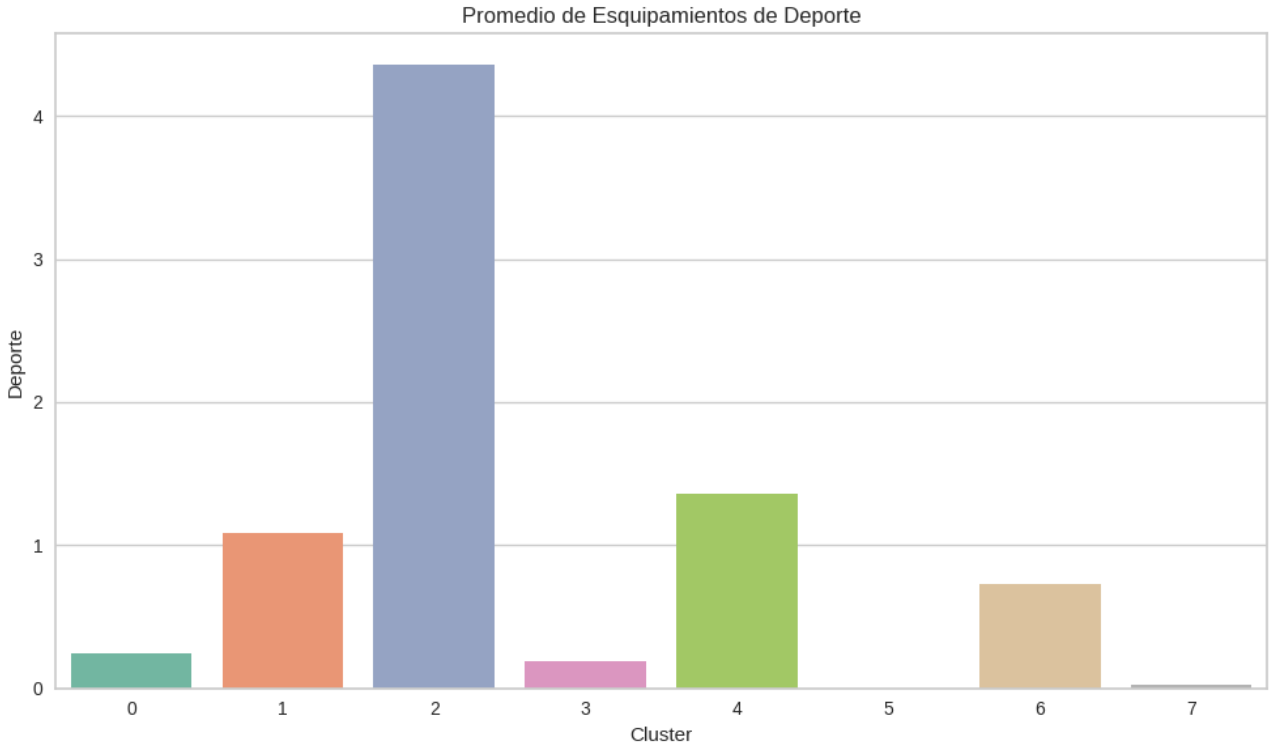

Gráfico 10. Equipamientos de Deporte

Su vocación educativa, de seguridad y bienestar social se explica por:

- Presencia de instituciones educativas con gran relevancia para la ciudad, como el Comfandi Miraflores, San Bosco, Santa Librada y la Gran Colombia, en el caso de la educación.
- En el caso de bienestar social (Gráfico 10), se da esta configuración debido a que, al ser de los primeros barrios de la ciudad, se configuraron con sus respectivos salones comunales como centros de encuentro social (destacándose los de Libertadores y Alameda), así como la presencia de hogares geriátricos (Posada del Abuelo y Fundación Hogar Señor de los Milagros) y centros de rehabilitación (por su cercanía con zonas donde se concentra población en situación de calle, como el Calvario).
- En el caso de seguridad, se encuentra la Estación de Bomberos, el Gaula

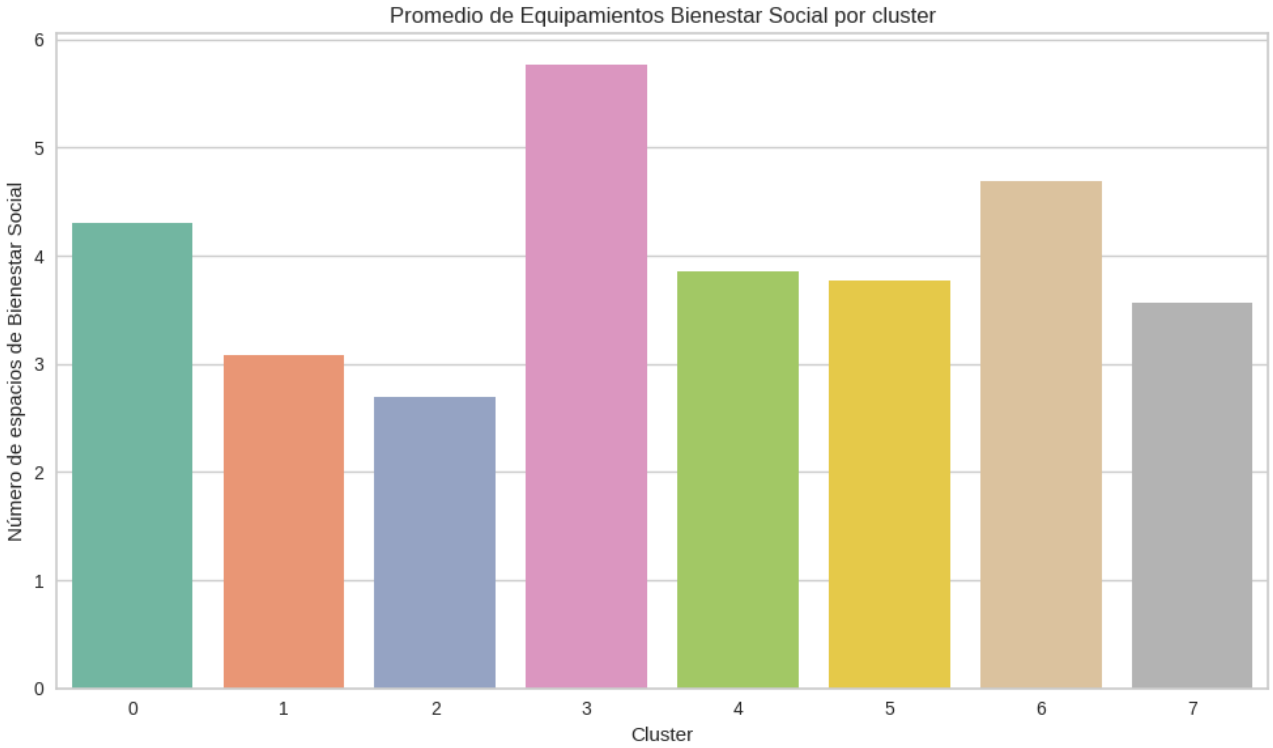

Gráfico 11. Equipamientos de Bienestar Social por Clúster

Por otro lado, su escasez de zonas verdes y espacio público se entiende en la medida que su proceso de consolidación fue de los primeros de la ciudad, momento histórico donde la norma urbana no hacía vinculante la generación de espacio público en su figura de cargas y beneficios para el desarrollo urbanístico. No obstante, gracias a su centralidad y al ser un paso obligado del centro administrativo con el sur de la ciudad, cuenta con una presencia equilibrada de rutas alimentadoras (Gráfico 12) y pretroncales (Gráfico 8) y la que más estaciones de MIO de dos vagones (Gráfico 13) presenta.

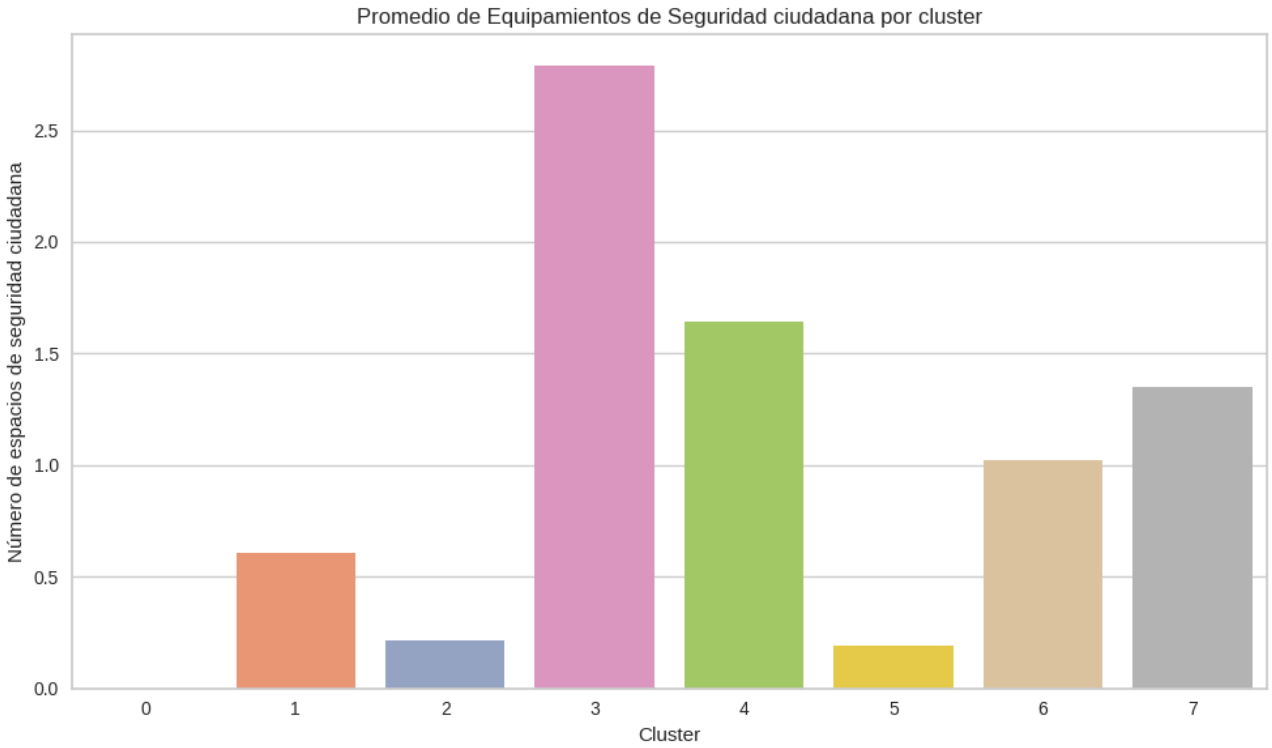

Gráfico 12. Equipamientos de Seguridad ciudadana por Clúster

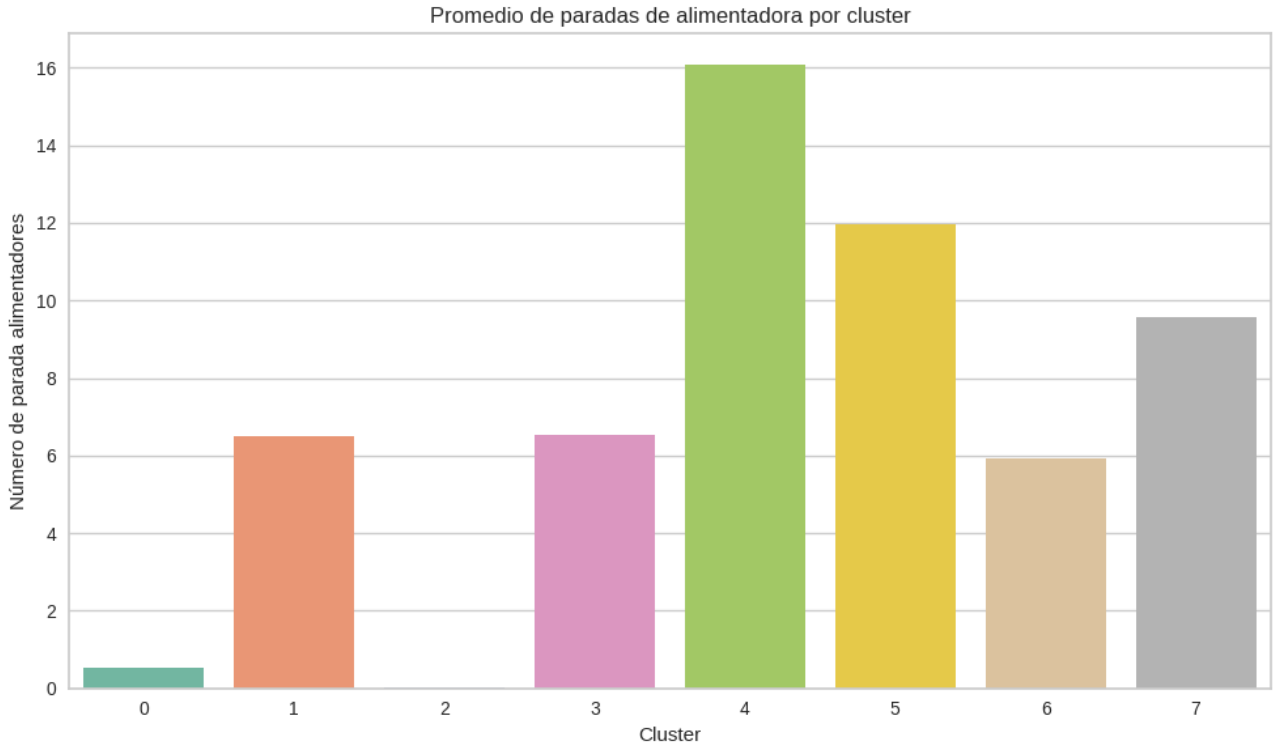

Gráfico 13. Paradas de alimentadores por Clúster

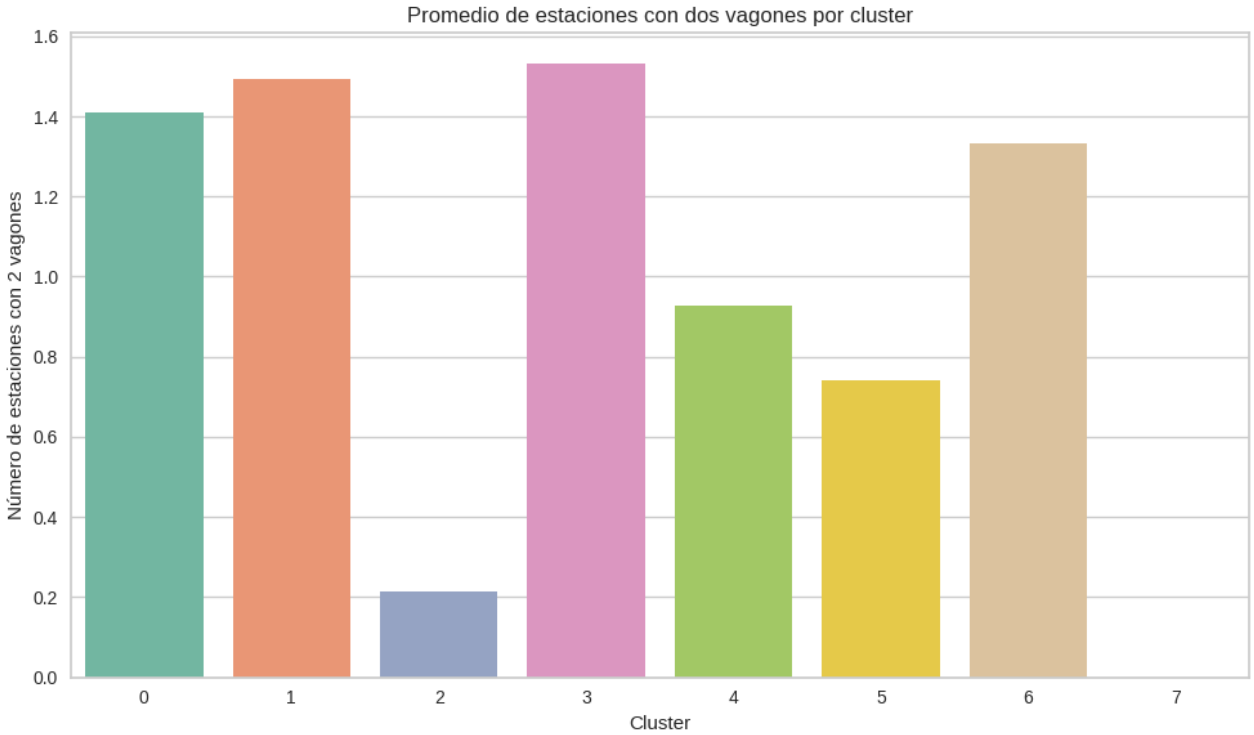

Gráfico 14. Estaciones con dos vagones por Clúster

En cuanto a su accesibilidad, tiene valores altos en elección (Gráfico 7), lo cual también se explica por su ubicación y función en la ciudad y en integración, entendida como la medida que determina qué tan fácil es llegar a un punto desde otros lugares de la red analizada.

Clúster 4:
Este clúster corresponde a manzanas del barrio El Lido, especialmente lotes vacíos donde se ubicaba la Rueda, así como la estación Cañaveralejo. Conserva cercanía con el barrio Tequendama, por lo que presenta valores medios en la cantidad de equipamientos de salud cercanos, mientras que es el que menos parques (Gráfico 14) y zonas verdes tiene. Así mismo, es el que más rutas alimentadoras tiene en su espectro espacial, lo cual podría explicarse por aquellas rutas que generan acceso al Sistema Integrado de Transporte desde la ladera a rutas

pretroncales (Gráfico 8), especialmente de la Calle 5.

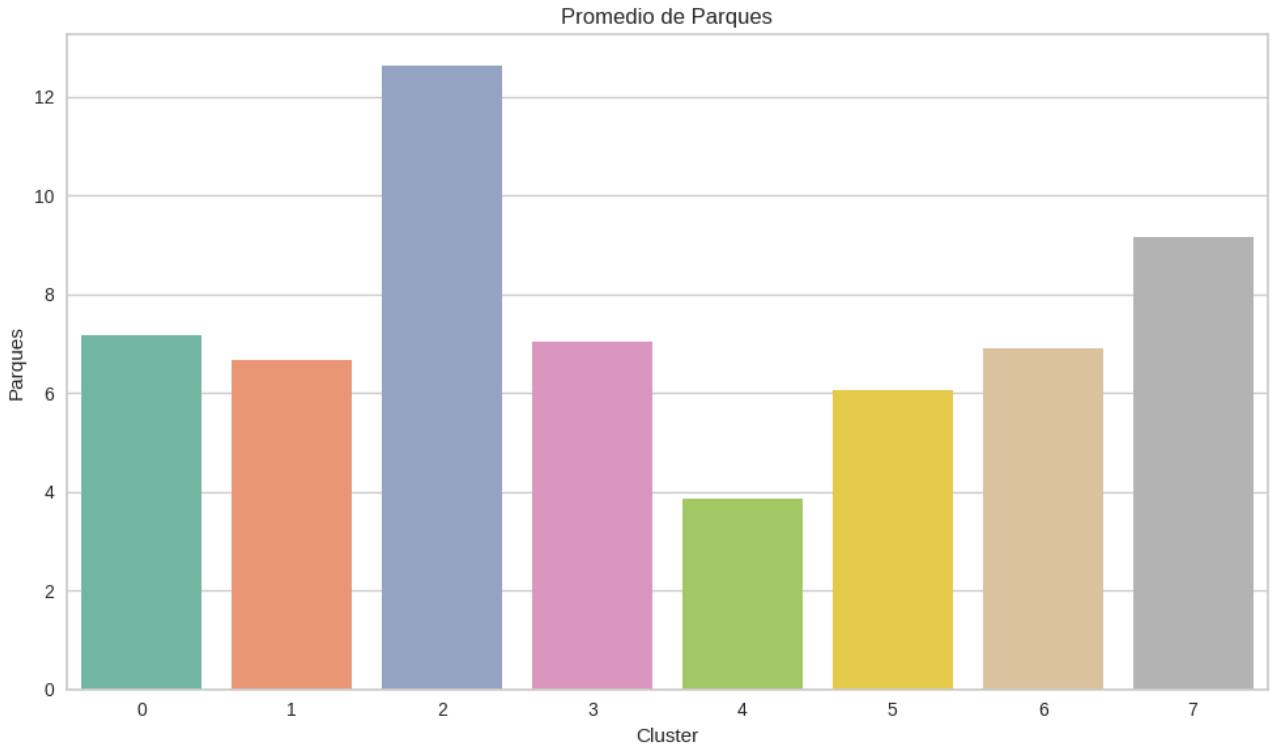

Gráfico 15. Parques por Clúster

Un elemento en común de estas manzanas es que la mayoría corresponden a lotes vacíos sin desarrollar o Asentamientos Humanos de Desarrollo Incompleto, por lo que las zonas verdes tienden a ser privadas y, al predominar mega equipamientos con este tipo de predios, no se han realizado compensaciones de espacio público en la zona.

Clúster 5:
Este clúster corresponde al centro administrativo de la ciudad (CAM y su área de influencia), por lo que, naturalmente predominan equipamientos de Administración Pública (Gráfico 15) (el mismo CAM, Instituto Geográfico Agustín Codazzi, Procuraduría General de la Nación, Gobernación del Valle, Asamblea Departamental, entre otros) y sus respectivas plazoletas (Gráfico 16). Así mismo, al ser de las zonas más tradicionales de la ciudad y, entendiendo que las ciudades latinoamericanas se configuraron en un principio en torno a iglesias y catedrales, es el clúster dónde más se concentran manzanas con cercanía a equipamientos de culto (Gráfico 17), como la Iglesia La Merced, el Templo Sagrado Corazón de Jesús, la Iglesia de San Francisco, la Catedral de San Pedro o el Convento de San Joaquín.

Debido a su centralidad geográfica y administrativa, es de los clústeres con mejores indicadores de integración de la ciudad y, además, de los más compactos en la medida que la mayoría de las manzanas se encuentran cerca a la media para esta medición.

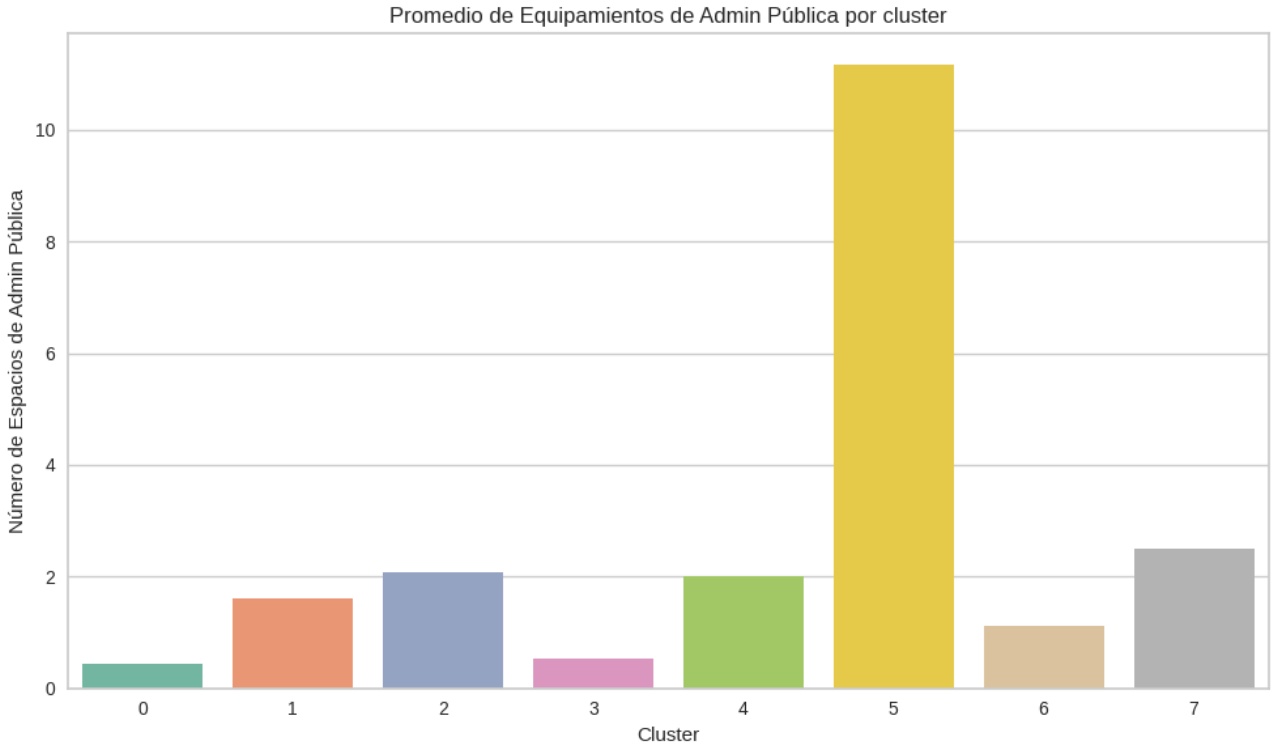

Gráfico 16. Equipamientos de Administración Pública por Clúster

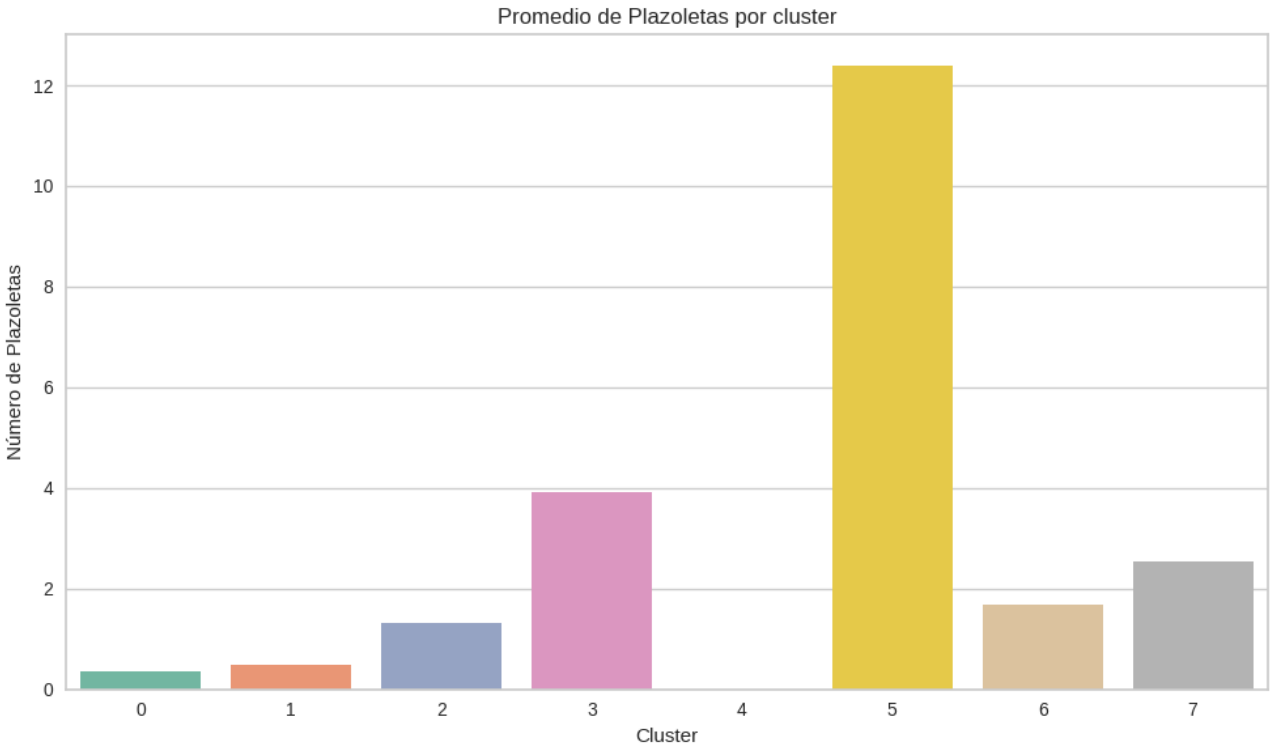

Gráfico 17. Plazoletas por Clúster

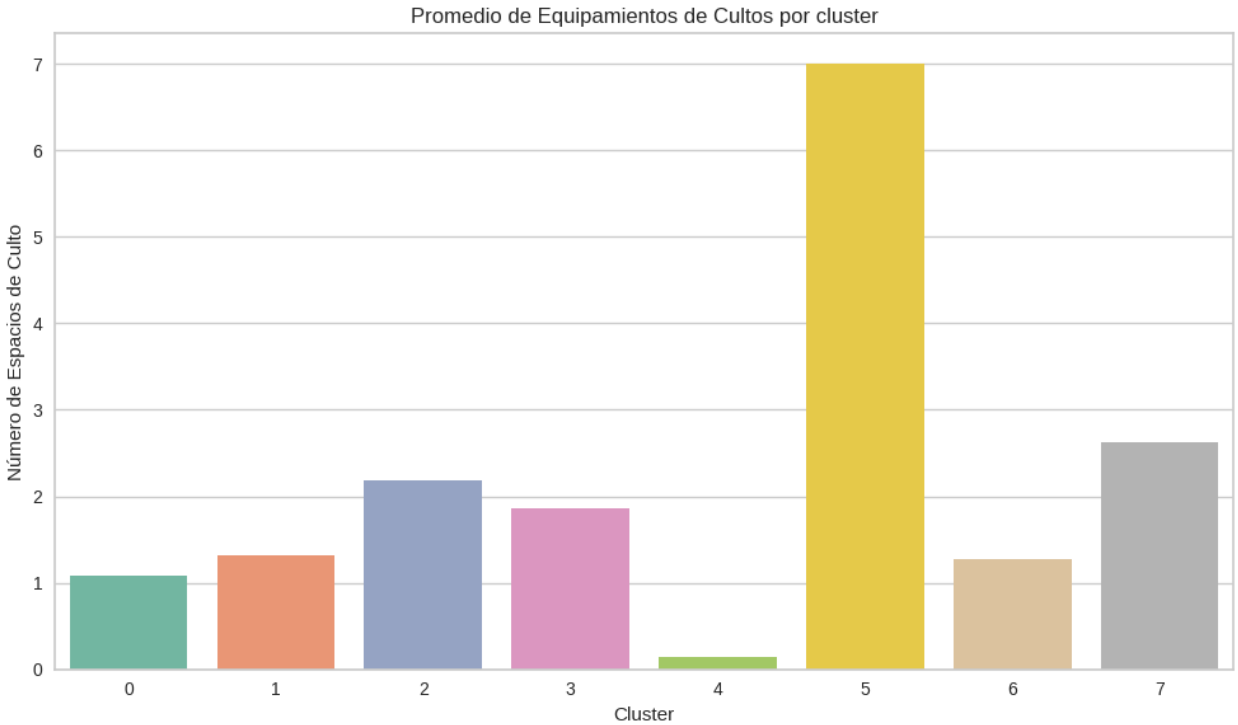

Gráfico 18. Equipamientos de culto por clúster

Clúster 6:

Este clúster, siendo el más disperso en términos espaciales, corresponde a manzanas que, en términos de cercanía a equipamientos de salud, educación, espacio público y bienestar social, es uno de los más equilibrados en comparación con los promedios de los demás clústeres. Sin embargo, el factor diferenciador de este clúster son las métricas de Sintaxis Espacial (Gráficos 18 y 19), ya que es el que menos integración, conexión axial y elección presentan y, en esencia, en su mayoría son equipamientos.

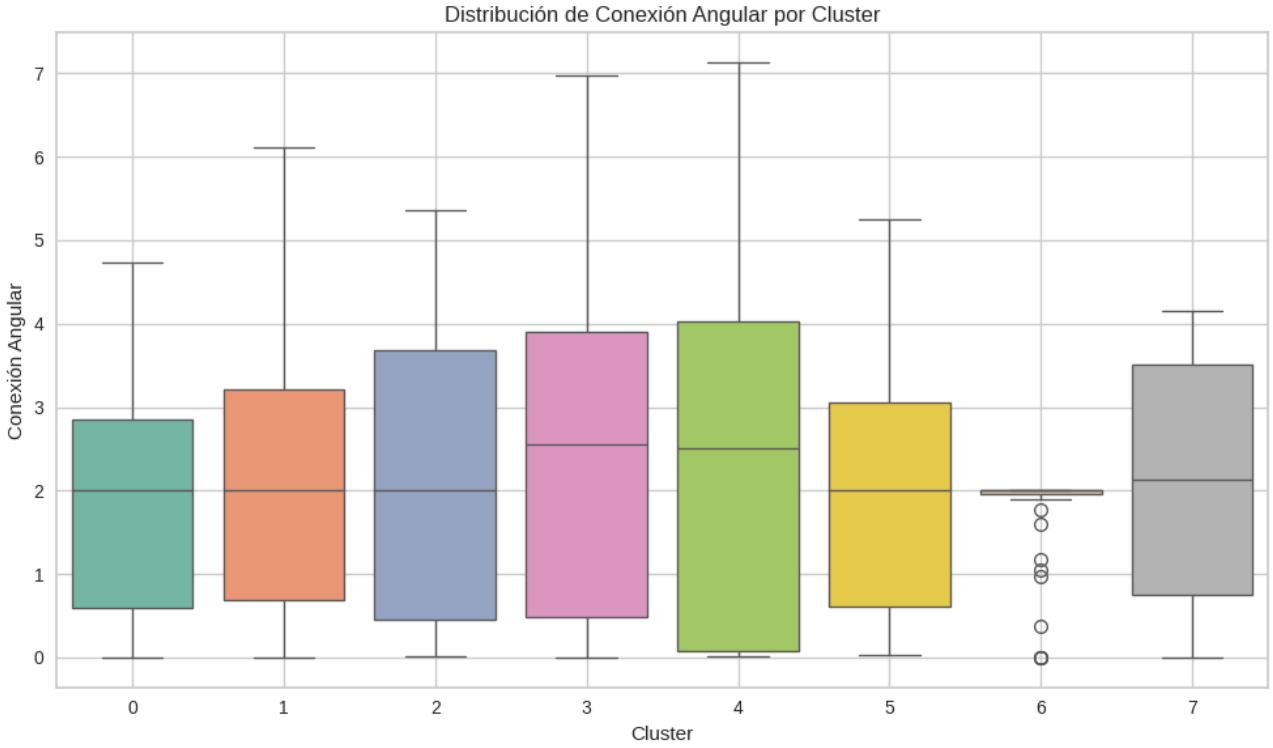

Gráfico 19. Conexión Angular por Clúster

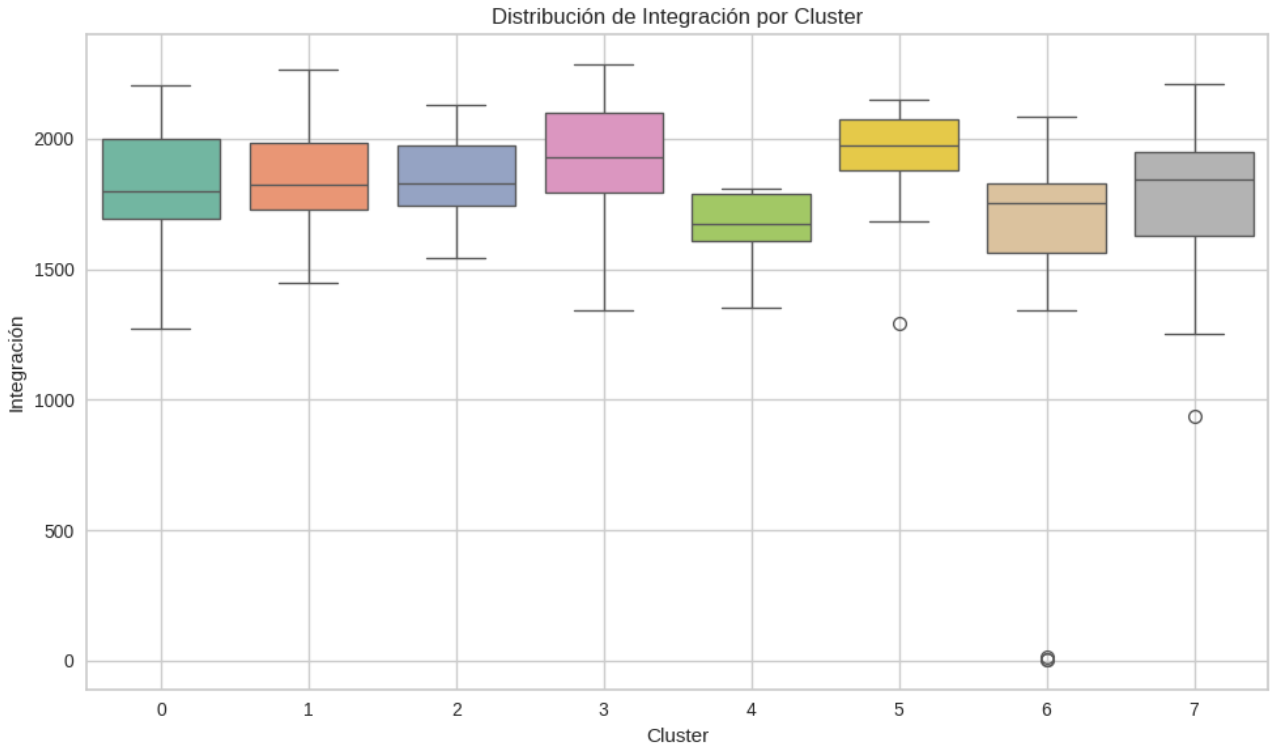

Gráfico 20. Integración por Clúster

Para entender por qué estos equipamientos, a diferencia de los demás analizados, presentan dicho comportamiento en torno a estos indicadores, se observa que son manzanas amplias (como el Estadio Pascual Guerrero, Biblioteca Departamental, Colegio Stella Maris, Corporación Universitaria Minuto de Dios o la Clínica San Fernando), lo que típicamente se traduce en una menor densidad de conexiones peatonales y vehiculares internas. A pesar de su proximidad a vías principales, su configuración espacial interna limita su accesibilidad y su capacidad de servir como conectores urbanos. Además del tamaño, la estructura vial también afecta los resultados de estas métricas, como es el caso de manzanas relativamente más pequeñas como el CAI de la Loma de la Cruz o la Cruz Roja.

Clúster 7:
Este clúster corresponde, primordialmente al barrio San Antonio y parte de San Cayetano, los cuales son unos de los barrios residenciales más tradicionales de la ciudad. En este sentido, sus manzanas presentan baja cercanía con equipamientos de salud o deporte, de manera que sus características de proximidad con la administración pública o equipamientos de culto, se explica más por su relación con el clúster 5 que por dinámicas internas de su configuración espacial y funcional.

Así mismo, es de los que menos integración presenta, lo cual se explica por su topografía

montañosa y estructura vial, la cual responde a dimensiones y trazados planteados en épocas incipientes del desarrollo urbano de la ciudad. No obstante, se puede decir que dicho aspecto se complementa, al ser uno de los clústeres con mayor cantidad de rutas alimentadoras en su radio de influencia.

5. Reconocimiento del impacto de las dinámicas de configuración urbana (OE 4)

Ejercicio ANOVA y varianza de los centroides:

Para identificar si existían diferencias estadísticamente significativas entre los grupos definidos por los clústeres, se aplicó un análisis de varianza (ANOVA) unidireccional a las variables numéricas del conjunto de datos. Primero, se seleccionaron todas las columnas de tipo numérico, excluyendo la variable categórica 'Cluster', que representa la clasificación de los grupos. Para cada variable numérica se agruparon los datos según el clúster correspondiente, y se aplicó la prueba de ANOVA utilizando la función f_oneway del módulo scipy.stats, que permite contrastar la hipótesis nula de igualdad de medias entre grupos.

Los resultados obtenidos incluyeron el estadístico F y el valor p para cada variable. Posteriormente, se clasificaron las variables como significativas ($p < 0.05$) o no significativas, y se ordenaron por el valor p para resaltar aquellas variables con mayor evidencia de diferencias entre clústeres.

Las variables con mayor significancia estadística fueron:

- Salud ($F = 296.05$, $p \approx 6.08e\text{-}169$)
- Administración Pública ($F = 188.61$, $p \approx 6.68e\text{-}133$)
- Elemento Plazoleta ($F = 171.15$, $p \approx 1.20e\text{-}125$)
- Administración Justicia y Convivencia ($F = 124.30$, $p \approx 3.21e\text{-}103$)
- Seguridad Ciudadana ($F = 117.08$, $p \approx 2.74e\text{-}99$)
- Estaciones de 2 Vagones ($F = 115.75$, $p \approx 1.52e\text{-}98$)
- Deporte ($F = 100.18$, $p \approx 1.88e\text{-}89$)
- Culto ($F = 93.65$, $p \approx 2.15e\text{-}85$)
- MIO - Servicio Pretroncal ($F = 75.29$, $p \approx 4.88e\text{-}73$)

Estas variables presentan diferencias significativas en sus promedios entre los diferentes clústeres, lo que sugiere que están fuertemente asociadas con la agrupación realizada. En total, 31 variables resultaron significativas ($p < 0.05$), mientras que otras, como Consultoría de Gestión, Conexión Angular y Servicios de Apoyo a las Empresas, no mostraron diferencias estadísticamente significativas entre grupos ($p > 0.05$).

Este análisis permitió identificar los factores que contribuyen a la diferenciación entre los clústeres y proporciona una base sólida para interpretar las características distintivas de cada grupo. Además, para identificar cuáles variables contribuyeron en mayor medida a la diferenciación entre los grupos resultantes del análisis de clústeres, se estimó la importancia

relativa de cada una con base en la varianza entre los centroides.

Primero, se extrajeron las coordenadas de los centroides obtenidos mediante el algoritmo k-means, excluyendo la variable de agrupamiento ('Cluster'). Luego, se calculó la varianza de cada variable en estos centroides: una mayor varianza indica que esa variable toma valores significativamente distintos entre clústeres, lo cual sugiere que tuvo un mayor peso en la segmentación.

Las variables se ordenaron de mayor a menor según su varianza entre centroides. Este análisis permitió establecer un ranking de importancia que aporta evidencia sobre qué dimensiones fueron más relevantes en la configuración de los grupos.

El ejercicio muestra que las variables con mayor peso en la diferenciación de los clústeres son principalmente aquellas relacionadas con funciones públicas, infraestructura urbana y servicios esenciales. La variable Administración de Justicia y Convivencia presentó la mayor varianza (0.1228), lo que indica que fue el factor más determinante en la separación de los grupos. Le siguen estaciones de dos vagones (0.0885), asociada al sistema de transporte masivo, y Salud (0.0780), relacionada con la presencia de servicios de atención médica. También resultaron relevantes Culto (0.0680), que puede reflejar diferencias en la localización de espacios religiosos entre grupos, y Administración Pública (0.0627), lo que sugiere una asociación entre la gestión institucional y la configuración territorial.

En el otro extremo, algunas variables mostraron una varianza mínima entre centroides, lo cual indica que no contribuyeron significativamente a la diferenciación entre grupos. Entre estas se encuentran actividades comerciales, todas con valores de varianza por debajo de 0.001. Este patrón sugiere que la segmentación territorial responde más a la distribución de servicios públicos, nodos de movilidad y equipamientos urbanos que a actividades comerciales especializadas o de menor escala.

Análisis de Regresiones:

Dado lo anterior, y con el objetivo de complementar el análisis de los datos y plantear de manera más concreta el papel de los establecimientos comerciales en las dinámicas de configuración urbana, se realizaron regresiones que exploraran la hipótesis de que los equipamientos influyen en la configuración comercial de las manzanas. Con el fin de definir el mejor tipo de regresión se corrió un ciclo que evaluara la desviación y los grados de libertad al correr la regresión poisson, si el p-value resultaba mayor a 0.05 se debía utilizar una binomial negativa. Al final todas las variables resultaron válidas para una regresión poisson, que además es coherente con el tipo de datos manejados pues corresponde a conteos discretos. En otras palabras, este tipo de regresión permitiría conocer de manera aproximada en cuánto pueden

aumentar los locales comerciales según la presencia de los equipamientos.

Luego de definir el tipo de regresión se tomó cada comercio como variable dependiente y se corrió la regresión multivariada con los equipamientos, con el fin de priorizar variables se hizo este mismo ejercicio aplicando en las regresiones el método stepwise combinado o bidireccional, que arrojaría las variables con mayor significancia estadística. El siguiente cuadro muestra los resultados de la regresión usando el stepwise:

Variable Dependiente (Comercio)	Variable Independiente (Equipamiento)	P-Value
Actividades de apoyo diagnóstico	Salud	1.24504e-25
Actividades de apoyo terapéutico	Salud	1.24504e-25
	Conexión Axial	0.000224195
Actividades de consultoría de gestión	Choice (Elección)	0.00320939
	Elemento Plaza	0.00254471
Actividades de estaciones vías y servicios complementarios	Salud	2.97158e-18
	Vagones_1 (Estaciones con 1 Vagón)	3.47036e-05
Actividades de hospitales y clínicas con internación	Salud	2.97158e-18
	Recreación	0.000501195
Actividades de la práctica médica sin internación	Salud	5.39333e-24
	Vagones_0	0.00140363
Actividades de la práctica odontológica	Salud	6.32448e-17
Actividades de las agencias de viaje	Elemento Plaza	9.25212e-08
Atividades inmobiliarias realizadas a cambio de una retribución	Elemento Plaza	2.48834e-06
Actividades inmobiliarias realizadas con bienes propios o arrendados	Elemento Plaza	3.99512e-15
Actividades jurídicas	Vagones 1	1.01399e-24
	Elemento Plaza	0.000239542
Alojamiento en hoteles	Elemento Plaza	0.000348346
Comercio al por mayor de productos farmacéuticos medicinales	Salud	8.07505e-06
	Deporte	0.00869913
Comercio al por menor de artículos de ferretería	Vagones 3	8.43026e-06
	Abastecimiento de Alimentos	0.00091244
	Vagones 1	0.000976761
	Integración Total	0.00598336
Comercio al por menor de artículos y utensilios de uso doméstico	Bienestar Social	5.7502e-05
	Culto	0.000628543
Comercio al por menor de bebidas y productos del tabaco	Vagones 3	1.81167e-06
	Elemento Plaza	0.000418833
Comercio al por menor de carnes incluye aves de corral	Abastecimiento de Alimentos	2.59831e-07

Comercio al por menor de electrodomésticos y gasodomésticos	Abastecimiento de Alimentos	7.80135e-07
	Vagones 1	1.29862e-05
Comercio al por menor de libros, periódicos y otros	Vagones 1	3.44959e-17
	Conectividad	0.00277479
Comercio al por menor de otros artículos domésticos	Abastecimiento de Alimentos	2.51774e-08
	Integración total	0.00841239
Comercio al por menor de otros productos alimenticios	Educación	5.95559e-05
	Vagones 2	0.00997153
Comercio al por menor de otros productos nuevos en establecimientos	Culto	0.00636199
Comercio al por menor de prendas de vestir y sus accesorios	Elemento Plaza	1.2751e-05
Comercio al por menor de productos agrícolas para el consumo	Abastecimiento de Alimentos	5.00939e-08
Comercio al por menor de productos farmacéuticos y medicinales	Salud	1.264e-10
	Educación	0.00112944
	Cultura	0.00739759
Comercio al por menor en establecimientos no especializados	Bienestar Social	3.25862e-06
	Vagones 3	0.00054622
	Salud	0.0089976
Comercio al por menor en establecimientos no especializados	No tuvo resultados	
Comercio de motocicletas y de sus partes piezas y accesorios	Vagones 3	1.49952e-11
	Elección (Choice)	0.00334162
Comercio de partes piezas autopartes y accesorios lujos	Vagones 3	1.36002e-06
	Tipo de Servicio: Pretroncal	1.60072e-05
	Administración de Justicia y Convivencia	0.00166649
Confección de prendas de vestir excepto prendas de piel	Seguridad Ciudadana	0.00125416
Elaboración de productos de panadería	Elección (Choice)	0.00827457
Expendio a la mesa de comidas preparadas	Cultura	2.3788e-07
	Conectividad	0.00267732
Expendio de bebidas alcohólicas para el consumo dentro del establecimiento	Elemento Plaza	0.000263962
	Seguridad Ciudadana	0.000228384
Expendio de comidas preparadas en cafeterías	Cultura	0.000389504
	Tipo de servicio: Alimentadora	0.000623245
Formación para el trabajo	Vagones 2	0.00562326
Mantenimiento y reparación de vehículos automotores	Vagones 3	4.21669e-12
	Bienestar Social	4.48376e-06
	Tipo de Servicio Pretroncal	0.000186
	Salud	0.000318959
	Vagones 2	5.77403e-05
	Salud	3.6041e-18

Otras actividades de atención de la salud humana	Deporte	0.00390969
Otras actividades de servicio de apoyo a las empresas	No tuvo resultados	
Otros tipos de alojamiento para visitantes	Cultura	2.01738e-05
	Elemento Plaza	2.96382e-05
Otros tipos de expendio de comidas preparadas	No Tuvo Resultados	
Peluquería y otros tratamientos de belleza	Salud	0.00207932
Publicidad	No tuvo resultados	

Tabla 8. Resultados Regresión Poisson usando Stepwise

En primer lugar, es relevante apuntar a los establecimientos comerciales que no tuvieron resultados, que fueron: Comercio al por menor en establecimientos no especializados, Otras actividades de servicio de apoyo a las empresas, Otros tipos de expendio de comidas preparadas y Publicidad. En comparación con los otros establecimientos, los que no obtuvieron resultados tenían una baja variabilidad y en general poca presencia en el territorio seleccionado para el análisis.

Ahora bien, se puede ver que el equipamiento Salud tiene la mayor cantidad de asociaciones significativas, los casos más relevantes son las Actividades de apoyo diagnóstico y las Actividades de apoyo terapéutico con un p-value de 1,25e-20 para ambas relaciones. Lo anterior apunta una significancia alta en la relación estadística y puede interpretarse diciendo que los equipamientos de salud atraen este tipo de actividades comerciales, que tiene sentido intuitivamente pues ambas están vinculadas con la atención de pacientes y servicios médicos en general. Otros resultados con relevancia estadística vinculados a los equipamientos de salud son Actividades médicas sin internación, Actividades odontológicas, Otras actividades de atención a la salud humana y Comercio de productos farmacéuticos y medicinales.

A continuación, se presenta una tabla que resume los equipamientos con mayor impacto o más asociaciones arrojadas por el ejercicio de las regresiones:

Equipamiento	# de asociaciones	Observación
Salud	12	Altamente asociado a actividades de salud y servicios conexos. Impacto fuerte y consistente.
Elemento Plaza	10	Aparece como significativo en varias actividades comerciales de diferente naturaleza. Podría tener relación con la prevalencia de plazas en la zona del centro.

Vagones 3 (Transporte)	6	Relacionado con comercio minorista de comida y salud.
Abastecimiento de Alimentos	5	Tiene apariciones recurrentes alrededor de actividades vinculadas con temas domésticos y ventas minoritarias de víveres.
Vagones 1 (Transporte)	4	Efecto más débil pero aún significativo para algunas actividades minoristas.

Tabla 9. Resumen equipamientos con mayor impacto según las regresiones

Por otro lado, se encuentra que como variables independientes los equipamientos de Administración de Justicia son los de menor incidencia, pues sólo hacen presencia una vez, estando vinculados con establecimientos de Comercio de partes piezas autopartes y accesorios lujos, sin embargo, esto puede estar relacionado con el bajo número de estos equipamientos en el sector estudiado.

Este ejercicio permite complementar el análisis de los resultados de la clusterización pues, mientras el agrupamiento de los datos permitió reconocer la zonificación en la distribución del territorio y encontrar patrones relevantes para la planeación urbana, las regresiones arrojan luces sobre los vínculos particulares entre los equipamientos y los locales comerciales, en cierta medida explicando por qué se dieron ciertas distribuciones dentro de la clusterización.

6. Conclusiones y trabajos futuros

En conclusión, después de haber realizado el entrenamiento, optimización y análisis de tres algoritmos de agrupación diferentes, se encontró que el algoritmo K-Means con una configuración de K=8 es el que mejor se ajusta a las particularidades del caso del corredor de la calle 5 en Cali. Lo anterior se corresponde a las investigaciones y trabajos ya citados que se apoyan de este mismo algoritmo debido a su robustez cuando se manejan conjuntos de datos discretos. Además, el funcionamiento interno de K-Means facilitó la lectura de los resultados, en tanto la asignación de centroides permite conocer las características promedio de las manzanas en cada clúster, esto reconociendo que si bien las métricas utilizadas para evaluar el algoritmo (Davies-Bouldin, Calinski-Harabasz y Silhouette) no fueron sobresalientes, se considera que el algoritmo tuvo un desempeño aceptable en tanto permitió realizar análisis y lecturas coherentes con la realidad del territorio.

Ahora bien, aunque el objetivo principal de este proyecto era identificar patrones comerciales en el sector, lo que arrojó el algoritmo es que no hay dichos patrones determinantes en la configuración comercial del corredor de la calle 5. Es decir, no se encontraron patrones definidos por la distribución de los establecimientos comerciales a lo largo del corredor, a excepción de aquellos relacionados con temas de salud, que gravitaron en función de la oferta de este servicio en el barrio Tequendama. Ahora bien, fueron los equipamientos, elementos asociados a la movilidad de la ciudad y, en menor medida, la accesibilidad y conectividad de las manzanas, las variables que más determinaron la configuración de patrones espaciales en el corredor estudiado.

Esto se corroboró a través de un análisis ANOVA de la varianza de centroides, al encontrar que los establecimientos comerciales constituyen variables que tienen poca incidencia en la definición de vocaciones y que, a lo sumo, basándose también en las regresiones realizadas, son parte del resultado de la presencia de variables que sí impactan la especialización del territorio, como los equipamientos urbanos.

También es importante apuntar la relevancia que tuvo incluir en las variables de estudio la manera como las personas recorren la ciudad, basándonos en el tamaño de las manzanas y su topografía, utilizando las métricas del Space Syntax. Conocer cómo las manzanas se integran a todo el entramado del sector fue importante para identificar por ejemplo cómo algunos equipamientos relevantes para la ciudad no tienen una buena conectividad y están en cierta medida aislados desde una perspectiva de movilidad. Este fue el caso del clúster número 6, compuesto principalmente por manzanas con equipamientos relevantes, pero con métricas de Space Syntax más bajas al promedio.

Por otro lado, resultaría valioso poder incluir la variable temporal en futuros trabajos con el fin de identificar si la planificación urbana y ordenamiento territorial de la ciudad han tenido un rol de disponer el espacio físico o, al contrario, han tenido una naturaleza primordialmente reactiva frente a dinámicas territoriales que responden a procesos orgánicos de la sociedad. Así mismo, la inclusión de datos sociodemográficos puede resultar enriquecedor para el análisis y un complemento fundamental en la medida que no solo se aborda la expresión territorial de las decisiones y preferencias de la población, sino las variables humanas, culturales e históricas que impactan en sí mismas la configuración de dichas inclinaciones.

Finalmente, la determinación de las distancias de sólo 500 metros para definir la relación de las manzanas con los equipamientos respondió más a dinámicas de ciudades que tienden a ser compactas y policéntricas o a estándares ideales de la planificación urbana, el ejercicio arrojó nociones sobre la especialización y definición de vocaciones en el corredor priorizado. En este sentido, la identificación de algunos clústeres plenamente especializados en sectores como la salud o el deporte, permitió encontrar ciertos desequilibrios territoriales y de accesibilidad que la ciudad no cumple en virtud de dichos estándares en tanto, idealmente, la mayoría de la población debería contar con cercanía a distintos bienes y servicios ofrecidos por una ciudad, lo cual no se observa en este corredor a pesar abarcar barrios y comunas con una mejor cobertura de oferta de bienes y servicios públicos en comparación a otras zonas de la ciudad. Por esto, una línea de trabajo que puede partir de la presente investigación podría centrarse en la creación de una metodología que permita cuantificar y establecer el comportamiento y los impactos de estas brechas territoriales en este corredor o, incluso, la ciudad como un sistema completo.

7. Referencias bibliográficas

[1] M. Castels, La Cuestión Urbana, México DF: Grupo Editorial Siglo Ventiuno, 2014.

[2] A. Novick, «Historias del Urbanismo/Historias de la Ciudad. Una revisión de la bibliografía,» *Seminarios de crítica,* nº 137, 2004.

[3] Banco de Desarrollo de América Latina (CAF), Crecimiento urbano y acceso a oportunidades: un desafío para América Latina, Bogotá: CAF, 2017.

[4] R. Romero, «UNA REVISIÓN DE LAS TRANSFORMACIONES URBANAS EN LAS CIUDADES LATINOAMERICANAS CONTEMPORÁNEAS,» *Revista Vivienda y Ciudad,* vol. 9, pp. 83-103, 2022.

[5] M. O. Smolka y C. Goytia, «Mercados de suelo (Urbano),» *The Wiley-Blackwell Encyclopedia of Urban and Regional Studies,* 2019.

[6] A. Sagi, A. Gal, D. Broitman y D. Czamanski, «An unsupervised machine learning approach to the spatial analysis of urban systems through neighbourhoods' dynamics,» *Land Use Policy,* vol. 144, 2024.

[7] J. Reades, J. De Souza y P. Hubbard, «Understanding urban gentrification through machine learning,» *Urban Studies,* vol. 56, nº 5, pp. 922-942, 2018.

[8] S. Carmona, «Algoritmos de clustering para la detección de posibles paradas de autobús,» *Informaticae Abstracta,* vol. 2, nº 1, pp. 4-24, 2024.

[9] BALLAERIKA, «datasciencesociety.net,» Data Science Socierty, 12 08 2023. [En línea]. Available: https://www.datasciencesociety.net/smart-cities-and-data-science-innovations-for-urban-planning-and-infrastructure. [Último acceso: 25 06 2024].

[10] F. Duarte y P. de Souza, «Data Science and Cities: A Critical Approach,» *Harvard Data Science Review,* vol. 3, nº 2, 2020.

[11] P. Tank, «planningtank.com,» Planing Tank, 01 07 2024. [En línea]. Available: https://planningtank.com/blog/applications-data-science-in-urban-planning. [Último acceso: 25 06 2024].

[12] Moldstud, «moldstud.com,» Moldstud, 02 02 2024. [En línea]. Available: https://moldstud.com/articles/p-data-science-in-urban-planning-smart-cities-and-sustainable-development. [Último acceso: 25 06 2024].

[13] U. C. d. Londres, «UCL Space Syntax - Overview,» [En línea]. Available: https://www.spacesyntax.online/overview-2/. [Último acceso: 16 abril 2025].

[14] J. Bermejo Tirado, «Leyendo los espacios: una aproximación crítica a la sintaxis espacial como herramienta de análisis arqueológico,» *Arqueología de la Arquitectura,* nº 6, pp. 47-62, 2009.

[15] R. Carlos, *Herramientas de Sintaxis espacial,* Buenos Aires: Universidad de Buenos Aires.

[16] M. Dawes y M. Ostwald, «Precise locations in space: An alternative approach to space syntax analysis using intersection points,» *Architecture Research,* vol. 3, nº 1, pp. 1-11, 2013.

[17] A. Mohamed, H. B. Khalil, T. M. Sobhy y F. Heba, «Space Syntax as a Vital Tool to Enhance Urban Spaces. Egyptian Review of Urban and Regional Planning,» *ERURJ,* vol. 2, nº 3, pp. 399-414, 2023.

[18] B. Vargas, «La sintaxis espacial como herramienta de análisis territorial en el medio rural,» de *VI Encuentro Latinoamericano de Metodolog´ıa de las Ciencias Sociales (ELMeCS)*, Buenos Aires, Memoria Académica, 2018, pp. 1-16.

[19] D. Montello, «The contribution of space syntax to a comprehensive theory of environmental psychology,» *Proceedings,* vol. 6, pp. 1-12, 2007.

[20] J. Boelaert y É. Ollion, «The Great Regression Machine Learning, Econometrics, and the Future of Quantitative Social Sciences,» *Revue française de sociologie,* vol. 59, nº 3, 2018.

[21] J. Grimmer, M. Roberts y B. Stewart, «Machine Learning for Social Science: An Agnostic Approach,» *Annual Review of Political Science,* 2021.

[22] M. Hindman, «Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences,» *ANNALS, AAPSS,* 2015.

[23] A. Beltrán, «Detección de Segregación: una propuesta basada en redes complejas y algoritmos de agrupamiento,» *Universidad de La Habana. Trabajo de diploma para optar al título de Licenciado en Ciencia de la Computación,* 2023.

[24] F. Muntagh y P. Legendre, «Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?,» *Journal of Classification,* vol. 31, pp. 274-295, 2014.

[25] A. Dongo, «DESCRIPCIÓN METODOLÓGICA DEL ANÁLISIS CLÚSTER UTILIZANDO EL ALGORITMO DE WARD,» *Trabajo Monográfico, Presentado para optar el título de Ingeniero Estadístico e Informático,* 2017.

[26] U. d. Oviedo, «unioviedo,» [En línea]. Available: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html. [Último acceso: 15 04 2025].

[27] J. Reades, J. DeSouza y H. Phil, «Understanding urban gentrification through machine learning,» *Urban Studies,* vol. 56, nº 5, pp. 922-942, 2018.

[28] L. Palafox y P. Ortiz-Monasterio, «Predicting gentrification in Mexico city using neural networks,» de *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.

[29] V. Delgadillo, «Desafíos para el estudo de desplazamientos sociales en los procesos de gentrificación,» *Contested Cities,* pp. 2-17, 2015.

[30] A. López-Gay, J. Sales-Favà, M. Solana, A. Fernández y A. Peralta, «Midiendo los procesos de gentrificación en Barcelona y Madrid: una propuesta metodológica.,» de *Proceedings: XIII International Conference on Virtual City and Territory: "Challenges and paradigms of the contemporary city"*, Barcelona, 2019.

[31] H. a. G. M. Rabiei-Dastjerdi, «Identifying patterns of neighbourhood change based on spatiotemporal analysis of airbnb data in Dublin,» de *2020 4th International Conference on Smart Grid and Smart Cities (ICSGSC)*, 2020.

[32] S. D. H. a. D. L. Han, «Exploring commercial gentrification using Instagram data,» de *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020.

[33] M. Martí-Costa, G. Durán y A. Marulanda, «Entre la movilidad social y el desplazamiento. Una aproximación cuantitativa a la gentrificación en Quito,» *Revista invi,* vol. 31, nº 88, pp. 131-160, 2016.

[34] E. Bournazou, «Cambios socioterritoriales e indicios de gentrificación. Un método para su medición,» *Academia XXII,* nº 12, pp. 46-59, 2015.

[35] T. Huang, T. Dai, Z. Wang, H. Yoon, H. Sheng, A. Y.Ng, R. Rajagopal y J. Hwang, «Detecting Neighborhood Gentrification at Scale via Street-level Visual Data,» de *IEEE International Conference on Big Data*, 2022.

[36] M. De Nadai y B. Lepri, «The economic value of neighborhoods: Predicting real estate prices from the

urban environment,» de *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, 2018.

[37] A. G. Aguilar y P. Mateos, «Diferenciación sociodemográfica del espacio urbano de la Ciudad de México,» *EURE,* vol. 37, nº 110, pp. 5-30, 2011.

[38] J. Vargas, «¿Urbanización sin Desarrollo?,» de *Crecimiento urbano y acceso a oportunidades: un desafío para América Latina*, Bogotá, CAF, 2017, pp. 19-65.

[39] M. d. R. Navarrete, Z. Borjas y H. Escobar, «La transformación urbana y las actividades económicas terciarias,» de *EMPRESAS, ACTORES SOCIALES E INSTITUCIONES EN LA ORGANIZACIÓN PRODUCTIVA DEL TERRITORIO Y LA INNOVACIÓN PARA EL DESARROLLO LOCAL*, Ciudad de México, Universidad Nacional Autónoma de México, 2018, pp. 272-292.

[40] L. C. Barrera, «Transformación de la ciudad a pequeñas dosis. Renovación urbana 'lote a lote' en Bogotá (2008-2018),» *Revista Ciudades, Estados y Política,* vol. 8, nº 1, pp. 17-32, 2021.

[41] M. Arnaiz, B. Ruiz-Apilánez y J. M. Ureña, «El análisis de la traza mediante Space Syntax. Evolución de la accesibilidad configuracional de las ciudades históricas de Toledo y Alcalá de Henares,» *ZARCH,* nº 1, pp. 128-141, 2013.

[42] A. Cuza-Sorolla, M. Hernández-Aguilar y M. Barrera-Rojas, «Aplicación de polígonos Thiessen para la definición y análisis de áreas de influencia del sistema de salud en ciudades costeras del estado de Quintana Roo,» *Quivera Revista de Estudios Territoriales,* vol. 23, nº 1, pp. 49-71, 2021.

[43] S. d. P. d. Territorio, «Unidad de Planificación Urbana 10 - Estadio,» Departamento administrativo de Planeación Municipal, Santiago de Cali, 2017.

[44] V. F. Beebe, *Selección y valoración del mayor y mejor uso de terreno: Caso de propiedad en el barrio urbanización Tequendama de Cali.,* Trabajo de Grado para optar por el título de Magister en Finanzas, 2014.

[45] M. Kennedy y P. Leonard, «Dealing with Neighborhood Change: A Primer on Gentrification and Policy Choices,» *The Brookings Institution Center on Urban and Metropolitan Policy,* 2001.

[46] N. Smith, The new urban frontier: gentrification and the revanchist city, Londres: Routledge, 1996.

[47] M. Rubiales, «¿Medir la gentrificación? Epistemologías, metodologías y herramientas de investigación de carácter cuantitativo y mixto.,» *Contested Cities,* 2014.

[48] A. Hasan, «GRC Globalgroup,» 18 02 2024. [En línea]. Available: https://insights.grcglobalgroup.com/gentrification/. [Último acceso: 06 2024].

[49] G. Bridge, T. Butler y L. (. Lees, Mixed communities: Gentrification by stealth? (1st ed.), Bristol University Press, 2012.

[50] W. Thackway, «Building a predictive machine learning model of gentrification in Sydney,» *Cities,* vol. 134, 2023.

[51] L. Ilic, M. Sawada y A. Zarzelli, «Deep mapping gentrification in a large Canadian city using deep learning and Google Street View,» *PLoS ONE,* vol. 14, nº 3, 2019.

[52] R. Figueroa, «A housing-based delineation of gentrification: a small area analysis of regina, Canada,» *Geoforum,* vol. 26, nº 2, pp. 225-236, 1995.

[53] A. Rowald, «Measuring gentrification and displacement in greater London,» *Urban Studies,* vol. 37, nº 1, pp. 149-165, 2000.

[54] I. Gould y K. O'Regan, «Reversal of fortunes? Lower-income urban neighbourhoods in the US in the 1990,» *Urban Studies,* vol. 45, nº 4, pp. 845-869, 2008.

[55] A. Podagrosi, I. Vojnovic y B. Pigozzi, «The diversity of gentrification in Houston's urban renaissance: From cleansing the urban poor to supergentrification,» *Environment and Planning A: Economy and Space,* vol. 43, nº 8, pp. 1910-1929, 2011.

[56] M. Ye, I. Vojnovic y G. Chen, «The landscape of gentrification: exploring the diversity of "upgrading" processes in Hong Kong, 1986–2006,» *Urban Geography,* vol. 36, nº 4, pp. 471-503, 2015.

[57] J. Yoo, «Identifying Gentrification Using Machine Learning,» *Census Bureau,* 2023.

[58] J. Eshtiyagh, B. Zhang, Y. Sun, L. Wu y Z. Wang, «A graph-based multimodal framework to predict gentrificatio,» *arXiv,* 2023.

[59] N. Naik, J. Philipoom, R. Raskar y C. Hidalgo, «Streetscore-predicting the perceived safety of one million streetscapes,» *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* 2014.

[60] T. Gebru, J. Kranuse, Y. Wang y L. Fei-fei, «Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States,» *PNAS,* vol. 114, nº 50, 2017.

[61] Y. Alejandro y L. Palafoz, «Gentrification Prediction Using Machine Learning,» *Advances in Soft Computing. Springer International Publishing,* pp. 187-199, 2019.

[62] A. d. S. d. Cali, «Cali.gov,» 16 05 2025. [En línea]. Available: https://www.cali.gov.co/boletines/publicaciones/186848/asi-avanza-el-plan-de-manejo-y-proteccion-con-el-que-cali-busca-conservar-el-patrimonio-de-su-centro-historico/. [Último acceso: 18 05 2025].