

Santiago de Cali, 07 de Junio del 2024

Dr. Diego Luis Linares

Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado **“ANÁLISIS DE LA RED DE TRANSPORTE DEL SISTEMA DE TRANSPORTE MASIVO DE CALI (MIO) EN TÉRMINOS DE TRANSBORDO”** el cual fue realizado por los estudiantes **CESAR ALEJANDRO MARÍN ZAMORA- JHON ALEXANDER GUZMAN ZUÑIGA** con código **8974496- 8975317** pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de **MIGUEL ANGEL VARGAS VALENCIA** identificado con C.C. 1144029180, como Co-Director **JUAN CARLOS MARTINEZ ARIAS** identificado con CC 7549319

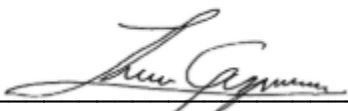
El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.



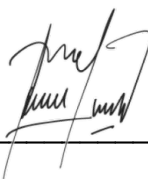
CESAR ALEJANDRO MARÍN ZAMORA
CC 1130606900 de Cali



MIGUEL ANGEL VARGAS VALENCIA
CC. 1144029180 de Cali



JHON ALEXANDER GUZMAN ZUÑIGA
CC 1144089048 de Cali



JUAN CARLOS MARTINEZ ARIAS
CC 7549319 de Armenia



Pontificia Universidad
JAVERIANA
Cali

ANÁLISIS DE LA RED DE TRANSPORTE DEL SISTEMA DE TRANSPORTE MASIVO DE CALI (MIO) EN TÉRMINOS DE TRANSBORDO

*Cesar Alejandro Marín Zamora
Jhon Alexander Guzman Zúñiga*

0

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

Director

Dr. Miguel Ángel Vargas Valencia

Codirector

Mgr. Juan Carlos Martínez Arias

**FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, Junio de 2024**

FICHA RESUMEN

TÍTULO: ANÁLISIS DE LA RED DE TRANSPORTE DEL SISTEMA DE TRANSPORTE MASIVO DE CALI (MIO) EN TÉRMINOS DE TRANSBORDO

1. ÁREA DE TRABAJO: Transporte, redes
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Investigación
3. ESTUDIANTE(S): Cesar Alejandro Marín Zamora - Jhon Alexander Guzman Zúñiga
4. CORREO ELECTRÓNICO: cesalmar@javerianacali.edu.co
gzijhon@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Cr 76#14C -35(3173691933) - Calle 2b# 71 a 11 (3192595352)
6. DIRECTOR: Miguel Ángel Vargas Valencia
7. VINCULACIÓN DEL DIRECTOR: Externo
8. CORREO ELECTRÓNICO DEL DIRECTOR: miguel.vargas@unad.edu.co
9. CODIRECTOR (Si aplica): Juan Carlos Martínez Arias
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): No aplica
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5): Web Scraping, Clustering, Teoría de redes, Network Analysis, Medidas de centralidad de una red
13. FECHA DE INICIO: 1/mayo/2023
14. DURACIÓN ESTIMADA (En meses): 10 meses
15. RESUMEN: *En el 2008 en la Ciudad de Cali se implementó un sistema de transporte masivo (MIO), que desde la fecha a la actualidad ha presentado inconvenientes para cubrir su demanda con eficiencia, generando colapso en sus estaciones e inconformidades en sus pasajeros. A partir de lo anterior, se planteó este proyecto de investigación donde se realizó un análisis a la red de transporte del sistema MIO en términos de transbordo, a través de técnicas de recolección y procesamiento de datos como web scraping, clustering aglomerativo, teoría de redes, Network Analysis, medidas de centralidad e intermediación de una red, encontrando como resultado la caracterización de la red de transporte MIO con relación a los transbordos que la conforman, mediante la identificación de las rutas y los puntos de transbordos más utilizados e importantes de la Ciudad, el reconocimiento del patrón de movilidad de los pasajeros con relación al uso de transbordos y la eficiencia del sistema de transporte MIO; información que se podrá utilizar en futuras investigaciones para crear planes de mejora.*

TABLA DE CONTENIDO

1. INTRODUCCIÓN	7
2. DEFINICIÓN DEL PROBLEMA	8
2.1. Planteamiento del problema	10
2.2. Formulación del problema.....	11
3. OBJETIVOS DEL PROYECTO	12
3.1. Objetivo general	12
3.2. Objetivos específicos	12
4. MARCO TEÓRICO Y ANTECEDENTES	13
4.1 Marco teórico	13
4.1.3 Web scraping:	13
4.1.4 Clustering (agglomerative):	14
4.1.5 Algoritmo de naive	15
4.1.6 Distancia de Haversine	16
4.1.7 Función de enlace.....	16
4.1.8 Teoría de redes, nodos, links, conexidad:	17
4.1.9 Network analysis (análisis de redes):.....	17
4.1.9.1 Medidas de centralidad de una red:	17
4.1.1 Evaluación de modelos	19
4.2 Antecedentes	21
5. METODOLOGÍA.....	23
6. RESULTADOS.....	25
6.1 Datos de las rutas y paraderos del sistema MIO.....	25

6.2	Clasificación de las plataformas de parada georreferenciadas	27
6.3	Evaluación del modelo	28
6.4	Red de datos para el estudio del flujo de pasajeros.....	29
6.5	Centralidad	31
6.6	Intermediación	33
6.7	Cercanía	34
7.	CONCLUSIONES Y TRABAJOS FUTUROS	37
7.1	Conclusiones.....	37
7.2	Trabajos futuros	38
8.	REFERENCIAS	39

TABLA DE CONTENIDO DE FIGURAS

Fig. 1 Congestión MIO en el año 2011 [3]	8
Fig. 2 Congestión MIO en el 2019	9
Fig. 3 Modelo de obtención de los datos en el webscrapping	25
Fig. 4 Datos simplificados por rutas y coordenadas	26
Fig. 5 Ubicación de plataformas y paraderos SITM, 2023	26
Fig. 6 Ubicación de plataformas y paraderos con clúster aglomerativo de 300 metros, 2023	27
Fig. 7 Tamaño de los clúster	28
Fig. 8 Red bipartita entre conjunto de rutas y el conjunto de puntos de transbordo	29
Fig. 9 Red de transbordos con una densidad de 0.084617	30
Fig. 10 Centralidad	31
Fig. 11 Centralidad de vector propio: Eigenvalor	32
Fig. 12 Intermediación	33
Fig. 13 Cercanía	34
Fig. 14 Frecuencia de valores de los índices por tipo de índice	35

1. INTRODUCCIÓN

La ciudad de Santiago de Cali, ubicada en Colombia, es una de las más pobladas del país y ha experimentado un crecimiento demográfico significativo en los últimos años. Como resultado, la demanda de transporte público se ha visto afectada, lo que ha llevado a la implementación de un sistema de transporte masivo (SITM) llamado MIO en 2008. Este sistema utiliza autobuses articulados que transitan por carriles exclusivos y se conectan con redes de autobuses pre troncales y alimentadores para cubrir las necesidades de movilidad de la ciudad [1].

Desde la implementación del SITM MIO se ha identificado que el servicio no ha logrado cumplir con su demanda especialmente durante las horas pico. A pesar de las soluciones planteadas en el pasado, como el uso de vehículos adicionales y la creación de más estaciones, la congestión y la dificultad para abordar el sistema continúan siendo un problema para los ciudadanos que ha generado inconformidades en muchos de los usuarios que se transportan en este sistema [2].

Por lo tanto, en este trabajo se planteó hacer un análisis de la agrupación de los puntos de transbordo por medio de un modelo de aprendizaje automático y de una posterior descripción del diseño de la estructura de la red de rutas del sistema de transporte en términos de transbordos para así determinar su eficiencia y posteriormente determinar la que tan eficiente es el sistema actual y explorar posibles mejoras que puedan tener un impacto positivo en la calidad de vida de los habitantes de la Ciudad de Cali.

Para el análisis se construyó y analizó una red mediante medidas de centralidad, intermediación y agrupamiento para identificar los puntos más importantes en el sistema de transporte. Para alcanzar este objetivo, se emplearon técnicas de aprendizaje automático no supervisado para agrupar los datos. Como resultado, se identificaron los puntos de transbordo más relevantes en un mapa que muestra el sistema de transporte del MIO en Cali, considerando los tres tipos de importancia en la red: intermediación, grado y centralidad.

Gracias a estos datos, se podrán diseñar intervenciones para optimizar el sistema de transporte, en busca de reducir la congestión del sistema y la mejora en calidad de vida de los residentes de la ciudad.

2. DEFINICIÓN DEL PROBLEMA

Desde que se implementó el SITM en la Ciudad de Santiago de Cali se viene identificando que el servicio no logra cumplir 100% con su demanda, presentando congestiones en los sitios donde estos pueden ser abordados.

En el 2011 hay reporte de una noticia en el diario EL PAIS de Cali, donde una gran número de usuarios presentan quejas por la saturación de diversas estaciones en horario pico, generando que los pasajeros tengan que realizar largas esperas desde veinte (20) hasta cuarenta (40) minutos, ya sea para abordar o para realizar transbordos a las diferentes rutas para llegar a su destino, generando que en muchas ocasiones no puedan abordar el servicio ya que por su alta demanda este llega congestionado, se afirma que algunas rutas del sistema MIO pueden llegar a transportar hasta 150.000 pasajeros al día [2].

Para solucionar esto, se plantea por la administración el uso de vehículos de reserva y la creación de más estaciones en horario pico para lograr cubrir la demanda de pasajeros [3]. Sin embargo, ocho años después, en el 2019, se encuentran publicaciones realizadas a través de la red social de Twitter (hoy X) donde se identifica que los usuarios continúan reportando gran dificultad para abordar al sistema de transporte por alta demanda y congestión del servicio [4].



Fig. 1 Congestión MIO en el año 2011 [3]



Fig. 2 Congestión MIO en el 2019

Este problema se puede deber en parte a la ineficiencia del sistema en la red para realizar los trasbordos del sistema.

Para el año 2024, en cuanto a la congestión del sistema y las quejas de los usuarios, el MIO ha enfrentado críticas y desafíos relacionados con la capacidad de transporte y la satisfacción de los usuarios. La congestión en ciertas rutas y horas pico ha sido un problema recurrente, especialmente en áreas clave de la ciudad. Esto ha llevado a quejas de los usuarios sobre tiempos de espera prolongados, autobuses abarrotados y dificultades para acceder al sistema en momentos de alta demanda.

Además, se han reportado problemas relacionados con la puntualidad y frecuencia de los autobuses, lo que ha generado frustración entre los usuarios que dependen del sistema para sus desplazamientos diarios. Estos desafíos han contribuido a una percepción negativa en algunos sectores de la población, que demandan mejoras en la eficiencia y calidad del servicio ofrecido por el MIO [4].

2.1. Planteamiento del problema

El transporte público urbano es una necesidad básica para los habitantes de una ciudad [5]. Esta necesidad es impactada negativamente cuando el sistema de transporte público tiene deficiencias en los tiempos de traslado de un punto de la ciudad a otro. Se hace énfasis en los tiempos perdidos en los puntos de transbordo con la espera de la conexión del siguiente recorrido que complementa la ruta. El SITM MIO cuenta con una amplia red de rutas y paradas que conectan a los usuarios con diferentes puntos de la ciudad permitiéndoles llegar a sus diferentes destinos. Estos puntos de parada tienen una gran influencia en el servicio, el tiempo de viaje o de espera, ya que si son bien diseñadas las conexiones de la red se puede optimizar el uso de la red. El tiempo de viaje se puede llegar a reducir si se tiene una correcta organización de los puntos de transbordo del sistema.

A pesar de la importancia del sistema de transporte MIO, actualmente no se cuenta con suficientes estudios para llegar a tener un análisis completo de la red de transporte en términos de transbordo de pasajeros en el sistema, siendo esta una de las posibles causas de que la solución a su congestión, que previamente han sido planteadas no haya logrado dar respuesta al requerimiento.

Teniendo en cuenta lo anterior se decide plantear la siguiente investigación, la cual se enfoca en evaluar la red de transporte del sistema MIO de la ciudad (Cali, Colombia) en el estado que se encuentra en el año 2023 en términos de transbordo de pasajeros. Para esto se obtuvieron los datos de la estructura de rutas y paraderos del sistema MIO, con lo cual se agruparon las diferentes plataformas de parada con georreferenciación de todas las paradas de todas las rutas de buses del sistema (troncales, pretroncales y alimentadores). Además, se construyó una red que permite estudiar el tránsito de los pasajeros entre los diferentes puntos de la ciudad y analizar la red mediante medidas de centralidad y agrupamiento. Esto con el objetivo de lograr la identificación de los puntos más importantes en el sistema de transporte en términos de transbordo.

Esta importancia se mide de tres formas referenciando a la intermediación, el grado y la centralidad. Para ello se utilizaron técnicas de análisis de redes y de aprendizaje automático no supervisado, con el fin de lograr una identificación con los datos obtenidos, para mejorar la eficiencia del sistema de transporte y la calidad de vida de los habitantes de la Ciudad de Cali.

2.2. Formulación del problema

La pregunta de investigación es: ¿Cómo evaluar la eficiencia del diseño de la red del sistema integrado de transporte masivo de Cali (SITM MIO) en términos de transbordo, mediante el análisis y clasificación de datos georreferenciados y Network Analysis aplicado a sus rutas y paraderos?

Las preguntas que se derivan a partir de esta son las siguientes:

1. ¿Cuál es la estructura y calidad de los datos obtenidos mediante Web Scraping de las rutas y paraderos del sistema MIO?
2. ¿Qué patrones o agrupamientos emergen al clasificar las diferentes plataformas de parada georreferenciadas utilizando métodos de aprendizaje de máquina no supervisados?
3. ¿Cómo se comparan y qué conclusiones se pueden extraer al evaluar la calidad y rendimiento del modelo de aprendizaje no supervisado utilizando métricas de evaluación pertinentes?
4. ¿Qué información relevante se puede derivar al analizar una red de datos que permita comprender el flujo de pasajeros entre diferentes áreas de la ciudad, considerando las rutas establecidas por el sistema de transporte MIO y la ubicación de los puntos de transbordo?

3. OBJETIVOS DEL PROYECTO

3.1. Objetivo general

Evaluar la eficiencia de los puntos de transbordos del MIO mediante el uso de clustering no supervisado de paraderos y plataformas con análisis de datos georreferenciales y métricas de análisis de redes para caracterizar la red de transporte MIO reconociendo el patrón de movilidad y la eficiencia del sistema actual de transporte durante el año 2023.

3.2. Objetivos específicos

1. Estructurar los datos de las rutas y paraderos del sistema MIO, mediante Web Scraping a su plataforma pública.
2. Clasificar las diferentes plataformas de parada georreferenciadas mediante un método de aprendizaje de máquina no supervisado, para detectar los puntos de transbordo en toda la ciudad.
3. Evaluar la calidad y rendimiento del modelo de aprendizaje no supervisado mediante la utilización de métricas de evaluación pertinentes, con el fin de determinar la eficacia del modelo en la tarea de agrupamiento.
4. Analizar una red de datos que facilite el estudio del flujo de pasajeros entre los distintos puntos de la ciudad, utilizando las rutas establecidas por el sistema de transporte MIO y los puntos de transbordo

4. MARCO TEÓRICO Y ANTECEDENTES

4.1 Marco teórico

Este marco teórico proporcionó una base sólida de conocimientos para llevar a cabo una investigación sobre la red de transporte del sistema MIO en la ciudad de Cali, Colombia, en términos de transbordo.

Las técnicas y herramientas utilizadas en este estudio permitieron obtener una comprensión detallada de la estructura y las propiedades de esta red, convirtiéndose en una herramienta útil para la toma de decisiones en el ámbito del transporte público en la ciudad. Para ello, se utilizaron diversas técnicas y herramientas, incluyendo web scraping, clustering aglomerativo, teoría de redes, análisis de redes y medidas de centralidad, temática que se describirá a continuación:

4.1.3 Web scraping:

Web scraping es una técnica de extracción automatizada de datos utilizada para obtener información de páginas web y almacenarla en una base de datos que posteriormente permite ser analizada.

Esta técnica consiste en el uso de herramientas automatizadas que comúnmente son programadas para hacer consultas en servidores web, a partir de un escaneo en la página se crea un archivo o documento en el que se recopila información de manera estructurada y que posteriormente es analizada y procesada, esto hace que sea una excelente herramienta para recolectar información masiva [6] [7]. La técnica es necesaria cuando no existe una base de datos estructurada en un documento, la información se encuentra públicamente expuesta, abierta y es éticamente permitido obtener esta información.

En este estudio, se utilizó web scraping en Python para obtener los datos de la estructura de rutas y paraderos del sistema MIO de la ciudad de Cali, Colombia, a partir de su plataforma pública.

4.1.4 Clustering (agglomerative):

El clustering o agrupamiento es una técnica de aprendizaje automático no supervisado utilizada para clasificar objetos o datos en grupos o clústeres según su similitud o cercanía [8].

En un algoritmo de agrupación aglomerativa, los objetos o puntos de datos se agrupan en clústeres de manera iterativa en función de su similitud. Este proceso comienza considerando que cada punto de datos está en su propio clúster individual y, en cada paso, se combinan los dos clústeres más cercanos en uno solo. Este proceso continúa hasta que todos los puntos de datos estén en un solo clúster grande.

Los clústeres aglomerativos se construyen de manera ascendente, lo que significa que comienzan con unidades individuales (puntos de datos) y se combinan gradualmente en clústeres más grandes a medida que avanzan en el proceso. Durante el proceso de agrupación, se calcula la similitud entre los clústeres y se fusionan los dos clústeres que tienen la mayor similitud.

Los resultados de un algoritmo de agrupación aglomerativa se pueden representar en forma de un dendrograma, que es un árbol jerárquico que muestra cómo se agrupan los elementos a diferentes niveles de similitud. El dendrograma permite la visualización de la jerarquía de los clústeres y la estructura de agrupación de los datos.

Esta es una técnica útil para explorar la estructura de los datos, identificar patrones de similitud y comprender cómo los elementos se relacionan en diferentes niveles de detalle. Se utiliza en una variedad de aplicaciones, como segmentación de clientes, análisis de texto, biología, geografía y más.

En este estudio, se utilizó una técnica de clustering aglomerativo para agrupar las diferentes plataformas de parada georreferenciadas del sistema MIO, con el fin de detectar los puntos de transbordo a lo largo de la ciudad.

Para los clústeres aglomerativos existe el proceso de asignación de valores para determinar la similitud de un clúster con otro. Para este caso se utiliza la medida de “centro distancia” para la cual el valor asignado es un registro de distancia, con esto el centro del clúster es el punto promedio y de manera similar a la utilizada en el algoritmo K-means el centro que esté más cerca al valor de distancia dado será en términos de distancia el clúster más cercano [19].

Este modelo se basa en realizar un proceso iterativo en el que en un principio cada punto de datos se encuentra de forma individual. Luego, en cada iteración, los clústeres más cercanos se fusionan hasta obtener un número deseado de clústeres o hasta que se cumpla un criterio de parada definido como por ejemplo 300 metros.

$D(X, Y)$ es la distancia entre los clústers X y Y de tal forma que:

$$D(X, Y) = \max_{x_i \in X, x_j \in Y} d(x_i, x_j) \quad (1)$$

$d(x_i, x_j)$ es la distancia entre $x_i \in X$ y $x_j \in Y$
 X y Y son pares de elementos (clústers)

4.1.5 Algoritmo de naive

El algoritmo de Complete Linkage en clustering aglomerativo es una forma de agrupar datos en clústers basándose en la mayor distancia entre puntos de diferentes clústers [9].

- 1) Inicialización: Comienza considerando cada punto como un clúster individual.
- 2) Cálculo de distancias: Calcula todas las distancias posibles entre los puntos utilizando alguna medida de distancia, como la distancia de haversine.
- 3) Encuentra la distancia máxima: Para cada par de clústers, calcula la distancia máxima entre un punto del primer clúster y otro del segundo clúster. Esto representa la distancia completa entre los dos clústers según el enlace completo.
- 4) Fusiona clústers: Une los dos clústers con la mayor distancia completa entre ellos, es decir, aquellos cuya distancia completa sea la máxima entre todos los pares de clústers.
- 5) Actualización de la matriz de distancias: Actualiza la matriz de distancias para reflejar la fusión de clústers y repite los pasos 3 y 4 hasta que quede un solo clúster o se cumpla algún criterio de parada.

4.1.6 Distancia de Haversine

La distancia de Haversine es una fórmula utilizada para calcular la distancia entre dos puntos en una esfera, como la Tierra. Esta distancia es útil para el ejercicio a realizar en este documento por lo que las distancias no corresponden directamente a las utilizadas en geometría euclidiana como tradicionalmente está determinado en los softwares.

Esta fórmula es útil en aplicaciones de geolocalización y navegación, donde hay que medir la distancia entre dos coordenadas geográficas, como latitud y longitud, incluso a grandes distancias.

La fórmula de Haversine se basa en el principio trigonométrico del mismo nombre, que es una función trigonométrica haversine: $(x) = \sin^2(x/2)$ [10].

4.1.7 Función de enlace

La función de enlace es la forma en la que se utilizarán los datos para calcular la distancia entre grupos o clústers a medida que se van fusionando durante el proceso de agrupamiento. Este tipo de agrupamiento comienza considerando cada punto como un clúster individual y luego se iterativamente los clústers más cercanos hasta que todos los puntos están en un solo clúster o se cumple algún criterio de parada [9].

Existen varios métodos de enlace que determinan cómo se calcula la distancia entre clústers durante este proceso de fusión. Algunos de los métodos de enlace más comunes son: enlace único, completo, promedio o Ward's.

Enlace único (Sencillo Linkage): La distancia entre dos clústers se define como la menor distancia entre un punto del primer clúster y un punto del segundo clúster. Este método tiende a producir clústers alargados.

Enlace completo (Complete Linkage): La distancia entre dos clústers se define como la mayor distancia entre un punto del primer clúster y un punto del segundo clúster. Esto tiende a producir clústers compactos y separados.

Enlace promedio (Average Linkage): La distancia entre dos clústers se define como el promedio de todas las distancias entre los puntos de los dos clústers. Este método busca clústers con una cohesión equilibrada.

Enlace de Ward (Ward's Linkage): Este método calcula la distancia entre clústers teniendo en cuenta cómo la suma de cuadrados se incrementaría al fusionar dos clústers en uno nuevo. Busca minimizar la varianza dentro de cada clúster.

4.1.8 Teoría de redes, nodos, links, conexidad:

La teoría de redes se refiere al estudio de las relaciones entre nodos o elementos de un sistema. En una red, los nodos son los elementos que están conectados entre sí por links o conexiones [8]. En este estudio, se hace uso de la teoría de redes para construir una red que permita estudiar el tránsito de los pasajeros entre los diferentes puntos de la ciudad, mediante las rutas establecidas por el sistema MIO.

4.1.9 Network analysis (análisis de redes):

El análisis de redes es una técnica utilizada para estudiar la estructura y las propiedades de una red [11]. En este estudio, se realizó un análisis de redes para caracterizar los puntos más importantes en el sistema de transporte en términos de transbordo, mediante medidas de centralidad y agrupamiento.

4.1.9.1 Medidas de centralidad de una red:

Las medidas de centralidad son utilizadas para evaluar la importancia de un nodo en una red. Existen diferentes tipos de medidas de centralidad, como la centralidad de intermediación, la centralidad de grado y la centralidad de cercanía [12]. En este estudio, se utilizaron diferentes medidas de centralidad para analizar la red de transporte del sistema MIO y caracterizar los puntos más importantes en términos de transbordo.

Centralidad:

Las medidas de centralidad de grado para una parada miden la cantidad de paradas que pueden alcanzarse sin hacer transbordo, es decir, cuántas conexiones tiene una parada con otras paradas utilizando una única ruta de autobús. Esta medida está relacionada con el número de conexiones entre paradas.

La centralidad de grado para un nodo v es la fracción de nodos a los que está conectado. Los valores de centralidad de grado se normalizan dividiendo por el grado máximo posible en un grafo simple $n-1$ donde $n=|V(G)|$ es el número de nodos en G . (Aquí, $V(G)$ representa el conjunto de nodos en el grafo G) [13].

Para ello, se puede utilizar la matriz de adyacencia $A = (a_{ij})$ del grafo G . La centralidad de grado para un nodo v se calcula mediante la siguiente fórmula:

$$Deg(v) = \frac{1}{n-1} \sum_{w \in V(G)} a_{vw} = \frac{|\{w : (v,w) \in G\}|}{n-1}$$

Intermediación:

La centralidad de intermediación mide la participación de un nodo en las rutas, centrándose en los trasbordos, y ayuda a identificar las paradas altamente necesarias para el traslado de pasajeros. Mide la influencia de una parada en el flujo de pasajeros. La centralidad de intermediación de un nodo es la suma de la fracción de todos los pares de caminos más cortos que pasan a través de él:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

donde V es el conjunto de nodos, σ_{ij} es el número de caminos más cortos entre el par de nodos i y j , y $\sigma_{ij}(v)$ es el número de esos caminos que pasan por algún nodo v que no sea i ni j . Si $i = j$, $\sigma_{ij} = 0$; y si $v = i$ o $v = j$, $\sigma_{ij}(v) = 0$ [14].

Cercanía:

La centralidad de cercanía mide qué tan fácilmente una parada puede alcanzar a las demás paradas. Ayuda a identificar las paradas con un mayor número de conexiones directas a otras paradas de la red.

La centralidad de cercanía de un nodo v es el recíproco de la suma de las distancias de los caminos más cortos desde v a todos los demás nodos. Dado que la suma de distancias depende del número de nodos en el grafo, la cercanía se normaliza dividiendo por la suma de las distancias mínimas posibles.

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)},$$

Es importante destacar que valores más altos de centralidad de cercanía indican una mayor centralidad.

Donde $d(v, u)$ es la distancia del camino más corto entre los nodos v y u , y n es el número de nodos en el grafo [15].

Centralidad Eigenvector:

La centralidad del valor propio mide la fuerza de una parada basada no solo en su propia conexión sino también en la conexión de sus conexiones. Mide qué tan bien conectada está una parada y qué tan bien conectadas están sus conexiones. Así, una parada puede ser importante por su conexión con paradas influyentes, aunque no lo sea.

La centralidad del vector propio calcula la centralidad de un nodo en función de la centralidad de sus vecinos. La centralidad del vector propio para el nodo es:

$$Ax = \lambda x$$

donde A la matriz de adyacencia del gráfico G con valor propio λ , [16]. Es importante remarcar que, según el teorema de Perron-Frobenius, existe una solución única y positiva si es el mayor valor propio asociado al vector propio de la matriz de adyacencia [17].

4.1.1 Evaluación de modelos

En el análisis de datos y la evaluación de la calidad de agrupamiento, se utilizan varios índices para medir la consistencia y la separación entre grupos. Entre estos índices se encuentran el Coeficiente de Silhouette, el Índice de Calinski - Harabasz y el Índice Davies - Bouldin, cada uno con su propia fórmula y significado en términos de la estructura de los conglomerados obtenidos en un proceso de agrupamiento. Estos índices son herramientas clave para evaluar la eficacia y la interpretación de los resultados de algoritmos de agrupamiento en análisis de datos.

4.1.10.1 Índice de Silhouette

El Coeficiente de Silhouette se define para cada muestra y se compone de dos puntuaciones: [18]

- a. La distancia media entre una muestra y todos los demás puntos de la misma clase.
- b. La distancia media entre una muestra y todos los demás puntos del siguiente grupo más cercano.

se expresa como:

$$S = \frac{b-a}{\max(a,b)} \quad (2)$$

4.1.10.2 Índice de Calinski – Harabasz

El índice es la relación entre la suma de la dispersión entre conglomerados y la dispersión dentro de los conglomerados para todos los conglomerados (donde la dispersión se define como la suma de las distancias al cuadrado) [19].

La puntuación es mayor cuando los conglomerados son densos y están bien separados, lo que se relaciona con el concepto estándar de conglomerado.

Para un conjunto de datos E de tamaño n que se ha agrupado en k clústers, el índice Calinski-Harabasz se define como:

$$S = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \quad (3)$$

Donde $tr(B_k)$ es la traza de la matriz de dispersión entre grupos y $tr(W_k)$ es la traza de la matriz de dispersión al interior del grupo, donde:

$$W_k = \sum_{q=1}^k \sum_{x \in C_q} (x - C_q)(x - C_q)^t \quad (4)$$

$$B_k = \sum_{q=1}^k n_q (C_q - C_E)(C_q - C_E)^t \quad (5)$$

En el cual C_q es el conjunto de datos del clúster q, c_q Es el centro del clúster q, c_E Es el centro del clúster E y n_q el número de puntos en el clúster q

4.1.10.3 Índice Davies – Bouldin

Un índice de Davies-Bouldin más bajo se relaciona con un modelo con una mejor separación entre los grupos. Este índice significa la "similitud" promedio entre conglomerados, donde la similitud es una medida que compara la distancia entre conglomerados con el tamaño de los conglomerados mismos. Cero es la puntuación más baja posible. Los valores más cercanos a cero indican una mejor partición.

El índice se define como la similitud promedio entre cada grupo. [20]

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} R_{ij} \quad (6)$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad (7)$$

C_i para $i = 1, \dots, k$ y el más similar C_j . la similaridad es definida por la medida R_{ij}
 s_i , el promedio de la distancia entre cada punto del clúster i y el centroide de ese mismo clúster
 d_{ij} , la distancia de los centroides de los clúster i y j

4.2 Antecedentes

El uso de bases de datos es una estrategia que se puede implementar para identificar necesidades en el uso de Sistemas Integrados de Transportes Masivo (SITM) ya que estas se encargarán de generar información continuamente actualizada que permita plantear posibles mejoras para el uso de los SITM, sin dejar de lado el desarrollo económico favoreciendo que la población logre acceder a sus sitios de empleo y mejorando su calidad de vida, a su vez debe promover la inclusión social, y reducir los daños al medio ambiente [21]; con base a lo anterior se realiza una búsqueda de investigaciones previas donde se utilicen bases de datos como propuesta de mejoras en el transporte masivo.

En una investigación realizada en el año 2019 por Vera et al., mencionan la implementación de big data como estrategia de planificación del transporte urbano, alimentando esta base de datos mediante el uso de dispositivos móviles conectados a redes de internet; evidenciando que se logra tener información continuamente actualizada respecto a cantidad de pasajeros, tiempo de destino, punto de partida, punto de llegada y rutas utilizadas; sin embargo la información no logra ser 100% efectiva ya que mencionan determinado sesgo de información dado por los posibles pasajeros que no cuentan con dispositivos móviles conectados a redes de internet y el poco personal capacitado para el uso y la interpretación de bases de datos [21]. Por otra parte, en un artículo publicado por Fox en el 2010 realizado en Londres (Inglaterra). Se utiliza una base de datos anclada a un sistema de reconocimiento automático de matrículas (Automatic Number Plate Recognition, ANPR sus siglas en ingles), donde logra estimar el número de viajes entre origen y destino en función del tiempo, con esto logran concluir que muchos de los viajes realizados se les puede plantear uso de vías alternas logrando liberar tráfico para viajes con trayectos más prolongados [22].

Así mismo, en el 2023 se publicó un artículo donde realizan un estudio de caso realizado en Budapest, Hungría donde crean una base de datos encargada de analizar la cobertura

(paradas y acceso a la población), oferta (viajes por día por km², frecuencia de servicio por km² y paradas por km²) y diversidad (conexiones a diferentes destinos) del sistema de transporte de dicha ciudad, logrando concluir que con la uso de big data se puede combinar una cantidad grande de medidas para el sistema de transporte y así mismo se pueden construir modelos predictivos para los SITM, logrando identificar y clasificar los sitios que necesitan mejoras, determinar acciones que se pueden implementar como posibles mejoras y predecir los posibles impactos de las acciones de mejora [23].

Con base en todo lo anterior se puede afirmar que las investigaciones previamente descritas son de gran aporte para el desarrollo de esta investigación, ya que todas muestran el uso de base de datos como estrategia para el análisis del funcionamiento del servicio de transporte, logrado con estas llegar a conclusiones que permiten medir el rendimiento del servicio y plantear posibles oportunidades de mejora para la calidad de vida de los ciudadanos en el desarrollo de sus actividades de la vida diaria; sin embargo no se logran encontrar investigaciones similares realizadas respecto al sistema de transporte masivo de la ciudad de Santiago de Cali, es por esto que se plantea esta propuesta buscando identificar y analizar la forma en que opera el sistema MIO midiendo diferente variables como distancias de desplazamiento, números de transbordo, efectividad de los transbordos, distancia en las plataformas, entre otras y así poder proponer en futuras investigaciones actividades de mejora que permitan optimizar la red de transporte.

5. METODOLOGÍA.

- 1) Se realizó web scrapping en la página del sistema de transporte masivo MIO. Se logró identificar la fuente de información para la construcción de una tabla de datos con todas las rutas posibles del sistema, luego de esto se depuró la información y se calculó las distancias de haversine entre las paradas, estaciones y puntos de transbordo del sistema.
- 2) Se desarrolló un modelo de agrupación de paradas, estaciones y puntos de transbordo con clustering aglomerativo con la función de enlace complete en un rango máximo de 300 metros. Para esto se usó el paquete AgglomerativeClustering de scikit-Learn versión 1.4.2 donde se especifican los siguientes parámetros:

N_CLÚSTERS: Se refiere al máximo número de clúster a encontrar por lo que se deja como default: None

METRIC: Se refiere a la métrica a utilizar, dado que se va a usar la matriz de distancias de haversine se utiliza la opción precomputada

DISTANCE_THRESHOLD: La distancia máxima para los clústers, en este caso es de 300 metros o 0.3km

LINKAGE: complete

El algoritmo no supervisado de clustering Agglomerative, se puede entrenar con diferentes tipos de "linkage" o agrupamientos, en este caso se utilizó la distancia máxima, debido que se requiere agrupar conjuntos con una distancia máxima de 300 metros la cual es la distancia caminable para realizar transbordos en una ciudad como Cali.

Para el entrenamiento de este modelo se utilizó una matriz de distancias previamente calculada.

Dado que los puntos son georreferenciados de tipo latitud, longitud, se calculó la distancia entre cualquier par de ellos utilizando la distancia Haversine.

- 3) Se realizó la evaluación del modelo propuesto utilizando tres indicadores diferentes: el Coeficiente de Silhouette, el Índice de Calinski - Harabasz y el Índice Davies - Bouldin. Estos indicadores permitieron medir y analizar la eficacia y la calidad del modelo propuesto.
- 4) Se realizó el análisis de la red resultante mediante tres medidas de eficiencia distintas: centralidad, intermediación y cercanía. Estas medidas permitieron evaluar la calidad de la red actual y ofrecer comentarios sobre su estado actual, así como sugerencias para posibles mejoras.
- 5) Se realizaron conclusiones sobre los resultados obtenidos

6. RESULTADOS

A continuación, se presentan los resultados obtenidos a partir del análisis de datos relacionados con el sistema de transporte masivo MIO. En primer lugar, se detallan los datos recopilados sobre las rutas y paraderos del MIO, seguido de la clasificación de las plataformas de parada georreferenciadas. Posteriormente, se analiza la evaluación del modelo utilizado y se presentan los hallazgos clave obtenidos a través de la red de datos para el estudio del flujo de pasajeros.

6.1 Datos de las rutas y paraderos del sistema MIO

Se realizó web scrapping en la página <https://servicios.siuur.com.co/buscarutas> en la cual se encuentra toda la información de rutas del sistema de transporte masivo de la ciudad de Cali, se identificó que el Query que permitió obtener la información era a través del siguiente enlace <https://servicios.siuur.com.co/buscarutas/name.php>, usando librerías de Python como: beautifulsoup donde se obtuvieron como resultado los datos en formato json como se observa en la Figura 3. Estos datos fueron consolidados y sometidos a un proceso de extracción, transformación y carga (ETL) el cual permitió simplificar los datos por ruta y sus respectivas coordenadas de parada como se logra visualizar en la figura 4.

Unnamed: 0	ROWNUM	LINE	DESCRIPTION	STOPID	SHORTNAME	LONGNAME	DECIMALLATITUDE	DECIMALLONGITUDE	ORIENTATION	LINEVA	STOPCHECK	SENTIDO
0,0,1,A17E,Bochalema - Universidades,511505,C48K109,Cl 48 entre Kr 109 y 111,3.35701,-76.51638556,1,2,Normal Transit,Ida												
1,0,2,A17E,Bochalema - Universidades,511506,K112C46,Kr 112 entre Cl 46 y 44,3.35613889,-76.51771833,1,2,Normal Transit,Ida												
2,0,3,A17E,Bochalema - Universidades,511507,K112C44,Kr 112 entre Cl 44 y 43,3.35621028,-76.5192275,1,2,Normal Transit,Ida												
3,0,4,A17E,Bochalema - Universidades,511387,C42K112,Cl 42 entre Kr 112 y 111,3.35687,-76.52036333,1,2,Normal Transit,Ida												
4,0,5,A17E,Bochalema - Universidades,511386,C42K111,Cl 42 entre Kr 111 y 109,3.35841278,-76.52028583,1,2,Normal Transit,Ida												
5,0,6,A17E,Bochalema - Universidades,511409,K109C43,Kr 109 entre Cl 43 y 44,3.358595,-76.51839806,1,2,Normal Transit,Ida												
6,0,7,A17E,Bochalema - Universidades,511410,K109C46,Kr 109 entre Cl 46 y 48,3.35849194,-76.5171475,1,2,Normal Transit,Ida												
7,0,8,A17E,Bochalema - Universidades,511411,C49K109,Cl 49 entre Kr 109 y 111,3.3579575,-76.51495556,1,2,Normal Transit,Ida												
8,0,9,A17E,Bochalema - Universidades,511503,C49K111,Cl 49 entre Kr 111 y 112,3.35627028,-76.51500361,1,2,Normal Transit,Ida												
9,0,10,A17E,Bochalema - Universidades,511504,C48K111,Cl 48 entre Kr 111 y 109,3.35714361,-76.51591028,1,2,Normal Transit,Ida												
10,0,11,A17E,Bochalema - Universidades,511533,C48K101,Cl 48 entre Kr 101 y 99A,3.3664,-76.51623611,1,2,Normal Transit,Ida												
11,0,12,A17E,Bochalema - Universidades,511521,K99C481,Kr 99 entre Cl 48 y 45,3.36847111,-76.51739639,1,2,Normal Transit,Ida												
12,0,13,A17E,Bochalema - Universidades,511408,K99C451,Kr 99 con Cl 45,3.36846694,-76.51856556,1,2,Normal Transit,Ida												
13,0,14,A17E,Bochalema - Universidades,511100,K99C45,Kr 99 entre Cl 45 y 42,3.36847167,-76.51995694,1,2,Normal Transit,Ida												
14,0,15,A17E,Bochalema - Universidades,511226,K99C34,Kr 99 entre Cl 34 y 25,3.36762222,-76.52347167,1,2,Normal Transit,Ida												
15,0,16,A17E,Bochalema - Universidades,502312,UNIVCOM7,Universidades C4,3.36665083,-76.52821944,1,2,Normal Transit,Ida												
16,0,1,A17E,Bochalema - Universidades,502315,UNIVCO10,Universidades D4,3.3662175,-76.52801306,1,4,Normal Transit,Ida												
17,0,2,A17E,Bochalema - Universidades,511034,K99C25,Kr 99 entre Cl 25 y 34,3.36731444,-76.52330722,1,4,Normal Transit,Ida												
18,0,3,A17E,Bochalema - Universidades,511107,K99C42,Kr 99 entre Cl 42 y 45,3.36814056,-76.52006194,1,4,Normal Transit,Ida												
19,0,4,A17E,Bochalema - Universidades,511456,K99C48,Kr 99 entre Cl 45 y 48,3.36814667,-76.51731028,1,4,Normal Transit,Ida												

Fig. 3 Modelo de obtención de los datos en el webscrapping

	LINE	DECIMALLATITUDE	DECIMALLONGITUDE
0	A17E	3.357010	-76.516386
1	A17E	3.356139	-76.517718
2	A17E	3.356210	-76.519227
3	A17E	3.356870	-76.520363
4	A17E	3.358413	-76.520286

Fig. 4 Datos simplificados por rutas y coordenadas

Posteriormente se realizó una eliminación de duplicados para obtener todas las plataformas y paradas del sistema MIO obteniendo un total de 2111, de tal forma que quedó como se muestra en la Figura 5. Donde cada uno de los puntos azules representa un paradero, plataforma o una estación.

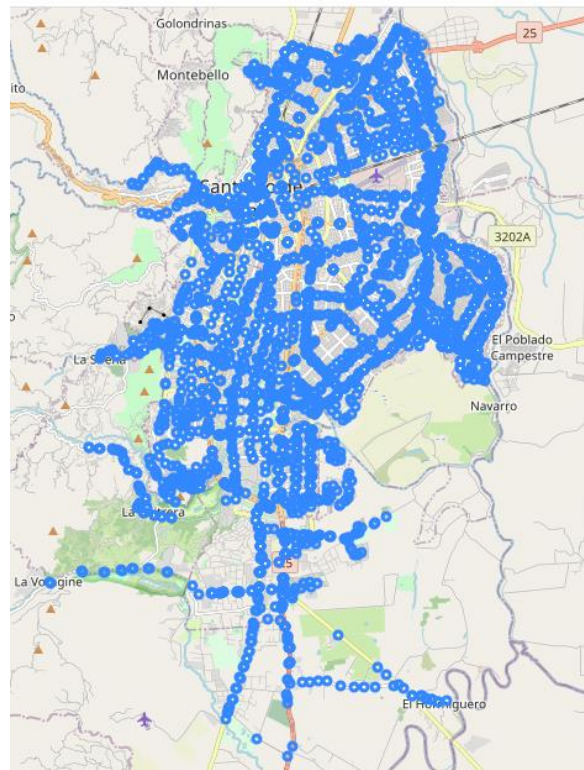


Fig. 5 Ubicación de plataformas y paraderos SITM, 2023

6.2 Clasificación de las plataformas de parada georreferenciadas

Una vez organizadas las 2111 paradas se agruparon en un rango de 300 metros con otros paraderos y estaciones, obteniendo como resultado 715 clústeres donde se tienen los puntos de mayor interés según las medidas de importancia, tal como se observa en la figura 6.

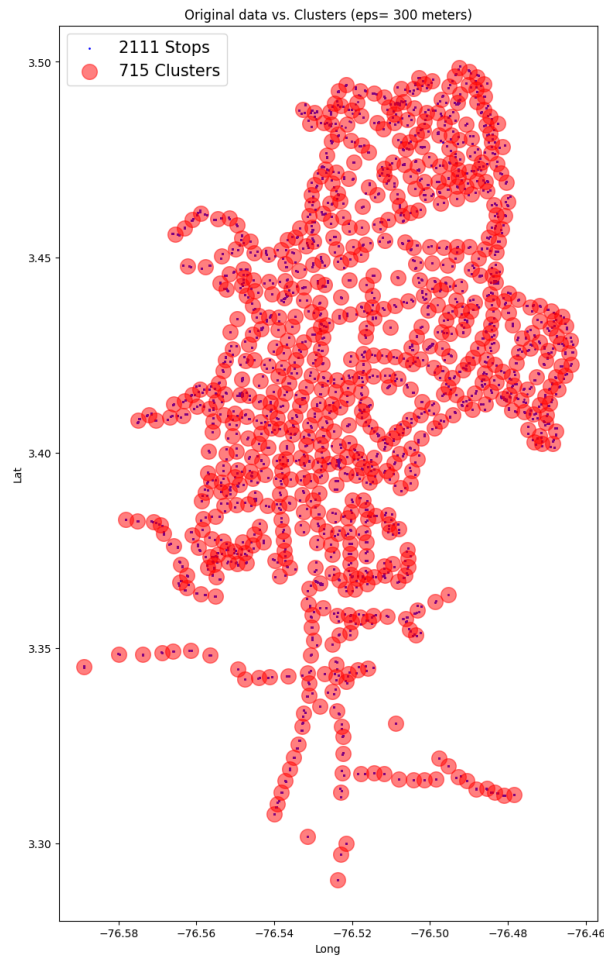


Fig. 6 Ubicación de plataformas y paraderos con clúster aglomerativo de 300 metros, 2023

Posteriormente se realiza medición de tamaño de los clústeres, donde si su tamaño equivale a 1 indica la presencia de una parada, si equivale 2 indica dos paradas y así sucesivamente, observando que la mayoría de los clústeres presentan un tamaño 2 seguidos por un tamaño 3 y 4 respectivamente, como se muestra en la figura 7.

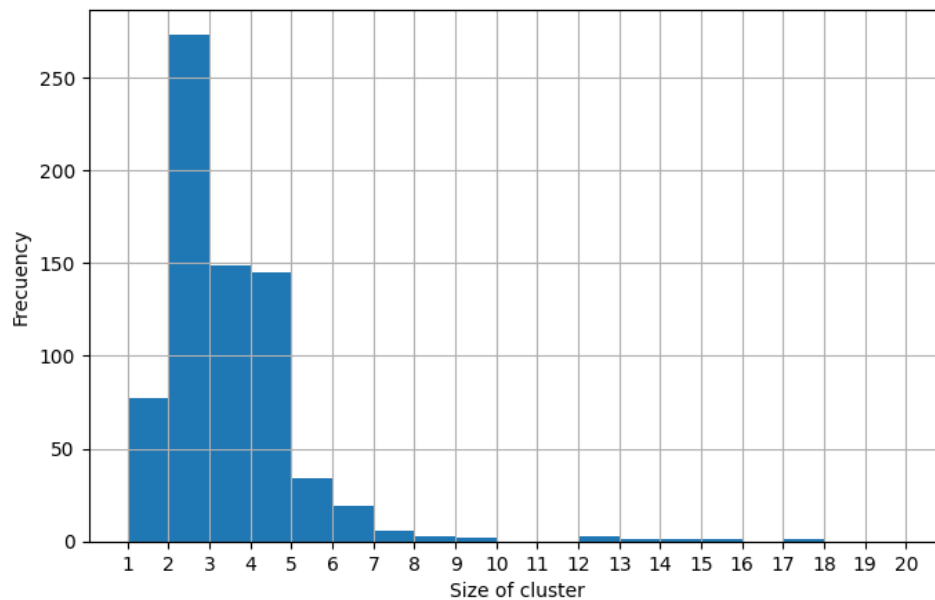


Fig. 7 Tamaño de los clúster

Se tienen como resultados preliminares que para realizar el transbordo más largo, se deben caminar 299.92 metros. En promedio, el máximo de metros que se debe caminar para realizar el transbordo más largo de cada clúster es de 145.54 metros y para realizar un transbordo entre cualquier cluster la distancia que se debe caminar es de 66.68 metros, en promedio.

6.3 Evaluación del modelo

El modelo se evaluó haciendo uso de diferentes indicadores que se describirán a continuación:

En primer lugar, se utilizó el indicador de Silhouette - Coefficient se obtuvo un resultado de **0.632** el cual se considera bueno; ya que este coeficiente debe tener un resultado entre -1 y 1. La puntuación está limitada entre -1 para agrupaciones incorrectas y 1 para agrupaciones muy densas. Las puntuaciones cercanas a cero indican grupos superpuestos. Un resultado de **0.63** significa que se obtuvieron grupos de plataformas muy cercanas, pero entre grupos que están relativamente alejados.

En segundo lugar, se aplicó el indicador Clinski – Harabasz con el que se obtuvo un resultado de **20001.2**. En esta evaluación la puntuación es mayor cuando los

conglomerados son densos y están bien separados, lo que se relaciona con el concepto estándar de conglomerado.

Para finalizar la evaluación del modelo se aplicó el indicador Davies – Bouldin con el cual se obtuvo un resultado de **0.46** este resultado se considera aceptable ya que este índice permite determinar la efectividad del clustering de acuerdo con la cercanía al valor cero.

Con todo lo anteriormente descrito se puede afirmar que el modelo utilizado para esta investigación es efectivo, ya que en los indicadores propuestos para la evaluación tiene un buen comportamiento

6.4 Red de datos para el estudio del flujo de pasajeros

Para la construcción de la RED, se usaron los datos generados en la fase de Clustering para la clasificación de plataformas. En esta fase se construyó una red bipartita entre el conjunto de rutas y el con conjunto de puntos de transbordo. Para graficar esta red fueron asignados atributos 0 a las rutas y atributo 1 a los puntos de transbordo quedando de la siguiente forma:

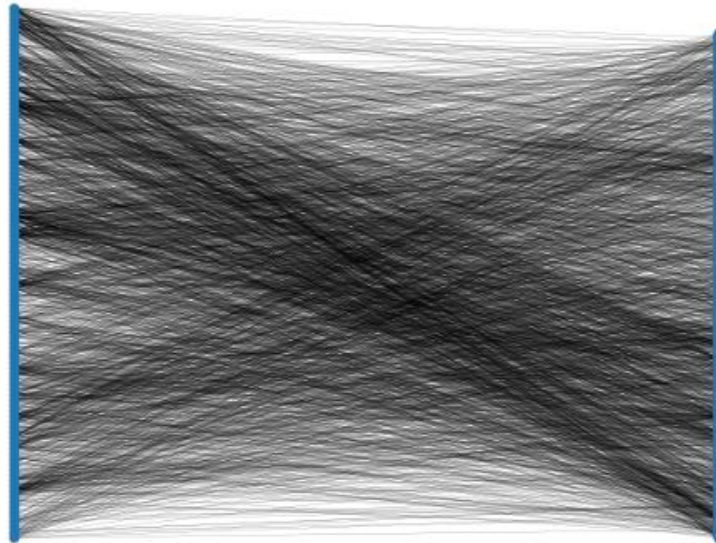


Fig. 8 Red bipartita entre conjunto de rutas y el conjunto de puntos de transbordo

En la figura 8 se observa al lado izquierdo el conjunto de nodos de las 101 rutas y a la derecha los 715 puntos de transbordo. Se logra evidenciar que la red se encuentra

totalmente conectada, lo cual permite que se pueda viajar de cualquier punto de la ciudad a otro realizando transbordos en los agrupamientos.

Para la red de puntos de transbordo se tiene una red donde los nodos son los puntos de transbordo dentro de la ciudad y el eje o link de un punto a otro, significa que se puede ir de un punto a otro en una misma ruta. Para la obtención de esta red se emplea el concepto de proyección de una red bipartita sobre uno de sus conjuntos, específicamente el conjunto de clusters. En este proceso, dos puntos del conjunto proyectado se conectan si existe al menos un punto del otro conjunto (rutas) que los conecte, es decir, si la ruta pasa por ambos puntos.

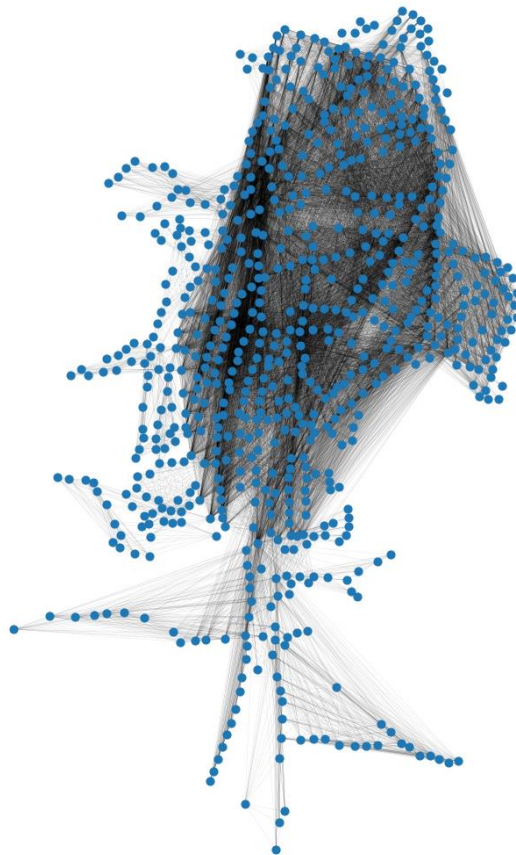


Fig. 9 Red de transbordos con una densidad de 0.084617

En la figura 9 se utilizó como atributo la posición del centroide. Se observa la conexión existente basada en la ubicación geográfica donde los nodos con mayor conexión se ven más saturados en líneas que los nodos con menos conexiones.

6.5 Centralidad

Una vez graficada esta red se procedió a medir la red de acuerdo a su centralidad identificando puntos críticos y/o influyentes dentro de la red, en escala de colores donde el color rojo indica la mayor influencia y el color azul indica la menor influencia en cada una de las mediciones.

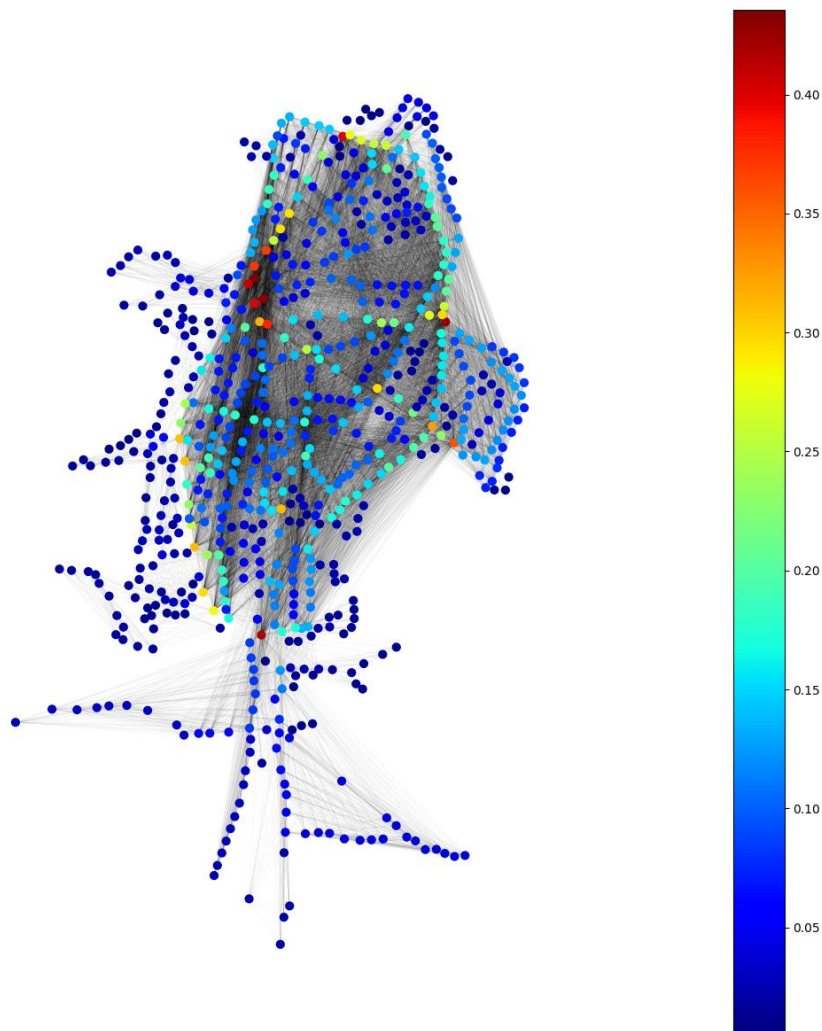


Fig. 10 Centralidad

En la figura 10 se evidencia las zonas con mayor concentración, se puede observar que hay pocas estaciones o puntos de parada de color rojo que muestran un alto índice de centralidad. Se observó mayor concentración de centralidad en la zona centro de la ciudad, seguido estaciones como Terminal Menga, Andrés Sanín, Nuevo latir mientras

que esta medida se encuentra en menor proporción hacia las zonas periféricas o residenciales de la ciudad.

6.5.1 Eigenvalor (centralidad de vector propio)

La medida de centralidad de valor propio evalúa la fuerza de una parada no solo en función de su propia conexión, sino también en función de la conexión de sus conexiones. Mide que tan bien está conectada una parada y que tan bien conectadas están sus conexiones. Por lo tanto, una parada puede ser importante debido a su conexión con paradas influyentes aunque en sí misma no sea influyente. La figura 11 permite observar la conexión de paradas más importantes dentro de la red.

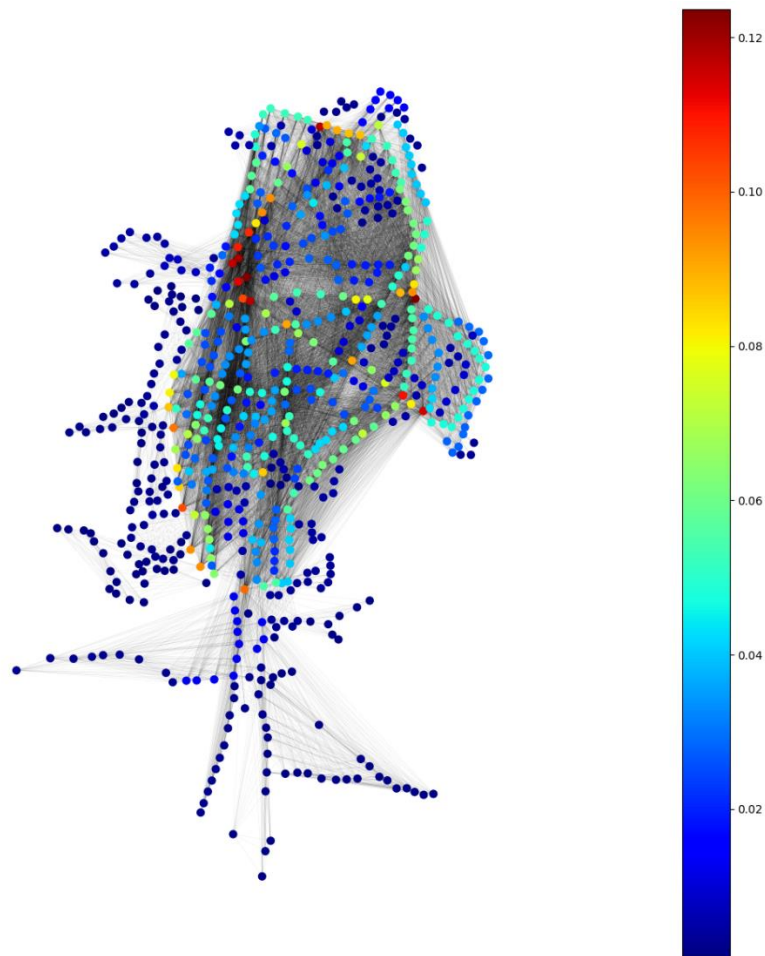


Fig. 11 Centralidad de vector propio: Eigenvalor

El valor propio más alto corresponde a la terminal Andrés Sanín, seguida de las terminales Universidades y Menga. También están la plataforma del centro de la ciudad. Para esta medición de igual forma se evidenció que las de mayor participación son las estaciones de la zona centro de la ciudad, quienes son de mayor relevancia en función de la conexión de sus conexiones.

6.6 Intermediación

Para la intermediación se identificaron las paradas de mayor conexión dentro de la red.

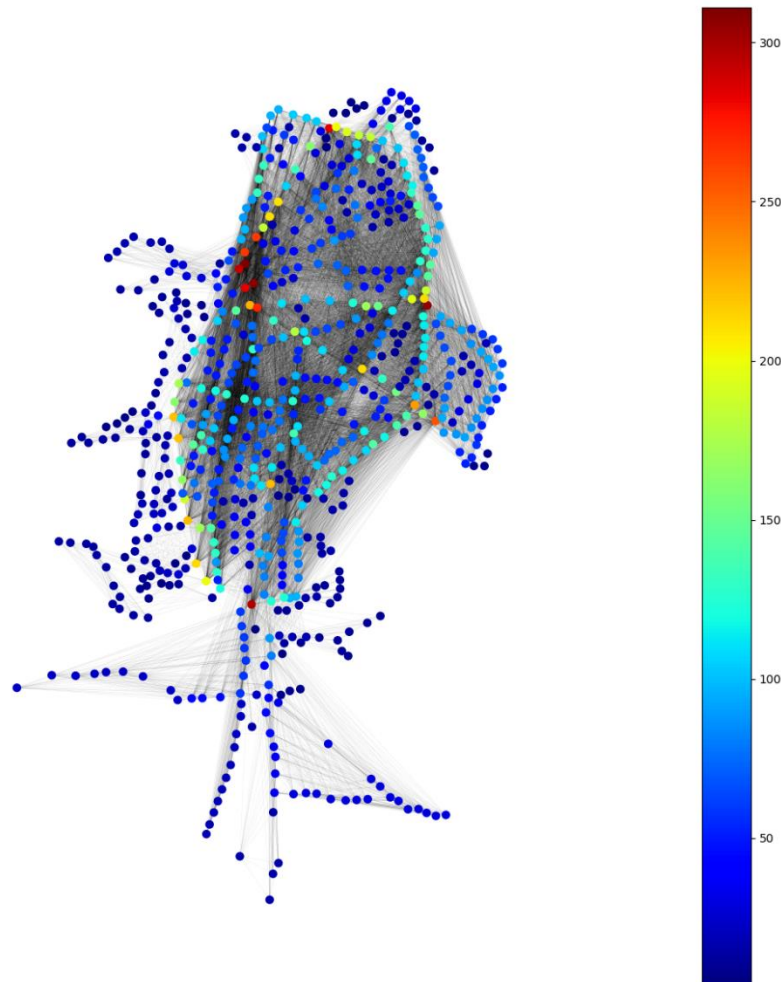


Fig. 12 Intermediación

Las paradas altamente necesarias para el traslado de pasajeros se encuentran en rojo, se puede ver que están principalmente en el centro de la ciudad por ser el eje de la red

mientras que se ve con menor relevancia en estaciones como universidades, Andrés Sanín y Menga en la periferia, al ser estaciones de inicio o fin dentro de la red.

6.7 Cercanía

En términos de cercanía se mide la capacidad que tiene una parada para alcanzar las otras paradas de la red. En la figura 13 se pueden observar las paradas con mayor índice en color rojo, seguidas de los puntos naranja y amarillo. Las de menos cercanía se encuentran en azul.

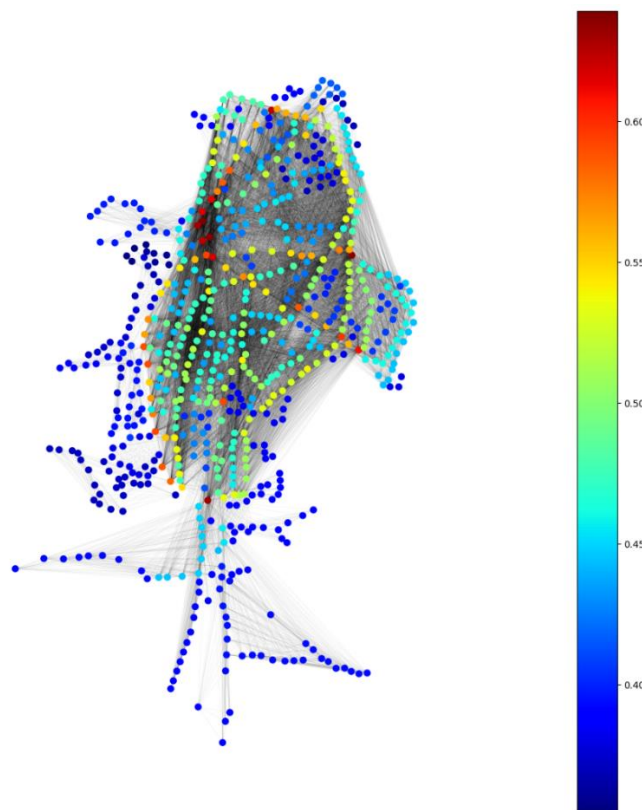


Fig. 13 Cercanía

Se puede observar en la figura 13 que hay una mayor cantidad de estaciones o paradas con un valor del índice alto que en anteriores índices, para este caso son más altas los valores por encima de 0.55. Este indicador permite concluir que en estas estaciones se está más cerca a todas las demás estaciones. Por ejemplo, en la estación San Pedro del centro se está más cerca de las demás estaciones porque tiene conexión directa a muchos más paraderos y estaciones que las demás, esto hace que sea más cercana en promedio.

Los valores de este indicador son mayores en la troncal de la calle quinta, la troncal del norte y la que conduce a la estación Andrés Sanfín.

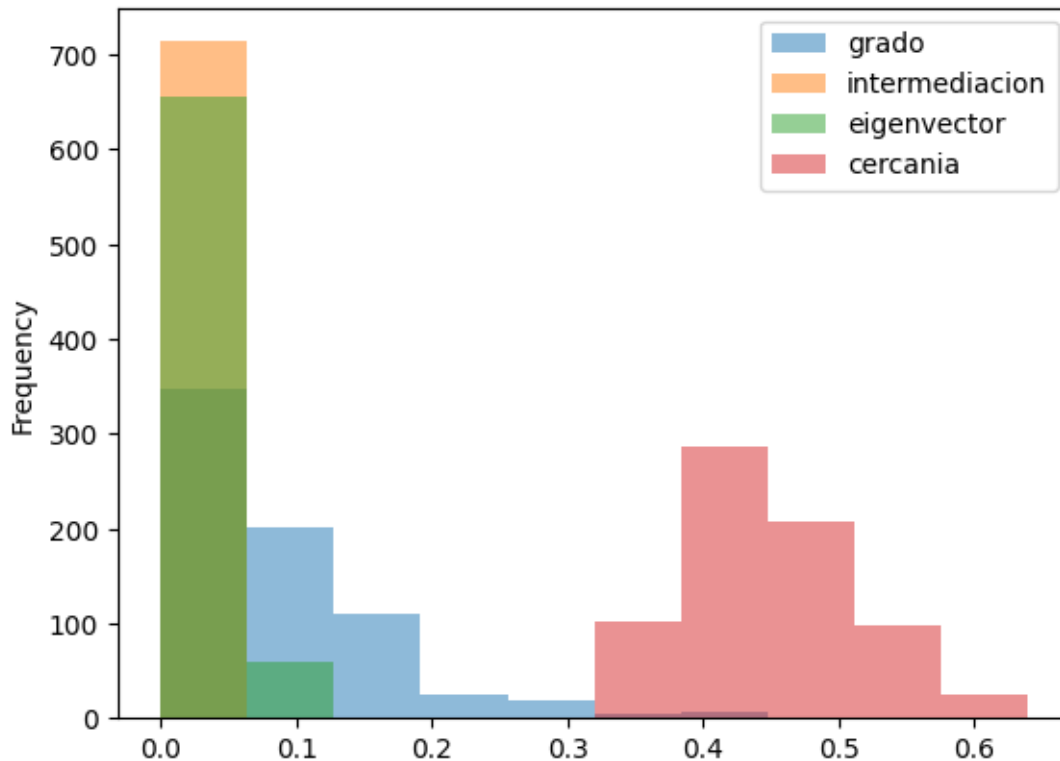


Fig. 14 Frecuencia de valores de los índices por tipo de índice

En la figura 14 se observa en el eje X el valor de los índices calculados, mientras que en el eje Y se observa el valor de la frecuencia de estos índices. Se puede concluir que los valores más altos se encuentran en el indicador de cercanía, seguido por el grado de centralidad y los valores más bajos se encuentran mayoritariamente en los índices de intermediación y de centralidad por eigenvector.

La figura 14 es un resumen de la investigación realizada. Se pueden ver los indicadores donde la red está mejor en la parte derecha y los que se deben mejorar en la parte izquierda.

Tras realizar los diversos análisis de las medidas de centralidad de la red se logra identificar que dentro de la red la zona de mayor significancia es la zona centro de la ciudad, seguido de terminales como Universidades, Menga, Andrés Sanín.

Se evidencia que en términos de cercanía (capacidad de alcanzar otras estaciones) las estaciones con mayor importancia son de la zona centro, Andrés Sanín, Menga y Universidades. En menor medida las estaciones de la Troncal de la Calle 5 y Corredor norte.

Por otra parte, las medidas calculadas permiten revelar que existen estaciones de gran tamaño como Calipso, Paso del comercio y Aguablanca que no poseen gran aporte a la red de transporte en ninguno de los indicadores calculados, lo que puede dar idea de que existe un contraste en las inversiones realizadas en infraestructura contra las pequeñas estaciones del centro que tienen una gran significancia dentro de la red pero con un menor tamaño y probablemente menor inversión.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 Conclusiones

- Este proyecto ha proporcionado una comprensión detallada de los patrones de transbordo en el sistema MIO. Esto incluye la identificación de rutas y puntos de transbordo clave, lo que permitirá a los administradores del sistema de transporte optimizar las rutas y mejorar la infraestructura en áreas críticas.
- En el desarrollo de este proyecto se ha obtenido información valiosa sobre cómo los pasajeros utilizan el sistema de transbordo, variable determinante para lograr entender las necesidades de movilidad de los usuarios y diseñar estrategias para satisfacer esas necesidades.
- Se logró evaluar de la eficiencia de la red del sistema de transporte masivo (MIO), proporcionando datos concretos sobre su rendimiento actual. Este análisis permitirá que las entidades locales puedan tomar decisiones sobre cómo mejorar la eficiencia operativa y la satisfacción del usuario.
- El estudio ha demostrado la utilidad de técnicas avanzadas de análisis de redes para comprender los sistemas de transporte urbano complejos. Esto resalta la importancia de incorporar enfoques interdisciplinarios y técnicas modernas para abordar los desafíos de movilidad urbana.
- Este estudio permitió identificar que el tamaño de la infraestructura de las estaciones o terminales del Sistema Integrado de Transporte Masivo (SITM) con respecto a la cobertura no es determinante dentro de la red evidenciada que estaciones de gran tamaño como Terminal Calipso, Terminal paso de comercio, entre otras. No son relevantes dentro de la red de transporte en comparación de pequeñas estaciones como las del centro de la ciudad.

7.2 Trabajos futuros

- La metodología y los datos obtenidos pueden ser utilizados como punto de partida para futuras investigaciones y para desarrollar planes de mejora a largo plazo en el sistema de transporte masivo de Cali.
- Utilizando los datos recopilados, se pueden desarrollar algoritmos de optimización para las rutas y los puntos de transbordo. Esto podría beneficiar el sistema al minimizar la congestión y reducir los tiempos de viaje para los pasajeros.
- Utilizar herramientas de simulación para crear escenarios futuros y evaluar cómo diferentes decisiones de planificación podrían afectar el sistema de transporte masivo y los patrones de transbordo.

8. REFERENCIAS

- [1] M. d. transporte, «SISTEMA INTEGRADO DE TRANSPORTE MASIVO PARA SANTIAGO DE CALI – MÍO (MASIVO INTEGRADO DE OCCIDENTE),» Departamento Nacional de Planeación, Bogotá, 2009.
- [2] E. País, «Usuarios del MÍO se quejan por demora en rutas y congestión en estaciones,» Noviembre 2011. [En línea]. Available: <https://www.elpais.com.co/california/usuarios-del-mio-se-quejan-por-demora-en-rutas-y-congestion-en-estaciones.html>. [Último acceso: 16 Abril 2023].
- [3] A. d. S. d. Cali, «Alcaldía de Santiago de Cali,» febrero 2011. [En línea]. Available: https://www.cali.gov.co/alcaldenlinea/publicaciones/37443/ms_estaciones_y_vehiculos_para_descongestionar_el_mio/. [Último acceso: 16 abril 2023].
- [4] J. A. Minota, «¿Hacia dónde va el MIO?: 15 años de funcionamiento del sistema de transporte caleño,» *El tiempo*, pp. <https://www.eltiempo.com/colombia/california/hacia-donde-va-el-mio-en-sus-15-anos-de-funcionamiento-en-cali-3325483>, 18 marzo 2024.
- [5] D. Hernandez, «Transporte público, bienestar y desigualdad: cobertura y capacidad de pago en la ciudad de Montevideo,» *Revista de la CEPAL*, nº 122, pp. 165-184, 2017.
- [6] C. O'Reilly, «Practical Web Scraping for data Science: Best Practices and Examples with Python,» 2013.
- [7] B. Zhao, «Web Scraping,» *Springer International Publishing AG*, pp. 1 - 3, 2017.
- [8] E. Alpaydin, *Introduction to Machine Learning*, Second edition, 2010.
- [9] S. T. Wierzchoń y M. A. Kłopotek, *Modern Algorithms of Cluster Analysis*, Warsaw, Polonia: Springer, 2018.
- [10] I. Diaz Ariel y R. J. Coppo, «Calculadora geoespacial,» de *Jornadas Argentinas de Informática e Investigación Operativa*, JAIIO 40, 2012.
- [11] N. Newman, *Networks: An introduction*, 2010.
- [12] S. Wasserman y K. Fausts, *Social Network Analysis: Methods and Applications.*, 1994.
- [13] NetworkX Developers, «NetworkX,» degree centrality — NetworkX 3.3 documentation, [En línea]. Available: <https://networkx.org/documentation/stable/reference/algorithms/generated/networkx>

- .algorithms centrality.degree centrality.html#networkx.algorithms centrality.degree centrality. [Último acceso: 2024].
- [14] U. Brandes, «On variants of shortest-path betweenness centrality and their generic computation,» *Social Networks*, vol. 30, nº 2, pp. 136-145, mayo 2008.
- [15] L. C. Freeman, «Centrality in social networks conceptual clarification,» *Social Networks*, vol. 1, nº 3, pp. 215-239, ene 1978.
- [16] P. Bonacich, «Power and Centrality: A Family of Measures,» *American Journal of Sociology*, vol. 92, nº 5, pp. 1170-1182, 1987.
- [17] M. Newman, *Networks*, Michigan: Oxford university Press, 2018.
- [18] P. J. Rousseeuw, «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,» *Journal Of Computational And Applied Mathematics*, vol. 20, pp. 53-65, nov 1987.
- [19] T. C. y. J. Harabasz, «A dendrite method for cluster analysis,» *Communications In Statistics. Theory And Methods/Communications In Statistics, Theory And Method*, vol. 3, nº 1, pp. 1-27, 1974.
- [20] D. D. y. D. W. Bouldin, «A cluster separation measure,» *EEE Transactions On Pattern Analysis And Machine Intelligence*, Vols. %1 de %2PAMI-1, nº 2, pp. 224-227, 1979.
- [21] J. Guitierrez, C. Benitez, j. Leañó, j. García, A. Condeco, C. Mojica, L. -schill, V. Adler, F. Vera, B. Moya y g. Romanillos, «BANCO INTERAMERICANO DE DESARROLLO,» Noviembre 2017. [En línea]. Available: <https://publications.iadb.org/es/como-aplicar-big-data-en-la-planificacion-del-transporte-urbano-el-uso-de-datos-de-telefonía-movil>. [Último acceso: 13 Abril 2023].
- [22] C. Fox, *Data Science for transport: A Self-Study Guide With Computer Exercises*, Reino unido: Springer International Publishing AG, 2018.
- [23] M. Maghrour, M. Atiullah, D. Esztergár y A. Torok, «A data-driven decision support tool for public transport service analysis and provision,» *Transport Policy*, nº 135, pp. 82-90, 2023.
- [24] D. N. d. Planeación, «¿En qué invierte el Estado Colombiano? Los grandes proyectos de inversión del Estado Comunitario en 2008,» Ministerio de transporte, Bogotá, 2009.
- [25] V. I. n. cali, «Twitter,» @Vivalasnoticias, Febrero 2019. [En línea]. Available: <https://twitter.com/METROCALI/status/1092403992192344064>. [Último acceso: 16 Abril 2023].

- [26] E. País, «El país,» Noviembre 2011. [En línea]. Available: <https://www.elpais.com.co/cali/usuarios-del-mio-se-quejan-por-demora-en-rutas-y-congestion-en-estaciones.html>. [Último acceso: 16 Abril 2023].
- [27] S. Ambrus, «BANCO INTERAMERICANO DE DESARROLLO,» 21 Septiembre 2017. [En línea]. Available: <https://blogs.iadb.org/ideas-que-cuentan/es/un-sistema-de-transporte-masivo-que-mejora-la-productividad/>. [Último acceso: 21 Abril 2023].
- [28] L. Perdomo, «Portafolio,» 22 Julio 2022. [En línea]. Available: <https://www.portafolio.co/economia/finanzas/como-les-ha-ido-a-los-sistemas-de-transporte-masivo-en-el-ultimo-ano-568512>. [Último acceso: 2 Mayo 2023].
- [29] F. Campo, F. Abarca y G. Serra, «Evaluación comparativa del nivel de desarrollo orientado al Transporte (dot) en torno a nodos de transporte de grandes ciudades: métodos complementarios de ayuda a la decisión,» *EURE (Santiago)*, vol. 45, nº 134, pp. 5 - 29, 2019.