UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

REVISTA UD Y LA GEOMÁTICA
https://revistas.udistrital.edu.co/index.php/UDGeo/index

UD y la GEOMÁTICA

INVESTIGACIÓN

Determinación de patrones de evolución de la congestión vial mediante el uso de técnicas de minería de datos espacial

Determination of patterns of evolution of road congestion using spatial data mining techniques.

Alexander Duvan Robles Mondragón[1]

Para citar este artículo: Robles-Mondragon, A. D. (2020). Determinación de patrones de evolución de la congestión vial mediante el uso de técnicas de minería de datos espacial. *UD y la Geomática*, (15), 5-15.
DOI: https://doi.org/10.14483/23448407.15250

Fecha de envio: 15 de julio de 2019 **Fecha de aceptación:** 12 de diciembre de 2019

RESUMEN

La congestión vial en carreteras urbanas tiene dos categorías: congestión recurrente (RC) y congestión no recurrente (NRC). La NRC es irregular, lo que significa que puede aparecer en cualquier momento y en cualquier lugar, y generalmente es causado por accidentes de tráfico, daños en la vía, control de tráfico temporal y otros eventos accidentales; por otro lado, la RC se produce con mayor frecuencia que la NRC, por lo general en un sitio, camino o área fijos durante las horas pico de la mañana o la tarde, y es causada por una alta demanda de tráfico, capacidad de tráfico insuficiente, señalización deficiente, infraestructura de tráfico inferior u otras condiciones relacionadas. Para la identificación de RC se utilizan procesos de minería de datos junto al algoritmo de *culstering* DBSCAN. Con esto se buscó identificar un patrón evolutivo, de tal forma que permitiera evaluar la movilidad y aquellos puntos de congestión sobre una vía específica de la ciudad de Bogotá.

Palabras clave: tráfico, congestión, patrón, recurrente, no recurrente, velocidad media, movilidad.

ABSTRACT

Road congestion on urban roads has two categories: Recurrent congestion (CR) and non-recurrent congestion (NRC), The NRC is irregular, which means it can appear at any time and anywhere, and is usually caused by traffic accidents , damage to the road, temporary traffic control and other accidental events, on the other hand, CR occurs more frequently than the NRC, generally in a fixed place, road or area during the morning or afternoon peak hours, and it is usually caused by high traffic demand, insufficient traffic capacity, poor signaling, lower traffic infrastructure or other related conditions. For the identification of RC, data mining processes are used together with the DBSCAN culstering algorithm, with this purpose it was sought to identify an evolutionary pattern, in such a way that it would allow to evaluate the mobility and those points of congestion on a specific road of the city of Bogotá.

Keywords: Traffic, congestion, pattern, recurrent, non-recurrent, average speed, mobility.

RESUMO

congestionamento do tráfego em vias urbanas tem duas categorias: o congestionamento recorrente (RC) e congestão não recorrente (NRC), o NRC é irregular, o que significa que ele pode aparecer a qualquer momento e em qualquer lugar, e é geralmente causada por acidentes de trânsito , danos à estrada, controle de tráfego temporário e outros eventos acidentais, por outro lado, CR ocorre com mais freqüência do que o NRC, geralmente em um lugar fixo, estrada ou área durante o horário de pico da manhã ou

1 Universidad del Rosario - Eninco S.A.S. Bogotá, Colombia. duvan.robles@urosario.edu.co - dalexrobles@gmail.com.

da tarde, e geralmente é causada por alta demanda de tráfego, capacidade de tráfego insuficiente, sinalização ruim, infra-estrutura de tráfego mais baixa ou outras condições relacionadas. Para identificação de RC processos de mineração de dados são usados pelo algoritmo culstering DBSCAN com este buscou-se identificar um padrão evolutivo, de modo que iria avaliar a mobilidade e esses gargalos em uma faixa especificada em Bogotá.

Palavras-chave: Tráfego, congestionamento, padrão, recorrente, não recorrente, velocidade média, mobilidade.

Introducción

La problemática causada por el crecimiento desordenado de las ciudades es típicamente afrontada por las entidades a cargo de la planeación de manera sectorial, lo que ha conducido a desarticulaciones funcionales de la vida urbana (Palau, s.f.), razón por la cual es necesario coordinar y ensamblar los procesos de cada sector, enfocándose hacia el mismo modelo de ciudad, estableciendo estrategias políticas y proyectos que obedezcan a inversiones globales del desarrollo y crecimiento de ciudad (Secretaria Distrital de Movilidad, 2018).

La congestión en redes viales representa uno de los mayores desafíos pues se comporta de formas distintas dependiendo la localización espacial, por lo que se han establecido divisiones por comportamiento y causa: el tráfico no recurrente se enfoca en incidentes viales que provocan congestión y el tráfico recurrente hace referencia a patrones regulares que generalmente ocurren en un sitio especifico, camino o área fija durante un periodo de tiempo que corresponde a horas pico; es común su ocurrencia debido a la alta demanda de transporte, capacidad de tráfico insuficiente o pobre infraestructura, entre otras. El tráfico recurrente afecta la ciudad en la escala regional por lo que se espera un patrón evolutivo; entonces identificar el patrón evolutivo permite diseñar estrategias de mejoramiento del tráfico que beneficien a la sociedad.

Como metodología para el estudio y solución del problema planteado se escogió el proceso de extracción del conocimiento *knowledge discovery in databases* (KDD), que implican las etapas de selección de datos, preprocesamiento y limpieza, minería de datos (determinación de patrones mediante algoritmos), interpretación y evaluación de resultados para llegar al conocimiento oculto en los datos.

Se aplicaron las tareas de minería de datos correspondientes al preprocesamiento de los datos, y para minería se consideró un algoritmo de agrupación que permitió la identificación de grupos representativos y no representativos, como fue el algoritmo DBSCAN, en la búsqueda por determinar los patrones evolutivos de congestión vial en el tramo estudiado como solución al problema planteado.

En la sección "Problema" se describe la congestión, las subdivisiones dentro de la misma y la complejidad de identificación, se especifica el marco espacial y el corredor vial a ser analizado.

En la "Recopilación e integración de los datos" se describen las tareas de minería de datos en la etapa de pre proceso, como son: tratamiento de datos atípicos, datos faltantes, y discretización de variables; labores necesarias para la preparación de los datos.

En "DBSCAN" se definió el algoritmo de agrupación seleccionado para encontrar la solución al problema planteado, en este caso DBSCAN se usó para detección de valores anómalos y la correspondiente agrupación de los datos.

Finalmente, "Interpretación y evaluación de resultados" detalla la identificación de los grupos más y menos representativos y sus respectivos porcentajes de composición.

Problema

La congestión que se produce en las redes de carreteras urbanas se puede dividir en dos categorías: congestión recurrente (RC, por su sigla en inglés) y congestión no recurrente (NRC, por su sigla en inglés) (Anbaroglu, Heydecker y Cheng, 2014; Luo, Hadiuzzaman, Fang y Qiu, 2015).

La NRC es irregular, lo que significa que puede aparecer en cualquier momento y lugar, y por lo general es causado por accidentes de tráfico, averías en la carretera, control de tráfico temporal y otros eventos accidentales (Gordon, 2015). La RC se produce con mayor frecuencia que la NRC, en su mayoría en un sitio, camino o área fijos durante las horas pico de la mañana o la tarde, y es causada por una alta demanda de tráfico, capacidad de tráfico insuficiente, control de señal deficiente, infraestructura de tráfico inferior u otras condiciones relacionadas (Ozbay y Ma, 2015).

En una red de carreteras urbanas, RC comienza en un punto específico (es decir, una intersección o un punto en el enlace de la carretera), luego se propaga a lo largo de una dirección y una ruta fijas. En última instancia, afecta el tráfico en una escala regional (Ji, Luo y Geroliminis, 2014; Ma, Yu, Wang y Wang, 2015). Sobre la base de estas características, la RC urbana muestra un patrón de evolución fijo: si se puede identificar el patrón de evolución de la congestión recurrente (RCEP, por su sigla en inglés), muchas partes interesadas del tráfico urbano podrían beneficiarse.

El diseño de soluciones de movilidad, apoyado en técnicas de análisis de datos que permiten determinar patrones en zonas urbanas que no son claramente identificables con análisis convencionales, plantea un nuevo escenario de decisiones apoyadas en minería de datos con posibilidad de generar algoritmos inteligentes que dinámicamente pueden dar un acercamiento a un escenario de movilidad lenta en un sector, según la dinámica de variables en el tiempo, por ejemplo: accidentes, protestas, eventos públicos, cierres viales, estado de vía, ocupación de parqueaderos, entre otros.

Marco espacial

Para este estudio se tomaron los datos de velocidad sobre el tramo vial definido sobre la avenida Boyacá entre la calle 170 y la diagonal 78 BIS S. Estos fueron compilados en el contrato de consultoría No. 1268-2016, cuya finalidad era la toma de información en vías principales de la ciudad, en aras del seguimiento y planeación del tránsito y el transporte en la ciudad de Bogotá. Esta información sin tratamiento fue publicada y es de libre acceso.

El corredor vial definido para este estudio fue de aproximadamente 32,47 km, conformado por cuatro calzadas de tráfico mixto y el recorrido se controló en 10 tramos. Esta vía es muy importante para la ciudad, pues permite evaluar la circulación vehicular en sentido norte-sur y viceversa. La figura 1 muestra los tramos escogidos para la manipulación de datos.

Preparación de datos

Para un ejercicio de minería de datos, la preparación de estos puede ser una de las actividades que consume más tiempo, adicionalmente es crítica, pues la calidad de los datos acumulados y su preparación son un factor influenciador en el éxito de los análisis finales. El propósito fundamental de la preparación es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o ser más accesible (Pyle, 1999).

Recopilación e integración de los datos

En la etapa de selección, una vez identificado el conocimiento relevante y prioritario, y definidas las metas del proceso KDD desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, del cual se toman todos o solo una muestra representativa, a partir de los cuales se realiza el descubrimiento. La selección varía según los objetivos (Pereira Timarán *et al.*, 2016).

Hay varias formas de capturar información relacionada con las redes de transporte, entre ellas se encuentran los sistemas de tráfico activo que controlan una variable clave que es la del límite de velocidad. Es usada por las entidades estatales para mejorar el tráfico y reducir los choques de automotores. Estos datos alimentados a un detector se clasifican en dos categorías:

- Captura fija: conteos, ocupación y velocidades medidas por sensores inductivos, magnéticos, microondas, infrarrojo, ultrasonido acústico, láser o procesadores de video.
- Datos de vehículo: trayectoria de vehículo muestreada a partir de GPS, geolocalización satelital o identificación automática.

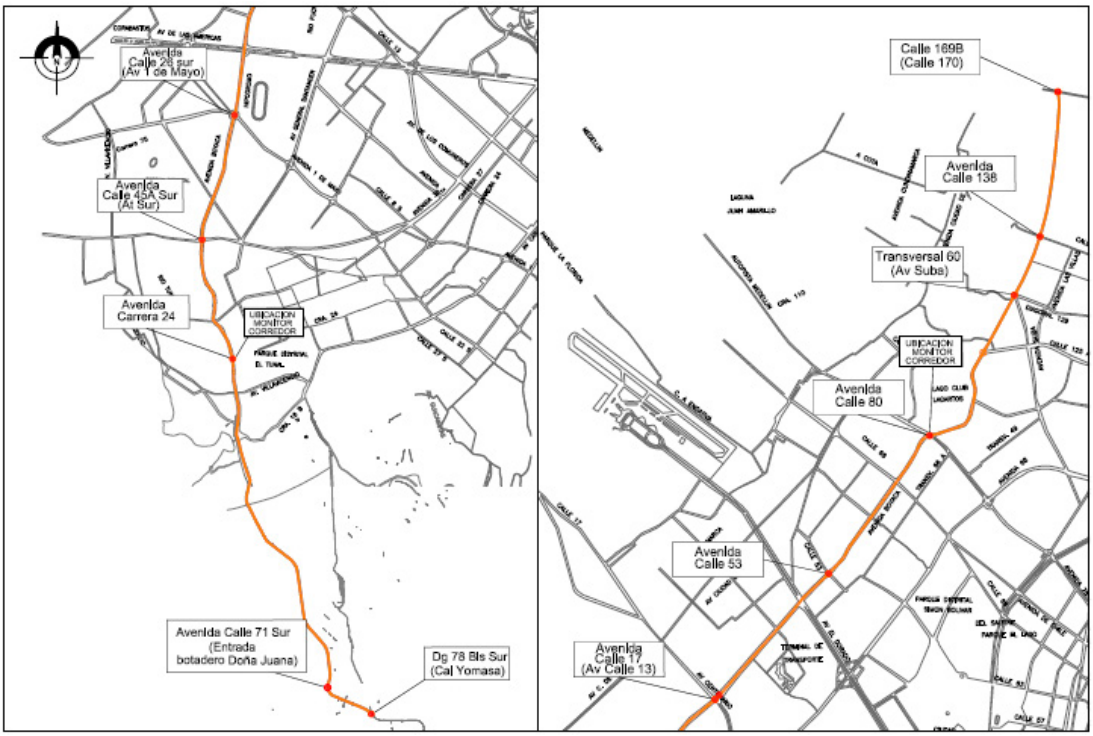

Figura 1. Tramos de estudio para toma de datos,

Fuente: (Secretaria de Movilidad, 2017)

En comparación con los detectores de tráfico tradicionales (por ejemplo, detectores de bucle, cámaras de tráfico, sensores de microondas de transporte remoto), la trayectoria de vehículo muestreada con GPS o el automóvil flotante tiene varias ventajas obvias: menor costo, mayor cobertura y mayor movilidad. Las técnicas basadas en datos se han desarrollado rápidamente y ahora se utilizan en numerosos sectores industriales (Yin, Li, Gao y Kaynak, 2015; Yin, Wang y Gao, 2016). Los datos manejados contemplan varios tipos de automóviles viajando alrededor de una red vial de la ciudad en un ciclo de 15 horas, incluidos taxis, autobuses y particulares. Estos automóviles flotantes equipados con GPS pueden recopilar y generar datos de movilidad periódicamente (cada 30 segundos, o cada 15 segundos, o 1 minuto), incluidos longitud, latitud, velocidad, títulos de los vehículos y marcas de tiempo (An *et al.*, 2016).

Se han acumulado grandes cantidades de datos de movilidad que podrían explotarse para caracterizar las condiciones de tráfico típicas por ciudad. En la última década, se ha realizado una amplia variedad de investigaciones de transporte basadas en datos de movilidad de automóviles flotantes, por ejemplo, dividiendo las horas de operación del autobús en intervalos de hora del día basados en datos de GPS del autobús (Bie, Gong y Liu, 2015), extrayendo la topología de un público red de transporte basada en datos de vehículos operativos públicos específicos (Fiori, Mignone y Rospo, 2016), construyendo un sistema de recomendación de taxi basado en datos GPS de taxi (Hwang, Hsueh y Chen, 2015) y prediciendo el tiempo de viaje de la ruta y el enlace según datos de taxis flotantes (Tulic, Bauer y Scherrer, 2015). La investigación relacionada con datos de movilidad de vehículos equipados con GPS es un tema relevante en el campo de los sistemas de transporte inteligente (ITS, por su sigla en inglés).

Para este estudio se seleccionaron los datos tomados por: 2 vehículos de tipo particular, 2 de transporte colectivo y 2 de transporte público individual, para los cuales se hicieron mediciones de tiempos de recorrido por el método de vehículo en movimiento, mediante GPS en los 10 tramos, donde se analizó la velocidad media de recorrido en transporte público colectivo (TPC), transporte público individual (TPI) y transporte particular (TP). La información sin tratamiento consta de 146.053 registros, los cuales se registraron durante 3 periodos de tiempo el martes 31 de enero de 2017; dichos periodos se detallan en la tabla 1.

Por lo general, los datos reales nunca se trabajan directamente, por lo que es usual encontrar inconsistencias, valores faltantes o datos erróneos, por lo que su tratamiento representa un reto en el análisis y exploración inicial, o en su preprocesamiento. La tabla 2 muestra los atributos correspondientes al conjunto de datos del estudio.

Tabla 1. Periodos de tiempo para la toma de datos

PERIODOS	HORA	PERIODO
HORARIO DEL PERIODO No. 1	6:00 – 9:00	AM
HORARIO DEL PERIODO No. 2	11:00 – 14:00	M
HORARIO DEL PERIODO No. 3	16:00 – 20:00	PM

Fuente: elaboración propia.

Tabla 2. Atributos del conjunto de datos del estudio

No.	ATRIBUTO	TIPO DE DATO
1	HORA PASO	Texto
2	VELOCIDAD KM/H	Numérico
3	TIEMPO DE ESPERA EN SEGUNDOS	Numérico
4	ALTURA (M)	Numérico
5	DISTANCIAS ACUMULADAS ENTRE INTERVALOS Y SUMATORIA DE DISTANCIA TOTAL ENTRE PUNTOS DE INTERES	Numérico
6	CONTROL	Texto
7	GPS	Numérico
8	LATITUD	Numérico (sexagesimal)
9	LONGITUD	Numérico (sexagesimal)
10	LATITUD	Numérico
11	LONGITUD	Numérico

Fuente: elaboración propia.

Tareas de preprocesamiento

La minería de datos forma parte de un proceso denominado *descubrimiento de conocimientos de KDD* en bases de datos (Han, Kamber y Pei, 2012). Hay una serie de técnicas de preprocesamiento –limpieza de datos, por ejemplo– que se pueden aplicar para eliminar el ruido y corregir inconsistencias, valores atípicos y valores perdidos. Los datos sin procesar a menudo se encuentran en formas inadecuadas para su procesamiento. Las tareas principales en el preprocesamiento se describen a continuación.

Limpieza de datos

En esta fase, aquellas entradas o registros faltantes, erróneos e incoherentes se eliminan de los datos. Es posible, una vez conocido el comportamiento de los datos, estimar los faltantes mediante cálculo. En este caso, para analizar los patrones de evolución del tráfico recurrente se realizaron varias tareas las cuales tienen que ver con la estructura del formato de datos crudos. Para el tema de interés aquí estudiado, las labores adelantas se exponen seguidamente.

Los datos crudos cuentan con promedios por tramo de control, haciendo que el atributo "Hora paso" contenga valores de tipo texto sin formato de tipo tiempo (hora), como se esperaría en la mayoría del conjunto de datos. Esta información de promedio por tramo se considera irrelevante, por lo cual se procede a eliminar. La modificación evidencia la presencia de datos repetidos o idénticos que no le aportan información al ejercicio, por lo que también se procede a su limpieza. Dicho ajuste en el conjunto de datos representa una reducción de 15.140 registros, dejando en el grupo original un total de 130.913 datos consistentes.

Reducción de características

En el conjunto de datos se encontró que no todas las columnas brindaban información importante o se encontraba repetida en un formato diferente, como fue el caso de la información geográfica (latitud y longitud), que se encontraba en formatos sexagesimal (x°x'x") y decimal. Para un adecuado análisis, los datos se suprimieron del conjunto. Otras variables que fueron suprimidas por no aportar información fueron "Control" y "GPS". En la tabla 3 se resaltan las variables suprimidas en el set de datos.

Datos atípicos

Para identificar los datos atípicos se procedió a referenciar la información geográfica; en este caso, los campos de latitud y longitud de tipo decimal fueron tomados para esta labor. En la figura 2 se observan zonas de puntos capturados aislados o fuera de la zona de estudio. Estos datos atípicos identificados en el componente espacial no se eliminaron en esta etapa. Se dejaron como insumo para ser tratados como ruido mediante el algoritmo DBSCAN.

Discretización de datos

La conversión de datos numéricos a datos categóricos es conocida como *discretización*, divide los datos numéricos en N rangos y les asigna un valor simbólico. Es posible que la discretización pierda cierta información, pero no representa un problema demasiado grande; por el contrario, es muy útil cuando el conjunto de datos es muy grande. Para este estudio se tomaron las mediciones y asignaron 3

Tabla 3. Variables suprimidas en el set de datos

No.	ATRIBUTO	TIPO DE DATO
1	HORA PASO	Texto
2	VELOCIDAD KM/H	Numérico
3	TIEMPO DE ESPERA EN SEGUNDOS	Numérico
4	ALTURA (M)	Numérico
5	DISTANCIAS ACUMULADAS ENTRE INTERVALOS Y SUMATORIA DE DISTANCIA TOTAL ENTRE PUNTOS DE INTERES	Numérico
6	CONTROL	Texto
7	GPS	Numérico
8	LATITUD	Numérico (sexagesimal)
9	LONGITUD	Numérico (sexagesimal)
10	LATITUD	Numérico
11	LONGITUD	Numérico

Fuente: elaboración propia.

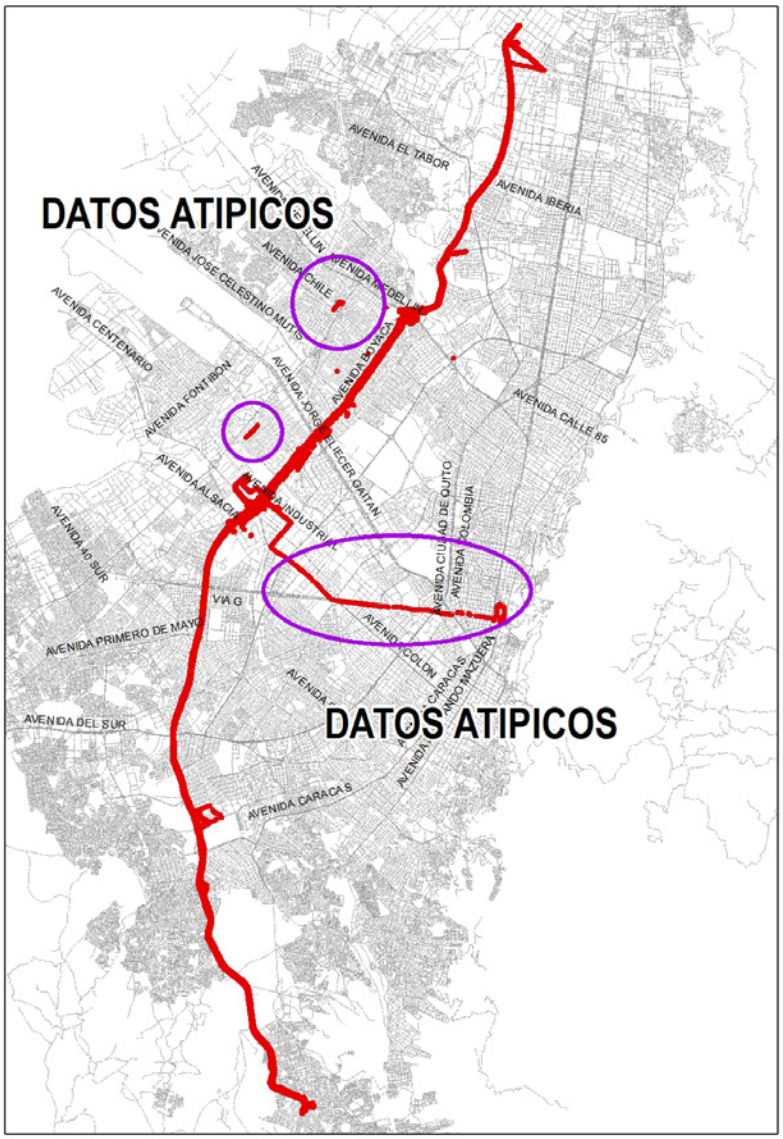

Figura 2. Identificación gráfica de datos atípicos

Fuente: elaboración propia.

rangos: mañana, mediodía-tarde y noche, para el atributo "Hora paso", como se muestra en tabla 4.

En la figura 3 se muestran las frecuencias correspondientes a los rangos de la variable discretizada.

Aplicación de minería

La minería de patrones frecuentes ha llegado mucho más allá de lo básico, debido a una investigación sustancial, numerosas extensiones del alcance del problema y amplios estudios de aplicación. Un patrón frecuente puede tener varias formas alternativas, incluido un patrón frecuente simple, un patrón cerrado o un patrón máximo. Los frecuentes también se pueden mapear en reglas de asociación u otras reglas basadas en medidas de interés. A veces también pueden interesar los patrones infrecuentes o raros (es decir, que ocurren raramente, pero son de importancia crítica), o negativos (es decir, que revelan una correlación negativa entre artículos) (Han, Kamber y Pei, 2012).

Selección del algoritmo

El conjunto de datos de este estudio está compuesto, en parte, por datos geográficos, lo que influyó en la selección

Tabla 4. Rangos de la variable discretizada "Hora paso"

RANGOS	HORA
Mañana	04:00 – 09:00
Medio día - Tarde	9:00 – 16:00
Noche	16:00 – 22:00

Fuente: elaboración propia.

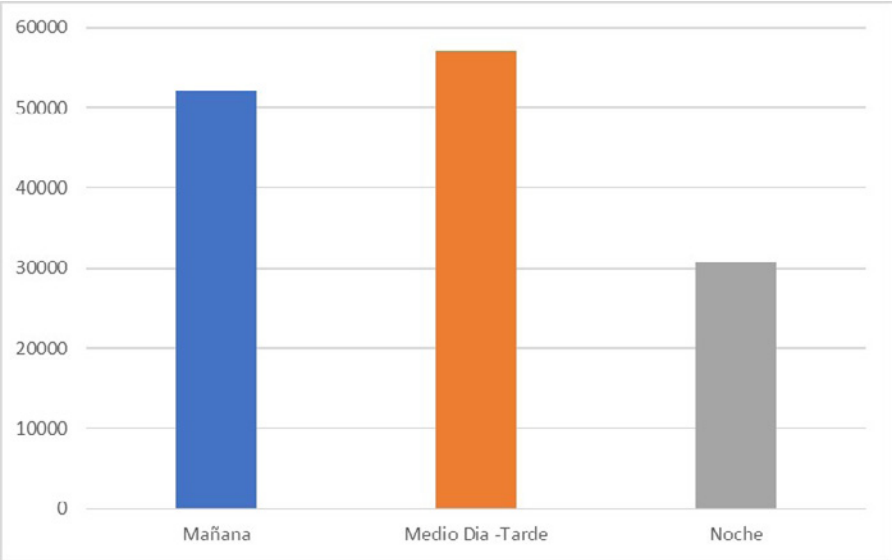

Figura 3. Frecuencia de los rangos de la variable discretizada

Fuente: elaboración propia.

del algoritmo a aplicar en la etapa de minería, por lo cual se escogió DBSCAN como el más apropiado. El algoritmo de *clustering* DBSCAN requiere solo un parámetro de entrada y ayuda al usuario a determinar un valor apropiado para ello. DBSCAN es eficiente incluso para grandes bases de datos espaciales. Este algoritmo (densidad *clustering* espacial basado en aplicaciones con ruido) está diseñado para descubrir los clústeres y el ruido en un espacio espacial (Ester, Kriegel, Jorg y Xu, 1996).

DBSCAN

El algoritmo de agrupación en clústeres DBSCAN que se basa en una noción de agrupaciones determinada por la densidad que está diseñada para descubrir agrupaciones de forma arbitraria. DBSCAN requiere solo un parámetro de entrada y ayuda al usuario a establecer un valor apropiado para él. En los puntos de un mismo clúster, su k-ésimo vecino debería estar más o menos a la misma distancia. En los puntos de ruido, su k-ésimo vecino debería estar más lejos. DBSCAN está diseñado para descubrir los clústeres y el ruido en una base de datos espacial (Han, Kamber y Pei, 2012). Idealmente, se deberían conocer los parámetros apropiados Eps y MinPts de cada grupo y al menos un punto del grupo respectivo. Luego, se podrían recuperar todos los puntos que son de densidad alcanzable desde el punto dado usando los parámetros correctos. Pero no hay una manera fácil de obtener esta información por adelantado para todos los grupos de la base de datos. Sin embargo, hay una forma simple y efectiva: la heurística para determinar los parámetros Eps y MinPts. Por tanto, DBSCAN usa valores globales para Eps y MinPts, es decir, los mismos para todos los grupos. Los parámetros de densidad del grupo "más delgado" son buenos candidatos para estos valores de parámetros globales que especifican la densidad más baja que no se considera ruido.

Interpretación y evaluación de resultados

Para la ejecución del algoritmo y tomar como ruido aquellos puntos poco concentrados, se tomó una distancia entre puntos de 10 metros, esto sabiendo que los datos se toman aproximadamente en intervalos de 8 a 15 segundos y un número mínimo por clúster de 20 individuos, esto permite ver las zonas con mayor cantidad de concentraciones y ubicar aquellos puntos de la ciudad de mayor tráfico recurrente.

Análisis de datos: mañana

DBSCAN provee una clusterización comparativamente mejor con otros métodos, además es apropiado para el tratamiento de datos geográficos, con lo cual es posible establecer gran cantidad de clústeres en la zona de estudio. En el caso de los datos del rango de tiempo de la mañana se obtuvo un total de 16 clústeres, aquellos que se consideran como ruido toman el valor de -1 con lo cual queda en evidencia el poder de discriminación y eliminación de datos atípicos del algoritmo de densidad. La tabla 5 detalla los clústeres con mayor cantidad de individuos.

La figura 4 muestra la totalidad de la zona de estudio (avenida Boyacá), donde se identifican los puntos donde posiblemente se concentra el tráfico en las primeras horas del día.

Tabla 5. Clústeres del rango "Hora paso-Mañana" con mayor cantidad de individuos

Mañana			
Cluster ID	**Frecuencia**	**Cluster ID**	**Frecuencia**
-1	47752	9	13
1	68	10	70
2	459	11	29
3	55	12	21
4	25	13	20
5	38	14	28
6	59	15	20
7	68	16	20
8	23		

Fuente: elaboración propia.

Figura 4. Zona de estudio: avenida Boyacá

Fuente: elaboración propia.

Análisis de datos: mediodía-tarde

Los clústeres generados en el rango de la tarde (figura 5) son más evidentes, pues en total se generaron 328 clústeres, esto se explica sabiendo que esta vía concentra la mayor cantidad de tráfico vehicular y conecta el sur con el norte, y viceversa.

Análisis de datos: noche

El rango de tiempo que comprende la noche, con las mismas condiciones de agrupamiento descritas en la Tabla 5, es útil para agrupar 72 clústeres de tráfico urbano. En la figura 6 se evidencia la congestión.

Comparación de zona

Una vez identificadas las zonas de mayor tráfico, en la figura 7 se aprecian los respectivos clústeres generados en cada periodo de tiempo para la zona más representativa: cruce de la avenida Boyacá con la calle 80, donde se evidencia un atascamiento en el tráfico.

Analizando el caso específico en la avenida Boyacá con calle 80, se puede ver una afectación baja en las horas de la mañana, cuando no es tan evidente la congestión. Hacia el rango definido como medio día se ve el progreso del patrón incrementando la longitud de afectación y el tiempo de transporte, este patrón de medio día es importante pues sobre este eje se mueve una gran parte de pasajeros y carga pesada hacia puntos de distribución.

Hacia la noche se aprecia la persistencia del tráfico en la zona que, aunque disminuye, sigue retrasando la movilidad y el regreso de las personas a sus hogares. Es posible que este punto en especial deba su congestión a un centro comercial en la zona, el cual genera mayor cantidad de vehículos de transporte público, público individual e individual particular.

Conclusiones

Hacer minería con bases de datos espaciales requiere de un amplio conocimiento de los algoritmos de agrupación

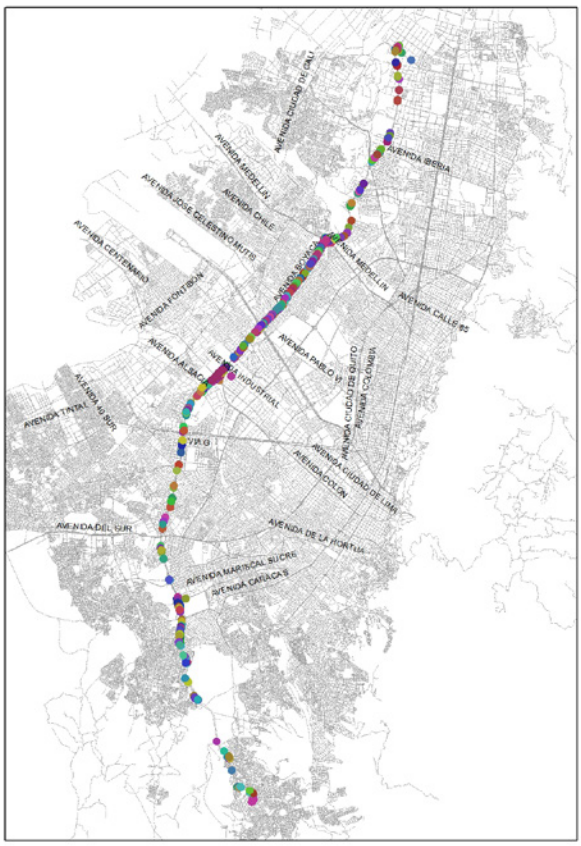

Figura 5. Clúster del rango "Hora paso-Tarde"

Fuente: elaboración propia.

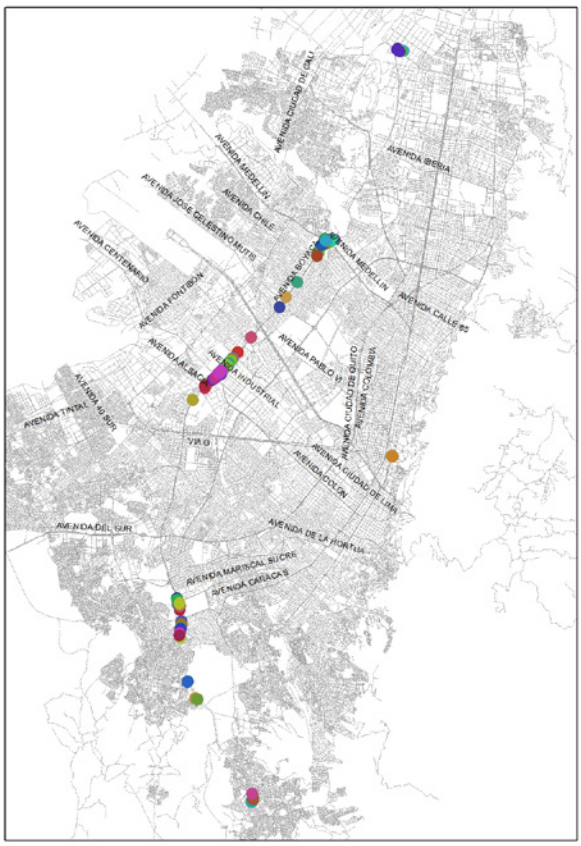

Figura 6. Clúster rango "Paso hora-Noche"

Fuente: elaboración propia.

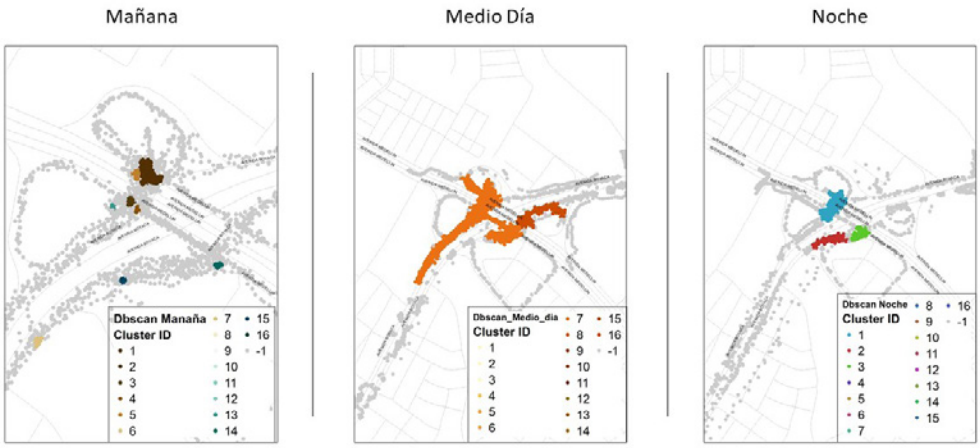

Figura 7. Clúster del cruce de avenida Boyacá con calle 80

Fuente: elaboración propia.

disponibles, ya que la selección del más adecuado para datos espaciales redundará en una adecuada tarea de extracción de conocimiento acorde con el problema evidenciado.

La etapa de preparación de datos para este estudio implicó una revisión detallada de los atributos de cada variable, con una evaluación de la conveniencia de mantener los formatos originales de los datos adquiridos o, por el contrario, la conversión a uno más conveniente para la interpretación de la información geográfica contenida en la correspondiente variable analizada. Por ejemplo, en la hora de la toma del dato fue necesario discretizar con el objeto de procesar adecuadamente la información por categorías. De igual manera, se eliminaron variables que representaban redundancia de información.

En este estudio se identifica la importancia de la aplicación del algoritmo DBSCAN, ya que permite determinar eficientemente los clústeres o núcleos de concentración de un evento a partir de los datos geográficos colectados, con los cuales es posible determinar los lugares de mayor congestión y concatenarlos con las horas de tráfico pico o valle.

Trabajos futuros podrían tomar varios ejes con centros que aglomeren gran cantidad de personas, como centros comerciales, e investigar la incidencia de este tipo de infraestructuras sobre vías principales. Esto se basa en la presente investigación en donde las intersecciones que contaban con centros comerciales agruparon un mayor número de clústeres en un rango.

Referencias bibliográficas

Secretaria Distrital de Movilidad. (2018). Plan Maestro de Movilidad - Movilidad y Desarrollo Sostenible V8. Bogotá: Secretaria Distrital de Movilidad.

Secretaria de Movilidad. (2017). Informe 2. Tiempos de recorrido en vehículo en movimiento. Bogotá: TPD ingeniería.

An, S., Yang, H., Wang, J., Cui, N. y Cui, J. (2016). Mining urban recurrent congestion evolution patterns from GPS-equipped vehicle mobility data. *Information Sciences*, 373, 515-526. DOI: https://doi.org/10.1016/j.ins.2016.06.033

Anbaroglu, B., Heydecker, B. y Cheng, T. (2014). Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, 48, 47-65. DOI: https://doi.org/10.1016/j.trc.2014.08.002

Bie, Y., Gong, X. y Liu, Z. (2015). Time of day intervals partition for bus schedule using GPS data. *Transportation Research Part C: Emerging Technologies*, 60, 443-456. DOI: https://doi.org/10.1016/j.trc.2015.09.016

Ester, M., Kriegel, H.-P., Jorg, S. y Xu, X. (1996). A Density-Based Clustering Algorithms for Discovering Clusters. *Kdd, 96*(34), 226-231. DOI: https://doi.org/10.1016/B978-044452701-1.00067-3

Fiori, A., Mignone, A. y Rospo, G. (2016). DeCoClu: Density consensus clustering approach for public transport data. *Information Sciences*, 328, 378-388. DOI: https://doi.org/10.1016/j.ins.2015.08.054

Gordon, R. (2015). *Intelligent transportation systems: Functional design for effective traffic management*. 2a. ed. Cham, Suiza: Springer. DOI: https://doi.org/10.1007/978-3-319-14768-0

Han, J., Kamber, M. y Pei, J. (2012). 10-Cluster Analysis: Basic Concepts and Methods. En J. Han, M. Kamber, & J. Pei (eds.), *Data Mining* (pp. 443–495). 3a. ed. Boston: Morgan Kaufmann. DOI: https://doi.org/ https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Hwang, R.H., Hsueh, Y.L. y Chen, Y.T. (2015). An effective taxi recommender system based on a spatio-temporal

factor analysis model. *Information Sciences*, 314, 28-40. DOI: https://doi.org/10.1016/j.ins.2015.03.068

Ji, Y., Luo, J. y Geroliminis, N. (2014). Empirical Observations of Congestion Propagation and Dynamic Partitioning with Probe Data for Large-Scale Systems. *Transportation Research Record: Journal of the Transportation Research Board, 2422*(1), 1–11. DOI: https://doi.org/10.3141/2422-01

Luo, Y., Hadiuzzaman, M., Fang, J. y Qiu, T.Z. (2015). Assessing the mobility benefits of proactive optimal variable speed limit control during recurrent and non-recurrent congestion. *Canadian Journal of Civil Engineering, 42*(7), 477–489. DOI: https://doi.org/10.1139/cjce-2013-0427

Ma, X., Yu, H., Wang, Y. y Wang, Y. (2015). Large-scale transportation network congestion evolution prediction using deep learning theory. *PLoS ONE, 10*(3), 1–17. DOI: https://doi.org/10.1371/journal.pone.0119044

Yang, H., Ozbay, K., Xie, K., & Ma, Y. (2016). Development of an Automated Approach for Quantifying Spatiotemporal Impact of Traffic Incidents. Transportation Research Board 95th Annual Meeting.

Palau, J.J. (s.f.). Análisis del transporte masivo y la movilidad en Bogotá. *Universidad & Empresa*, 15(24), 15-23.

Pyle, D. (1999). Data Preparation for Data Mining. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc

Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional* (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia. DOI: http://dx.doi.org/10.16925/9789587600490

Tulic, M., Bauer, D. y Scherrer, W. (2015). Link and Route Travel Time Prediction Including the Corresponding Reliability in an Urban Network Based on Taxi Floating Car Data. *Transportation Research Record: Journal of the Transportation Research Board, 2442*(1), 140-149. DOI: https://doi.org/10.3141/2442-15

Yin, S., Li, X., Gao, H. y Kaynak, O. (2015). Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics, 62*(1), 657–667. DOI: https://doi.org/10.1109/TIE.2014.2308133

Yin, S., Wang, G. y Gao, H. (2016). Data-Driven Process Monitoring Based on Modified Orthogonal Projections to Latent Structures. *IEEE Transactions on Control Systems Technology, 24*(4), 1480-1487. DOI: https://doi.org/10.1109/TCST.2015.2481318