



UNIVERSIDAD NACIONAL DE LA PLATA

TESIS DOCTORAL

---

**Análisis de trayectorias utilizando técnicas  
de minería de datos.**

---

*Autor:*  
Gary REYES

*Directora*  
Dra. Laura LANZARINI

*Codirector*  
Dr. Waldo Hasperué

*Tesis presentada en cumplimiento de los requisitos  
para obtener el grado de Doctor en Ciencias Informáticas*

Facultad de Informática - UNLP

5 de agosto de 2024

*«Si he logrado ver más lejos ha sido porque he subido a hombros de gigantes.»*

Isaac Newton

# Resumen

Gary REYES

*Análisis de trayectorias utilizando técnicas de minería de datos.*

El análisis de trayectorias de movimiento utilizando técnicas de minería de datos es un área de investigación relevante dada la gran cantidad de datos de localización que se generan actualmente. Esta tesis explora el uso de algoritmos no supervisados para descubrir patrones ocultos en conjuntos de datos de trayectorias. Específicamente, se desarrolla una metodología basada en agrupamiento espacio-temporal para identificar grupos de trayectorias similares. El énfasis estuvo puesto en analizar la densidad del flujo vehicular así como en la identificación de patrones de movimientos dentro de las trayectorias. Los casos de estudio analizados demuestran la efectividad del enfoque para revelar comportamientos recurrentes en los datos.

La aplicación de técnicas avanzadas de minería de datos ha desempeñado un papel fundamental en el análisis de trayectorias de movimiento. Diversas metodologías, además del agrupamiento espacio-temporal abordado en esta investigación, han contribuido significativamente al entendimiento de patrones subyacentes en conjuntos de datos de localización. Entre estas técnicas se destacan métodos de clasificación, regresión y análisis de series temporales. La clasificación se utiliza para identificar y etiquetar diferentes tipos de trayectorias, mientras que la regresión facilita la predicción de movimientos futuros. Asimismo, el análisis de series temporales permite comprender la evolución temporal de los desplazamientos. La aplicación de estas técnicas ha demostrado su eficacia en diversas aplicaciones, desde la detección de comportamientos anómalos hasta la predicción de patrones de movimiento en entornos urbanos.

En esta tesis, el enfoque gira en torno al agrupamiento como técnica principal para identificar patrones en el flujo de tráfico vehicular. La capacidad del agrupamiento para reunir trayectorias similares ha demostrado ser esencial para modelar comportamientos recurrentes en el tráfico cotidiano. La adaptación de algoritmos de agrupamiento a grandes volúmenes de datos multidimensionales y multifuente ha sido un aspecto crítico para mantener la utilidad de los conocimientos extraídos. La aplicación de esta técnica no solo permite la identificación de patrones, sino también el reconocimiento efectivo de diferentes tipos de movimientos, contribuyendo así a una comprensión profunda y detallada del flujo de tráfico en entornos urbanos complejos.

El análisis de flujos de tráfico en entornos urbanos presenta retos particulares debido a la complejidad de las redes viales, la densidad vehicular y los frecuentes eventos extraordinarios que afectan el tránsito.

El uso de información semántica proveniente de mapas que contienen información de la infraestructura vial es un recurso frecuentemente usado para perfeccionar la precisión de los algoritmos que identifican el flujo de tráfico vehicular. La riqueza de datos contenida en estos mapas, que incluye detalles sobre la topografía vial, ubicación de semáforos, capacidad de las carreteras, y la presencia de puntos de interés, proporciona un contexto valioso a la hora de efectuar el análisis del agrupamiento. En algunas investigaciones, este enfoque ha contribuido a la identificación de rutas cortas y situaciones problemáticas relacionadas al tráfico vehicular; también ha ayudado a revelar puntos de interés relevantes que pueden influir en los patrones de movilidad. La inclusión de información semántica puede potenciar la capacidad de los algoritmos para construir particiones acordes al contexto real de la red vial. En consecuencia, una integración de este tipo de información puede fortalecer la interpretación de los grupos identificados, aportando una comprensión más completa de los factores subyacentes que explican tanto patrones normales como anomalías en el tráfico urbano.

Los dispositivos GPS, son herramientas fundamentales en la actualidad para determinar la ubicación geográfica con precisión utilizando señales de satélites. Frecuentemente, los dispositivos GPS introducen errores aleatorios en las coordenadas registradas, lo que subraya la necesidad de emplear medidas de similitud robustas durante el agrupamiento. Además de la distancia de Hausdorff, que se destaca por su capacidad para manejar eficazmente datos con errores, otras métricas comunes como la distancia euclidiana y la distancia de Manhattan también desempeñan roles significativos en el análisis de trayectorias. La distancia euclidiana resulta útil para capturar relaciones de proximidad en entornos donde la precisión es menos crítica. Por otro lado, la distancia de Manhattan es eficaz para evaluar la similitud en entornos urbanos con patrones de movimiento más estructurados. Aunque estas métricas son valiosas en ciertos contextos, es esencial reconocer sus limitaciones frente a errores aleatorios, destacando la importancia de la elección cuidadosa de la métrica según las características específicas de los datos de trayectorias y los objetivos del agrupamiento.

Otra alternativa es aplicar algoritmos probabilísticos, que modelan la ubicación de cada punto como una distribución en lugar de un valor puntual. Aunque esto incrementa la complejidad computacional, permite cuantificar explícitamente el nivel de certeza sobre la verdadera posición de las trayectorias.

Es indispensable considerar tanto la cercanía espacial como temporal de las trayectorias durante la construcción de grupos. Esto permite detectar situaciones anómalas emergentes como la congestión vehicular y modelar la evolución dinámica del tráfico, incluyendo aquellas que están por desvanecerse.

Algunas técnicas que incorporan un dinamismo al análisis son el agrupamiento multidimensional, agrupamiento en ventanas deslizantes y métodos de flujos de datos, donde las particiones varían activamente a lo largo del tiempo.

La precisión de los algoritmos de agrupamiento no solo depende de la dimensionalidad de los datos, sino también de otros factores cruciales. Entre ellos, la calidad de la recolección de los datos, la selección adecuada de las métricas a usar y la consideración de posibles sesgos introducidos por errores de medición son fundamentales para obtener resultados precisos, especialmente cuando se trabaja con grandes volúmenes de información. Es esencial abordar estos aspectos para garantizar la robustez y la validez de los resultados obtenidos.

En cuanto a la escalabilidad, enfrentar eficientemente el procesamiento de grandes conjuntos de datos es un desafío crítico. La aplicación de técnicas como agrupamiento aproximado, por lotes y en memoria se presentan como soluciones viables para analizar grandes volúmenes de trayectorias de manera eficiente. Estas estrategias permiten mantener un equilibrio entre la complejidad computacional y la necesidad de obtener resultados precisos, abordando el obstáculo que representa el tamaño masivo de los datos.

Aunque el agrupamiento de trayectorias vehiculares urbanas es una tarea compleja, la investigación en curso y la incorporación de técnicas adaptadas de inteligencia artificial se presentan como alternativas prometedoras para obtener información valiosa sobre la dinámica del tráfico cotidiano a partir de los crecientes conjuntos de datos disponibles.

El análisis de trayectorias presenta tres líneas bien diferenciadas según si se trabaja con toda la trayectoria, segmentos de trayectorias o si se trabaja con las distintas ubicaciones vehiculares que las componen.

Las investigaciones relacionadas con el desarrollo de esta tesis comenzaron analizando segmentos de trayectorias y siguiendo esa línea se diseñaron distintas soluciones para segmentarlas primero y agruparlas después. La segmentación de trayectorias presenta varias dificultades que aún siguen siendo tema de estudio. Las primeras publicaciones generadas en el marco de esta tesis se refieren al uso de distintos criterios para realizar el proceso de segmentación y comparación de tramos de trayectorias. La calidad de los resultados depende de la manera en la que se realicen estos procesos. En lo referido al agrupamiento de segmentos de trayectorias, se definieron un par de estrategias, la primera identifica los cambios de dirección bruscos o significativos y la segunda genera segmentos con características similares respecto a su forma y dirección, logrando representaciones más cortas de información de las trayectorias para optimizar los requerimientos en el procesamiento.

En una segunda etapa, se focalizó en el análisis de flujos de datos correspondientes a ubicaciones GPS con el objetivo de identificar patrones en tiempo real.

Esto dió lugar a la contribución clave de esta tesis que consiste en el diseño e implementación de un método de agrupamiento dinámico de flujos de datos capaz de mantener actualizada la información referida al flujo del tránsito ayudando en

la detección de patrones de movimiento y en la rápida visualización de la densidad vehicular. Esto permite detectar cambios en los patrones de movilidad de forma adaptativa. Las observaciones de los resultados indican la existencia de características recurrentes en los patrones encontrados.

Otro aspecto importante explorado es la visualización de los agrupamientos de trayectorias utilizando técnicas que permitan representar gráficamente los grupos resultantes, esta información es esencial para la identificación de los patrones encontrados, la planificación a largo plazo y para la toma de decisiones.

Los resultados de estudios de usuarios realizados en investigaciones externas sugieren que este tipo de visualización incrementa la comprensibilidad de los patrones descubiertos.

Dentro del método propuesto, se emplea una técnica para el análisis de variabilidad en la densidad vehicular en diferentes zonas urbanas. Este valor proporciona una medida cuantitativa de la saturación del tráfico considerando la densidad actual de vehículos respecto a la densidad habitual.

Para concluir, esta tesis introduce novedosas contribuciones aplicando técnicas de minería de datos espacio-temporales al problema de análisis de conjuntos masivos de trayectorias. Los diferentes métodos desarrollados en el transcurso de esta tesis tienen utilidad en diversos dominios incluyendo transporte inteligente, social y análisis medioambiental.

La aplicación del método propuesto a datos de trayectorias de las ciudades de San Francisco, Roma y Guayaquil ha demostrado ser una valiosa aproximación para abordar problemas relacionados al tráfico urbano. Los resultados, basados en la recopilación y análisis de datos en tiempo real, destacan ventajas como precisión espacial y temporal, flexibilidad en distintos escenarios urbanos y la capacidad de reconocer diversas situaciones en el tráfico urbano. Sin embargo, se han identificado limitaciones, como la elección de un horario específico afectando la representatividad de los datos, la dependencia de datos en tiempo real y la influencia de la ubicación de las estaciones de monitoreo en la representatividad de los resultados.

En esta tesis el foco de las investigaciones estuvo puesto en el análisis del desplazamiento así como en la variabilidad de densidad vehicular en entornos urbanos.

En lo referido al análisis de las densidades vehiculares, el método propuesto ha contribuido en la detección de zonas con reiteradas fluctuaciones en su densidad vehicular dando cuenta de posibles escenarios de congestión. La detección temprana de estas situaciones es de gran ayuda al momento de organizar la planificación del transporte en las ciudades.

En general, se identificó una alta variabilidad en la densidad vehicular en áreas urbanas densamente pobladas en oposición a lo que ocurre en zonas con menor densidad de tráfico.

En relación a la detección de patrones distintivos relacionados con los cambios en la movilidad de las trayectorias, el método propuesto utiliza un agrupamiento dinámico de las ubicaciones GPS que las conforman. Como resultado se obtienen grupos

asociados a regiones con densidades específicas. La Figura 1 resume los resultados del análisis de densidad que fueron obtenidos al aplicar el método propuesto en los tres casos de estudio mencionados previamente. La densidad total refleja la cantidad de vehículos existentes en el área de cobertura del grupo, esta densidad abarca las áreas acumuladas de todos los microgrupos.

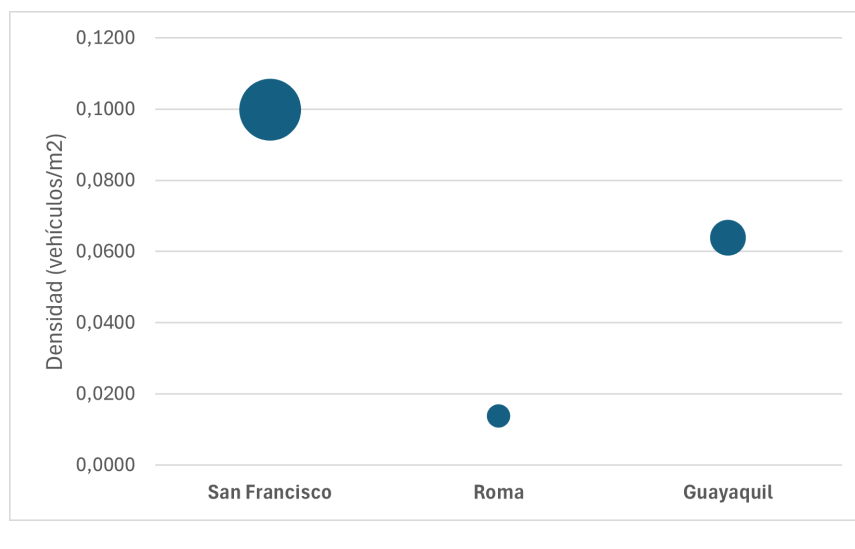


FIGURA 1: Resumen estadístico de densidades.

En la Figura 1 puede observarse que San Francisco se destacó por su densidad y desviación altas, Roma presenta una densidad y desviación bajas, Guayaquil presenta una densidad y una desviación medias. En Roma los grupos presentan poca expansión, los grupos de Guayaquil se han expandido significativamente más que Roma y en San Francisco se encuentra la mayor expansión. Esto posiblemente se dé por la infraestructura vial de las ciudades analizadas.

Por otro lado, observando el desplazamiento del centro de masa de cada grupo es posible establecer patrones de movimiento vehicular que ayudan a comprender la dinámica de tránsito urbano. Utilizando el método propuesto se almacenan las secuencias de estos centros de masa permitiendo su posterior reconstrucción.

Como resultado de la aplicación del método propuesto a los tres casos de estudio se obtuvieron desplazamientos medios de los centros de masa de los grupos y su desviación. Como se observa en la Figura 2(a) el desplazamiento medio del centro de masa de San Francisco es mayor con una alta desviación, esta alta dispersión causada por los valores máximos de desplazamiento origina en la gráfica valores muy alejados de la media, los desplazamientos de Guayaquil son parecidos a los de Roma, pero su desviación es mayor; Roma es la que presenta el menor desplazamiento y la menor desviación. La alta desviación de los desplazamientos puede ser provocada por el hecho de que, con el tiempo, los grupos identifican zonas densas alejadas del centro de masa previo. Esto se puede observar en los mapas que genera el método propuesto, los que permiten analizar con respecto al tiempo una posible dirección de los centros de masa.

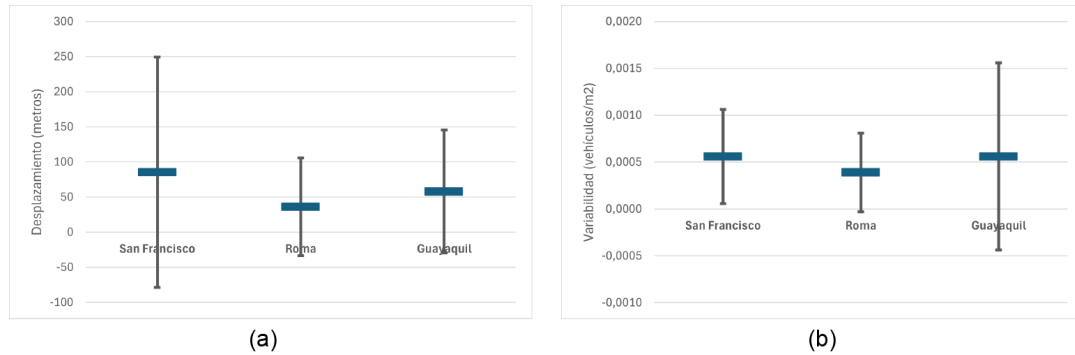


FIGURA 2: Resumen estadístico de desplazamientos (a) y variabilidades (b).

En lo referido al análisis de la variabilidad en la densidad de grupos vehiculares se identificaron fluctuaciones significativas con respecto a la densidad habitual en las zonas de estudio. La variabilidad está determinada mediante la diferencia entre la densidad actual y la histórica de las distintas zonas. Establecer adecuadamente el incremento de este valor es un indicador útil en la predicción de escenarios de congestión. La Figura 2(b) resume la variabilidad detectada en cada caso de estudio.

Se observa que San Francisco y Guayaquil presentan medias similares en las variabilidades, aunque Guayaquil muestra una mayor dispersión. Roma, en contraste, exhibe la menor variabilidad con una menor dispersión.

En base a los resultados obtenidos y luego de abordar el análisis de trayectorias vehiculares desde distintos puntos de vista puede afirmarse que el método de agrupamiento dinámico de flujos de datos definido en esta tesis es una herramienta sumamente útil para identificar, en tiempo real, los movimientos de grupos de vehículos y sus fluctuaciones de densidad. Su capacidad de adaptarse a diversas condiciones y la precisión en la detección de grupos vehiculares hacen de este método una herramienta valiosa para la gestión eficiente del tráfico y la planificación urbana. Los resultados que pueden obtenerse con este método, acompañados de mapas interactivos con capacidad de ser reproducidos en tiempo real, facilitan la comprensión de la dinámica del tráfico en diferentes contextos.



# Agradecimientos

A Laura Lanzarini, mi directora de tesis, quien con su acertada orientación, rigurosidad y discusión crítica permitió que esta tesis llegara a buen término. Gracias por sus conocimientos, guía, acompañamiento, y sobretodo por su aliento a no desmayar en los momentos difíciles a lo largo de todo este tiempo.

A Waldo Hasperué, César Estrebow y Aurelio F. Bariviera por colaborar con sus ideas, asesoramiento y experiencias en el desarrollo de esta tesis.

A los docentes e investigadores de la Universidad Nacional de La Plata y de la Universidad de Guayaquil, de manera particular a los tesistas de grado, junto a quienes he investigado y han sido partícipes de los logros alcanzados.

A mis familiares por su comprensión, comunicación constante y apoyo incondicional durante todo el proceso dentro de mis estudios doctorales.

A la Universidad Nacional de La Plata por la educación de excelencia y calidad que se plasmó durante mi formación doctoral.

A la Universidad de Guayaquil por la ayuda económica brindada para la realización de esta investigación. Gracias por su ayuda y confianza en mí depositada.

A la Universidad Bolivariana del Ecuador por el financiamiento otorgado para poder difundir los resultados de esta investigación.

A todos, mis más sinceros agradecimientos.



A mis Hijos Bruno, Arianna,  
Romina y Garito, con amor.



# Índice general

<b>Agradecimientos</b>	<b>IX</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Motivación . . . . .	2
1.2. Objetivos . . . . .	8
1.3. Alcances y limitaciones . . . . .	9
1.4. Contribuciones . . . . .	10
1.5. Publicaciones . . . . .	11
1.5.1. Artículos de Revistas . . . . .	11
1.5.2. Capítulos de libros y congresos con referato . . . . .	12
1.5.3. Participaciones en Workshops . . . . .	12
1.5.4. Publicaciones complementarias . . . . .	13
1.6. Organización de la Tesis . . . . .	14
<b>2. Agrupamiento de trayectorias</b>	<b>15</b>
2.1. El Sistema de Posicionamiento Global (GPS) . . . . .	15
2.2. Las trayectorias y sus campos de investigación . . . . .	16
2.2.1. Categorización general de las trayectorias . . . . .	18
2.2.2. Campos de investigación activos de trayectorias . . . . .	18
2.3. Minería de datos para el análisis de trayectorias . . . . .	20
2.3.1. Preprocesamiento de trayectorias . . . . .	21
Segmentación de trayectorias . . . . .	22
Interpolación . . . . .	23
Detección de valores atípicos . . . . .	23
2.3.2. Clasificación de trayectorias . . . . .	23
2.3.3. Extracción de patrones en trayectorias . . . . .	24
Patrones de movimientos en conjunto . . . . .	24
Patrones secuenciales . . . . .	25
Patrones periódicos . . . . .	25
2.4. Agrupamiento de trayectorias . . . . .	26
2.4.1. Breve estudio bibliométrico de algoritmos de agrupamiento de trayectorias . . . . .	28
2.4.2. Medidas de similitud . . . . .	34
2.4.3. Clasificación de algoritmos . . . . .	37
Algoritmos de agrupamiento particionales . . . . .	37

Algoritmos de agrupamiento basados en distribución . . . . .	38
Algoritmos de agrupamiento jerárquicos . . . . .	39
Algoritmos de agrupamiento basados en densidad . . . . .	39
Algoritmos de agrupamiento basados en cuadrículas . . . . .	40
Algoritmos de agrupamiento basados en grafos . . . . .	40
Algoritmos basados en modelos . . . . .	40
Algoritmos basados en lógica difusa . . . . .	41
Algoritmos de flujo de datos . . . . .	41
2.4.4. Modalidades de procesamiento . . . . .	43
Primer grupo: Modalidad estática y dinámica . . . . .	43
Segundo grupo: Modalidades por lotes (Batch) y en línea (On-line) . . . . .	44
2.5. Mecanismo de olvido . . . . .	44
2.6. Conclusiones . . . . .	45
<b>3. Identificación de patrones . . . . .</b>	<b>47</b>
3.1. Método de agrupamiento de trayectorias GPS para la toma de decisiones en sistemas de transporte inteligente . . . . .	48
3.1.1. Segmentación de las trayectorias . . . . .	49
3.1.2. Agrupamiento de subtrayectorias . . . . .	49
3.1.3. Métricas de calidad para agrupamiento de trayectorias . . . . .	51
3.1.4. Consideraciones relevantes . . . . .	51
3.2. Propuesta de un método de agrupamiento de trayectorias de vehículos basado en pivotes . . . . .	52
3.2.1. Agrupamiento de subtrayectorias . . . . .	53
3.2.2. Consideraciones relevantes . . . . .	57
3.3. Análisis de flujo vehicular utilizando grupos . . . . .	58
3.3.1. Definición de celdas para resumir información de trayectorias . . . . .	59
3.3.2. Agrupamiento adaptativo y dinámico de trayectorias . . . . .	59
3.3.3. Visualización de rangos de velocidades . . . . .	61
3.3.4. Consideraciones relevantes . . . . .	61
3.4. Agrupación dinámica de trayectorias de vehículos . . . . .	62
3.4.1. Consideraciones relevantes . . . . .	63
3.5. Método de procesamiento de flujo de datos para el agrupamiento de trayectorias . . . . .	64
3.5.1. Paso 1: Uso de un buffer . . . . .	66
3.5.2. Paso 2: Uso de índices para crear celdas con información de resumen . . . . .	66
3.5.3. Paso 3: Uso de una técnica de agrupamiento . . . . .	67
3.5.4. Paso 4: Visualización de resultados . . . . .	67
3.5.5. Consideraciones relevantes . . . . .	67
3.6. Conclusiones . . . . .	69

<b>4. Variabilidad del Flujo Vehicular</b>	<b>71</b>
4.1. Método de identificación de variabilidad en el flujo del tráfico . . . . .	73
4.1.1. Módulo principal de nivel 1: Agrupamiento dinámico de ubi- caciones . . . . .	74
Formación de microgrupos . . . . .	75
Olvido porcentual de datos basado en el tiempo de ingreso . . . . .	76
4.1.2. Módulo principal de nivel 2: Agrupamiento por proximidad. . . . .	78
Sub-módulo complementario 1: Preprocesamiento de áreas de tráfico históricas . . . . .	79
Agrupamiento por proximidad . . . . .	81
Sub-módulo complementario 2: Visualizador de resultados . . . . .	85
4.2. Resultados obtenidos . . . . .	89
4.2.1. Datos utilizados . . . . .	89
4.2.2. Parametrizaciones iniciales . . . . .	90
4.2.3. Análisis de resultados obtenidos en los casos de estudio . . . . .	90
Aspecto a analizar 1: Desplazamiento. . . . .	90
Aspecto a analizar 2: Densidad. . . . .	94
Aspecto a analizar 3: Variabilidad. . . . .	95
4.2.4. Distribución acumulada de variabilidades . . . . .	98
4.2.5. Validación del agrupamiento . . . . .	100
Indicadores de calidad aplicada a los microgrupos del agrupa- miento de nivel 1 . . . . .	101
Indicadores de calidad aplicada a los grupos del agrupamiento de nivel 2 . . . . .	102
4.2.6. Análisis de los tiempos de ejecución . . . . .	103
4.3. Discusiones . . . . .	105
Acerca de los grupos resultantes en el agrupamiento . . . . .	105
Acerca de la distribución acumulada de variabilidades . . . . .	108
Acerca de los tiempos de ejecución resultantes . . . . .	109
4.4. Conclusiones . . . . .	109
<b>5. Conclusiones y trabajos futuros</b>	<b>111</b>
5.1. Conclusiones . . . . .	111
5.2. Trabajos futuros . . . . .	112
<b>A. Diagramas de procesos por componentes</b>	<b>115</b>
A.1. Módulo principal de nivel 1: Agrupamiento dinámico de ubicaciones . . . . .	115
A.2. Sub-módulo complementario 1: Preprocesamiento de áreas de tráfico históricas . . . . .	117
A.3. Módulo principal de nivel 2: Agrupamiento por proximidad. . . . .	118
A.4. Sub-módulo complementario 2: Visualizador de resultados . . . . .	120

<b>B. Evaluación de una rejilla para la identificación de patrones de congestión del tráfico</b>	<b>121</b>
B.1. Introducción . . . . .	121
B.2. Método propuesto . . . . .	122
B.2.1. Definición de celdas . . . . .	122
B.2.2. Agrupación dinámica de celdas . . . . .	123
B.2.3. Evaluación de la congestión . . . . .	123
B.2.4. Visualización de agrupaciones . . . . .	125
B.3. Resultados obtenidos . . . . .	126
B.4. Conclusiones . . . . .	129
<b>C. Metodología para la Identificación de Congestiones Vehiculares Basada en Agrupación Dinámica</b>	<b>131</b>
C.1. Introducción . . . . .	131
C.2. Materiales y métodos . . . . .	132
C.2.1. Paso 1: Cargar Información de la Red Vial . . . . .	133
C.2.2. Paso 2: Procesamiento de datos . . . . .	133
C.2.3. Paso 3: agrupación basada en la distancia . . . . .	134
C.2.4. Paso 4: Clasificación de los grupos según su evaluación de congestión . . . . .	136
C.2.5. Paso 5: Visualización de resultados . . . . .	139
C.3. Resultados obtenidos . . . . .	140
C.4. Conclusiones . . . . .	146
<b>Bibliografía</b>	<b>149</b>



# Índice de figuras

1.	Resumen estadístico de densidades. . . . .	VII
2.	Resumen estadístico de desplazamientos (a) y variabilidades (b). . . .	VIII
1.1.	Clasificación de las técnicas de agrupamiento de trayectorias. . . . .	4
2.1.	Órbitas de los satélites GPS. Fuente: <a href="https://www.gps.gov/">https://www.gps.gov/</a> . . . . .	16
2.2.	Principales áreas de estudio en la minería de datos para el análisis de trayectorias. Fuente: (Zheng 2015) . . . . .	21
2.3.	Ejemplo de segmentación de una trayectoria. Fuente: (Yao et al. 2023) .	22
2.4.	Ejemplo de diferentes patrones encontrados sobre las trayectorias conjuntas. Fuente: (Mazimpaka y Timpf 2016) . . . . .	25
2.5.	Extracción de patrones secuenciales en trayectorias. Fuente: (Zheng 2015) . . . . .	26
2.6.	Ejemplo de un agrupamiento de trayectorias: (a) trayectorias espaciales; (b) segmentos de trayectorias agrupadas. Fuente: (Zheng y Zhou 2011) . . . . .	27
2.7.	Diagrama estratégico de las palabras clave del autor, generado con bibliometrix. Fuente: Scopus. . . . .	33
2.8.	Mapa de nubes de palabras en títulos y resúmenes (recuento completo), generado con VOSviewer. Fuente: Scopus. . . . .	33
3.1.	Representación del proceso de recálculo del centroide mediante un pivote . . . . .	51
3.2.	Reticulación inicial y centroides de subtrayectorias. . . . .	53
3.3.	Área de trazado del pivote. . . . .	54
3.4.	Ángulo formado por una subtrayectoria respecto del eje de abscisas. .	55
3.5.	Pivote con (a) ángulo medio negativo y (b) positivo. . . . .	56
3.6.	Posiciones utilizadas para determinar el centroide. . . . .	56
3.7.	Metodología propuesta . . . . .	58
3.8.	Funcionamiento de segunda etapa de algoritmo dinámico. (A) Identificación de microgrupos directamente conectados y (B) Expansión resultante . . . . .	60
3.9.	Mapa a una evolución. (A) Capas correspondientes a los grupos seleccionados. (B) Visualización de marcadores . . . . .	61
3.10.	Esquema general del procesamiento del método por microlotes . . . .	65
3.11.	Diferencias entre los métodos usados en la conformación de celdas. . .	67

3.12. Visualización de resultados de agrupamientos. . . . .	68
4.1. Diagrama correspondiente al método propuesto . . . . .	74
4.2. Elementos que componen un microgrupo. . . . .	75
4.3. Información asociada a un microgrupo. . . . .	76
4.4. Porcentaje de olvido y relevancia de los datos por cada unidad de tiempo transcurrido para un valor de lambda establecido en 5 %. . . . .	78
4.5. Selección de la ubicación inicial en el nuevo grupo. . . . .	83
4.6. Área de expansión del grupo. . . . .	83
4.7. Expansión iterativa de un grupo. . . . .	84
4.8. Ubicación del nuevo centro de masa. . . . .	84
4.9. Grupos proyectados sobre el mapa. . . . .	87
4.10. Distribución acumulada de variabilidades positivas. . . . .	88
4.11. Grupo extraído en la ciudad de Roma. . . . .	91
4.12. Grupo extraído en la ciudad de San Francisco. . . . .	92
4.13. Grupo extraído en la ciudad de Guayaquil. . . . .	93
4.14. Ejemplos de los distintos niveles de variabilidad que se pueden obtener. . . . .	96
4.15. Ejemplo de variabilidad positiva. . . . .	97
4.16. Ejemplo de variabilidad negativa. . . . .	97
4.17. Ejemplo de variabilidad mixta. . . . .	98
4.18. Distribución acumulada por celdas en la ciudad de Roma. . . . .	99
4.19. Distribución acumulada por celdas en la ciudad de Guayaquil. . . . .	100
4.20. Distribución acumulada por celdas en la ciudad de San Francisco. . . . .	101
4.21. Ejemplo de diferentes complejidades de redes. Una sola carretera (a), cuatro carreteras (b) y seis carreteras (c). . . . .	105
4.22. Variabilidad de los flujos de tráfico en diferentes entornos urbanos. . . . .	107
A.1. Pasos del módulo principal de nivel 1. . . . .	116
A.2. Pasos del sub-módulo complementario 1 . . . . .	117
A.3. Pasos del módulo principal 2 . . . . .	119
A.4. Pasos del sub-módulo complementario 2 . . . . .	120
B.1. Proporciones según el nivel de las celdas . . . . .	124
B.2. Grupos proyectados sobre el mapa . . . . .	125
B.3. Evaluación de celdas congestionadas usando promedios históricos . . . . .	126
B.4. Distribución de la congestión por promedios históricos . . . . .	128
C.1. Pasos de la metodología propuesta. . . . .	133
C.2. Enrutamiento e interpolación de una trayectoria. . . . .	135
C.3. Elementos que conforman un grupo. . . . .	136
C.4. Segmentos de carretera analizadas en el area de un grupo. . . . .	137
C.5. Tramos identificados en una región de la red vial. Los segmentos con diferentes colores representan el flujo vehicular y los segmentos negros no muestran ningún flujo vehicular. . . . .	138

C.6. Grupos proyectados en una parte del mapa. . . . . 140  
C.7. Grupo proyectado en la rejilla. . . . . 141



# Índice de tablas

2.1. Principales áreas de investigación asignadas a los trabajos de la muestra. Fuente: Scopus. . . . .	30
2.2. Diez países de autores correspondientes. Fuente: Scopus. . . . .	30
2.3. Diez principales citas totales por país. Fuente: Scopus. . . . .	31
2.4. Las diez fuentes más relevantes. Fuente: Scopus. . . . .	31
2.5. Palabras clave principales. Fuente: Scopus. . . . .	32
2.6. Resumen de las medidas de similitud más comunes. . . . .	37
4.1. Resumen estadístico de densidades. . . . .	94
4.2. Resultados del coeficiente de Silhouette. . . . .	102
4.3. Resultados del índice DBCV. . . . .	103
4.4. Tiempos de ejecución (minutos) del método propuesto. . . . .	104
4.5. Densidades de los grupos obtenidos. . . . .	106
4.6. Desplazamientos de los centros de masa de los grupos (en metros). . .	106
B.1. Tiempos de ejecución . . . . .	127
B.2. Matriz de contingencia basada en la coincidencia de resultados de la evaluación del flujo . . . . .	128
C.1. Matriz de confusión para la ciudad de San Francisco utilizando 45 segundos de olvido y sin tolerancia. . . . .	141
C.2. Matriz de confusión para la ciudad de San Francisco usando 45 segundos de olvido y tolerancia de 0,2. . . . .	142
C.3. Matriz de confusión para la ciudad de San Francisco utilizando 60 segundos de olvido y sin tolerancia. . . . .	142
C.4. Matriz de confusión para la ciudad de San Francisco utilizando 60 segundos de olvido y tolerancia de 0,2. . . . .	143
C.5. Resultados de precisión en situaciones de congestión. . . . .	144
C.6. Ejemplo de la evolución de un grupo. . . . .	145
C.7. Ejemplo del dinamismo de los segmentos de carreteras en un grupo del ciclo 5. . . . .	146



## Capítulo 1

# Introducción

Los avances en las tecnologías de localización y comunicación inalámbrica en los últimos años han dado lugar al incremento de los Servicios Basados en la Localización (LBS), lo que ha generado un gran volumen de datos de trayectorias espaciales que representan la movilidad de diversos objetos, como personas, vehículos, animales, fenómenos naturales y otros.

Hoy en día en las vías de las ciudades circulan un gran número de vehículos equipados con GPS. Por ejemplo, muchos taxis han sido provistos de un sensor GPS, que les permite informar sobre su ubicación con una marca de tiempo a un centro de datos con una determinada frecuencia. La información de estas trayectorias se puede utilizar para la asignación de recursos, la gestión de la seguridad, y análisis de tráfico.

El congestionamiento vial es una problemática relacionada al análisis del tráfico, cada vez más inevitable en las áreas urbanas modernas, resultado del aumento constante de vehículos y el crecimiento poblacional. Esta situación ha generado un incremento significativo en el flujo vehicular, lo que a su vez ha llevado a la aparición de congestiones recurrentes en las principales vías de las ciudades.

Estos embotellamientos no solo causan retrasos en los tiempos de viaje, sino que también afectan la productividad económica, generan altos niveles de contaminación ambiental y contribuyen al deterioro de la calidad del aire. La falta de infraestructura adecuada y el crecimiento constante del parque automotor han agravado aún más la congestión del tráfico, resultando en pérdidas económicas considerables y un impacto negativo en el medio ambiente. En este contexto, se hace imperativo desarrollar estrategias eficaces para identificar zonas de alta densidad vehicular que puedan ocasionar congestiones y comprender los patrones de flujo vehicular, con el objetivo de tomar decisiones informadas que contribuyan a una gestión más efectiva del tráfico y mejoren la movilidad urbana. Esta necesidad se ha vuelto apremiante para las autoridades de transporte y planificadores urbanos, quienes buscan implementar medidas para mitigar el problema y optimizar la circulación en las ciudades.

Entendiendo la importancia de contar con estrategias que ayuden a caracterizar el flujo vehicular, esta tesis propone una solución basada en un método de agrupamiento dinámico de flujos de datos capaz de identificar el incremento de las variabilidades de densidades en el flujo vehicular en tiempo real.

A continuación, se desarrollan la motivación, objetivos, alcances, limitaciones. Además, se detallan las publicaciones científicas derivadas de este trabajo y la organización del documento en capítulos.

## 1.1. Motivación

El incremento constante de la población en zonas urbanas y el consiguiente aumento en la cantidad de vehículos plantean desafíos significativos en la gestión del tráfico en entornos urbanos (Soumia Goumiri y Djahel 2023).

Con el creciente volumen de datos de tráfico disponibles, se hace necesario implementar enfoques analíticos avanzados para extraer información relevante y tomar decisiones informadas.

Este ámbito de gestión debe abordar aspectos cruciales, como el impacto ambiental y la seguridad vial, y es fundamental para la mejora de la calidad de vida de los ciudadanos.

Los avances tecnológicos actuales han facilitado la captura, representación y almacenamiento de información relacionada con las trayectorias vehiculares, gracias al uso de sensores, sistemas de navegación GPS y aplicaciones móviles. Estos datos, combinados con técnicas de análisis de Big Data y algoritmos de inteligencia artificial, permiten identificar patrones de flujo vehicular, detectar áreas de alta congestión y predecir futuros puntos críticos de congestión (Zhou et al. 2021). La integración de estas herramientas en sistemas de gestión de tráfico inteligente proporciona una base sólida para tomar decisiones oportunas y efectivas para mejorar la movilidad urbana y reducir los impactos negativos del congestionamiento vial. Esto es debido a que la gestión eficiente del tráfico permite mejorar la fluidez en las vías, reducir los tiempos de desplazamiento y minimizar las emisiones de contaminantes.

La identificación precisa de zonas densas capaces de provocar congestiones, es crucial para la implementación efectiva de medidas de alivio del tráfico. En este sentido, la aplicación de métodos avanzados juega un papel fundamental para lograr una detección precisa y confiable de áreas con altos niveles de congestión.

En los últimos años, los avances en tecnologías de recolección de datos y análisis han brindado nuevas oportunidades para abordar el problema del congestionamiento vial y la identificación de zonas altamente densas.

La comprensión de los datos de tráfico y el uso de técnicas de Inteligencia Artificial (IA) se han vuelto fundamentales para abordar la identificación de patrones y zonas de congestión de manera eficiente y precisa (Mazimpaka y Timpf 2016).

El uso de técnicas de Inteligencia Artificial (IA) para abordar la problemática de congestión en el tráfico ofrece diversas ventajas, pero también presenta desafíos que deben ser considerados para lograr resultados efectivos en la mejora de la movilidad urbana.

Entre las principales ventajas se destaca la capacidad de procesar y analizar grandes volúmenes de datos de tráfico en tiempo real. En este sentido, la capacidad de



predicción de estas técnicas también es valiosa, ya que permite anticipar situaciones de alta congestión y tomar medidas preventivas para evitar embotellamientos (Kim y Mahmassani 2015).

Sin embargo, utilizar técnicas de Inteligencia Artificial para abordar la congestión en el tráfico también presenta desafíos importantes. Uno de los principales desafíos es la necesidad de contar con una infraestructura de datos adecuada. Los sistemas de Inteligencia Artificial requieren datos de calidad y en tiempo real para funcionar de manera óptima, por lo que es fundamental tener sensores y dispositivos de recolección de datos confiables y una infraestructura de comunicaciones robusta.

Otro desafío es la complejidad técnica y el costo asociado con la implementación de soluciones basadas en Inteligencia Artificial. La construcción de modelos de Inteligencia Artificial y la programación de algoritmos avanzados requiere conocimientos especializados y recursos significativos. Además, el mantenimiento y actualización continua de estos sistemas también implica costos adicionales.

Además, es crucial abordar cuestiones de privacidad y seguridad en el manejo de datos de tráfico. El uso de técnicas de Inteligencia Artificial implica la recopilación y análisis de grandes cantidades de información, lo que puede plantear preocupaciones sobre la protección de la privacidad de los usuarios y la seguridad de los datos.

A pesar de estos desafíos, el potencial de las técnicas de Inteligencia Artificial para abordar la congestión en el tráfico es prometedor. Con un enfoque integral y la colaboración entre expertos en transporte y especialistas en Inteligencia Artificial, es posible superar los obstáculos y aprovechar plenamente los beneficios de estas tecnologías para mejorar la movilidad urbana y reducir las situaciones que pueden provocar congestión vial.

La combinación de estrategias de agrupamiento con otras metodologías de análisis proporciona una visión integral y sistémica del flujo vehicular en diversos contextos (Zhang et al. 2016; Erdelić et al. 2021; Kim y Mahmassani 2015).

Investigaciones recientes han integrado técnicas de aprendizaje automático y minería de datos para descubrir patrones ocultos y anomalías en el flujo de tráfico (Kamble y Kounte 2020; Sun, Chen y Sun 2019; Shahraki et al. 2022), así como para predecir el tráfico, sin embargo, estos no llegan a adaptarse por completo a diferentes escenarios urbanos y algunos carecen de funcionalidades de análisis temporal que les permitan ajustar las predicciones en función de cambios temporales y patrones emergentes en tiempo real.

Para el análisis de datos, el uso de técnicas de agrupamiento se ha vuelto un elemento esencial para descifrar patrones ocultos y estructuras inherentes en conjuntos de datos diversos. Se han analizado investigaciones previas que utilizan estas técnicas en el análisis de datos y se han identificado algunas similitudes en la forma como realizan el agrupamiento de datos, estas similitudes se pueden expresar mediante la clasificación de estas técnicas, la Figura 1.1 presenta esta clasificación.

Es fundamental distinguir entre el agrupamiento en general y el agrupamiento de trayectorias. Mientras que el agrupamiento agrupa datos según su similitud, el

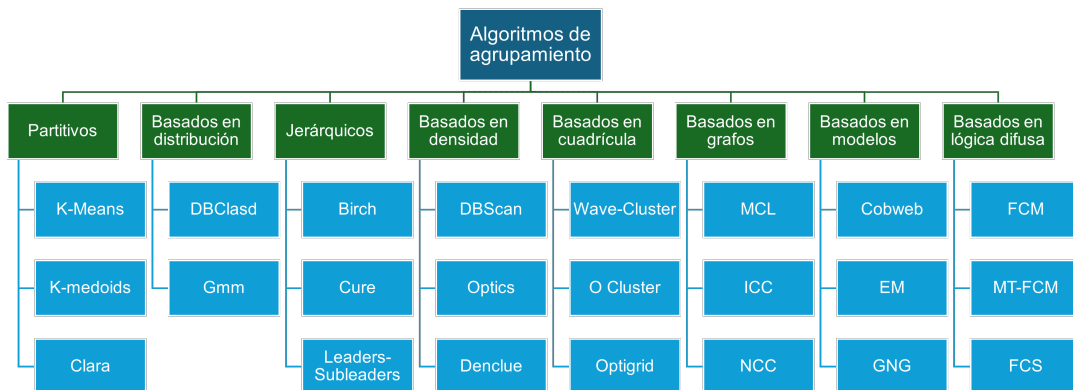


FIGURA 1.1: Clasificación de las técnicas de agrupamiento de trayectorias.

agrupamiento de trayectorias se enfoca en datos que representan movimientos en un espacio determinado. No todas las técnicas de agrupamiento son adecuadas para el análisis de trayectorias debido a la naturaleza secuencial y continua de estos datos. Por lo tanto, la anterior clasificación debe interpretarse con esta distinción en mente, destacando la necesidad de seleccionar métodos que se alineen con las características específicas del conjunto de datos y de los objetivos estudiados.

Se ha identificado que en trabajos anteriores (Zhang, Tangwongsan y Tirthapura 2017; Mu et al. 2023; Anil Meera y Wisse 2021) se han hecho adaptaciones a algunos de estos algoritmos mencionados para que posean la versatilidad de operar tanto con enfoques estáticos como dinámicos. Originalmente, estos algoritmos no estaban diseñados para manejar trayectorias, pero con la adaptación a enfoques dinámicos se abre la posibilidad de desarrollar adaptaciones específicas para trayectorias principalmente utilizadas en entornos de evolución constante, dado que ahora pueden procesar datos de forma continua y dinámica.

En un ámbito específico como es el agrupamiento de flujos de datos, han surgido algoritmos especializados para enfrentar los desafíos asociados con la dinámica y continuidad de la información. Estos algoritmos, están específicamente adaptados para el procesamiento continuo de datos y representan avances significativos en el campo del agrupamiento, proporcionando herramientas eficaces para el análisis en tiempo real de flujos de información dinámicos.

Entre ellos, se destacan los siguientes:

- CluStream, que utiliza microgrupos para resumir y ajustarse dinámicamente a la evolución temporal de los datos, facilitando la detección de tendencias.

- StreamKM++, una extensión del K-means, implementa técnicas de inicialización eficientes para centroides, siendo apropiado para conjuntos de datos en constante movimiento.
- DenStream, por su parte, se enfoca en la identificación de grupos densos mediante el uso de microgrupos, adaptándose en tiempo real a variaciones en la densidad de los datos.
- DyClee, se destaca como un enfoque dinámico para el seguimiento de entornos en evolución, ajustando grupos de manera adaptable a cambios continuos en la información.

Con el uso de estas técnicas se ha permitido no solo la detección eficiente de grupos en tiempo real, sino también la adaptación continua a cambios en la distribución y densidad de los datos. Esto es principalmente a que la aplicación de algoritmos de agrupamiento en flujos de datos proporciona una herramienta esencial para el análisis y la extracción de patrones significativos en entornos dinámicos.

Esta tesis se enfoca en el procesamiento de trayectorias vehiculares. Una trayectoria se define formalmente como una serie cronológicamente ordenada de puntos geoespaciales generados por un objeto en movimiento a través del espacio geográfico. En un contexto de tráfico vehicular, los vehículos a medida que se desplazan por las rutas, generan trayectorias vehiculares.

Uno de los métodos ampliamente utilizados es el análisis de datos de trayectorias que son generadas por vehículos, mediante técnicas de agrupamiento. Estos métodos permiten dividir el flujo vehicular en grupos homogéneos, identificando patrones y comportamientos similares en el tráfico, lo que facilita la identificación de áreas con altos niveles de congestión sean estos persistentes o emergentes.

Las técnicas de agrupamiento resultan invaluable para representar estos flujos de datos de manera efectiva, permitiendo la identificación de patrones de tráfico, la organización de datos en grupos basados en similitudes y la posible predicción de tendencias futuras en el tráfico urbano. Algunas de las técnicas de agrupamiento de trayectorias son T-Optics, Tra-Dbscan, TraClus, Tc-Fdbscan, DyTRA, cada una se basa en características particulares, sin embargo, esta tesis se enfoca en analizar el flujo dinámico del tráfico, adaptándose a las condiciones cambiantes y proporcionando resultados en tiempo real.

En situaciones específicas, como el análisis de tráfico, estos algoritmos de agrupamiento dinámico pueden adaptarse para abordar algunas complejidades inherentes a un entorno en constante cambio. Aunque estos algoritmos de agrupamiento dinámico pueden adaptarse para analizar la evolución temporal de flujos vehiculares, enfrentan desafíos significativos frente al método presentado en esta tesis. Este último se destaca por su capacidad para gestionar eficazmente grandes volúmenes de datos en tiempo real, así como por su precisión en la identificación de patrones complejos de tráfico en entornos urbanos densos y su capacidad para adaptarse a

eventos poco usuales, como obras viales y eventos especiales. En base a estas ventajas, se destaca que los algoritmos dinámicos son los más apropiados para el análisis de flujos de datos, ya que pueden capturar y modelar eficientemente las variaciones y tendencias a lo largo del tiempo, proporcionando así una visión acertada de los patrones en los conjuntos de datos con origen dinámico.

El autor de esta tesis ha considerado el uso de una técnica de agrupamiento dinámico que procese flujos de datos, con énfasis en el tratamiento de datos de trayectorias vehiculares como base para el desarrollo de su propuesta. Esta decisión está fundamentada en la destacada capacidad de estas técnicas para abordar los desafíos específicos presentes en el agrupamiento de flujos de datos. A diferencia de métodos convencionales como los antes mencionados que tienden a enfrentar dificultades en la identificación precisa de patrones complejos en tráfico urbano, el método propuesto en esta tesis se distingue por su capacidad para adaptarse y captar de manera efectiva las variaciones o fluctuaciones repentinas en las condiciones de tráfico.

Se ha observado una dualidad de enfoques en el análisis de flujos de datos de trayectorias, con un enfoque tanto en la identificación de situaciones que pueden provocar congestión del tráfico como en la aplicación de algoritmos de agrupamiento (Azimi y Zhang 2010; Rempe, Huber y Bogenberger 2016a; Shang, Yu y Xie 2022). La convergencia de estas perspectivas representa un área de investigación prometedora, ya que ofrece un enfoque efectivo para el análisis del flujo vehicular, por lo que el autor elabora su propuesta bajo estos enfoques.

Los siguientes artículos, disponibles en bases de datos científicas, analizan patrones de tráfico y han sido una referencia clave para esta investigación:

- Segmented Trajectory Clustering-Based Destination Prediction in IoVs (Wang et al. 2020).
- A Hybrid Method for Traffic State Classification Using K-Medoids Clustering and Self-Tuning Spectral Clustering (Shang, Yu y Xie 2022).
- Categorizing Freeway Flow Conditions by Using Clustering Methods (Azimi y Zhang 2010).
- Spatio-Temporal Congestion Patterns in Urban Traffic Networks (Rempe, Huber y Bogenberger 2016b).

En base a lo dicho anteriormente y al análisis de las anteriores investigaciones se pueden reconocer algunas limitaciones inherentes a estos métodos, ya que podrían influir directamente en la precisión y aplicabilidad de los resultados obtenidos. La comprensión detallada de estas limitaciones es esencial para mejorar la validez y robustez de los diferentes algoritmos. A continuación, se listarán algunas de los aspectos que inciden en la efectividad de algunos algoritmos de agrupamiento de flujo de datos al procesar datos de trayectorias vehiculares:

- Sensibilidad al ruido. Algunas trayectorias con ruido pueden afectar la calidad de los resultados, especialmente en algoritmos de agrupamiento que no manejan bien puntos atípicos.
- Dimensionalidad. Los algoritmos pueden tener dificultades en conjuntos de datos que poseen una alta dimensionalidad, esto puede afectar a su desempeño.
- Visualización de resultados. La visualización y posterior interpretación de los grupos puede ser un reto para los usuarios.
- Requisitos computacionales. Algunos algoritmos, pueden requerir una cantidad considerable de recursos computacionales, limitando su aplicabilidad en entornos con restricciones de hardware, esto se da principalmente en los que utilizan aprendizaje profundo.
- Variabilidad en datos de trayectorias. Las diferencias en la longitud y frecuencia de muestreo de las trayectorias pueden afectar la generalización de los resultados.
- Variabilidad a lo largo del tiempo. La variación en patrones de movimiento a lo largo del tiempo puede ser desafiante para algunos algoritmos, especialmente si no consideran la dinámica temporal de las trayectorias.
- Uso de datos históricos. La disponibilidad de datos históricos puede ser aprovechada en la capacidad de los algoritmos para identificar patrones a largo plazo y entender la evolución de trayectorias.
- Información de la infraestructura vial. La disponibilidad de esta información puede ser importante para que determinados algoritmos consideren el contexto vial en su análisis.

Por otra parte, un insumo importante a la hora de analizar trayectorias es contar con información vial detallada. Las redes de carreteras, representaciones gráficas y estructuradas de las conexiones viales en una región, son fundamentales en este contexto. Estas redes no solo comprenden la disposición física de carreteras y caminos, sino que también incorporan información relevante sobre su jerarquía, tipos de vías, límites de velocidad permitida, interconexiones y elementos clave como semáforos e intersecciones. Su utilidad abarca diversos campos, desde la planificación urbana hasta la gestión del tráfico y la logística. En la planificación urbana, las redes de carreteras han ayudado a diseñar sistemas viales eficientes y sostenibles, incluyendo la planificación de rutas que minimizan eventos que derivan en congestión y el impacto ambiental. En relación a la gestión del tráfico, han permitido comprender y optimizar los flujos vehiculares mediante el conteo de vehículos y la velocidad del tráfico para identificar patrones característicos de situaciones como la congestión y áreas problemáticas.

La importancia de contar con información detallada sobre la infraestructura vial radica en su capacidad para enriquecer el análisis de trayectorias. Al entender la complejidad de la red de carreteras, incluyendo detalles como el tipo de vías y las restricciones de velocidad, se pueden comprender de manera precisa los movimientos registrados por los vehículos. Aunque el método propuesto no procesa redes de carreteras, es importante tener en cuenta que existe una gran diferencia entre analizar una trayectoria en una autopista principal y en una carretera local con semáforos.

La integración de datos de infraestructura vial proporciona una base relevante que permite evaluar cómo el entorno vial influye en los desplazamientos. Verificar y considerar estos datos podría mejorar significativamente los análisis de tráfico, permitiendo una comprensión más profunda de las dinámicas vehiculares y optimizando la gestión del tráfico. Incorporar datos detallados sobre la red vial podría ser una estrategia clave para mejorar la precisión y la eficacia de los análisis de tráfico.

Considerando las restricciones observadas en los algoritmos examinados y la necesidad crucial de realizar un análisis del flujo vehicular para detectar áreas de variabilidad alta en base a la densidad vehicular, se plantea este como un desafío central a abordar.

## 1.2. Objetivos

El objetivo general de esta investigación es desarrollar una técnica adaptativa basada en un algoritmo de agrupamiento de flujo de datos que procese datos de trayectorias e índices vehiculares históricos que permita la identificación de las diversas fluctuaciones en las densidades vehiculares de zonas urbanas.

Para lograr este propósito, se plantean los siguientes objetivos específicos:

- Explorar y analizar diversos algoritmos y métodos de agrupamiento de flujo de datos aplicados al análisis de datos de tráfico urbano.
- Identificar información geoespacial y de tráfico que se puede obtener de las trayectorias de origen vehicular, así como también índices vehiculares históricos que sean relevantes para medir las fluctuaciones en las densidades vehiculares en las ciudades.
- Desarrollar un método que identifique zonas de variabilidad de densidad vehicular que se adapte a las variaciones temporales del tráfico
- Implementar un índice que registre las zonas con densidades altas de manera acumulada en las ciudades y que sirva para la toma de decisiones.
- Analizar el método propuesto mediante casos de estudio donde se identifiquen las características presentes en las tendencias de las variabilidades.

### 1.3. Alcances y limitaciones

La representación adecuada de datos de tráfico y trayectorias generada principalmente por vehículos es un aspecto crucial para la identificación precisa de zonas de altas densidades vehiculares. La elección de una representación adecuada influye directamente en la calidad y eficacia de los métodos utilizados para analizar y detectar áreas con altos niveles de tráfico.

Una de las formas más comunes de representar datos de tráfico es mediante el uso de mapas y visualizaciones geoespaciales. Estos mapas muestran la ubicación y movimiento de los vehículos en tiempo real, lo que permite a los analistas y planificadores de transporte tener una visión general del flujo vehicular en una determinada área urbana (Lou y Cheng 2020). La representación geoespacial de datos facilita la identificación de puntos críticos y áreas de concentración de tráfico, lo que es valioso para la detección de zonas irregulares.

Otra forma de representación es a través del uso de gráficos y diagramas que muestran tendencias y patrones en los datos de tráfico. Estos gráficos pueden incluir información sobre la velocidad, densidad vehicular y tiempos de viaje en diferentes puntos de la ciudad. La representación visual de estos datos permite identificar rápidamente áreas con altas variabilidades recurrentes de la densidad vehicular y evaluar la evolución del tráfico a lo largo del tiempo.

La representación de trayectorias vehiculares es esencial para entender el flujo vehicular en áreas específicas. Estas trayectorias brindan información sobre las rutas más utilizadas y los puntos de mayor concurrencia. Su visualización en mapas y gráficos permite identificar patrones de movilidad y áreas donde los vehículos tienden a detenerse o desacelerar con frecuencia.

Es importante destacar que la representación de datos de tráfico debe ser dinámica y actualizarse en tiempo real para identificar oportunamente zonas con alta variabilidad en la densidad del tráfico. Los sistemas de monitoreo y gestión de tráfico deben recopilar, procesar y mostrar datos en tiempo real, permitiendo a los analistas tomar decisiones y reaccionar rápidamente ante situaciones de congestión. Estas herramientas son esenciales para comprender el flujo vehicular y mitigar el aumento de densidad vehicular en áreas urbanas.

La identificación de zonas con características especiales es crucial para abordar la problemática del aumento de densidad vehicular de manera efectiva. El uso de técnicas como el agrupamiento, (Han, Kamber y Tung 2001; Choong et al. 2016), es útil para detectar zonas de congestión recurrente y áreas con variaciones dinámicas en la densidad vehicular.

El agrupamiento dinámico se distingue por su capacidad para identificar patrones cambiantes de congestión en el tráfico vehicular, este enfoque captura y entiende las fluctuaciones del tráfico a lo largo del tiempo, siendo esencial para detectar cambios durante diferentes horas del día o días específicos.

Al aplicar estas técnicas se pueden analizar factores como la infraestructura vial, densidad de población, eventos especiales y disponibilidad de transporte público. Esto proporciona una comprensión más detallada del tráfico urbano y orienta el diseño de soluciones específicas para cada situación.

Aunque valioso para comprender la movilidad urbana, el uso de técnicas de agrupamiento enfrenta desafíos como la eficiencia computacional y la escalabilidad, requiriendo algoritmos eficientes y técnicas de procesamiento distribuido para garantizar resultados precisos y oportunos.

La investigación presenta las siguientes limitaciones:

- La manipulación y análisis de grandes volúmenes de datos de tráfico en tiempo real es compleja y requiere de recursos computacionales significativos, por lo que se ha utilizado un entorno centralizado para el procesamiento de la técnica.
- Se requiere a priori información del tráfico histórico como estrategia para mejorar la precisión de los patrones habituales considerando que la identificación de zonas de alta variabilidad vehicular debe ser en tiempo real.
- La precisión y fiabilidad de los datos de tráfico en tiempo real pueden ser afectadas por factores externos, como fallas en los sensores de monitoreo o errores en la transmisión de datos, lo que podría generar resultados inexactos en la identificación de zonas con diversas variabilidades.
- La implementación de estrategias de mitigación en tiempo real basadas en datos de tráfico puede enfrentar obstáculos regulatorios y de privacidad, especialmente en lo que respecta a la recopilación y el uso de datos de localización de vehículos, lo que podría limitar la efectividad de la estrategia de procesamiento adoptada.

## 1.4. Contribuciones

Esta investigación presenta los siguientes aportes significativos:

- El desarrollo y fundamentación de una técnica adaptativa basada en un algoritmo de agrupamiento dinámico que procesa datos de trayectorias de origen vehicular e índices vehiculares históricos permitiendo la identificación de zonas con variabilidades altas en densidades del tráfico vehicular.
- La implementación de un algoritmo de agrupamiento dinámico para la identificación de variabilidades altas en tiempo real, que puede ser aplicado en Sistemas Inteligentes de Transportación.
- El desarrollo de un método que identifica variabilidades altas en la densidad vehicular con capacidad para adaptarse a las variaciones temporales del tráfico vehicular.



- La propuesta de un índice que permite registrar las zonas con variabilidades altas que sirva para la toma de decisiones.

## 1.5. Publicaciones

Esta investigación está avalada por las siguientes publicaciones científicas del autor:

### 1.5.1. Artículos de Revistas

- Gary Reyes, Roberto Tolozano-Benites, Laura Lanzarini, César Estrebou, Aurelio F. Bariviera, y Julio Barzola-Monteses. «Method for the Identification and Classification of Zones with Vehicular Congestion». *ISPRS International Journal of Geo-Information*, 2024.
  - Impact Factor: 3,4 Cuartil 1 Computers in Earth Sciences.
  - SJR: 0,74 Cuartil 1 Computers in Earth Sciences.
- Gary Reyes, Roberto Tolozano-Benites, Laura Lanzarini, César Estrebou, Aurelio F. Bariviera, y Julio Barzola-Monteses. «Methodology for the Identification of Vehicle Congestion Based on Dynamic Clustering». *Sustainability* 15(24):16575. doi: 10.3390/su152416575, 2023.
  - Impact Factor: 3,9.
  - SJR: 0,664. Cuartil 1 Geography, Planning and Development.
- Gary Reyes, Christopher Crespo, Oscar León-Granizo, Wellington Bazán, y Richard Horta. «Propuesta de método de extracción de ubicaciones georreferenciales de una red de carreteras para el análisis de trayectorias GPS». *Investigación, Tecnología e Innovación* 14(16):1-15. doi: 10.53591/iti.v14i16.1465, 2022.
  - Indexada en Latindex.
- Gary Reyes, José Roldán, Angélica Macias, Francisco Cordova, y Oscar León-Granizo. «Identificación de patrones de congestión vehicular utilizando algoritmos de agrupamiento de trayectorias basados en densidad». *Investigación, Tecnología e Innovación* 14(17):1-15. doi: 10.53591/iti.v14i17.1473, 2022.
  - Indexada en Latindex.
- Gary Reyes, Laura Lanzarini, Cesar Estrebou, y Aurelio Fernandez Bariviera. «Dynamic Grouping of Vehicle Trajectories». *Journal of Computer Science and Technology* 22(2):e11. doi: 10.24215/16666038.22.e11, 2022.
  - Impact Factor: 0,79 Cuartil 4 Artificial Intelligence.
  - SJR: 0,18 Cuartil 4 Artificial Intelligence.

- Gary Reyes, Laura Lanzarini, Waldo Hasperué, y Aurelio F. Bariviera. «Proposal for a Pivot-Based Vehicle Trajectory Clustering Method». *Transportation Research Record* 2676(4):281-95. doi: 10.1177/03611981211058429, 2022.
  - Impact Factor: 1,7 Cuartil 2 Civil and Structural Engineering.
  - SJR: 0,62 Cuartil 2 Civil and Structural Engineering.
- Gary Reyes, Laura Lanzarini, Waldo Hasperué, y Aurelio F. Bariviera. «GPS Trajectory Clustering Method for Decision Making on Intelligent Transportation Systems». *Journal of Intelligent & Fuzzy Systems*. 38(5):5529-35. doi: 10.3233/JIFS-179644, 2020.
  - Impact Factor: 2 Cuartil 2 Engineeringe (miscellaneous).
  - SJR 0,37. Cuartil 2 Engineeringe (miscellaneous).

### 1.5.2. Capítulos de libros y congresos con referato

- Gary Reyes, Laura Lanzarini, César Estrebou, Aurelio Bariviera, y Victor Maquilón. «Evaluation of a Grid for the Identification of Traffic Congestion Patterns». Pp. 277-90 en *Technologies and Innovation, Communications in Computer and Information Science*, Cham: Springer Nature Switzerland, 2023.
- Gary Reyes, Laura Lanzarini, César Estrebou, y Aurelio Bariviera. «Data Stream Processing Method for Clustering of Trajectories». Pp. 151-63 en *Technologies and Innovation*. Vol. 1658, *Communications in Computer and Information Science*, Cham: Springer International Publishing, 2022.
- Gary Reyes, Laura Lanzarini, César Armando Estrebou, y Víctor Maquilón. «Vehicular Flow Analysis Using Clusters». En XXVII Congreso Argentino de Ciencias de la Computación (CACIC), 2021.

### 1.5.3. Participaciones en Workshops

- Waldo Hasperué, Estrebou César, Camele Genaro, Rucci Enzo, Ronchetti Franco, Eguren Sebastián, Castillo David, Reyes Zambrano Gary, Lanzarini Laura, y Fernandez Bariviera Aurelio. «Procesamiento inteligente de grandes volúmenes de datos». En XXVI Workshop de Investigadores en Ciencias de la Computación (WICC 2024). UNP. Chubut. Puerto Madryn. Abril 2024
- Waldo, Hasperué, Estrebou César, Camele Genaro, Rucci Enzo, Ronchetti Franco, Castillo David, Reyes Zambrano Gary, Lanzarini Laura, y Fernandez Bariviera Aurelio. «Sistemas inteligentes en el uso de aplicaciones de bioinformática y sistemas embebidos». En XXV Workshop de Investigadores en Ciencias de la Computación (WICC 2023). UNNOBA. Junín. Prov. Bs.As. Abril 2023

- Hasperué, W., C. Estrebou, G. Camele, P. López, M. Peña, G. Reyes Zambrano, L. Lanzarini, A. Fernandez Bariviera, y M. Cerrada. «Procesamiento inteligente de la información. Aplicaciones en bioinformática, trayectorias vehiculares, mantenimiento preventivo industrial y sistemas embebidos». En XXIV Edición del Workshop de investigadores en Ciencias de la Computación, 2022.
- Hasperué, W., C. Estrebou, G. Camele, P. López, P. Jimbo Santana, G. Reyes Zambrano, L. Lanzarini, y A. Fernandez Bariviera. 2021. «Procesamiento inteligente de grandes volúmenes de información y de flujos de datos». En Proceedings of the XXIII Workshop de Investigadores En Ciencias de La Computación (WICC 2021, Chilecito, La Rioja), 2021.
- Laura Lanzarini, Waldo Hasperué, César Armando Estrebou, Augusto Villa Monte, Patricia Rosalía Jimbo Santana, G. Reyes Zambrano, Genaro Camele, P. López, Julieta Pilar Corvi, y Aurelio Fernández Bariviera. «Sistemas inteligentes: aplicaciones en minería de datos y big data». En XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz), 2020.
- Laura Lanzarini, Waldo Hasperué, Augusto Villa Monte, Patricia Jimbo Santana, Gary Reyes Zambrano, Julieta Pilar Corvi, Aurelio Fernández Bariviera, y José Ángel Olivas Varela. «Minería de datos, minería de textos y Big Data». En XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan), 2019.

#### 1.5.4. Publicaciones complementarias

A continuación, se listan otras publicaciones del autor relacionadas con el procesamiento y análisis de trayectorias, que contribuyen de manera complementaria a esta investigación:

- Gary Reyes, Vivian Estrada, Roberto Tolozano-Benites, y Victor Maquilón. «Batch Simplification Algorithm for Trajectories over Road Networks». ISPRS International Journal of Geo-Information 12(10):399. doi: 10.3390/ijgi12100399, 2023.
- Gary Reyes, Victor Maquilón, y Vivian Estrada. «Relationships of Compression Ratio and Error in Trajectory Simplification Algorithms». Pp. 140-55 en Technologies and Innovation, Communications in Computer and Information Science. Cham: Springer International Publishing, 2021.
- Gary Reyes, Vivian Estrada. «Comparison analysis on noise reduction in GPS trajectories simplification». En LACCEI Inc, 2021
- Gary Reyes, «GPS trajectory compression algorithm». Pp. 57-69 en Computer and Communication Engineering: First International Conference, ICCCE 2018, Guayaquil, Ecuador, October 25–27, 2018, Proceedings 1. Springer, 2019.

- Gary Reyes, Antonio Cedeño. «Algoritmo GR-B para la compresión de trayectorias vehiculares». Revista Espacios, 2018.
- Gary Reyes, Raul Nassib, y Raul Hidalgo Veliz. «Vehicular Swept Path Analysis Using K-Means Algorithm». International Journal of Applied Engineering Research 11:9536-38, 2016.

## 1.6. Organización de la Tesis

La tesis está estructurada en cinco capítulos:

En el capítulo 1 se presenta la introducción la cual contempla la situación problemática que constituye la base de la investigación, así como sus objetivos alcances y contribuciones.

En el capítulo 2 se presenta el estado del arte que contiene los fundamentos teóricos de la investigación, se realiza un análisis de los principales algoritmos de agrupamiento de trayectorias resaltando sus características, principales ventajas y desventajas.

En el capítulo 3 se explora el desarrollo de las investigaciones que identifican patrones analizando flujos vehiculares basados en agrupamientos, que han dado origen a cada uno de los módulos y sub-módulos del método propuesto.

En el capítulo 4 se describe la propuesta de un método de agrupamiento dinámico de dos niveles que permite la identificación de zonas con diversas variabilidades en la densidad vehicular. Además, se analizan casos de estudio, correspondientes a distintas ciudades con condiciones de tráfico variables para identificar las características recurrentes en las distintas situaciones respecto a los aspectos de desplazamiento, densidad y variabilidad.

Finalmente, el Capítulo 5 presenta las conclusiones derivadas del estudio y posibles direcciones para futuras investigaciones.

## Capítulo 2

# Agrupamiento de trayectorias

**E**n este capítulo, se realiza un repaso breve pero integral de los conceptos ampliamente utilizados en el análisis de trayectorias espacio-temporales. Se aborda la esencia de términos clave y enfoques relevantes, proporcionando una comprensión más profunda de las bases teóricas que sustentan el estudio. Es importante destacar que, dentro de este repaso, se seleccionaron cuidadosamente ciertos conceptos que desempeñan un papel fundamental en la conceptualización y formulación del método propuesto en esta tesis. La elección de estos elementos se basa en su pertinencia y contribución directa a la investigación en cuestión, marcando una conexión significativa entre la revisión teórica y la aplicación práctica en el contexto de este trabajo de investigación.

### 2.1. El Sistema de Posicionamiento Global (GPS)

El Sistema de Posicionamiento Global, también conocido como GPS, es un sistema de navegación por satélite desarrollado originalmente con fines militares que permite determinar la ubicación precisa de objetos o personas en cualquier punto del planeta. Funciona mediante una red de 24 satélites que orbitan la Tierra, cada uno equipado con relojes atómicos de alta precisión, como se puede observar en la Figura 2.1. Al captar las señales al menos cuatro satélites, un receptor GPS puede calcular su latitud, longitud y altitud en base al tiempo que tardan en llegar estas señales (Dziadczyk, Zabierowski y Napieralski 2007; Villegas 2021).

Hoy en día, el GPS tiene aplicaciones civiles muy diversas, como localizar teléfonos móviles, navegar en automóviles, barcos y aviones, realizar trabajo de topografía y muchas más con aplicaciones en diferentes campos como los industriales, comerciales y de investigación (Upreti y Kumar 2008; Mahmood Hussien et al. 2023). Su precisión de unos pocos metros ha hecho de este sistema una herramienta indispensable en actividades que requieren ubicación y seguimiento (Abulude, Akinnusotu y Adeyemi 2015). Con el desarrollo de técnicas mejoradas de posicionamiento, es probable que la gama de usos del GPS siga en aumento, potenciando su valor para la sociedad moderna.

El GPS no se limita a proporcionar coordenadas de latitud, longitud y altitud para ubicar puntos en la superficie terrestre; esta información de posicionamiento

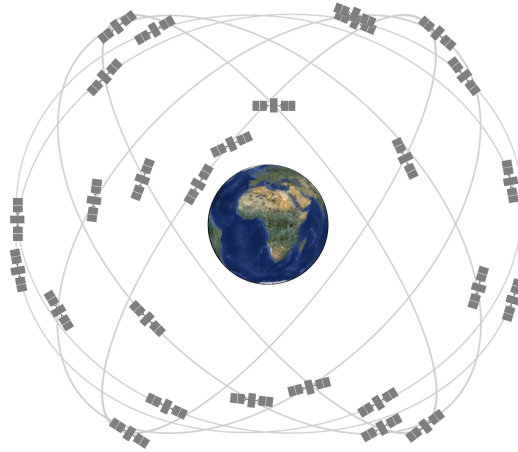


FIGURA 2.1: Órbitas de los satélites GPS. Fuente: <https://www.gps.gov/>

geográfico también resulta fundamental para el rastreo de objetos en órbita y la navegación espacial.

## 2.2. Las trayectorias y sus campos de investigación

Una trayectoria espacial se define como la representación de la ruta tridimensional seguida por un objeto en movimiento a través del espacio geográfico, enfocándose exclusivamente en la descripción de la ubicación espacial en puntos específicos. Por otro lado, una trayectoria espacio-temporal se define formalmente como una serie cronológicamente ordenada de ubicaciones geoespaciales generadas por un objeto en movimiento a través del espacio geográfico. Cada ubicación consiste en un conjunto de coordenadas espaciales y una marca de tiempo, y puede ser denotado como  $P = \{(x, y), t\}$ , donde  $x$  e  $y$  son las coordenadas espaciales y  $t$  es el tiempo. El estudio de trayectorias vehiculares se puede abordar analizando puntos individuales para captar aspectos momentáneos, segmentos de trayectorias para discernir patrones de movimiento a lo largo de rutas y trayectorias completas para obtener una visión global del recorrido y los comportamientos generales del tráfico. No obstante, la definición más comúnmente utilizada por los investigadores para una trayectoria completa es la que se representa como una secuencia  $TR = (P_1, P_2, \dots, P_n)$  (Yu et al. 2019b; Mazimpaka y Timpf 2016). En la propuesta de esta tesis, se adoptará esta misma formulación.

Las trayectorias espacio-temporales representan los movimientos de los objetos a través de un espacio físico, capturados a intervalos de tiempo discretos mediante tecnologías de geolocalización.

El concepto de trayectorias espacio-temporales es fundamental para el objetivo de analizar y modelar patrones de movimiento de distintos fenómenos en un espacio geográfico.

Durante el análisis e interpretación de trayectorias espacio-temporales, puede obtenerse información importante sobre los comportamientos y patrones de movimiento de los objetos monitorizados. El procesamiento y modelado adecuado de estas trayectorias permiten responder preguntas como cuáles son las rutas más comúnmente tomadas, cuáles son los destinos principales, cómo varían las velocidades, cuáles son los factores que influyen en los cambios de dirección, etc. Esta información resulta clave en una amplia variedad de aplicaciones prácticas, que incluyen planificación urbana, transporte inteligente, protección ambiental y análisis de riesgos, entre muchas otras (Reyes et al. 2020; Zheng 2015).

Es fundamental distinguir entre trayectorias espaciales y espacio-temporales. Mientras que las trayectorias espaciales se centran exclusivamente en la ubicación geoespacial del objeto a través del tiempo, las trayectorias espacio-temporales incorporan la dimensión temporal, brindando una perspectiva más holística del movimiento. En otras palabras, una trayectoria espacial describe la ruta física de un objeto en un espacio tridimensional, mientras que una trayectoria espacio-temporal añade la dimensión temporal, permitiendo no solo seguir la posición espacial sino también capturar cómo esa posición evoluciona a lo largo del tiempo.

Un término alternativo a las trayectorias espaciales es la expresión trayectorias GPS. Este término se refiere a la secuencia de puntos geoespaciales cuya ubicación se determina mediante el Sistema de Posicionamiento Global. Mientras que las trayectorias espaciales abordan la noción general de movimientos en un espacio geográfico, las trayectorias GPS resaltan que la precisión de la geolocalización es proporcionada por el sistema GPS. Este término se lo suele emplear en diversos campos, como la navegación, la planificación urbana, la gestión ambiental y otras áreas que se benefician de la comprensión desde un contexto geoespacial.

A lo largo de esta tesis, el término “trayectorias” será empleado específicamente para referirse a las “trayectorias espacio-temporales”, resaltando la inclusión de la dimensión temporal en la descripción de los movimientos de objetos en el espacio geográfico, destacando la consideración simultánea de la ubicación y el tiempo en el análisis de desplazamientos. En un contexto más específico, cuando se haga referencia exclusivamente a las “trayectorias espaciales”, se utilizará este término de manera explícita para diferenciarlas de aquellas que incorporan la dimensión temporal.

En este contexto, es fundamental destacar que cualquier técnica aplicada a las “trayectorias espaciales” puede extrapolarse y aplicarse a las “trayectorias espacio-temporales” debido a que comparten una estructura similar, añadiendo únicamente la dimensión temporal. Sin embargo, no todas las técnicas desarrolladas para las “trayectorias espacio-temporales” pueden ser directamente aplicadas a las “trayectorias espaciales”, ya que estas últimas carecen de la dimensión temporal, resaltando la necesidad de tener en cuenta las variaciones específicas al elegir y aplicar enfoques analíticos.

### 2.2.1. Categorización general de las trayectorias

Esta sección aborda la taxonomía de las trayectorias en función del origen de las movilidades, ofreciendo una comprensión generalizada de los distintos contextos y proporcionando un análisis sobre las características basadas en la clasificación de (Zheng 2015).

- Movilidad de personas. Las trayectorias de movimiento de personas registran desplazamientos individuales, ya sea por motivos recreativos y concretos cuando se organizan viajes, o bien de forma involuntaria en su vida cotidiana. Estos corresponden a datos que revelan patrones de comportamiento humano. Los seres humanos han registrado sus desplazamientos geográficos desde hace mucho tiempo, tanto de forma intencional mediante GPS para planificar viajes, como de forma pasiva a través de los teléfonos móviles que van dejando huella en las torres de telefonía a las que se conectan.
- Movilidad de vehículos de transporte. Las trayectorias de vehículos de transporte corresponden a itinerarios físicos de traslados individuales y colectivos. Estas trayectorias están sujetas a la infraestructura urbana y la logística local de las ciudades proporcionando información aplicable a nivel práctico principalmente para la planificación urbana o de servicios. En la actualidad se cuenta con gran cantidad de medios de transporte equipados con sistemas de geolocalización, como taxis, autobuses o barcos, que generan valiosos datos sobre recorridos y movimientos que tienen múltiples aplicaciones en ámbitos como la gestión del tráfico.
- Movilidad de animales. Las trayectorias de animales siguen patrones de comportamiento de la vida silvestre. Estos responden a patrones de comportamiento natural de las especies para adaptarse a su hábitat y ofrecen datos sobre la relación de la fauna con distintos entornos. La comunidad científica registra y analiza los movimientos y migraciones de especies animales buscando extraer conclusiones sobre su comportamiento y hábitat.
- Movilidad de fenómenos naturales. Estas trayectorias reflejan alteraciones ambientales y climatológicas de impacto a gran escala. Estos datos son útiles para entender y mitigar desastres, así como para estudios ecológicos. Del mismo modo, se monitorizan fenómenos y disturbios naturales como huracanes y corrientes oceánicas, con el fin de entender los cambios medioambientales, prevenir desastres, y proteger los ecosistemas.

### 2.2.2. Campos de investigación activos de trayectorias

Es importante destacar que cuando se habla de análisis de trayectorias no se limita solo al tráfico vehicular. En la actualidad, las trayectorias representan un área de investigación activa que abarca diversos campos y contribuye significativamente



al avance del conocimiento y la mejora de servicios en distintas disciplinas. Algunas áreas de investigación que se apoyan en el análisis de trayectorias son:

- Salud. En el ámbito médico, el análisis de trayectorias se utiliza en el monitoreo remoto de pacientes y personal médico mediante dispositivos y sensores portables con GPS. Esta aplicación no solo mejora los tiempos de respuesta en emergencias, sino que también posibilita nuevos estudios sobre movilidad humana, la efectividad de tratamientos y la predicción de brotes epidémicos.
- Seguridad. La seguridad se beneficia del análisis de trayectorias mediante el uso del posicionamiento global, aumentando la efectividad de respuestas ante desastres al localizar rápidamente a los afectados. Además, el modelado de amenazas utiliza trayectorias proyectadas para evaluar y planificar contramedidas ante eventos como tsunamis o deslaves.
- Clima. En campos como la climatología y geofísica, las trayectorias históricas de huracanes y el movimiento milimétrico anual de placas tectónicas, registradas vía satélite y coordenadas GPS, respectivamente, son esenciales. Estas contribuyen a mejores predicciones estacionales, modelos de cambio climático global y la comprensión de la dinámica terrestre.
- Infraestructura. En el campo de la infraestructura, el mapeo y modelado 3D con precisión centimétrica posibilitados por GPS son fundamentales. Esta tecnología facilita el diseño, construcción y mantenimiento de carreteras, puentes, redes eléctricas, hidráulicas y de comunicaciones (Varghese, Unnikrishnan y Poulouse 2013).
- Transporte. La movilidad y el transporte se ven transformados gracias a los sistemas globales y regionales de navegación de trayectorias por satélite. Estos se han vuelto indispensables para vehículos, optimización de rutas y flotas, así como para el control de tráfico aéreo y marítimo. Los datos agregados sobre localización y movimiento de vehículos alimentan modelos de gestión de tráfico que buscan mitigar la congestión en ciudades mediante estrategias de transporte inteligente (Markovi et al. 2018; Tang et al. 2018; Hu et al. 2018; Bathaee et al. 2018). Además, la gestión adecuada del tráfico también puede conducir a mejoras ambientales (Wang et al. 2021).

Como se puede observar, los datos de trayectorias registradas aportan inteligencia invaluable para la investigación interdisciplinaria. La expansión en la precisión, cobertura y asequibilidad de estas soluciones de posicionamiento, ayuda a impulsar beneficios en prácticamente todos los ámbitos de la actividad humana moderna.

### 2.3. Minería de datos para el análisis de trayectorias

La minería de trayectorias es un campo emergente dedicado al análisis de conjuntos de datos masivos que capturan el movimiento de objetos a través del tiempo y el espacio. Utilizando sofisticadas técnicas computacionales, esta disciplina extrae patrones ocultos, tendencias, detecta anomalías, reduce la incertidumbre y otras ideas valiosas sobre el comportamiento dinámico de entidades en constantes desplazamientos (Zheng 2015).

Las trayectorias reflejan una descripción dinámica sobre el mundo en movimiento. Es entonces que la minería de estos rastros es crucial para comprender mejor el funcionamiento de sistemas complejos e interconectados como ciudades, ecosistemas o la atmósfera. Este análisis permite predecir a dónde se dirigen distintos actores y cómo interactúan entre ellos, facilitando la planificación y coordinación en distintos ámbitos.

Desde optimizar redes de transporte hasta mitigar el impacto ambiental de huracanes, la minería de datos aplicada a trayectorias revoluciona la capacidad para administrar dinámicas en constante cambio. Detectando similitudes entre trayectorias aparentemente inconexas es posible revelar vínculos ocultos y oportunidades para la colaboración entre disciplinas.

Sin embargo, importantes obstáculos permanecen a la hora de aplicar estas técnicas. Integrar variables contextuales de distintos tipos, el escalado de conjuntos de datos masivos o preservar la privacidad son sólo algunos de los desafíos.

Con el objetivo de contextualizar el enfoque del presente trabajo, se presenta una rápida visión general del campo de la minería de datos de trayectorias, tomando la Figura 2.2 como referencia esencial para mostrar las principales áreas de estudio. Este vasto dominio puede dividirse en métodos primarios y secundarios, cada uno desempeñando un papel fundamental en el análisis y la comprensión de la información temporal (Mazimpaka y Timpf 2016).

Dentro de los métodos primarios, se puede observar cómo la agrupación y clasificación de trayectorias permite categorizar las bases de datos masivas de acuerdo a propiedades intrínsecas del movimiento registrado. Por ejemplo, mediante el agrupamiento se identifican conjuntos de trayectorias similares, mientras que con la clasificación se asignan etiquetas como “vehículos lentos” o “peatones rápidos”. Esto permite descubrir patrones subyacentes y sienta las bases para investigaciones posteriores, al revelar grupos de comportamientos similares entre multitudes de trayectorias aparentemente inconexas.

Complementando estos métodos primarios, se presentan técnicas secundarias que ahondan en patrones espacio-temporales tanto dentro como entre categorías. En estas técnicas se destacan las que detectan configuraciones frecuentes, la identificación de anomalías que resaltan casos atípicos, y la predicción anticipada de comportamientos futuros en base a tendencias históricas.

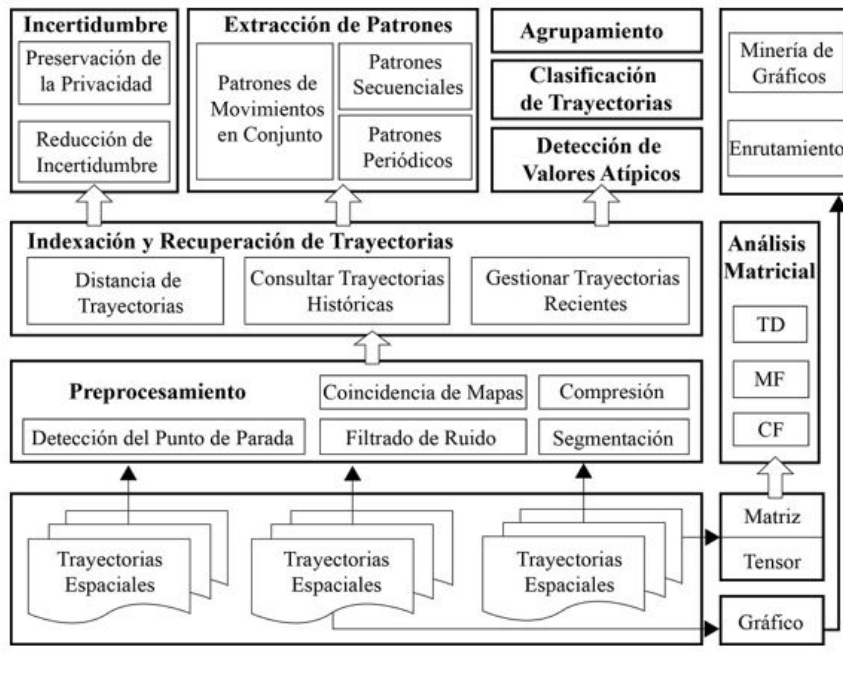


FIGURA 2.2: Principales áreas de estudio en la minería de datos para el análisis de trayectorias. Fuente: (Zheng 2015)

Si bien se enfatiza la utilidad individual de estas técnicas, se considera vital resaltar cómo su aplicación combinada potencia exponencialmente su valor analítico. Por ejemplo, mediante la clasificación de trayectorias de conductores según edad, y posterior minería de patrones etarios, los planificadores de tránsito pueden personalizar señales de tráfico considerando las distintas conductas generacionales.

Así, más allá de sus capacidades individuales, comprender estas metodologías como un conjunto integrado de herramientas para desentrañar la complejidad espacio-temporal, destaca su enorme potencial para informar decisiones en prácticamente cualquier dominio dinámico moderno.

### 2.3.1. Preprocesamiento de trayectorias

El preprocesamiento de los datos de trayectorias constituye una etapa fundamental previa a la aplicación de métodos de minería de datos. Implica una serie de tareas destinadas a transformar los datos crudos de localización en secuencias de trayectorias listas para ser analizadas. Esto incluye la limpieza de valores anómalos, la interpolación temporal para obtener una frecuencia de muestreo uniforme, la segmentación para dividir trayectorias largas en partes significativas, y la generalización espacial para reducir el ruido manteniendo los rasgos característicos. Un buen preprocesamiento filtra errores, resalta la información relevante y prepara los datos para la obtención eficiente de patrones y conocimiento útil sobre el movimiento y comportamiento de las entidades móviles. La investigación en esta área busca

desarrollar técnicas cada vez más automatizadas y parametrizables para adaptarse a las necesidades y particularidades de diversos conjuntos de datos de trayectorias.

### Segmentación de trayectorias

La segmentación de trayectorias consiste en particionar una trayectoria en sub-trayectorias significativas según algún criterio específico. Es un paso clave de pre-procesamiento cuando se pretende analizar trayectorias extensas para extraer patrones o aplicar técnicas de minería de datos. Permite enfocarse en partes de interés en lugar de hacerlo en trayectorias completas, a menudo excesivamente largas y variables. Un ejemplo de esto puede visualizarse en la Figura 2.3 en la que se observa que una extensa trayectoria puede estar compuesta por varios segmentos más pequeños.

Existen distintos enfoques, desde ventanas deslizantes sean del tipo espaciales o temporales (Etemad et al. 2020; Chen, Ji y Wang 2014), hasta métodos que detectan puntos críticos como paradas, giros o cambios de velocidad (Yuan et al. 2013). La selección del método de segmentación y sus parámetros depende en gran medida de la aplicación y los objetivos buscados. Un buen balance evita la sobresegmentación y la subsegmentación. Un exceso de segmentos dificulta identificar partes significativas, mientras que muy pocos no resaltan adecuadamente los cambios de comportamiento. Se requiere más investigación para desarrollar técnicas de segmentación versátiles, parametrizables y más guiadas por la semántica de las actividades subyacentes detectable en los datos.

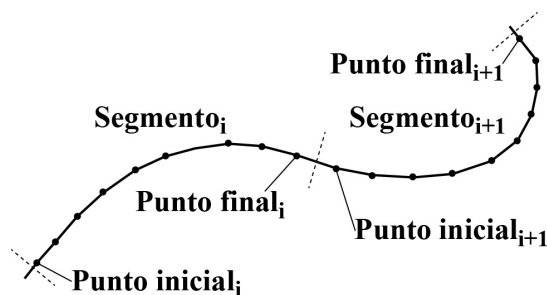


FIGURA 2.3: Ejemplo de segmentación de una trayectoria. Fuente: (Yao et al. 2023)

Dentro de la literatura, se pueden encontrar investigaciones que han abordado la segmentación de trayectorias como una etapa previa al proceso de agrupamiento. Por ejemplo, (Lee, Han y Whang 2007) han propuesto un enfoque que utiliza la segmentación para mejorar el rendimiento de los algoritmos de agrupamiento y obtener clústeres de mayor calidad. De manera similar, (Mao et al. 2017; Yuan et al. 2017) también presentaron métodos que incorporan la segmentación de trayectorias antes de la agrupación, lo que ha demostrado ser beneficioso para el análisis de trayectorias.

## Interpolación

La interpolación espacio-temporal es una técnica clave en el preprocesamiento de datos de trayectorias. Se utiliza para reconstruir posiciones en momentos específicos a partir de una secuencia discreta de ubicaciones muestreadas. Su aplicación permite obtener trayectorias uniformemente muestreadas, facilitando su análisis posterior y comparación. Existen múltiples métodos de interpolación, desde simples aproximaciones lineales hasta modelos físicos del movimiento o splines complejos (Etemad et al. 2019; Su et al. 2023). La selección del método adecuado depende de factores como la regularidad del movimiento o el nivel de suavizado requerido. Un reto importante es lograr un equilibrio entre un exceso de simplificación, que distorsiona los datos originales, y un sobreajuste al ruido, que no filtra adecuadamente la trayectoria subyacente. Se requieren técnicas más robustas frente a diferentes tipos de movimiento y parametrizaciones que permitan adaptar el grado de reconstrucción necesario según la aplicación. La interpolación inteligente de trayectorias continúa siendo un área activa de investigación.

## Detección de valores atípicos

La presencia de valores atípicos es común en conjuntos de datos de trayectorias debido a errores en la captura de posicionamiento o comportamientos anómalos de algunas entidades móviles. Su no tratamiento distorsiona el análisis del movimiento típico y dificulta descubrir patrones significativos. La detección de estos valores atípicos permite filtrar o corregir aquellos puntos de trayectoria que se desvían significativamente del movimiento esperado. Se aplican pruebas estadísticas, agrupamiento espacio-temporal o aprendizaje supervisado para identificar observaciones que no se ajustan a modelos del movimiento normal. Sin embargo, los valores atípicos no siempre implican errores, por lo que su eliminación puede suprimir eventos extraordinarios de interés. Se requieren técnicas más avanzadas que distingan entre anomalías y comportamientos infrecuentes válidos. La investigación en esta área busca mejorar la precisión de detección, reducir falsos positivos y brindar herramientas de análisis exploratorio de valores atípicos antes de descartarlos.

### 2.3.2. Clasificación de trayectorias

La clasificación de trayectorias consiste en asignar categorías o etiquetas a trayectorias individuales según sus características, permitiendo identificar distintos tipos de movimiento o comportamiento. Se emplean diversos algoritmos de aprendizaje supervisado que aprenden un modelo de clasificación a partir de un conjunto previamente etiquetado, para luego aplicar dicho modelo a nuevas trayectorias (Endo et al. 2016).

Entre los desafíos presentes en el proceso de clasificación está la selección y extracción de características con alto poder discriminativo, tales como origen-destino, duración, velocidad promedio, cambios de dirección, etc (Yang et al. 2018). También

es clave elegir adecuadamente el clasificador considerando factores como la precisión, interpretabilidad y escalabilidad frente al tamaño y dimensionalidad de los datos (Arrais De Freitas et al. 2021). Algunas áreas emergentes incluyen métodos profundos para aprendizaje automático de características, técnicas de votación para combinar múltiples clasificadores, y enfoques generativos que modelan cada clase para mejor comprensión del comportamiento. En la actualidad, las investigaciones centradas en la clasificación de trayectorias buscan lograr clasificadores más flexibles y adaptables frente a diversos tipos de movimiento (Leite Da Silva, May Petry y Bogorny 2019).

### 2.3.3. Extracción de patrones en trayectorias

La obtención de patrones permite descubrir comportamientos de **movimiento frecuentes, recurrentes o emergentes** en conjuntos de trayectorias. Algunos autores mencionan las siguientes tres categorías: patrones de movimientos en conjuntos, patrones secuenciales y patrones periódicos. Mediante estas técnicas se extraen patrones espacio-temporales caracterizando rutas, horarios, localizaciones críticas y transiciones entre actividades. Esto posibilita una mejor comprensión de flujos, interacciones y actividades subyacentes. Los retos incluyen manejar adecuadamente la complejidad, variabilidad e incertidumbre de los datos. Es necesario evitar sobreajuste y distinguir patrones genuinos de casuales. También es clave lograr un balance entre simplicidad, para facilitar interpretación humana, y eficacia en la captura de datos con la capacidad de indentificar particularidades que describan adecuadamente subculturas o subpoblaciones. Algunas de estas categorías exploran técnicas emergentes como integración con semántica contextual, aprendizaje profundo o procesamiento en paralelo distribuido, para escalar a conjuntos de datos masivos preservando utilidad de los patrones minados.

A continuación se explica brevemente las categorías de patrones analizados utilizando técnicas de minería de trayectorias:

#### Patrones de movimientos en conjunto

El descubrimiento de patrones de movimiento en un conjunto de trayectorias permite identificar comportamientos y flujos característicos de las entidades móviles analizadas. Existe una amplia diversidad de tipos de patrones, desde individuales como paradas frecuentes o caminos preferidos, hasta colectivos como convergencia, encuentros o dispersiones desde/hacia localizaciones críticas. Su detección requiere de técnicas flexibles ante la inherente variabilidad e incertidumbre del movimiento real. Es clave encontrar un equilibrio adecuado entre patrones lo suficientemente específicos para ser informativos, pero también lo suficientemente generales para capturar comportamientos comunes, considerando posibles excepciones. Interpretar adecuadamente el significado y utilidad de los patrones sigue siendo uno de los principales retos. Se continúa trabajando para desarrollar algoritmos más eficientes

y parametrizables, que integren información contextual y sean explicables para facilitar validación experta cuando sea necesario. Un ejemplo de diferentes patrones que se pueden identificar según la distribución de las trayectorias se puede observar en la Figura 2.4, los patrones mayormente estudiados son los clasificados como bandadas (a), los convoyes (b), y los enjambres (c).

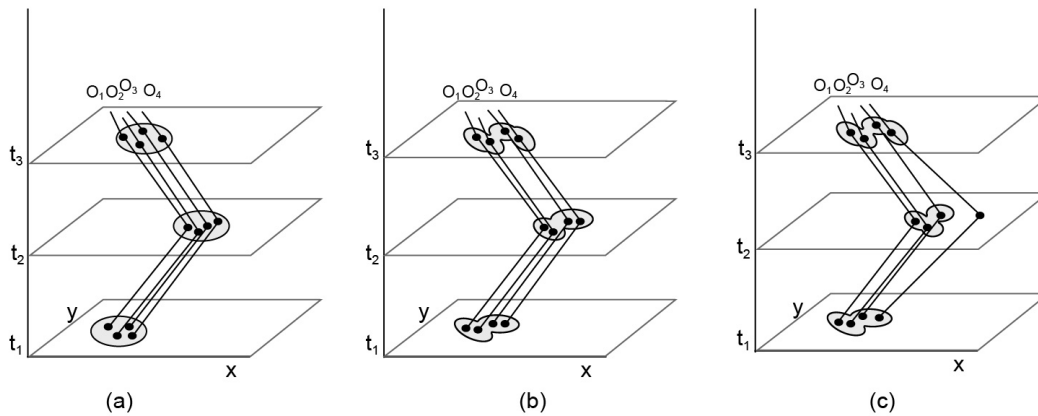


FIGURA 2.4: Ejemplo de diferentes patrones encontrados sobre las trayectorias conjuntas. Fuente: (Mazimpaka y Timpf 2016)

### Patrones secuenciales

La extracción de patrones secuenciales es un campo de investigación focalizado en la identificación de secuencias compartidas en trayectorias; esta se enfoca en la detección de movimientos recurrentes de objetos a lo largo del tiempo. Este proceso implica la búsqueda de patrones secuenciales en una única trayectoria o en conjuntos de ellas, donde se define un patrón como una sucesión específica de ubicaciones compartidas por múltiples objetos en intervalos temporales. La Figura 2.5 ilustra lo antes dicho. La relevancia de estos patrones se determina mediante un umbral de soporte, indicando la frecuencia con la que la secuencia común se presenta en un cúmulo de datos. Este enfoque, aunque es beneficioso para áreas como las recomendaciones de viaje y predicciones de ubicaciones, plantea desafíos en la comparación de ubicaciones, especialmente en trayectorias vehiculares, donde los puntos se caracterizan por coordenadas que no se repiten exactamente en cada instancia del patrón, generando complejidades computacionales notables.

### Patrones periódicos

Las técnicas de minería de datos utilizadas para detectar patrones periódicos se concentran en descubrir y comprender las pautas recurrentes presentes en las trayectorias de objetos en movimiento. Su funcionamiento parte del reconocimiento de que los objetos móviles exhiben con frecuencia comportamientos periódicos, como las actividades repetitivas de las personas o las migraciones estacionales de

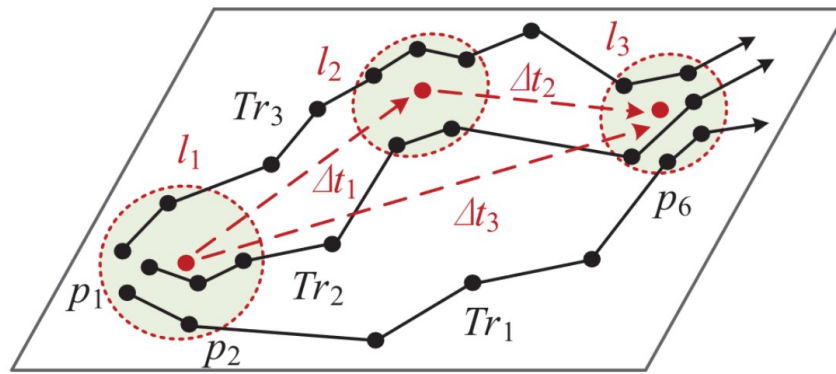


FIGURA 2.5: Extracción de patrones secuenciales en trayectorias.

Fuente: (Zheng 2015)

animales. Estos patrones, al proporcionar explicaciones concisas a lo largo de extensos historiales de movimiento, desempeñan un papel crucial en la compresión de datos de trayectorias y en la capacidad de predecir futuros movimientos. Para abordar esta complejidad se han desarrollado métodos específicos para cada tipo de desafío. A modo de ejemplo cabe mencionar la adaptación de algoritmos eficientes para recuperar patrones máximos y la consideración de comportamientos periódicos en entornos del mundo real, donde la presencia de períodos entrelazados y ruido espacio-temporal demanda enfoques más sofisticados. Enfoques más complejos brindan soluciones para detectar ubicaciones de referencia mediante algoritmos de agrupamiento basados en densidad y otros utilizan la aplicación de técnicas matemáticas de mayor complejidad junto con la autocorrelación para calcular períodos en cada ubicación de referencia. Este campo de estudio ha evolucionado a lo largo de las investigaciones extendiéndose hacia la detección de patrones periódicos en conjuntos de datos incompletos y dispersos, ampliando así su aplicabilidad a diversas situaciones del mundo real.

## 2.4. Agrupamiento de trayectorias

El agrupamiento o clusterización es una técnica de minería de datos ampliamente utilizada en el análisis de trayectorias. Permite identificar grupos de trayectorias similares según distintas características espacio-temporales, como cercanía, dirección, destino, paradas y velocidad, tal como se puede observar en la Figura 2.6. Esto posibilita describir distintos tipos o categorías de movimientos existentes en un conjunto de datos, facilitando la caracterización, interpretación y comparación de comportamientos. Los principales retos incluyen seleccionar medidas de distancia/similitud apropiadas según la aplicación, determinar el número óptimo de grupos e identificar agrupamientos con significado real más allá de similitudes geométricas entre trayectorias. También es clave lograr resultados explicables y fácilmente



interpretables en términos del fenómeno en estudio. Enfoques emergentes incorporan información de contexto, aprendizaje semisupervisado, detección automática de cantidad de grupos y técnicas de ensamblado para mejorar robustez, utilidad y comprensión de los agrupamientos de trayectorias obtenidos.

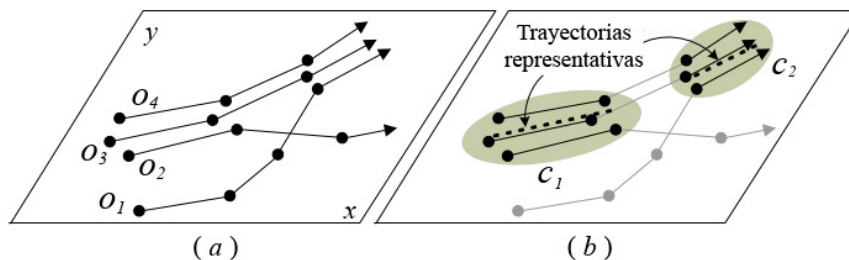


FIGURA 2.6: Ejemplo de un agrupamiento de trayectorias: (a) trayectorias espaciales; (b) segmentos de trayectorias agrupadas. Fuente: (Zheng y Zhou 2011)

Algunas investigaciones utilizan algoritmos de agrupamiento tradicionales, como el algoritmo k-means (Jain 2009) o el algoritmo DBSCAN (Ester et al. 1996), adaptados para el análisis de trayectorias en sus respectivos contextos y enfoques (Wu et al. 2024; Zheng y Liu 2024; Liu et al. 2023).

El agrupamiento de trayectorias de origen vehicular permite caracterizar y modelar el flujo de tráfico. Al identificar grupos de trayectorias similares, es posible describir distintos patrones de movimiento y así comprender mejor la dinámica compleja del tráfico. Mediante el análisis de grupos se pueden determinar rutas típicas, identificar cuellos de botella, encontrar anomalías o estudiar impactos de cambios en la infraestructura.

El autor de esta tesis ha evaluado la aplicabilidad del agrupamiento de trayectorias de origen vehicular como medio para caracterizar y modelar el flujo de tráfico. El autor reconoce la relevancia de identificar grupos de trayectorias similares, por lo que se ha inclinado hacia la consideración de una técnica de agrupamiento dinámico. Con este enfoque específico busca procesar flujos de datos, con un énfasis particular en el tratamiento de datos de trayectorias vehiculares.

Previamente se ha resaltado la importancia de las técnicas de agrupamiento de trayectorias y su capacidad para caracterizar y modelar el flujo de tráfico.

Algunas de las técnicas de agrupamiento tradicionales han sido adaptadas por otros investigadores y se han utilizado con éxito en diferentes situaciones. En este contexto, se explorará la aplicación de técnicas de agrupamiento especialmente adaptadas al análisis de información del flujo vehicular, ya que su aplicación no solo permite identificar de manera precisa patrones en flujos de datos de trayectorias, sino también comprender relaciones significativas, evidenciando su utilidad en diversas situaciones de análisis de datos espaciales y temporales.

El flujo vehicular se define como el desplazamiento continuo de vehículos a lo largo de una red vial durante un período de tiempo determinado. Esta dinámica

se caracteriza por la interacción compleja entre vehículos, regulaciones de tráfico y condiciones de la infraestructura vial. La velocidad, la densidad y el flujo son parámetros clave frecuentemente utilizados para cuantificar y describir el movimiento de vehículos en una determinada área.

Frecuentemente se hace uso de la incorporación de información contextual como señales de tránsito, capacidad vial o eventos extraordinarios, junto con atributos espacio-temporales de las trayectorias, lo cual posibilita descubrir ciertas relaciones y factores que se incluyen en la complejidad del flujo vehicular. Por ejemplo, las modificaciones a estas técnicas de agrupamiento las han vuelto eficaces para identificar patrones de congestión en áreas urbanas con mucho tráfico o para analizar cómo eventos climáticos afectan el flujo vehicular en zonas altamente urbanas.

Los retos más frecuentes asociados al análisis de flujos de datos vehiculares incluyen la escalabilidad ante grandes volúmenes de datos (Zhou et al. 2015), lidiar con ruido e incertidumbre (Yuan et al. 2018), y lograr grupos representativos, distinguibles y fácilmente interpretables en términos del fenómeno de tráfico. Se exploran técnicas avanzadas de agrupamiento, segmentación previa de trayectorias y métodos de múltiples algoritmos para obtener agrupamientos robustos y óptimos para el análisis y modelado de flujos vehiculares.

El análisis del flujo vehicular en las grandes ciudades se ha vuelto relevante debido al constante aumento en el volumen de tráfico, lo que causa problemas en la circulación vehicular (Jain 2009). Para abordar este problema, se ha utilizado la técnica de agrupamiento de datos para identificar características comunes entre instancias del mismo problema (Madhulatha 2012). A lo largo del tiempo, se han propuesto mejoras para superar las limitaciones de algunas técnicas, como la inicialización de algoritmos de agrupamiento (Bahmani et al. 2012).

Es importante destacar que se han realizado esfuerzos de investigación en esta área (Hu et al. 2020; Lou y Cheng 2020). Las técnicas de agrupamiento han demostrado un buen desempeño en el análisis de las trayectorias de origen vehicular, aunque la parametrización sigue siendo un desafío interesante debido a que son técnicas no supervisadas que combinan medidas de distancia y densidad para controlar la construcción de los grupos.

Como punto de partida y teniendo el foco puesto en el uso de técnicas de agrupamiento para flujos de ubicaciones GPS de trayectorias, se realizó un breve estudio bibliométrico que se explica a continuación.

#### **2.4.1. Breve estudio bibliométrico de algoritmos de agrupamiento de trayectorias**

El objetivo de este estudio bibliométrico es presentar un análisis breve y objetivo sobre el panorama actual de la investigación en el ámbito de los algoritmos de agrupamiento de trayectorias. A través de una revisión de la literatura relevante, en este estudio bibliométrico se propone examinar la evolución, tendencias y contribuciones

más significativas en este campo de estudio. Al emplear herramientas bibliométricas, se pretende identificar las publicaciones más influyentes, autores destacados, así como determinar patrones y áreas emergentes de investigación.

El estudio se centra en la evaluación bibliométrica de los algoritmos de agrupamiento de trayectorias GPS, analizando 559 artículos de Scopus para ofrecer una perspectiva actualizada de las tendencias en esta área de investigación. Se destaca la importancia de la investigación bibliométrica en el contexto científico contemporáneo, que ha experimentado un notable crecimiento, especialmente desde la fundación del Instituto de Información Científica por Eugene Garfield en la década de 1960. Este enfoque cuantitativo examina diversos aspectos de la producción científica, como la autoría, publicación, citas y contenido, utilizando grandes bases de datos bibliográficas como Scopus. La indexación de artículos permite no solo medir el impacto y la relevancia en un campo científico, sino también proporcionar información valiosa para instituciones académicas y consejos de investigación, facilitando la toma de decisiones en la contratación de personal y el diseño de estrategias de investigación.

Se ha analizado la metadata de todos los artículos indexados en la base de datos bibliográfica Scopus que realizan algoritmos o métodos para agrupamiento de trayectorias GPS. La muestra generada ha sido filtrada manualmente para excluir todos los artículos que no sean parte del campo de estudio. Se proporcionará información útil sobre las principales revistas que están interesadas en publicar artículos acerca de este tema en particular, así como la evolución que ha tenido su campo científico con el paso del tiempo. Además, se discuten otros aspectos como los autores más citados, las áreas en que más se publican estos artículos, el número de publicaciones por año, los diagramas estratégicos sobre el impacto de los temas, la evolución temática, entre otros.

Se trabajó con la metadata bibliográfica (bibliographic metadata) de los artículos que se encuentran indexados en la base de datos bibliográfica Scopus. Por lo tanto, se seleccionaron solamente los artículos en inglés que desarrollen o investiguen acerca de la “agrupamiento de trayectorias GPS”. Scopus alberga un total de 559 documentos de la muestra, publicados en 333 fuentes (revistas, libros, etc), durante el período de 2002-2023. Estos documentos fueron (co) escritos por 1 416 personas, siendo solo 11 documentos de autoría única. La media de autores por documentos es 3,87. Al realizar el análisis se observó que se concentraban en dos áreas de investigación principales: Ciencias de la Computación e Ingeniería. Scopus asigna artículos indexados a una o más áreas de investigación. Los 559 artículos de la muestra fueron asignados a diversas áreas de investigación dando un total de 1 094, es decir que pertenecen a más de un área específica. Las cinco principales áreas de investigación se muestran en la Tabla 2.1.

En Scopus los primeros años tuvieron la acogida de pocos artículos relacionados con el “agrupamiento de trayectorias GPS”, aunque en la última década la cantidad de artículos que se publicaron se ha incrementado posiblemente por la acogida de

TABLA 2.1: Principales áreas de investigación asignadas a los trabajos de la muestra. Fuente: Scopus.

Áreas de investigación	Registros	% de 1094
Ciencias de la computación	391	35,74 %
Ingeniería	176	16,09 %
Ciencias sociales	125	11,43 %
Matemáticas	123	11,24 %
Tierra y ciencias planetarias	69	6,31 %
Total de las 5 principales áreas de investigación	884	80,80 %

la comunidad científica. El total de registros de la muestra indican un crecimiento promedio por año del 15,6 % desde el 2002 hasta el 2023.

### Distribución geográfica del autor correspondiente

La Tabla 2.2 muestra a China como el principal país cuyos autores han publicado más documentos, seguido por USA como el segundo país que más documentos publicados tiene. Los diez primeros países acumulan el 53,9 % de los artículos publicados relacionados con “agrupamiento de trayectorias GPS”.

TABLA 2.2: Diez países de autores correspondientes. Fuente: Scopus.

País	Artículos	Frecuencia
China	203	36,3 %
USA	29	5,2 %
India	16	2,9 %
Italia	13	2,3 %
Corea	11	2,0 %
Portugal	8	1,4 %
Japón	6	1,1 %
Australia	5	0,9 %
Francia	5	0,9 %
Alemania	5	0,9 %
Total 10 países	301	53,9 %

La Tabla 2.3 muestra los principales países ordenados por el número total de citas. El promedio de citas de todos los artículos es de 21,92. China y USA, son los dos países con más artículos publicados y citas totales, con un promedio de 19,30 y 34,60 respectivamente. A pesar de que China es el primer país en términos de artículos publicados, tiene el segundo promedio más bajo de citas por artículo entre los países líderes. También es importante destacar que USA es el país con el promedio más alto de citas por artículo, lo que puede utilizarse como un denominador común en la importancia científica media o la calidad de los artículos. Además, alrededor del 40 % de los artículos publicados, están realizados por autores de varios países. El país que más colabora internacionalmente con otros países es China, donde el 30,8 % de los trabajos son de este tipo.

### Principales fuentes de publicación

TABLA 2.3: Diez principales citas totales por país. Fuente: Scopus.

País	Total de citas	Citas promedio de artículos
China	3908	19,30
USA	1004	34,60
Turquía	400	400,00
Italia	202	15,50
Hong Kong	173	34,60
Suiza	173	34,60
Grecia	167	33,40
España	156	39,00
Francia	127	25,40
Australia	118	23,60
Total (Todos los países)	7138	21,92

TABLA 2.4: Las diez fuentes más relevantes. Fuente: Scopus.

Fuentes	# Artículos	Tipo
Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	38	Book Series
ISPRS International Journal of Geo-Information	17	Journal
IEEE Access	15	Journal
GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems	13	Conference Proceedings
ACM International Conference Proceeding Series	12	Conference Proceedings
International Journal of Geographical Information Science	11	Journal
IEEE Transactions on Intelligent Transportation Systems	8	Journal
International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives	8	Journal
Jiaotong Yunshu Xitong Gongcheng Yu Xinxi/Journal of Transportation Systems Engineering and Information Technology	7	Journal
Transactions in GIS	7	Journal

La Tabla 2.4 muestra las diez fuentes principales que publican artículos relacionados con “algoritmos de agrupamiento de trayectorias”. Las tres primeras según la cantidad de publicaciones son Lecture Notes in Computer Science (LNCS) (incluida su subserie Lecture Notes en Inteligencia Artificial, LNAI y Apuntes de Conferencias en Bioinformática, LNBI) es una serie de actas de congresos que publica los últimos avances en investigación en todas las áreas de la informática. ISPRS International Journal of Geo-Information que es una revista internacional de acceso abierto revisada por pares sobre geoinformación. IEEE Access que es una importante revista

multidisciplinaria de acceso abierto.

### Palabras claves principales.

La Tabla 2.5 muestra las diez palabras clave más utilizadas en los artículos de agrupamiento de trayectorias GPS. Scopus proporciona los datos de las palabras clave (Author Keywords), que son proporcionadas por los autores originales. Las dos palabras clave de autor más frecuentes son “clustering” y “trajectory”. Se observa que las principales Keywords engloban todo lo que tiene que ver con las temáticas trayectorias, agrupamiento y datos gps, además, de que son utilizados en el proceso de la minería de datos.

TABLA 2.5: Palabras clave principales. Fuente: Scopus.

Palabras clave de autor	# Artículos
clustering	67
trajectory	34
trajectory clustering	27
gps	26
gps trajectory	24
dbscan	22
data mining	21
gps data	21
gps trajectories	20
big data	14

### Diagrama estratégico de las palabras claves.

En el diagrama estratégico se pueden observar los temas que están emergiendo, son tendencia y están o han desaparecido de un campo de investigación mediante el análisis de las palabras claves. Cuando se utiliza el análisis de palabras conjuntas para cartografiar la ciencia, se obtienen grupos de palabras clave (y sus interconexiones). Estos grupos se consideran temas. Cada tema de investigación obtenido en este proceso se caracteriza por dos parámetros (“densidad” y “centralidad”) (Cobo et al. 2011). En la Figura 2.7 se observa el diagrama estratégico pertenecientes a las palabras claves del autor de Scopus, su cuadrante superior derecho parcialmente contiene el tema “mobility” considerado como un grupo de subtemas parcialmente bien desarrollados e importantes para el campo de investigación de los métodos para agrupamiento de trayectorias GPS. Su cuadrante superior izquierdo cuenta con los temas “vehicle trajectory”, parcialmente los tema “mobility” y “urban computing” que son de importancia para el campo de investigación, el primero se acerca a los temas bien desarrollados mientras que el segundo a los temas emergentes. En este último se pueden llegar a encontrar subtemas que están pocos desarrollados, no son tomados en cuenta, son emergentes o han desaparecidos. Por último, su cuadrante inferior derecho contiene temas fundamentales como el “clustering” y “trajectory clustering” señalando que los temas relacionados al agrupamiento de trayectorias son temas importante pero que está poco desarrollado.

### Gráficos de citas, fuentes y autores

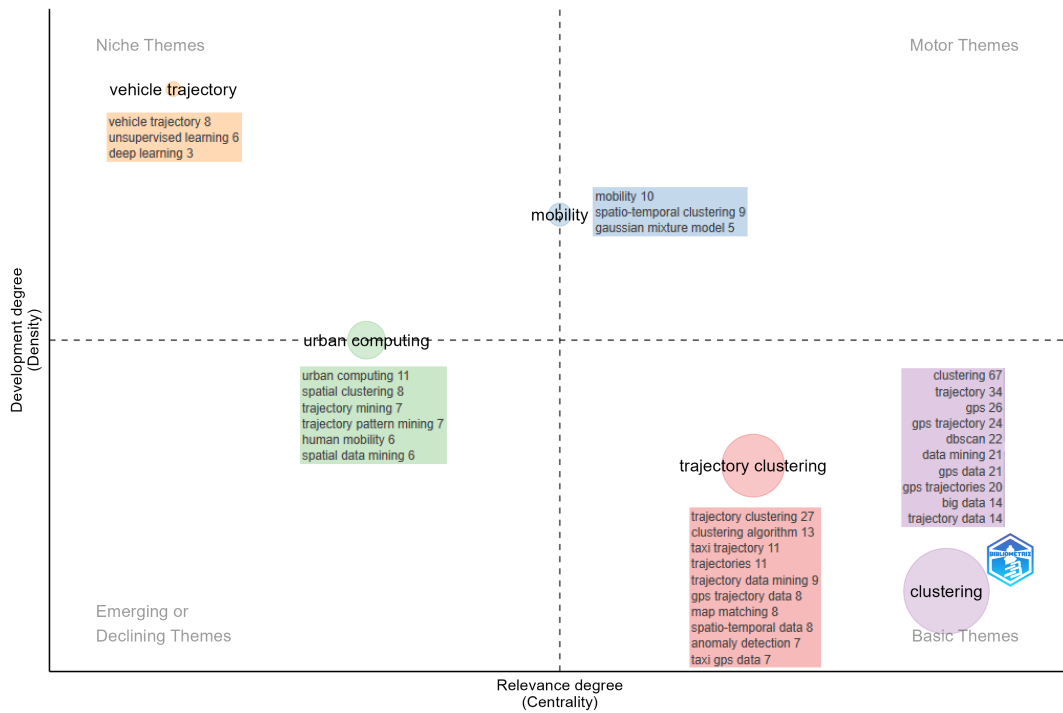


FIGURA 2.7: Diagrama estratégico de las palabras clave del autor, generado con bibliometrix. Fuente: Scopus.

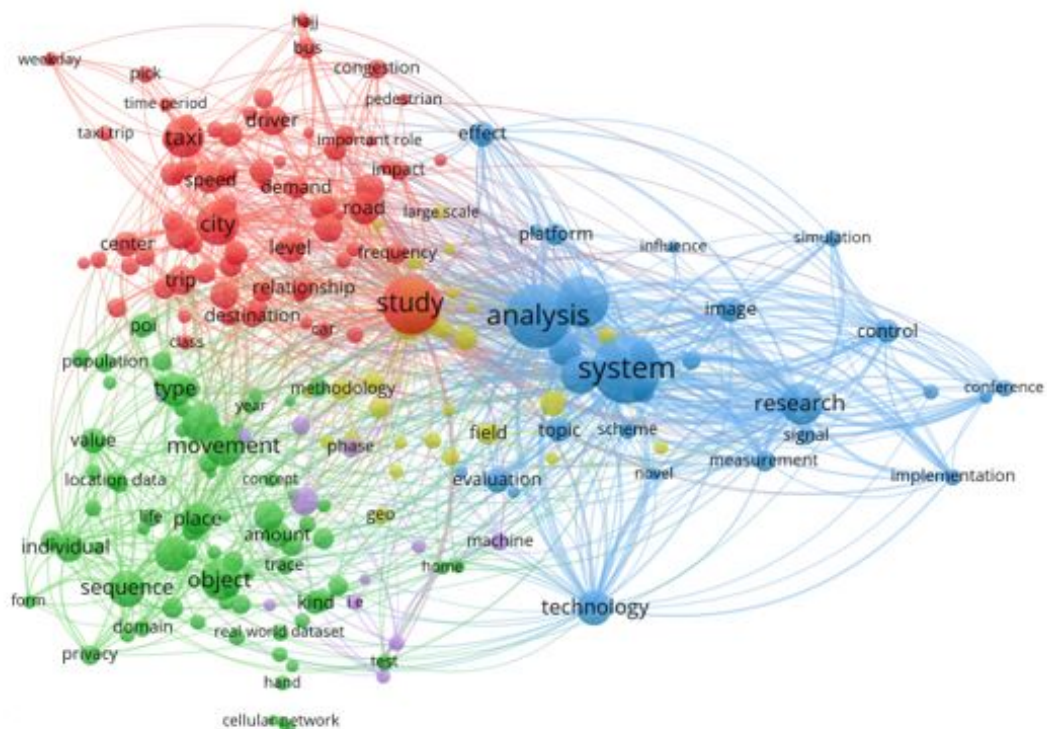


FIGURA 2.8: Mapa de nubes de palabras en títulos y resúmenes (recuento completo), generado con VOSviewer. Fuente: Scopus.

Las siguientes figuras fueron generadas utilizando la herramienta de software VOSviewer que permite crear mapas basados en red, permitiendo la visualización y la exploración. La Figura 2.8 representa el mapa de nubes con las palabras que son relevantes en los artículos. El mapa muestra la cantidad de veces en que las palabras aparecen en los artículos y que tanta es la relación que existe entre ellos. En mapa se divide por grupos, la parte (azul) tiene una concentración de la palabra carretera, que a su vez se relaciona con la palabra análisis, sistema, tecnología e investigación. En la parte (roja) se observan palabras que están relacionadas con estudios realizados sobre carreteras, ciudades y taxis con sus diferentes aplicaciones refiriéndose principalmente a la planificación urbana o urbanismo. En la parte (verde, amarilla y morada) hacen alusión al movimiento de los objetos, entre sus palabras están tipo, objetos, secuencia y movimiento.

#### **Conclusiones del estudio bibliométrico.**

El análisis realizado evidencia que el estudio de agrupamiento de trayectorias GPS es una combinación entre el urbanismo y los efectos de la circulación vehicular en distintas vías. Cabe destacar que esto no sería posible sin los sistemas de posicionamiento global. Este breve estudio bibliométrico, que abarcó 559 artículos de la base de datos Scopus, ha sido fundamental para comprender las relaciones entre palabras clave, autores, citas, entre otros atributos.

A través del estudio, se identificaron trabajos, métodos y autores relevantes que sirven de base para el desarrollo de esta tesis. Además, se destacaron las revistas, congresos y conferencias más relevantes en la temática, lo que es crucial tanto para difundir los resultados como para identificar contribuciones significativas. En relación con el capítulo 3, el estudio resaltó la importancia de la identificación de patrones en el análisis de flujos vehiculares, subrayando su relevancia para mejorar el tráfico urbano.

Para examinar información más detallada, el análisis completo puede encontrarse en (Reyes et al. 2024).

#### **2.4.2. Medidas de similitud**

Las medidas de similitud permiten cuantificar qué tan parecidas o distintas son dos trayectorias, mediante funciones matemáticas que mapean sus características a valores numéricos (Bian et al. 2018). Estas medidas son esenciales en diversas técnicas de minería y aprendizaje automático aplicadas al análisis de trayectorias, como agrupamiento y clasificación. Las técnicas de agrupamiento tradicionales han sido adaptadas según las necesidades para funcionar con diversas medidas de similitud, en el ámbito investigado se exploran aquellas diseñadas especialmente para trayectorias. Las diversas adaptaciones han permitido que técnicas como el agrupamiento o la clasificación sean más eficientes según las características deseadas y las situaciones adecuadas (Choong et al. 2016; Kim y Mahmassani 2015).



La selección de la métrica de similitud más apropiada depende de diversos factores, como la naturaleza de los datos, los objetivos del análisis y las características particulares del caso de estudio. Es fundamental evaluar cada métrica en función de su idoneidad para capturar las relaciones entre los elementos del conjunto de datos y su capacidad para proporcionar resultados significativos para el análisis en cuestión.

Algunas de las medidas de similitud más utilizadas son la distancia Euclidiana (Bian et al. 2018), que cuantifica la separación geométrica, midiendo la longitud del camino más corto entre dos puntos en un espacio, como la distancia directa entre dos ubicaciones en un mapa. Esta métrica es especialmente útil en contextos donde se necesita evaluar la similitud entre objetos representados por vectores numéricos.

Por otro lado, la distancia de Manhattan (Yang 2019), también denominada distancia de la ciudad, cuantifica la suma de las diferencias absolutas entre las coordenadas de los puntos. Esta métrica calcula la suma de las diferencias absolutas entre las coordenadas de los puntos, lo que implica medir la distancia entre ellos moviéndose solo en líneas verticales y horizontales, como si se estuviera recorriendo las calles de una ciudad en un plano cartesiano.

La distancia de Hausdorff (Chen et al. 2011) es una métrica más avanzada, esta evalúa la máxima discrepancia entre dos conjuntos. Esta distancia considera la similitud como la mayor de las distancias mínimas entre puntos de un conjunto al punto más cercano en el otro conjunto. Es especialmente útil en aplicaciones donde se requiere una evaluación precisa de la similitud entre conjuntos de datos, como en el análisis de imágenes, reconocimiento de patrones o en la detección de objetos.

La distancia de Bhattacharyya (Schweppe 1967) es una métrica que evalúa qué tan cerca están dos distribuciones de probabilidad entre sí que están asociadas con las direcciones de puntos discretizados. Esta métrica es particularmente útil en aplicaciones donde se necesitan evaluar las similitudes entre distribuciones de probabilidad, como en el análisis de datos multivariados o en problemas de clasificación donde las características están representadas como distribuciones de probabilidad.

Otra frecuentemente utilizada es la distancia de Frechet (Eiter y Mannila 1994) que evalúa la similitud entre dos curvas al tener en cuenta tanto la ubicación como el orden temporal de los puntos. La distancia de Frechet considera la forma en que las dos curvas se mueven juntas en relación con el tiempo y la distancia entre los puntos correspondientes a lo largo de sus trayectorias. Esta medida es especialmente útil en aplicaciones donde se necesitan evaluar la similitud entre curvas, como en el análisis de trayectorias de movimiento o en la comparación de señales temporales.

Por otra parte está la Dynamic Time Warping (DTW) (Byoung-Kee Yi, Jagadish y Faloutsos 1998) es un método de alineación de secuencias que busca encontrar una correspondencia óptima entre dos trayectorias y medir la similitud sin tener en cuenta las longitudes y el orden temporal. Es una técnica bien conocida para encontrar una alineación óptima entre dos secuencias dadas, que pueden ser dependientes del tiempo, bajo ciertas restricciones. Su principal característica es que permite comparar dos secuencias de manera flexible, incluso si tienen longitudes diferentes o si

los elementos no están temporalmente alineados de manera directa.

La distancia Longest Common Subsequence (LCSS) (Wei et al. 2014) tiene como objetivo encontrar la subsecuencia común más larga entre todas las secuencias, y la longitud de esta subsecuencia podría considerarse como la similitud entre dos trayectorias arbitrarias, incluso si las trayectorias difieren en longitud o en la ubicación exacta de los puntos. En casos específicos esta distancia ha demostrado proporcionar un rendimiento superior sin importar la longitud de las trayectorias, lo que la hace robusta y efectiva en una variedad de aplicaciones. Al no centrarse en la longitud de las trayectorias, la distancia de LCSS puede capturar la similitud entre secuencias de manera más precisa, lo que la convierte en una opción valiosa para la comparación de trayectorias en campos como el análisis de movimiento, reconocimiento de patrones y procesamiento de señales temporales.

Es crucial en el análisis de datos GPS distinguir entre varios enfoques fundamentales: la comparación de ubicaciones GPS, el análisis de segmentos de trayectorias, el estudio de distribuciones de probabilidad, así como el análisis de secuencias y curvas. La comparación de ubicaciones GPS implica la evaluación de coordenadas geográficas en puntos específicos en el tiempo, mientras que el análisis de segmentos de trayectorias se centra en tramos continuos de movimientos entre ubicaciones sucesivas. El estudio de distribuciones de probabilidad examina la frecuencia o probabilidad de ocurrencia de eventos, como las ubicaciones visitadas o los tiempos de permanencia en áreas específicas. Además, el análisis de secuencias se enfoca en la identificación y comprensión de patrones secuenciales en los movimientos, mientras que el análisis de curvas permite modelar y comprender la forma y la tendencia de los datos de ubicación a lo largo del tiempo o en función de otras variables. Estas distinciones son fundamentales para comprender la complejidad de los datos de ubicaciones lo que permite obtener patrones complejos, entender la dinámica temporal y modelar fenómenos espaciales con mayor precisión.

Una breve descripción de las principales aplicaciones y la complejidad computacional requerida por cada métrica de similitud recientemente descrita, se pueden observar en la Tabla 2.6.

Algunas medidas de similitud utilizan otras características como el ángulo o la longitud relativa entre trayectorias, y algunas no tradicionales están basadas en formas o deformaciones (Besse et al. 2016; Magdy et al. 2015). Otras medidas más avanzadas consideran también factores como la correlación temporal de los puntos, la sincronización en eventos relevantes o la verificación de patrones secuenciales compartidos.

Diversas investigaciones desarrollan nuevos algoritmos de agrupamiento que calculan las similitudes a partir de medidas de similitud adaptadas a sus respectivos contextos, tal es el caso de (Yu et al. 2019a) que calculan la similitud entre dos trayectorias utilizando múltiples características de los datos.

Por mencionar algunas de estas adaptaciones, se tienen algoritmos como Improved DBScan (Luo et al. 2017) que mejora la versión tradicional mediante su propio

TABLA 2.6: Resumen de las medidas de similitud más comunes.

Medida de similitud	Principales aplicaciones	Complejidad computacional
Distancia euclidiana	Trayectorias que ocurren al mismo tiempo	$O(n)$
Distancia manhattan	Datos con dimensiones independientes	$O(n)$
Distancia hausdorff	Conjuntos de puntos multidimensionales	$O(mn)$
Distancia bhattacharyya	Distribuciones de probabilidad	$O(n)$
Distancia frechet	Conjuntos de series temporales	$O(mn)$
Distancia DTW	Series temporales	$O(mn)$
Distancia LCSS	Trayectorias arbitrarias con longitudes diferentes	$O(mn)$

método de medición similitud basado en densidad, introduciendo el concepto de capacidad de movimiento y la teoría de campos de datos. Otras técnicas, como la utilizada en el algoritmo Tra-DBScan (Liu et al. 2012), utiliza el algoritmo DBScan (Ester et al. 1996) y agrega una fase de segmentación de trayectorias que emplea la distancia de Hausdorff como medida de similitud, trabajando en un marco de partición y agrupamiento.

Ninguna métrica es óptima en todos los casos, por lo que se recomienda evaluar y seleccionar las más alineadas a los objetivos específicos de cada aplicación. La selección adecuada de la métrica de similitud es crucial para obtener resultados precisos en el análisis de trayectorias.

### 2.4.3. Clasificación de algoritmos

La clasificación de los diversos algoritmos de agrupamiento para trayectorias se basa principalmente en dos criterios fundamentales: la formación de los grupos y la naturaleza de la distribución proporcionada como entrada, ya sea de ubicaciones GPS discretas o continuas (Saxena, Goyal y Mittal 2015). Esta clasificación tomó en cuenta la diversidad de enfoques para estructurar los conjuntos y la variabilidad en la representación de datos de trayectorias, ambas características han sido constituidas como elementos clave para la categorización eficaz de los algoritmos.

#### Algoritmos de agrupamiento particionales

Este tipo de algoritmos se centra en dividir el conjunto de datos en un número predeterminado de grupos o particiones. La idea fundamental es asignar cada ubicación GPS de datos a un grupo de manera que la partición global del conjunto de

datos sea coherente y represente patrones subyacentes en los datos. Algunos ejemplos notables de algoritmos de agrupamiento particionales incluyen los algoritmos k-means, k-medoids y el algoritmo de agrupamiento espectral.

En el algoritmo k-means, por ejemplo, se especifica de antemano el número de grupos ( $k$ ) que se desea formar. El algoritmo busca iterativamente ajustar los centroides de los grupos y asignar ubicaciones de datos a los grupos correspondientes de manera que la suma de las distancias al cuadrado entre los elementos y los centroides sea minimizada. Este proceso se repite hasta que la convergencia se alcanza o se cumple un criterio predefinido.

Los algoritmos de agrupamiento particionales son eficaces cuando se conoce de antemano la cantidad deseada de grupos y se busca una partición clara y no solapada de los datos. Sin embargo, pueden ser sensibles a la elección inicial de los centroides y pueden no ser ideales para conjuntos de datos con estructuras más complejas o grupos de tamaños y formas variables. La elección de algoritmos particionales dependerá de la naturaleza específica del problema y de los requisitos del análisis de datos.

### **Algoritmos de agrupamiento basados en distribución**

Los algoritmos basados en la distribución buscan agrupar datos considerando la distribución subyacente de las ubicaciones en el espacio de características. A diferencia de algunos algoritmos que asignan ubicaciones a grupos según ciertos centroides o densidades locales, los algoritmos basados en la distribución consideran la forma general de la distribución de los datos en el espacio.

Uno de los ejemplos más destacados en esta categoría es el algoritmo de mezclas gaussianas (Gaussian Mixture Model, GMM). El algoritmo GMM el cual asume que los datos son generados por una mezcla de varias distribuciones gaussianas. En lugar de asignar directamente elementos a grupos, el algoritmo GMM modela la distribución de los datos como una combinación ponderada de múltiples distribuciones gaussianas. Cada componente gaussiano representa un grupo potencial, y la contribución relativa de cada componente se ajusta durante el proceso de entrenamiento.

Es importante destacar que el algoritmo GMM puede aplicarse a cualquier conjunto de datos representados como vectores numéricos, lo que lo hace versátil para una variedad de aplicaciones de análisis de datos, como puede ser el análisis de trayectorias en aplicaciones de movilidad.

Este enfoque basado en la distribución es útil cuando los datos pueden tener una estructura más compleja o cuando la naturaleza de los grupos no es clara y puede variar en diferentes regiones del espacio de características. Los algoritmos basados en distribución son especialmente útiles cuando se busca capturar la variabilidad y las relaciones complejas entre los datos.

### Algoritmos de agrupamiento jerárquicos

Los algoritmos de agrupamiento jerárquico construyen una jerarquía de grupos utilizando una estructura de árbol que representa la relación de agrupamiento entre los datos a diferentes niveles de detalle. Estos métodos pueden ser aglomerativos, empezando con elementos individuales y fusionando grupos cercanos, o divisivos, comenzando con un grupo único y dividiéndolo en subgrupos.

Un ejemplo específico de algoritmo de agrupamiento jerárquico es BIRCH, el cual es un algoritmo que utiliza una estructura de árbol para organizar los datos de manera jerárquica. En lugar de almacenar todas las ubicaciones GPS en cada nivel del árbol, BIRCH utiliza un resumen compacto llamado "Clustering Feature" para representar regiones densas de datos en el espacio de características. Este resumen facilita la construcción eficiente de la jerarquía de grupos y ayuda a gestionar grandes conjuntos de datos.

El algoritmo BIRCH-T es una extensión del popular algoritmo BIRCH diseñada específicamente para agrupamiento incremental por lotes de conjuntos de trayectorias. Utiliza una estructura de árbol balanceado que absorbe información estadística de grupos de trayectorias en cada lote, permitiendo un procesamiento muy eficiente.

### Algoritmos de agrupamiento basados en densidad

Este tipo de algoritmo de agrupamiento identifica regiones densas de elementos en el espacio de características para formar grupos. A diferencia de los algoritmos particionales que requieren la especificación previa del número de grupos, los algoritmos de agrupamiento basados en densidad pueden descubrir grupos de forma automática y son capaces de manejar conjuntos de datos con estructuras más complejas y tamaños de grupos variables.

Uno de los ejemplos más notables de algoritmo de agrupamiento basado en densidad es DBSCAN. Este algoritmo define un grupo como una región densa de elementos que está separada de otras regiones densas por áreas más dispersas o de menor densidad. La idea central es que los elementos dentro de un grupo deben estar cercanos unos de otros y debe haber suficientes elementos cercanos a cada uno de los elementos analizados para considerarlos parte del grupo.

Existen variaciones de los algoritmos DBSCAN y OPTICS que se han propuesto específicamente para el análisis de trayectorias, tales como, T-OPTICS que es una adaptación de OPTICS para trayectorias que ordena las secuencias de elementos usando distancias temporales y espaciales entre trayectorias, este permite el análisis en diferentes niveles de granularidad. Otro algoritmo es TRACCLUS que es pionero en el agrupamiento de trayectorias basado en densidad, este usa una métrica de distancia longitudinal para identificar segmentos densos de trayectorias cercanas como grupos.

### Algoritmos de agrupamiento basados en cuadrículas

Los algoritmos de agrupamiento basados en cuadrícula organizan el espacio de las características analizadas en una cuadrícula y asignan las ubicaciones GPS a celdas específicas. Este enfoque simplifica el problema de agrupamiento al discretizar el espacio de características en una cuadrícula regular.

El algoritmo Optimal-Grid es un ejemplo de esta técnica de agrupamiento, está diseñado para agrupar grandes cantidades de datos de alta dimensionalidad y manejar el ruido inherente (Hinneburg y Keim 1999). Se basa en la construcción de una partición óptima de cuadrícula mediante el cálculo de hiperplanos de partición adecuados para cada dimensión. Sus principales ventajas incluyen una sólida base matemática, alta efectividad y eficiencia en comparación con otros algoritmos, incluso con conjuntos de datos extensos.

### Algoritmos de agrupamiento basados en grafos

Los algoritmos basados en grafos modelan las ubicaciones GPS como nodos en un grafo, y las relaciones entre los elementos se representan mediante conexiones o aristas. Estos algoritmos identifican grupos al analizar la conectividad y la estructura de estos grafos.

Un ejemplo común de algoritmo de agrupamiento basado en grafos es el algoritmo de corte espectral. En este método, se construye una matriz de similitud entre las ubicaciones GPS, y luego se representa como un grafo ponderado donde los nodos son los elementos y las aristas representan las similitudes entre ellos. El corte espectral busca dividir el grafo en subgrafos disjuntos, que pueden interpretarse como grupos.

### Algoritmos basados en modelos

Los algoritmos basados en Modelos utilizan modelos estadísticos o probabilísticos para representar la estructura de los datos y asignar elementos a grupos en función de la adecuación a estos modelos. Estos algoritmos buscan ajustar un modelo a los datos y utilizan la información derivada de dicho modelo para realizar el agrupamiento.

Un ejemplo específico es el algoritmo de Expectation-Maximization (EM) utilizando comúnmente en problemas de agrupamiento cuando se asume que los datos pueden ser generados por un conjunto de distribuciones probabilísticas, como las distribuciones gaussianas. El EM busca encontrar los parámetros que maximizan la probabilidad de que los datos observados provengan de esos modelos probabilísticos.

En la literatura, existen diferentes enfoques basados en modelos para el análisis de trayectorias, que incluyen técnicas de agrupamiento y algoritmos basados en la red vial (Zhou et al. 2018; Lee, Han y Whang 2007; Hong, Chen y Mahmassani

2018; Feng, Bai y Xu 2019). Algunos investigadores han propuesto métodos que utilizan redes neuronales para clasificar datos y mejorar la precisión de los resultados (Zhang et al. 2019). La combinación de técnicas de agrupamiento y redes neuronales ha mostrado su eficacia en el análisis de trayectorias en regiones urbanas y en la optimización del tráfico en zonas congestionadas (Xiong, Shang y Bian 2017).

Asimismo, se ha propuesto un modelo de trayectoria mejorado (Yu et al. 2019a) y un nuevo algoritmo de agrupamiento que maximiza la similitud entre trayectorias mediante una medida que tiene en cuenta múltiples características de los datos. Ferreira et al. (Ferreira et al. 2012) presentan una nueva técnica de agrupamiento de trayectorias que utiliza campos vectoriales para representar los centros de los grupos y proponen una definición de similitud entre trayectorias. Por otro lado, Reyes et al. (Reyes-Zambrano et al. 2020) presentan un método de agrupamiento de trayectorias para datos vehiculares que utiliza información angular para segmentar las trayectorias y una función de similitud guiada por pivotes.

### Algoritmos basados en lógica difusa

Los algoritmos basados en lógica difusa utilizan conceptos de lógica difusa para asignar grados de pertenencia de las ubicaciones GPS a varios grupos en lugar de asignarlos de manera binaria a un solo grupo. Estos algoritmos permiten la existencia de grados de pertenencia difusos, lo que significa que una ubicación puede pertenecer parcialmente a varios grupos simultáneamente.

Un ejemplo específico de algoritmo de agrupamiento basado en lógica difusa es el algoritmo Fuzzy C-Means (FCM). En FCM, cada ubicación tiene grados de pertenencia asociados con respecto a todos los grupos, representando la probabilidad o grado de pertenencia difusa.

### Algoritmos de flujo de datos

Los algoritmos de flujo de datos se especializan en el procesamiento continuo de datos a medida que llegan, a su vez, agrupan constantemente los nuevos datos adaptándose a la naturaleza dinámica de la información en tiempo real. Estos métodos son esenciales para aplicaciones que requieren respuestas rápidas a cambios en la distribución de los datos y son capaces de identificar y actualizar grupos de manera eficiente a medida que se produce la evolución de los datos.

Un ejemplo de este tipo de algoritmos es DyClee, el cual es un algoritmo diseñado para el agrupamiento dinámico en flujos de datos continuos. Se adapta a cambios en la distribución de los datos y puede descubrir y actualizar grupos a medida que los nuevos datos llegan. Este tipo de algoritmos son cruciales en entornos donde los datos evolucionan con el tiempo, como en aplicaciones de monitoreo continuo o análisis de transmisiones de datos en tiempo real.

En esta misma categoría se tiene el algoritmo D3CAS (Molina y Hasperué 2018), este es un algoritmo de agrupamiento basado en densidad implementado en Spark

Streaming, que se beneficia de la eficiencia del framework distribuido Spark para el procesamiento de flujos de datos, destacando su capacidad dinámica para seleccionar automáticamente el número de grupos y su superioridad en la calidad de los resultados en comparación con otros algoritmos similares como CluStream.

Los algoritmos de agrupamiento de flujos de datos más influyentes como DenStream, StreamKM++, StreamLS y D-Stream extienden técnicas tradicionales de agrupamiento para operar con flujos de datos secuenciales en movimiento. DenStream utiliza un modelo incremental basado en densidad para detectar y hacer seguimiento de regiones densas cambiantes en tiempo real. StreamKM++ optimiza k-means para seleccionar centroides representativos y actualizar grupos ante nuevos datos de forma eficiente. StreamLS construye un agrupamiento robusto alrededor de puntos medoides evolutivos en el flujo de datos. Finalmente, D-Stream particiona y fusiona grupos dinámicamente mediante test estadísticos sobre ventanas deslizantes para capturar el concept drift. Todos equilibran estabilidad del modelo, adaptabilidad rápida e interpretabilidad en un escenario inherentemente cambiante y altamente restrictivo.

Entre los algoritmos de agrupamiento de flujos de datos adaptados a datos de trayectorias, TraClus y moreClust destacan por extender modelos incrementales de agrupamiento con actualizaciones eficientes para detectar y hacer seguimiento de nuevos patrones de movimiento en el tiempo. TraClus incorpora información espacial y temporal tanto en su métrica de distancia entre trayectorias como en la evolución adaptativa de grupos densos de trayectorias similares bajo flujos continuos de datos. Por su parte, moreClust refina esta idea con una política de encadenamiento múltiple que permite fusionar y dividir agrupaciones de trayectorias dinámicamente mejorando la captura del concept drift. Ambos algoritmos son especialmente aplicables en escenarios con múltiples objetos en movimiento, donde detectar y reaccionar a cambios rápidos en los flujos de datos requiere de métodos en línea robustos para comprender la evolución subyacente de los patrones de movilidad.

Abordando el agrupamiento de trayectorias, también se encuentra el agrupamiento en micro-lotes como un enfoque intermedio entre el agrupamiento por lotes y el agrupamiento online. Este método combina la actualización periódica de datos, similar al enfoque online, con la eficiencia de procesamiento de conjuntos más pequeños, a diferencia del agrupamiento por lotes. Operando en fragmentos actualizados a intervalos regulares, el agrupamiento en micro-lotes permite adaptarse a cambios temporales significativos, manteniendo la relevancia de los resultados sin la complejidad computacional constante asociada con el agrupamiento online. Este enfoque es especialmente valioso en el análisis de trayectorias, donde se busca un equilibrio entre la actualización de resultados y la eficiencia en el procesamiento de datos, el autor de la presente investigación lo evidencia en (Reyes et al. 2023b).



A pesar de las características particulares mencionadas en los algoritmos de agrupamiento presentados, estos también presentan determinadas deficiencias específicas como en algunos casos, la sensibilidad a la inicialización de parámetros y su dependencia de la configuración previa de los datos. Por ejemplo, K-means y el Modelo de Mezcla Gaussiana (GMM) son especialmente vulnerables a la selección inicial de centroides y componentes, respectivamente, lo que puede conducir a resultados subóptimos. Otros algoritmos como DBSCAN y T-OPTICS requieren ajustes precisos de parámetros, tales como distancia y densidad, que pueden ser complejos de determinar y afectan la calidad del agrupamiento. Además, métodos como TRADBSCAN y TRACLUS están diseñados para datos de segmentos de trayectorias en general, lo que limita su aplicabilidad a otros contextos más específicos. Los algoritmos que manejan datos dinámicos, como D3CAS y D-STREAM, enfrentan retos adicionales en la implementación y ajuste, y pueden no ser adecuados para datos con características estáticas o no estructuradas.

#### 2.4.4. Modalidades de procesamiento

A continuación se explorarán las diversas formas en que se pueden abordar y manipular los datos. Principalmente habrá un enfoque en dos grupos, el primero presentado por las categorías principales: Estáticos y Dinámicos, y segundo presentado por los métodos de procesamiento: Lotes y En línea. Este análisis permitirá comprender cómo cada modalidad impacta en la gestión de datos y su aplicabilidad en diferentes contextos. A través de esta exploración, se identificarán las fortalezas y limitaciones de cada enfoque, proporcionando una guía para elegir la modalidad más adecuada según las necesidades específicas de cada proyecto.

##### Primer grupo: Modalidad estática y dinámica

**Estática.** Se caracteriza por abordar el proceso de agrupamiento en un único paso, sin tener en cuenta posibles variaciones o modificaciones en los datos en el futuro. Este enfoque implica que los algoritmos generan resultados estáticos que no se ajustan a cambios temporales. Un ejemplo representativo de esta modalidad es el algoritmo k-means, que asigna los datos a grupos de manera determinística sin considerar la evolución potencial de los datos con el tiempo. La ventaja de este enfoque radica en su simplicidad y predictibilidad, aunque puede no ser adecuado para conjuntos de datos dinámicos o sujetos a cambios.

**Dinámica.** Se caracteriza por su capacidad para adaptar los grupos a medida que los datos evolucionan con el tiempo. Este enfoque implica un proceso iterativo donde los grupos pueden ajustarse, fusionarse o dividirse conforme a la llegada de nuevos datos o cambios en los existentes. Algoritmos en línea, como MiniBatch K-Means, son ejemplos que ilustran esta modalidad, ya que permiten ajustes continuos en los grupos a medida que se actualizan los datos. Este enfoque es especialmente valioso en situaciones donde la naturaleza de los datos es dinámica, ya que facilita la

captura de patrones emergentes y la adaptación a cambios temporales en el conjunto de datos.

La elección entre este primer grupo de modalidades depende de la estabilidad temporal de los datos y la necesidad de un agrupamiento que evolucione con el tiempo.

### **Segundo grupo: Modalidades por lotes (Batch) y en línea (Online)**

**Por lotes (Batch).** La modalidad por lotes, se refiere a enfoques que procesan el conjunto de datos completo en una sola iteración. Bajo esta modalidad, el algoritmo realiza el agrupamiento después de haber considerado todos los datos de entrada, asignando cada elemento a un grupo en función de la estructura global del conjunto de datos. Este enfoque es eficiente para conjuntos de datos estáticos y de tamaño moderado, ya que no requiere ajustes continuos a medida que llegan nuevos datos. Sin embargo, puede ser menos apropiado en entornos dinámicos o cuando los datos se generan de manera incremental, ya que implica el procesamiento de todo el conjunto de datos en cada iteración.

**En línea (Online).** Contrastando con la modalidad de lotes, la modalidad en línea se refiere a enfoques que actualizan los grupos a medida que se reciben los datos, permitiendo una adaptación continua a la evolución del conjunto de datos. En este enfoque, el agrupamiento se realiza de manera incremental, lo que significa que los grupos pueden ajustarse, expandirse o contraerse conforme nuevos datos se incorporan al sistema. Algoritmos como MiniBatch K-Means son ejemplos de implementaciones en línea, ya que permiten la actualización continua de los grupos sin necesidad de procesar el conjunto de datos completo en cada iteración. Esta modalidad es especialmente beneficiosa en situaciones donde los datos cambian con el tiempo o en aplicaciones que requieren respuestas en tiempo real, ya que permite adaptarse de manera eficiente a cambios dinámicos en los datos.

La elección entre uno de este segundo grupo de modalidades dependerá de las características específicas del conjunto de datos y de los requisitos de la aplicación.

## **2.5. Mecanismo de olvido**

El olvido de datos espacio-temporales se manifiesta como un fenómeno intrínseco en el procesamiento de trayectorias, donde la información tiende a debilitarse o desvanecerse con el transcurso del tiempo, particularmente cuando no se refuerza ni se utiliza con regularidad. Este mecanismo es esencial para optimizar la eficiencia del sistema de almacenamiento, posibilitando la constante actualización y adaptación a nuevas trayectorias.

La finalidad fundamental del olvido espacio-temporal radica en su capacidad para potenciar los métodos adaptativos al centrarse en los datos más actuales descartando aquellos que han perdido relevancia con el tiempo. Desde esta perspectiva,

la optimización de la memoria surge como un resultado de la priorización de la relevancia temporal en este proceso.

Dada la limitación de recursos en el procesamiento de datos de trayectorias, la eliminación progresiva de información no esencial o menos relevante libera espacio para la incorporación de nuevos datos. Este proceso selectivo garantiza que la memoria esté continuamente disponible para almacenar información más pertinente y actualizada.

Adicionalmente, el olvido contribuye a la eficacia del proceso de recuperación de la memoria. Al mantener únicamente la información más relevante y actualizada, se facilita la localización y recuperación de datos cruciales cuando se requiere. Este principio de selectividad en la retención de información mejora la eficiencia en el procesamiento cognitivo al evitar la sobrecarga de datos y favorecer la toma de decisiones fundamentada en la información más reciente y pertinente.

En términos prácticos, el olvido espacio-temporal se establece como un mecanismo adaptativo para garantizar que la memoria de trayectorias sea un recurso eficiente y efectivo en la gestión de la información. Aunque pueda ser interpretado como una limitación, su función es esencial para la capacidad general del sistema para aprender, adaptarse y responder de manera dinámica a un entorno en constante cambio.

## 2.6. Conclusiones

La exploración detallada de técnicas de minería de datos ha destacado a los algoritmos de agrupamiento dinámico como herramientas esenciales para analizar datos de trayectorias y descubrir patrones ocultos.

Esto se debe a la capacidad innata de estos algoritmos para adaptarse a los cambios y variaciones en los patrones de movimiento de los vehículos, convirtiéndolos en una opción eficaz y versátil para el análisis de datos de trayectorias.

Los algoritmos de agrupamiento mencionados también presentan varias deficiencias que afectan su aplicabilidad y eficiencia en contextos específicos. Muchos requieren una estructura de datos particular, lo que puede limitar su uso en situaciones donde los datos no se ajustan a estos requisitos. Además, algunos enfoques no son adecuados para el análisis de trayectorias en tráfico urbano, ya que no manejan bien los datos de trayectorias o la dinámica del tráfico. La necesidad de información externa adicional y concreta también puede ser una limitación significativa, al igual que la falta de generalización en los tipos de datos, que puede llevar a incompatibilidades con la información del tráfico. En algunos casos, la complejidad en la implementación de ciertos algoritmos puede representar una barrera considerable para su adopción y uso efectivo en aplicaciones prácticas.

La dinámica del tráfico vehicular y la evolución temporal de las trayectorias son elementos fundamentales que deben tenerse en cuenta a la hora de procesar trayectorias, por lo que se seleccionó una técnica capaz de procesar flujos de datos como

la adecuada para la propuesta de esta tesis debido a que facilita un análisis continuo y en tiempo real que es esencial para capturar la naturaleza cambiante del tráfico vehicular.

La relevancia de emplear técnicas de agrupamiento dinámico para el procesamiento de trayectorias es fundamental para descubrir patrones ocultos en los movimientos realizados por vehículos. La aplicación de este tipo de algoritmos se constituye como un método crucial en la identificación de patrones similares formados a partir del análisis de trayectorias vehiculares, permitiendo el entendimiento de comportamientos recurrentes que podrían pasar desapercibidos en un análisis individual.

Además, el olvido de datos espacio-temporales también es un aspecto que cobra importancia en el análisis de trayectorias, donde la información tiende a ser menos importante con el tiempo si no se utiliza con frecuencia. Esta dinámica permite la adaptación de nuevas trayectorias y evita la saturación del sistema con datos antiguos, esto debido a que se prioriza los datos más recientes y desecha aquellos menos relevantes con el tiempo.

## Capítulo 3

# Identificación de patrones

**E**n este capítulo se aborda la identificación de patrones basado en algoritmos de agrupamiento de trayectorias. En este apartado se establece la base de los métodos investigados que han aportado con contribuciones significativas para el desarrollo de esta investigación, donde se abordan las complejidades inherentes a la identificación de patrones en las diversas trayectorias examinadas. En este contexto, se inspeccionan algunos de los factores determinantes que influyen en la identificación de dichos patrones, revelando conexiones y particularidades que contribuyen a una comprensión más precisa de los eventos en cuestión. La estructuración de esta sección responde a la necesidad de establecer las bases metodológicas relevantes para la identificación de trayectorias vehiculares.

El análisis meticuloso de las diversas trayectorias revela patrones intrínsecos, destacando la compleja red de factores que influyen en su configuración. La aplicación de técnicas de agrupamiento es adecuada para hallar las similitudes entre las trayectorias estudiadas, permitiendo establecer una base que podría utilizarse en futuras investigaciones y propuestas para profundizar en estas técnicas enfocados a las trayectorias de origen vehicular. Este capítulo, establece un fundamento sólido en el funcionamiento de algunas de la variantes basadas en técnicas de agrupamiento enfocado a trayectorias vehiculares, además se destaca la importancia de la identificación de patrones en el estudio de trayectorias y su aplicación práctica mediante técnicas de agrupamiento.

A continuación, se describen cinco publicaciones realizadas por el autor de esta tesis, por medio de las cuales han quedado documentados los inicios de la solución aquí propuesta. Cada una de ellas aborda aspectos clave para comprender las trayectorias de los vehículos, utilizando diversas técnicas de agrupamiento con el fin de revelar patrones significativos. Los dos primeros trabajos, descritos en las secciones 3.1 y 3.2, estuvieron enfocados en el agrupamiento de segmentos de trayectorias. En esta primera etapa, el objetivo fue identificar las zonas comunes entre trayectorias. Se propuso una mejora en la conformación de los grupos introduciendo el concepto de pivote como guía durante el proceso de actualización del modelo analizando las ubicaciones iniciales a fin de promover una cobertura uniforme por partes de los grupos en la zona de interés. Estos pivotes guiaron el proceso de actualización del

modelo. Hasta aquí, el análisis no estaba enfocado en el congestionamiento vehicular sino en el reconocimiento de las zonas por las que transitaban los vehículos. Por lo tanto, con el objetivo de comenzar a analizar el flujo vehicular, se desarrolló un tercer trabajo centrado en la característica de velocidad de los vehículos en circulación. Se adaptó un algoritmo de agrupamiento dinámico y se representaron los patrones obtenidos a través de mapas interactivos dinámicos con el objetivo de facilitar la categorización del flujo vehicular en distintos horarios. La sección 3.3 describe esta propuesta.

La cuarta investigación llamada Agrupación Dinámica de Trayectorias de Vehículos investiga la identificación de patrones dinámicos observados de los vehículos que se agrupan en entornos cambiantes de tráfico a lo largo del tiempo. Se presta atención a posibles cambios de velocidad y cambios en las densidades de vehículos.

El quinto trabajo es denominado Método de Procesamiento de Flujo de Datos para el Agrupamiento de Trayectorias, se enfoca en la identificación continua de patrones a través del procesamiento de flujos de datos. Proporciona una perspectiva dinámica y en tiempo real de la evolución de los patrones, clasificándolos según la presencia del patrón. Esto podría revelar cambios temporales en la densidad del tráfico, incluyendo períodos de mayor o menor congestión.

### **3.1. Método de agrupamiento de trayectorias GPS para la toma de decisiones en sistemas de transporte inteligente**

En el mundo de los sistemas de transporte inteligente, la tarea de identificar patrones en las complejas trayectorias para entender los patrones ocultos en las rutas de los vehículos es esencial. Se han examinado varias técnicas fundamentales de la línea de investigación, destacando los que tienen su aplicación en el agrupamiento de trayectorias.

Se considera el empleo del método K-means como base del agrupamiento, junto con la métrica de similitud de distancia de Hausdorff y el concepto de pivote para recálculo del centroide, cada técnica ha sido seleccionada con precisión para el desarrollo de un análisis significativo en el contexto de las trayectorias.

La aplicación de K-means está respaldada por su capacidad para distribuir conjuntos de datos en grupos coherentes, se plantea como una herramienta clave para el análisis detallado de las complejas trayectorias vehiculares en los entornos urbanos debido a su robustez y eficacia en diversos estudios. Estas técnicas son elegidas por su capacidad para mejorar la precisión y relevancia de los agrupamientos de trayectorias, permitiendo un análisis más efectivo de los datos, el autor de la presente investigación lo evidencia en (Reyes et al. 2020).

Un aspecto importante de cualquier técnica de agrupamiento es la correcta inicialización de los centros iniciales. En este caso, se propone hacer una cuadrícula del plano que cubra todas las subtrayectorias. La distancia de Hausdorff se utiliza como

medida de similitud entre subtrayectorias (Gao y Leung 2002). Como resultado, el método devuelve el conjunto de grupos formados.

Los grupos se asignan inicialmente mediante la técnica de reticulado, lo que constituye otra diferencia significativa con los métodos analizados debido a que, la selección de centroides es un paso crítico en los métodos de agrupamiento, los procesos siguientes se realizan sobre este reticulado.

### 3.1.1. Segmentación de las trayectorias

Siguiendo con la investigación y habiendo definido las técnicas de cuadrícula y pivote que mejoran la precisión en la agrupación de trayectorias, se aborda un paso fundamental el cual es la segmentación de trayectorias. Este paso es fundamental, ya que se busca encontrar similitudes entre segmentos de trayectorias que posean orientaciones similares centrándose especialmente en mantener una orientación estable. Cada segmento, definido por la estabilidad en la orientación, permite capturar tramos de trayectorias con cambios mínimos o insignificantes en la dirección. Este enfoque se alinea estratégicamente con el objetivo de analizar con mayor precisión los patrones de movimiento, evitando la pérdida de información en cambios abruptos de dirección.

Se emplea un proceso basado en ángulos, donde se calcula el ángulo  $\alpha$  formado por el segmento de línea que conecta los dos primeros puntos de las trayectorias. Este ángulo actúa como referencia para la dirección de la subtrayectoria en formación. Luego, se evalúan los ángulos que surgen entre puntos sucesivos, como el ángulo formado por el segmento de línea que conecta el segundo y tercer punto.

La comparación entre este nuevo ángulo y  $\alpha$  es crucial. Si la diferencia entre los dos ángulos es menor que un valor umbral definido previamente, conocido como tolerancia angular, el punto en cuestión se agrega a la subtrayectoria actual. Esta subtrayectoria continúa creciendo mientras los ángulos sigan dentro de la tolerancia angular, lo que permite capturar segmentos con cambios suaves en dirección.

Si la diferencia supera la tolerancia angular, se concluye la subtrayectoria actual y se inicia una nueva. El proceso se repite para todos los puntos de la trayectoria, generando un conjunto de subtrayectorias que representan los diferentes segmentos estables y coherentes. Este conjunto de segmentos se convierte en la entrada para el siguiente paso, el cual está conformado por un método de agrupamiento de subtrayectorias GPS, el Pseudocódigo 1 muestra el algoritmo que resume este proceso.

### 3.1.2. Agrupamiento de subtrayectorias

Continuando con la estrategia de agrupamiento de subtrayectorias, se procede a definir los centroides, estos son definidos como el conjunto de puntos pivote que representa el centro geométrico de un grupo específico de subtrayectorias. Los puntos de cada centroide son calculados a partir de las posiciones de las subtrayectorias en el grupo.

**Pseudocódigo 1** Pseudocódigo del método

**Input:** Trayectoria  $T = (P_1, P_2, \dots, P_n)$  formada por ubicaciones GPS y el valor  $THRESHOLD$  representando la tolerancia angular

**Output:** La lista de subtrayectorias  $L$

```

1: Calcular el ángulo  $\alpha$  determinado por  $P_1$  y  $P_2$ 
2: Añadir  $P_1$  y  $P_2$  a la subtrayectoria.
3: for  $i = 3$  to  $n$  do
4:   Calcular el ángulo  $\beta$  determinado por  $P_{i-1}$  y  $P_i$ .
5:   if  $(\beta - \alpha) < THRESHOLD$  then
6:     Añadir  $P_i$  a la subtrayectoria.
7:   else
8:     Añadir subtrayectoria a  $L$ )
9:     Calcular el ángulo  $\alpha$  determinado por  $P_i$  y  $P_{i+1}$ 
10:    Inicializar una nueva subtrayectoria formada por  $P_i$  y  $P_{i+1}$ 
11:     $i = i + 1$ 
12:   end if
13: end for

```

Luego se emplea una técnica de agrupamiento partitivo con un enfoque winner-take-all. Los ejemplos se asignan a los centroides más cercanos y luego se actualizan estos centros. La investigación presenta una mejora al algoritmo k-means en la que se propone una estrategia de inicialización de centroides mediante una cuadrícula, se divide el área cubierta por las trayectorias en sectores uniformes y se establecen 4 centroides en cada sector, los centroides cortan el sector a la mitad, con ángulos incrementales de 45 grados. Estas posiciones iniciales buscan capturar diversas inclinaciones presentes en los grupos de subtrayectorias. La asignación de subtrayectorias a los centroides se basa en la distancia de Hausdorff para segmentos de línea. Posteriormente, se procede al recálculo de los centroides.

El recálculo de los centroides es de vital importancia ya que desempeña un papel esencial en la solidez del análisis. Este procedimiento involucra el ajuste de la posición de los centroides, garantizando que los centroides se adapten con precisión a las características distintivas de las subtrayectorias agrupadas.

La Figura 3.1 representa el proceso de recálculo del centroide utilizando el pivote donde  $T_1$ ,  $T_2$  y  $T_3$  representan las trayectorias y  $P_1$  representa el primer punto del pivote a analizar. El cálculo de la distancia de todos los puntos de la trayectoria  $T_1$  al punto  $P_1$  se realiza utilizando la distancia euclidiana, y se selecciona el punto con la distancia más corta, denotado como  $A_1$ . Este punto se toma como referencia para calcular la distancia de todos los puntos de la trayectoria  $T_2$  al punto  $A_1$ . Como resultado de esta operación, se selecciona el punto con la distancia más corta y se denota como  $B_1$ , este punto se toma como referencia para calcular la distancia de todos los puntos de la trayectoria  $T_3$  a  $B_1$  y se selecciona el punto con la distancia más corta, denotado como  $C_1$ . Este proceso se lleva a cabo para todas las trayectorias y la información analizada se almacena en forma de  $(A_1, B_1, C_1, \dots, Z_1)$ . Esta información permite el cálculo del valor medio de los datos almacenados para longitud y latitud,



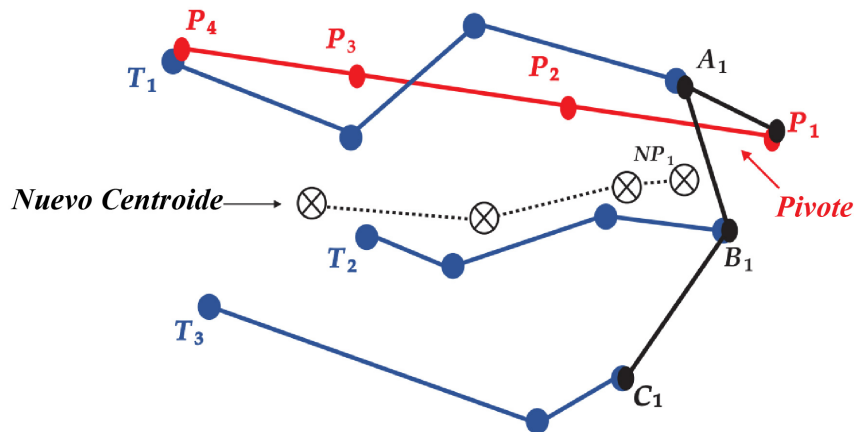


FIGURA 3.1: Representación del proceso de recálculo del centroide mediante un pivote

construyendo así el primer punto, denominado  $NP_1$ , del nuevo centroide. Los puntos restantes del centroide se calculan de la misma manera, pero comenzando en los puntos restantes del pivote ( $P_2, P_3, \dots, P_n$ ).

### 3.1.3. Métricas de calidad para agrupamiento de trayectorias

El recálculo del centroide puede contener graves errores si se utilizara una trayectoria seleccionada aleatoriamente como trayectoria elegida para iniciar el análisis, y se tratara de una trayectoria con muy pocos puntos, es decir, la trayectoria menos similar del grupo. Con el fin de evitar o reducir estos errores, se propone utilizar el pivote (trazado según las características de las trayectorias que componen un grupo) como trayectoria inicial o guía, para realizar los cálculos de los puntos del nuevo centroide, como se muestra en la Figura 3.1. El trazado del pivote se construye teniendo en cuenta dos aspectos importantes de todas las trayectorias que componen un grupo: el desplazamiento del grupo de trayectorias, determinando si el pivote se trazará vertical u horizontalmente, y el valor medio de la cantidad de puntos de todas las trayectorias del grupo.

### 3.1.4. Consideraciones relevantes

En el análisis de un novedoso método de agrupamiento de trayectorias basado en k-means con criterios angulares, se resaltaron elementos clave para investigaciones futuras. La elección del umbral para la segmentación y la estrategia de inicialización de centroides mediante una cuadrícula se presentaron como consideraciones cruciales. Se denotó la necesidad de explorar enfoques más sólidos para la selección de umbrales y estrategias de inicialización, que reflejaron de manera más precisa la distribución real de las trayectorias en el área de interés, destacando como una dirección esencial para futuros desarrollos.

El método innovador, basado en información angular y técnicas de agrupamiento basadas en pivotes, mostraron resultados prometedores al segmentar trayectorias en subtrayectorias más cortas, evitando cambios bruscos en la dirección y generando agrupamientos representativos. Aunque se reconocen limitaciones, como la dependencia de la distancia de Hausdorff y la subóptima cuadrícula para la inicialización de centroides, estos hallazgos sugirieron un avance significativo en el análisis de trayectorias GPS y la identificación de patrones de congestión. La investigación futura se orienta hacia la mejora de la selección de umbrales y métricas de similitud, así como la exploración de estrategias de inicialización de centroides más adaptativas, con el objetivo de consolidar y generalizar la aplicabilidad del método en diversos contextos.

Este avance se ha convertido en un elemento clave para las propuestas posteriores, donde se ha extendido la atención al proceso de inicialización, la segmentación y la selección de umbrales, consolidando un proceso más robusto en el análisis de trayectorias. Además, cabe destacar que el método propuesto ha analizado la implementación mejoras en la métrica de similitud, enfocándose en perfeccionar la precisión de la evaluación de la proximidad entre los datos de las trayectorias.

### **3.2. Propuesta de un método de agrupamiento de trayectorias de vehículos basado en pivotes**

Para una mejora del método, el autor ha introducido mejoras sustanciales, como se expone en (Reyes et al. 2021a). La focalización principal radica en la segmentación, permitiendo el uso de subtrayectorias más cortas con cambios de dirección menores, particularmente enfocándose en cambios bruscos en la dirección o velocidad de los vehículos.

En cuanto a la determinación de la posición de los centroides iniciales, se implementó una estrategia que emplea una retícula regular llamada cuadrícula, esta asegura una distribución uniforme en el área de estudio.

La cuadrícula se configura como un entramado que abarca la extensión del área de estudio. Su implementación asegura una distribución uniforme sobre el territorio donde se despliegan las trayectorias y desempeña un papel crucial al establecer la posición inicial de los representantes medios de cada grupo. Se destaca la importancia de considerar que las dimensiones de esta red inciden directamente en la granularidad espacial del proceso de agrupamiento, siendo este un aspecto que requiere definición en función de las características específicas del conjunto de datos a analizar.

Los centroides iniciales se calculan creando una subtrayectoria rectilínea, basada en las posiciones medias de las subtrayectorias y los vértices en relación con cada cuadrícula de celdas en el conjunto de cuadrículas. Las subtrayectorias iniciales del centroide se trazan diagonalmente en cada celda, como se muestra en la Figura 3.2.

Estas posiciones iniciales no son las únicas posibles. Realizando experimentos colocando estos centroides perpendiculares, como se muestra en la Figura 3.2, se detectó que no tuvo un impacto significativo en los resultados generales. Por tanto, dentro de esta investigación, se puede decir que el uso de cualquier diagonal produce resultados similares.

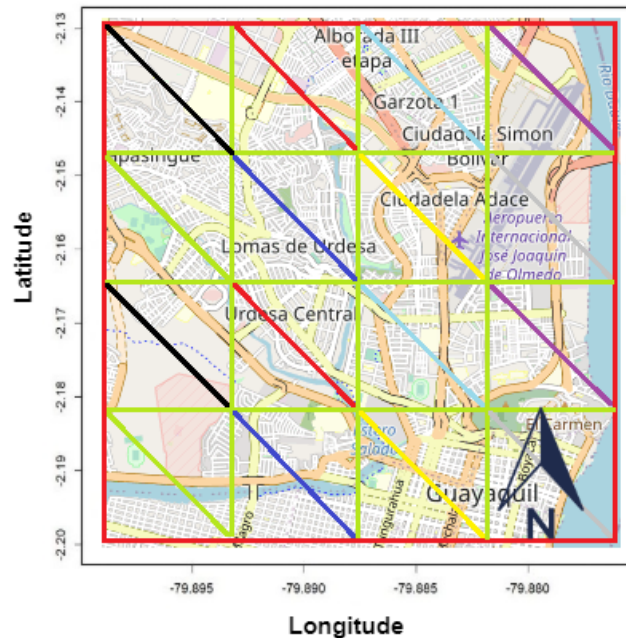


FIGURA 3.2: Reticulación inicial y centroides de subtrayectorias.

### 3.2.1. Agrupamiento de subtrayectorias

El anterior procedimiento de cálculos por subtrayectorias sentó las bases para una adaptación del algoritmo K-means especializada en subtrayectorias. Al igual que en el trabajo anterior, el desarrollo consistió en la integración de pivotes como una estrategia que reconfigura la representación del centroide respecto a las subtrayectorias del grupo, por lo que el algoritmo propuesto es denominado K-pivote. Aunque visualmente lineal, este enfoque aprovecha la colocación precisa de los pivotes en cada celda, atrayendo el centroide de manera efectiva hacia el área más representativa. Este refinamiento equilibra la sencillez visual con una eficacia mejorada en la identificación de patrones en el análisis de subtrayectorias.

Se utilizó la distancia de Hausdorff como medida de similitud para asignar cada subtrayectoria al conglomerado más cercano. La métrica de Hausdorff mide la distancia entre dos subtrayectorias de un espacio métrico definido por (Rockafellar y Wets 2009).

Su uso es el resultado de la necesidad de comparar elementos compuestos por varias posiciones geográficas. El proceso comienza con la asignación de cada subtrayectoria al centroide más cercano. La ubicación inicial de los centroides se indicó

inmediatamente después de la definición de la red como se indica en la configuración de la red. Como resultado de este proceso de asignación se forma la primera distribución de las subtrayectorias en grupos. A partir de aquí, las subtrayectorias se procesan en el contexto de cada grupo. A diferencia del trabajo anterior, se utiliza un proceso basado en cuadrícula de celdas para dibujar el pivote, por lo que es necesario identificar el área en la que se encuentran la mayoría de las subtrayectorias del grupo, así como la dirección principal o representativa de estas subtrayectorias.

El área trazada contendrá el pivote confinado, se obtiene organizando las posiciones de las subtrayectorias de cada grupo en orden creciente en dos listas diferentes, una para latitud y otra para longitud. Las dos listas que contienen todas las posiciones de la subtrayectoria se dividen por igual, obteniendo así cuatro listas. Luego se calcula el valor promedio de cada una de estas listas, generando las siguientes métricas: mitad superior de longitud, mitad inferior de longitud, mitad superior de latitud y mitad inferior de latitud.

En las figuras referenciales, el eje horizontal está asociado con la longitud, mientras que el eje vertical está asociado con la latitud.

Estos valores determinan cuatro posiciones:  $Pt_1$  (mitad inferior de latitud, mitad inferior de longitud);  $Pt_2$  (mitad superior de latitud, mitad inferior de longitud);  $Pt_3$  (mitad inferior de latitud, mitad superior de longitud);  $Pt_4$  (mitad superior de latitud, mitad superior de longitud), como se muestra en la Figura 3.3.

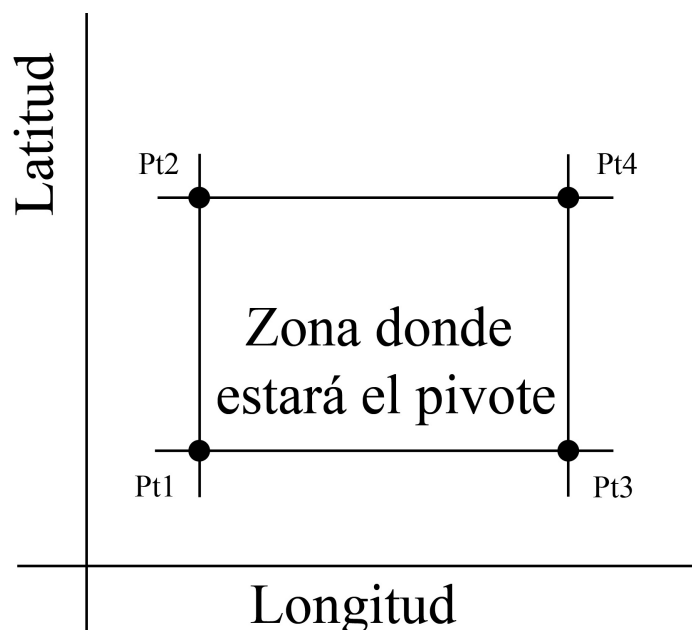


FIGURA 3.3: Área de trazado del pivote.

De esta forma, el pivote se traza en el área donde se encuentran la mayoría de las subtrayectorias del cúmulo. Los ángulos de la subtrayectoria del grupo deben promediarse para obtener la dirección del pivote.

La Figura 3.4 ejemplifica el ángulo  $\alpha$  correspondiente a una subtrayectoria, con respecto al eje de abscisas. Tenga en cuenta que, para cada subtrayectoria, solo se considera su posición inicial.

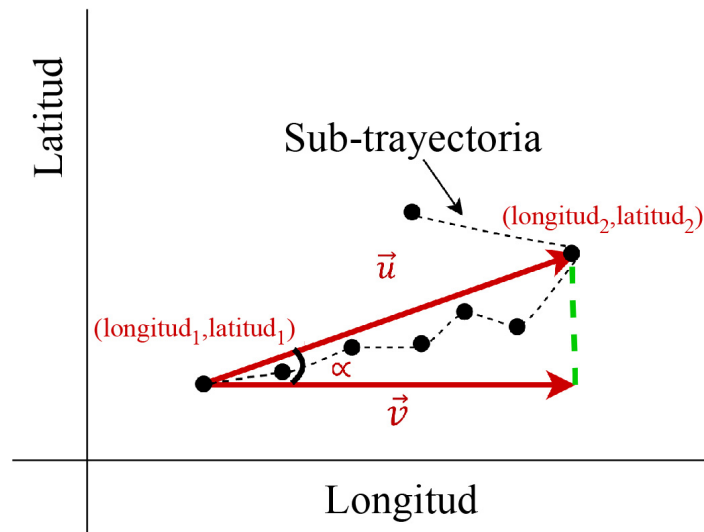


FIGURA 3.4: Ángulo formado por una subtrayectoria respecto del eje de abscisas.

El ángulo  $\alpha$  se calcula según la Ecuación 3.1, medido en el intervalo  $[-\pi, \pi]$

$$\alpha = \arccos \left( \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| |\vec{v}|} \right) \quad (3.1)$$

Finalmente, se promedian los ángulos de todas las subtrayectorias de un grupo, obteniendo así el ángulo  $j$ , que define la dirección del pivote dentro del área de trazado. Este ángulo puede ser positivo o negativo, permitiendo dibujar el pivote en correspondencia con la dirección y ángulo de las subtrayectorias que forman el grupo.

La traza del pivote depende del signo del ángulo. Si el ángulo medio ( $j$ ) es positivo, se calcula la posición media del segmento inferior de la superficie construida. Por el contrario, si el ángulo medio ( $j$ ) es negativo, se calcula la posición media del segmento superior. En cada caso, a partir de esa posición media  $(longitud_1, latitud_1)$ , se dibuja el pivote como se ilustra en la Figura 3.5.

Luego, la Ecuación 2 determina las posiciones del pivote, donde  $(longitud_1, latitud_1)$  es la posición media encontrada previamente.

$$latitud_1 = \tan \varphi (longitud_1 - longitud_2) + latitud_2 \quad (3.2)$$

Cada pivote está definido por un número dado de posiciones pertenecientes al segmento de línea dentro del área de trazado previamente definida. Este número coincide con el número medio de posiciones geográficas que componen las subtrayectorias del grupo correspondiente.

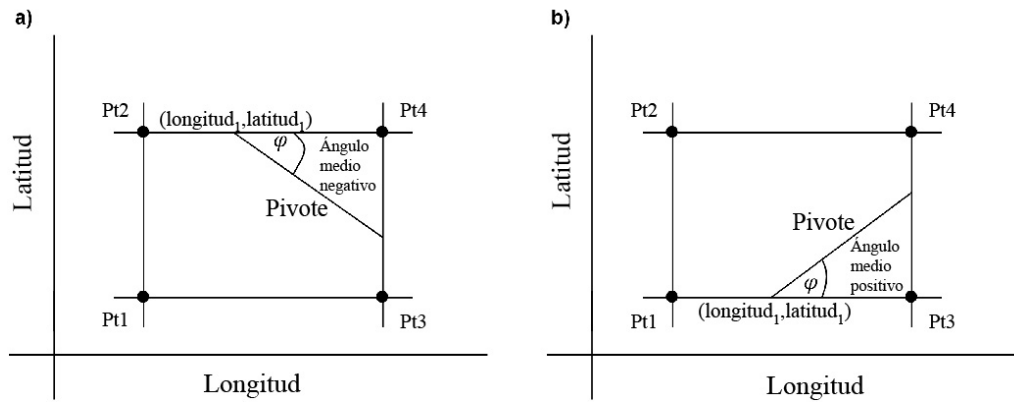


FIGURA 3.5: Pivote con (a) ángulo medio negativo y (b) positivo.

Una vez que se calcula el pivote, se usa como guía para identificar las posiciones de las subtrayectorias que, cuando se promedian, dan como resultado una nueva posición del centroide.

La Figura 3.6 muestra cómo se identifican las posiciones siguiendo la línea de pivote. En esta figura,  $T_1$ ,  $T_2$  y  $T_3$  representan subtrayectorias y  $P_1$  representa la primera posición de pivote que se procesará.

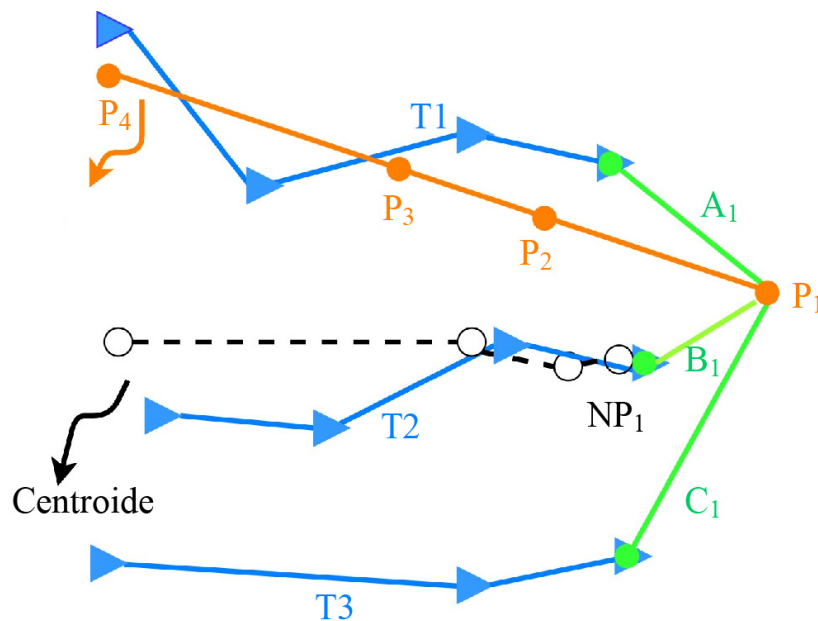


FIGURA 3.6: Posiciones utilizadas para determinar el centroide.

El proceso comienza con la identificación de la posición perteneciente a la trayectoria más cercana ( $T_1$ ) a las posiciones ( $P_1$ ) utilizando la distancia euclidiana que también se utiliza para la manipulación de datos y segmentación de trayectorias.

Esta posición se denota como  $A_1$  en la Figura 3.6. De manera similar, las posiciones  $B_1$  y  $C_1$  resultan ser las más cercanas a  $P_1$  para las trayectorias  $T_2$  y  $T_3$ , respectivamente. A medida que se identifica cada una de estas posiciones, sus valores

de latitud y longitud se almacenan para calcular su promedio, una vez que se han inspeccionado todas las subtrayectorias en el grupo. Este promedio corresponde a la primera posición del centroide actualizado ( $NP_1$ ). Las posiciones restantes de la subtrayectoria del centroide se calculan de la misma manera, considerando cada una de las otras posiciones de pivote. Por lo tanto, al final del proceso de actualización del centroide, el número de posiciones que componen el centroide coincide con el del pivote.

### 3.2.2. Consideraciones relevantes

En este estudio, se ha introducido la técnica de agrupamiento K-pivot, cuya contribución clave radica en el uso estratégico de un pivote para guiar el recálculo de centroides en la agrupación de subtrayectorias. Este enfoque innovador se destacó por su comparativa efectividad, evidenciada en la distribución uniforme de conglomerados en el área de estudio y resultados satisfactorios en términos de calidad de agrupamiento según el índice de Silhouette. La capacidad del método K-pivot para identificar trayectorias conglomeradas, especialmente indicativas de congestión vehicular, agregó un valor significativo al proporcionar información sobre la dirección del tráfico, esencial para la toma de decisiones en la gestión del flujo vehicular.

Para las futuras investigaciones, se sugiere una adaptación en la definición de la orientación del pivote, considerando otras características de la subtrayectoria que podrían influir en su dirección y ángulo. Además, la exploración de algoritmos de agrupamiento dinámico sin la necesidad de definir centroides iniciales se plantea como una vía prometedora para mejorar la versatilidad y la eficiencia del método. Estas propuestas son planteadas como puntos de enfoque para perfeccionar la técnica K-pivot y abordar posibles limitaciones identificadas durante la comparación con métodos alternativos.

En el siguiente trabajo, se consideró un proceso de inicialización dinámico y se exploró la adaptación de una técnica hacia un algoritmo diseñado para el procesamiento de flujos de datos para mejorar la versatilidad y la aplicabilidad del método en un contexto donde se analizan trayectorias a lo largo del tiempo. Asimismo, se planteó reenfocar la base de la técnica K-pivot mediante la implementación de mejoras destinadas a procesar flujos de datos de manera eficiente y mejorar el tratamiento de las trayectorias considerando su naturaleza. Para la propuesta de esta tesis se explora la posibilidad de utilizar una métrica que aproveche la información inherente a las trayectorias y proponer una alternativa como medida indicativa de posible congestión, con el objetivo de perfeccionar la identificación de patrones específicos dentro del tráfico.

### 3.3. Análisis de flujo vehicular utilizando grupos

A raíz de las limitaciones identificadas en el enfoque original del método K-pivot, se ha explorado una nueva dirección que implica la utilización de un algoritmo dinámico como base. Este algoritmo base posibilita la inicialización y mejora continua de los grupos formados, buscando así superar las restricciones previamente identificadas. Paralelamente, se ha decidido adoptar una metodología delineada por el autor de la tesis para el análisis del flujo vehicular en el tráfico, como se presenta en el trabajo de (Reyes et al. 2021b). Esta metodología está representada en la Figura 3.7, sigue tres pasos cruciales: el primer paso se enfoca en la representación precisa de los datos de las trayectorias en el área de interés; el segundo paso emplea una adaptación de un algoritmo de agrupamiento dinámico para identificar relaciones clave; y finalmente, el tercer paso implica la creación de recursos interactivos para visualizar los resultados de manera efectiva. Cada uno de estos pasos se aborda de manera detallada a continuación.

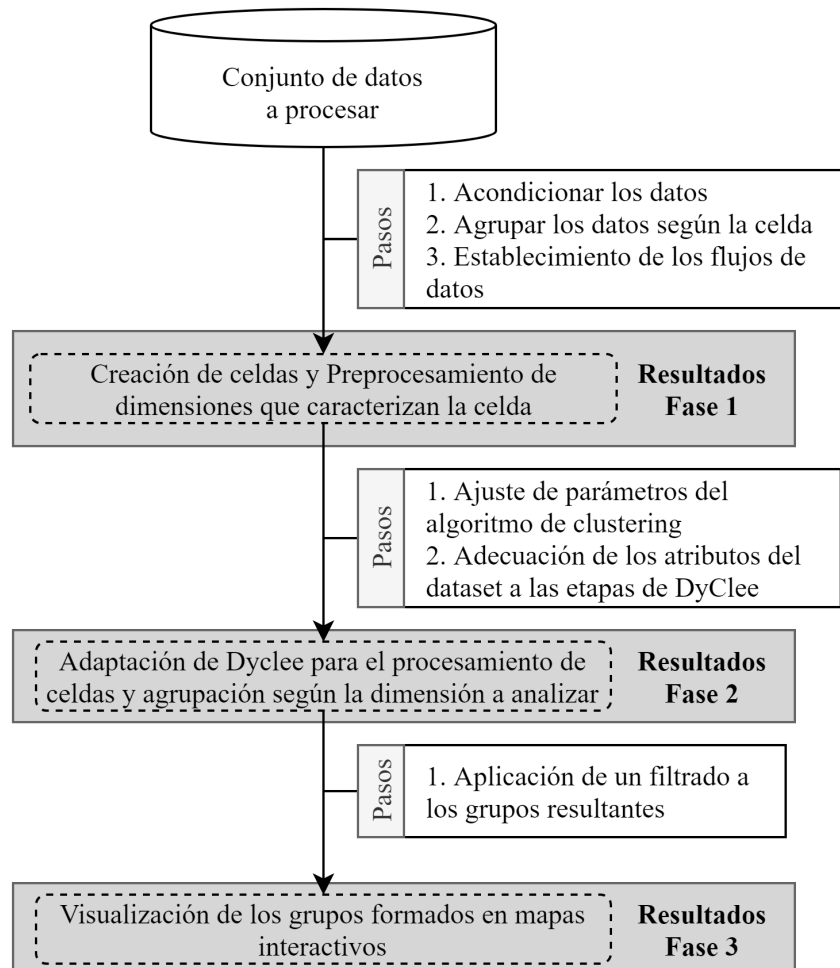


FIGURA 3.7: Metodología propuesta



### 3.3.1. Definición de celdas para resumir información de trayectorias

Iniciando el proceso, el primer paso consiste en brindar una representación precisa de los datos que conforman las trayectorias. Esto requiere, en primer lugar, definir claramente el área de interés geográfico al cual pertenecen las trayectorias objeto de análisis. Posteriormente, esta área se divide de manera uniforme en celdas para facilitar la posterior manipulación de los datos.

Este procedimiento implica la subdivisión sistemática del área de estudio en unidades más pequeñas y uniformes, conocidas como celdas, con el objetivo de optimizar el análisis de los puntos en las trayectorias. Esta estrategia conlleva varias ventajas, en primer lugar, la segmentación en celdas proporciona una representación estructurada del área de estudio al dividirla en regiones más pequeñas y organizadas, lo que facilita la identificación y seguimiento de patrones específicos. Además, esta técnica facilita la gestión y análisis eficiente de grandes conjuntos de datos, ya que cada celda representa una unidad independiente que puede ser evaluada por separado. Otra ventaja significativa es que la segmentación en celdas simplifica la implementación de algoritmos y técnicas de análisis, permitiendo una interpretación más clara de los resultados.

El tamaño de cada celda dependerá de la precisión con la que se desee analizar el flujo vehicular. En este trabajo se utilizaron celdas de  $200\text{ m}^2$ . Este es un detalle importante a tener en cuenta ya que la información analizada corresponde a un resumen de lo que sucede en cada celda durante un período específico.

La metodología analiza colectivamente lo que ocurre en cada celda, en lugar de considerar las trayectorias individuales de cada vehículo, facilitando así el análisis y la visualización. En cuanto al flujo vehicular representado en cada celda, se abordó en un modo por lotes en períodos de 3 minutos, aunque la adaptabilidad del análisis permite períodos más cortos, como un minuto, según las necesidades específicas. Cada período se considera una evolución, ya que el algoritmo dinámico actualiza el agrupamiento de manera secuencial con la incorporación de cada bloque de datos. En cada evolución, se introduce un flujo de datos para realizar cálculos específicos en cada celda y obtener información característica de estas áreas.

### 3.3.2. Agrupamiento adaptativo y dinámico de trayectorias

En el segundo paso de la metodología, se implementa una versión adaptada de un algoritmo dinámico (Barbosa Roa, Travé-Massuyès y Grisales-Palacio 2019), este algoritmo se caracteriza por tener un enfoque unificado para rastrear entornos en evolución, consta de dos etapas: una de agrupamiento basado en distancias y otra basada en densidad. En la primera etapa, los datos alimentan incrementalmente al algoritmo de agrupamiento basado en distancias, generando microgrupos. Luego, el algoritmo basado en densidad analiza estos microgrupos para formar los grupos finales. La adaptación está enfocada para procesar las trayectorias vehiculares dentro de celdas.

En este contexto, cabe destacar que la primera fase de este agrupamiento dinámico, centrada en la construcción de microgrupos, se ajustó considerando las velocidades de las ubicaciones GPS de cada trayectoria presentes en cada una de las celdas, en lugar de utilizar directamente las ubicaciones GPS y su densidad, como se propuso en el artículo original. Así, durante un período determinado, cada celda se representa por la velocidad promedio de las secciones de trayectoria que alberga. Este enfoque implica recortar las trayectorias de manera apropiada, considerando que variaciones de velocidad pueden resultar en una falta de registro o en un número limitado de ubicaciones GPS para vehículos que viajan a velocidades muy altas al atravesar una celda. Además, es importante considerar que las velocidades deben promediarse teniendo en cuenta los vehículos y no el número de ubicaciones registradas. Además cada celda representa un patrón que está dada por el centroide, para esta investigación este se define como el valor promedio de las velocidades de los puntos vehiculares que se encuentran dentro de una celda específica durante un período de tiempo dado. En cuanto al tamaño de los microgrupos, el valor del parámetro “relative size” especifica el tamaño relativo del parámetro “hipercaja” con respecto al área a procesar. Es decir, a medida que disminuye su valor, aumenta el número de microgrupos y viceversa.

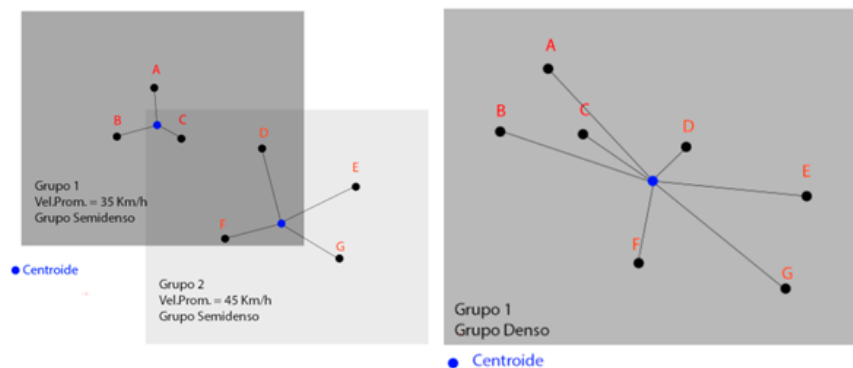


FIGURA 3.8: Funcionamiento de segunda etapa de algoritmo dinámico. (A) Identificación de microgrupos directamente conectados y (B) Expansión resultante

El algoritmo de agrupamiento dinámico en su segunda fase analiza las densidades de los microgrupos formados y los clasifica en dos categorías: densos y semi-densos. A partir de los microgrupos densos comienza a unirse a los que están directamente conectados. Dos microgrupos se considerarán conectados directamente si la distancia máxima a la que pueden encontrarse los centroides de dos microgrupos con densidades similares no excede el valor del parámetro hipercaja. Es decir, a medida que aumente el valor del parámetro hipercaja, el número de grupos será menor y por tanto, los rangos de velocidad en la zona de interés tendrán mayor amplitud. La Figura 3.8 ilustra el funcionamiento de esta etapa.

### 3.3.3. Visualización de rangos de velocidades

Concluida la formación de cada grupo en el registro, se genera un historial para cada celda incorporada en dichos grupos, documentando datos de trayectoria en el período de tiempo analizado. Aunque cada resultado se visualiza en tiempo real en un mapa interactivo, el registro correspondiente permite reconstruir situaciones previas. Cada agrupación culmina en la creación de un mapa interactivo, ofreciendo un análisis gráfico y dinámico de la información clave de cada celda.

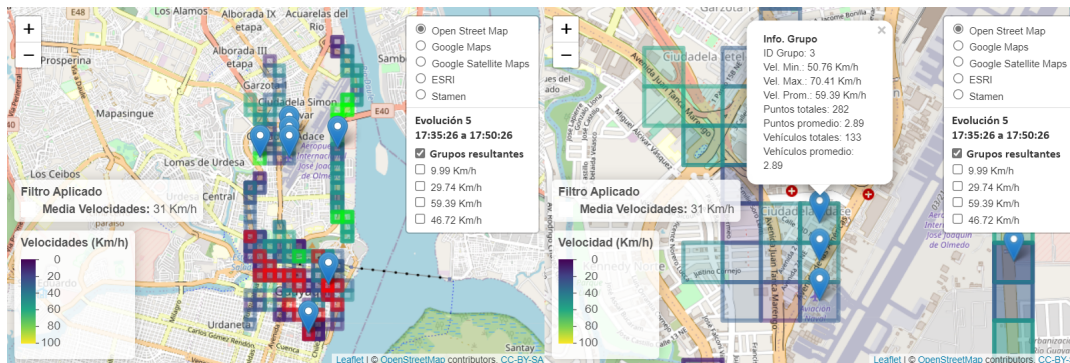


FIGURA 3.9: Mapa a una evolución. (A) Capas correspondientes a los grupos seleccionados. (B) Visualización de marcadores

Este registro posibilita la reconstrucción de todos los mapas desde el inicio del análisis del flujo vehicular, proporcionando una rápida visualización del estado del tráfico. En la interfaz de cada mapa, se incorporan controles para seleccionar capas y leyendas de referencia que facilitan la interpretación de los resultados representados. Se generan dos tipos de mapas: uno que refleja la última evolución o período de tiempo analizado, y otro que abarca todas las evoluciones.

En el mapa correspondiente a una evolución específica, cada grupo se representa en capas distintas, permitiendo al usuario seleccionar una o varias capas para filtrar la información de interés. Con el uso de colores diferentes para cada capa, se facilita la visualización y distinción entre múltiples grupos mostrados simultáneamente. Además, el mapa brinda la opción de mostrar marcadores que revelan información tanto del grupo como de la celda seleccionada, como se ilustra en la Figura 3.9 (B).

Por otro lado, el mapa que engloba todas las evoluciones presenta visualmente cada etapa realizada, asignando un color único a cada evolución. En cualquier momento, el usuario puede elegir una o varias evoluciones según el análisis que desee realizar, añadiendo flexibilidad y precisión al proceso.

### 3.3.4. Consideraciones relevantes

Este estudio, fundamentado en un conjunto de datos de tamaño limitado y duración específica, destacó la necesidad imperante de abordar las consideraciones sobre

la representatividad de los resultados. Una validación exhaustiva en diversos contextos urbanos se reveló como esencial para garantizar la aplicabilidad de la metodología propuesta. Por otro lado, la aplicación del algoritmo para la identificación dinámica de grupos de velocidad y la incorporación de mapas interactivos como herramienta de visualización emergieron como innovaciones valiosas. Este enfoque novedoso planteó bases para futuras investigaciones, destacando la importancia de optimizar continuamente el algoritmo y mejorar la interfaz de mapas interactivos para una gestión más eficaz del tráfico. Sin embargo, las limitaciones, como el tamaño limitado del conjunto de datos y la dificultad en la adaptabilidad del algoritmo en entornos con flujos de datos más rápidos, señalaron la necesidad crítica de abordar estas deficiencias para fortalecer la aplicabilidad y robustez de la metodología en escenarios más amplios y complejos.

En este sentido, en el próximo trabajo se tuvo la intención de tomar como base la metodología actual y extenderla a áreas más extensas, buscando generalizar su aplicabilidad a diferentes entornos urbanos. Este nuevo enfoque se centró en superar las limitaciones previas, especialmente la adaptabilidad del algoritmo en escenarios con flujos de datos de mayor volumen, consolidando así la eficacia y versatilidad de la metodología propuesta. También se planificó, abordar la gestión de resultados considerados como ruido o atípicos, lo que añadirá una capa de robustez al método. La propuesta de esta tesis se realiza mediante un procesamiento basado en flujos de datos y análisis a nivel de puntos, incorporando características específicas de los datos agrupados.

### 3.4. Agrupación dinámica de trayectorias de vehículos

Con el objetivo de mejorar la calidad de la metodología previamente descrita, el investigador extiende la evaluación de los resultados identificados como atípicos en el estudio previo, estos se presentaron como resultados con densidades inferiores a la media de densidades con respecto a los demás grupos. En este contexto, se reconoce la importancia de abordar la necesidad de una comprensión más profunda y un manejo más efectivo de los grupos atípicos, considerándolos como zonas que podrían albergar patrones ocultos en el análisis.

Consciente de esta necesidad, se propone la implementación de un proceso adicional específicamente diseñado para el tratamiento de los resultados categorizados como atípicos. Esta adición al método busca fortalecer la robustez y confiabilidad de la metodología empleada, como se detalla en el trabajo del autor de la presente investigación (Reyes et al. 2022c). La ejecución de este proceso adicional tiene lugar después de la generación de los resultados, donde la información de las celdas que componen cada grupo se exporta y guarda.

De manera adicional, se almacena un resumen de los resultados de los grupos obtenidos en la ejecución realizada. Antes de pasar a la visualización de los resultados, se lleva a cabo el procesamiento de los grupos de baja densidad característicos

de los grupos con una cantidad de ubicaciones inferior a la media de ubicaciones registrada en todos los demás grupos. En este paso, se utilizan todos los resúmenes de los grupos, los cuales se ordenan según su velocidad. Posteriormente, se excluyen los grupos con una sola ubicación GPS que, a su vez, no tienen grupos contiguos de baja densidad.

Luego, se seleccionan aquellos grupos de baja densidad que se encuentran directamente conectados con las ubicaciones GPS de otros grupos de baja densidad contiguos para ser considerados como un solo grupo. Esta unión se calcula ponderadamente según el número de celdas para determinar la velocidad promedio del nuevo grupo, y se totaliza la información del número de ubicaciones GPS, celdas y vehículos. Este proceso contribuye a una gestión más óptima de los resultados ya que favorece la identificación de patrones ocultos o emergentes, garantizando una representación que cubre la diversidad de patrones identificados en el análisis.

#### **3.4.1. Consideraciones relevantes**

En este proyecto, se destacó la necesidad de abordar la identificación dinámica de características del flujo vehicular en entornos urbanos en constante evolución. La propuesta metodológica se fundamentó en la representación de trayectorias de origen vehicular en celdas y su procesamiento a través de un algoritmo distintivo, respaldado por la utilidad de mapas interactivos.

La metodología aplicada en el manejo de grupos de baja densidad ha demostrado ser eficaz al mejorar la identificación de patrones ocultos adicionales, permitiendo una diversidad de los resultados del análisis al integrar cuidadosamente los grupos contiguos de baja densidad.

A pesar de los avances significativos, se reconocieron limitaciones sustanciales, como la necesidad de optimizar el tiempo de procesamiento y abordar posibles variables no contempladas, señalando la importancia de una evaluación crítica de estas deficiencias. Se ha expuesto que se debe mejorar la eficiencia y aplicabilidad de la metodología, mediante mejoras tangibles o técnicas como la incorporación incremental de datos.

Además, se plantea que la propuesta de esta tesis adopte un procesamiento de flujos constantes de datos de trayectorias, priorizando la continuidad y adaptabilidad en la identificación dinámica de características del flujo vehicular en entornos urbanos cambiantes. Para los futuros trabajos, se proyecta tomar como base la metodología actual para mejorar y optimizar el notable consumo de recursos asociado a la implementación actual. Se explorarán estrategias específicas, como la adopción de un enfoque en micro-lotes, con el objetivo de perfeccionar la eficiencia de la metodología permitiendo así una gestión más efectiva de los datos.

### 3.5. Método de procesamiento de flujo de datos para el agrupamiento de trayectorias

Los planteamientos anteriores han sido enfocados a procesar datos por lotes, por lo que esta sección explora la adaptación enfocada al procesamiento incremental de flujos de datos para agrupaciones dinámicas de trayectorias, logrando un bajo consumo de memoria y un menor tiempo de procesamiento, lo que permite un análisis ágil del flujo vehicular en un momento dado. A medida que avanza la recopilación de datos de la trayectoria del GPS, la información en cada celda de la cuadrícula se actualiza para reflejar su velocidad promedio durante un período de tiempo determinado. Estas celdas se delimitan al inicio del proceso y su tamaño depende de la precisión deseada del análisis dentro del área de estudio. El procesamiento de esta nueva representación en celdas se analiza utilizando como base una metodología de agrupamiento dinámico de procesamiento por lotes (Reyes et al. 2021b), que fue adaptada en este trabajo para procesar datos en línea y a la que se le incorporó el manejo de una memoria buffer y el uso de índices en la creación de celdas, permitiendo reducir el consumo de memoria y el tiempo necesario para realizar los cálculos, el autor de la presente investigación lo hace evidente en (Reyes et al. 2022b). Como resultado, se pueden identificar zonas con características similares y se genera un mapa interactivo en tiempo real en el que se pueden observar los rangos de velocidad correspondientes al flujo vehicular actual y las zonas donde se producen. La comparación entre el método por lotes y el método presentado en este estudio ha permitido identificar las diferencias obtenidas para el consumo de memoria, el tiempo de ejecución utilizado y la evaluación del coeficiente de Silhouette de los grupos.

Este método puede ser utilizado, junto con otras herramientas, por los gestores de tráfico de una ciudad para planificar vías urbanas, detectar puntos críticos en el flujo de tráfico, identificar situaciones anómalas, predecir futuros comportamientos de movilidad, analizar el flujo vehicular, entre otros. El método se puede utilizar para caracterizar datos correspondientes a trayectorias generadas por un grupo de estudiantes de la Universidad de Guayaquil en Ecuador y datos históricos de taxis de la ciudad de Roma en Italia. Los resultados obtenidos permiten identificar, en cada ciudad, diferentes instantes de tiempo donde los vehículos tienen velocidades comunes, y mediante las medidas de tiempo de ejecución, memoria utilizada y coeficiente de Silhouette, se evidencia una mayor eficiencia que el método por lotes.

En esta propuesta se mejora el procesamiento por lotes (Reyes et al. 2021b) para trabajar en base a micro-lotes y se empleó un método para mejorar el rendimiento del proceso de creación de celdas en una rejilla, haciendo uso de un buffer e índices para optimizar el tiempo y la cantidad de memoria necesaria. Se propone un método que, a diferencia de los métodos tradicionales que procesan trayectorias (Hu et al. 2020), consolida la información de un área determinada y la transforma en celdas con información resumida. Esta información es procesada por un algoritmo de agrupamiento modificado para utilizar una dimensión diferente, en este estudio se

usó la velocidad, obteniendo grupos que reflejan una perspectiva diferente de los patrones. Para facilitar la exploración y visualización de los resultados se propone el uso de una herramienta interactiva.

Se establece un método adaptado para el procesamiento en línea o micro-lotes para el agrupamiento dinámico de trayectorias en grandes volúmenes de información basado en el procesamiento por lotes, utilizando un buffer de datos que reduce el consumo de memoria requerido para procesar grandes cantidades de información de trayectoria. Además, el método hace uso de índices para mejorar el procesamiento que transforma los datos del GPS en celdas que serán utilizadas por el algoritmo de agrupamiento. Los datos correspondientes a los puntos GPS se presentan como microlotes consecutivos de datos ordenados con respecto al sello de tiempo de cada registro.

A diferencia del procesamiento por lotes, cuyos datos requeridos para el procesamiento se obtienen previamente de algún repositorio, en el procesamiento en línea o por microlotes, los datos se reciben a medida que transcurre el tiempo requerido para la ejecución de la agrupación. La construcción del método contempla el uso de procesos secuenciales, los cuales son creados y ejecutados cuando se requiere durante el procesamiento. El primer paso del procesamiento es la recepción de los datos del GPS, los cuales son recogidos del repositorio de forma progresiva durante un determinado período de tiempo, que en este trabajo se fijó en tres minutos, y se almacenan en una memoria intermedia. El segundo paso se encarga de la transformación de estas ubicaciones GPS a información resumida en celdas utilizando índices para mejorar el rendimiento del procesamiento; luego, en un tercer paso, las celdas se agrupan de acuerdo con los rangos de velocidad generados durante el proceso de agrupamiento. El último paso es la generación de mapas interactivos para la visualización de resultados del agrupamiento. Un esquema general del procesamiento de microlotes se puede observar en la Figura 3.10, los procesos utilizados para el tratamiento de cada ciclo, que involucra el procesamiento de todos los pasos del método para un flujo de datos, se repiten iterativamente; para procesar cada conjunto de datos, se realizarán los ciclos necesarios hasta cubrir el período de tiempo requerido y se utilizará un procesamiento en tiempo real.

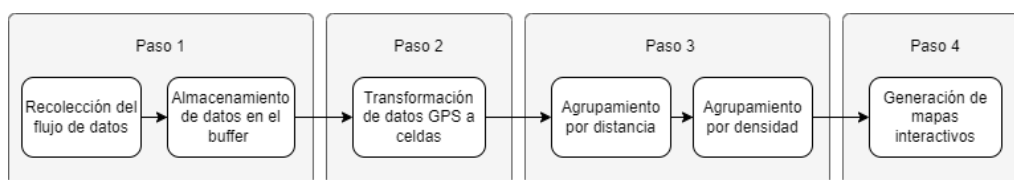


FIGURA 3.10: Esquema general del procesamiento del método por microlotes

### 3.5.1. Paso 1: Uso de un buffer

Una parte del procesamiento se encarga de almacenar en una memoria temporal de manera constante las ubicaciones GPS recolectados de algún repositorio y de acuerdo al período de tiempo que se esté analizando, para contemplar diferentes períodos de ejecución se hace uso de microlotes de memoria dentro del buffer que selecciona, organiza y almacena pequeños conjuntos de datos, este procesamiento permite que el buffer no requiera cargar en memoria la totalidad de los datos, ahorrando espacio y reduciendo considerablemente los cálculos necesarios para la generación de celdas en la rejilla. Cada dato GPS contiene su respectiva marca de tiempo que no se modifica, lo que asegura que cada dato se asigne a un microlote dentro del buffer. El tamaño de este buffer temporal puede contener hasta  $n$  microlotes, siendo flexible la cantidad de microlotes que se pueden almacenar temporalmente, esta capacidad requerida será menor si disminuyen los intervalos de recopilación de los flujos de datos para el buffer. La memoria de cada microlote en el buffer se libera cuando pasa al siguiente paso. Las ubicaciones GPS recopilados por el buffer se organizan y se ordenan según la marca de tiempo a su respectivo microlote, manteniendo de esta forma una coherencia en la recopilación y evitando errores en procesos futuros.

### 3.5.2. Paso 2: Uso de índices para crear celdas con información de resumen

Una celda es la representación de un conjunto de datos con información resumida sobre un área delimitada en la que se incluye la velocidad promedio de los vehículos y la cantidad total de vehículos de cada celda.

El proceso de creación de celdas se basa en un reticulado clásico, donde cada celda almacena registros de información GPS. Se generan índices de longitud y latitud para identificar previamente las celdas que contienen datos útiles, lo que mejora el acceso al permitir que el sistema se concentre únicamente en las celdas que efectivamente tienen información GPS y descartando las que están vacías. Con este enfoque se optimiza el análisis de los datos al evitar el procesamiento innecesario de celdas sin ubicaciones GPS asignadas.

El siguiente paso es conformar las celdas utilizando el conjunto de índices únicos, que determina las celdas que se analizarán para la transformación de los datos en celdas que contienen información de resumen. Durante la transformación se extrae la información común de los puntos GPS y se obtiene un resumen de estos datos, que formarán la información contenida en cada una de las celdas.

A diferencia del proceso de creación de celdas mediante el método por lotes cuya representación se puede ver en la Figura 3.11 (A), y que realiza un recorrido de celda a celda, el cual no es muy eficiente porque evalúa todas las celdas del área o rejilla, incluso evaluando celdas donde no se registran ubicaciones GPS en esa zona; utilizando el método de microlotes cuya representación se puede ver en la



Figura 3.11 (B), se evita la evaluación de celdas que no contienen datos, obteniendo un mejor rendimiento para este proceso.

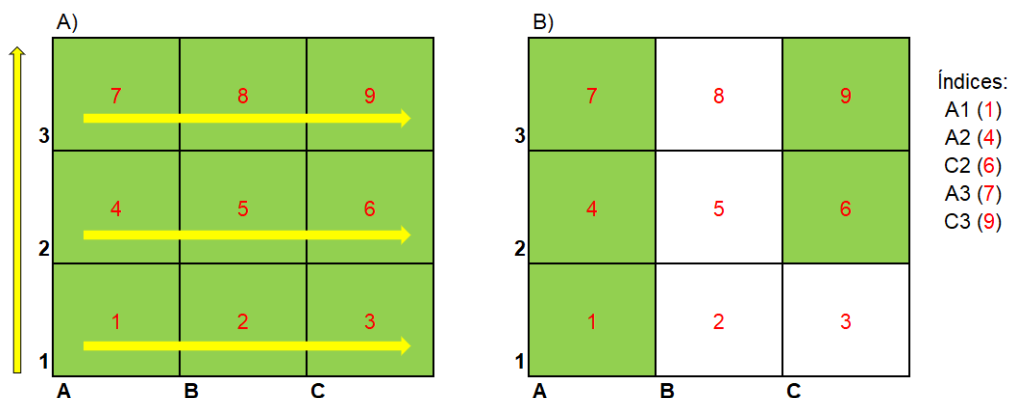


FIGURA 3.11: Diferencias entre los métodos usados en la conformación de celdas.

### 3.5.3. Paso 3: Uso de una técnica de agrupamiento

A continuación, las celdas se agrupan según su característica de velocidad.

Primero se aplica un agrupamiento por distancia que consiste en identificar agrupaciones cercanas en función de la velocidad de las celdas, la cual se obtiene calculando el promedio de velocidades de los puntos vehiculares existentes en cada celda; la asignación de una celda a algún grupo está determinada por la diferencia más pequeña de velocidades entre la celda y los grupos cercanos.

Después se realiza un agrupamiento por densidad en la cual se realiza una clasificación de los grupos existentes para categorizarlos en grupos densos o dispersos según la cantidad de puntos GPS contenidos en las celdas dentro de los grupos.

### 3.5.4. Paso 4: Visualización de resultados

Luego de obtener los agrupamientos de las celdas, se genera un mapa interactivo en el que se puede visualizar cada agrupamiento resultante para cada ciclo procesado y representado por una escala de colores de acuerdo a la velocidad de cada agrupamiento. La Figura 3.12 muestra la representación de las agrupaciones en un mapa interactivo.

### 3.5.5. Consideraciones relevantes

En este proyecto, se presentó un método para la agrupación dinámica de trayectorias de origen vehicular. Este enfoque se basó en el procesamiento de flujos de datos mediante una memoria buffer que gestiona varios microlotes de datos y la creación de celdas con índices, seguido de un análisis de agrupamiento dinámico que utiliza la dimensión de velocidad como criterio clave. Los resultados favorables

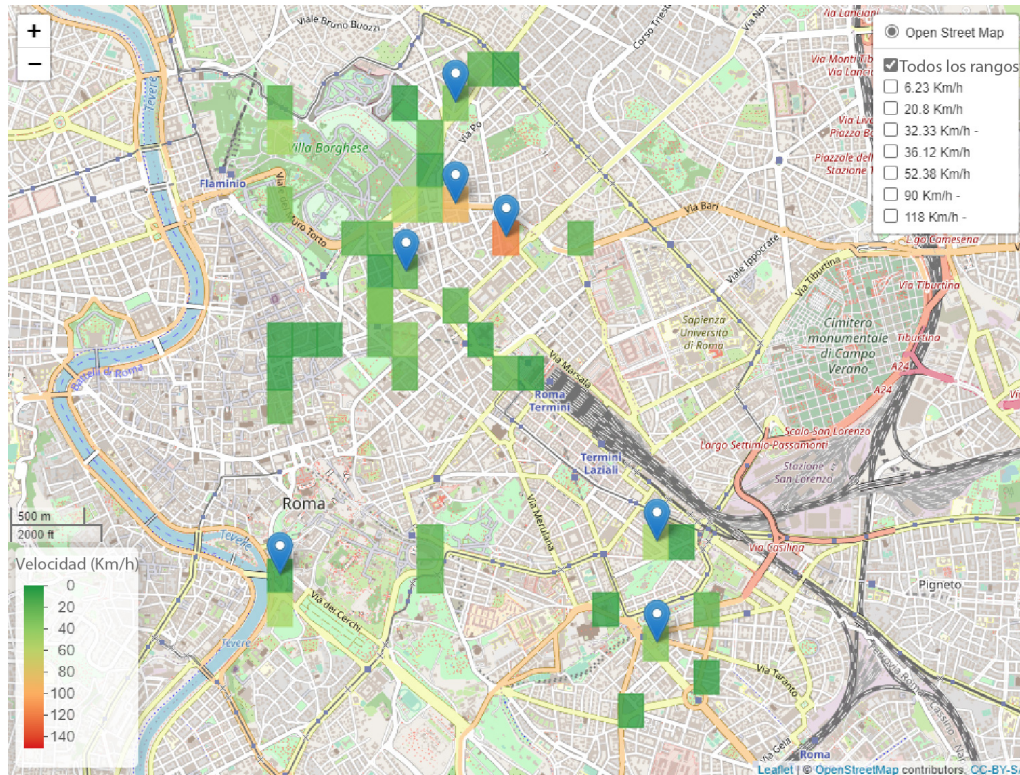


FIGURA 3.12: Visualización de resultados de agrupamientos.

respaldados por mediciones objetivas y visualización efectiva en mapas interactivos realizaron la utilidad potencial del método en la gestión del tráfico urbano. No obstante, se identificaron limitaciones significativas, como la sensibilidad a la cantidad de datos procesados y la necesidad de una calibración precisa de parámetros iniciales, así como la dependencia de información histórica. Estas limitaciones se considerarán para propuestas futuras, que sugieren explorar la implementación del método en plataformas de procesamiento en tiempo real y sistemas distribuidos mediante arquitecturas en paralelo. Estas extensiones buscan abordar las debilidades identificadas, mejorando la robustez y adaptabilidad del método a entornos dinámicos y desafiantes. En conjunto, este análisis proporcionó una base integral para la investigación futura en el desarrollo de enfoques avanzados en la gestión dinámica del tráfico, integrando innovación y consideraciones prácticas.

La propuesta de esta tesis se orienta hacia un enfoque de procesamiento continuo de puntos pertenecientes a trayectorias vehiculares, buscando maximizar la precisión en la identificación de patrones de congestión. Para mejorar la eficiencia entre los actuales procesos secuenciales, se implementa un esquema de procesamiento en paralelo, aprovechando arquitecturas distribuidas. Esta propuesta ha sido capaz de determinar diferentes categorías con patrones similares a partir de los grupos formados, pero la propuesta de esta tesis estará enfocada a analizar los patrones de variabilidad, dada la utilización de información inherente de las trayectorias, se optará por una métrica que permita determinar con precisión el nivel de variabilidad

presente en el flujo vehicular.

### 3.6. Conclusiones

Se han explorado y evaluado diversas metodologías para el análisis y agrupamiento de trayectorias GPS con el objetivo de comprender mejor el flujo vehicular y las dinámicas del tráfico. El primer estudio presentó un método innovador de agrupamiento basado en la segmentación de trayectorias GPS en subtrayectorias más cortas que evitan cambios abruptos en la dirección. Este método empleó un umbral definido y una cuadrícula para distribuir pivotes iniciales de manera uniforme en el área de interés, lo que ha demostrado ser efectivo en el análisis de trayectorias reales.

El segundo estudio definido por Reyes-Zambrano y otros (Reyes-Zambrano et al. 2020) introdujo la técnica de agrupamiento K-pivot, que utiliza pivotes para guiar el recálculo de centroides y se compara con variantes del algoritmo K-means y el método Tra-DBScan. La técnica K-pivot demostró una distribución uniforme de los grupos en el área de estudio y una calidad de agrupamiento superior, como lo indica el índice de silueta. Esta técnica fue capaz de identificar congestiones vehiculares y patrones de tráfico con mayor precisión, lo que facilitó la caracterización de distintos sectores del mapa y la toma de decisiones relacionadas con el tráfico.

El tercer estudio abordó el análisis dinámico del flujo vehicular utilizando una metodología que representa las trayectorias en celdas y aplica un algoritmo de agrupamiento dinámico. Los resultados mostraron que el método es eficaz para identificar velocidades comunes en diferentes momentos y para visualizar el flujo de tráfico mediante mapas interactivos.

El cuarto estudio está basado en la metodología definida en investigaciones previas (Reyes et al. 2021b) este validó las mejoras en la metodología de identificación dinámica del flujo vehicular en ciudades como Guayaquil, Roma y Beijing. Se destacó la utilidad de los mapas interactivos para representar rangos de velocidad y detectar áreas de congestión o situaciones atípicas.

Finalmente, el quinto estudio presentó un método para el procesamiento en tiempo real de flujos de datos vehiculares mediante un algoritmo de agrupamiento dinámico. El método ha mostrado ser eficiente en términos de consumo de memoria y tiempos de procesamiento, con resultados favorables en el análisis de trayectorias. No obstante, se identifican limitaciones relacionadas con la calibración de parámetros y la dependencia de datos históricos.

En conjunto, estos estudios han proporcionado una visión integral de técnicas de agrupamiento diversas en análisis de trayectorias GPS, han mostrado como factibles adaptar los métodos al entorno dinámico del tráfico vehicular.

La relación entre las características identificadas en los diversos patrones de agrupamiento de trayectorias vehiculares y las características variables del comportamiento vehicular es evidente. La convergencia de trayectorias en áreas críticas, resaltada en el primer patrón, refleja la acumulación de vehículos en zonas específicas, una característica clave de la congestión vehicular.

La presencia de cambios abruptos en la dirección o velocidad de los vehículos, señalada en el segundo patrón, se asemeja a comportamientos bruscos y frenados repentinos que son indicativos de condiciones anormales de tráfico.

Además, la identificación de patrones con velocidades similares en diferentes situaciones, como mencionado en el tercer patrón, sugiere áreas con circulación más fluida y otras con menor fluidez, revelando fluctuaciones en la densidad del tráfico, esto es una señal asociada a problemáticas como la congestión vehicular.

La agrupación dinámica de vehículos en entornos cambiantes, como indicado en el cuarto patrón, refleja la adaptación de los vehículos a las fluctuaciones de densidad en el tráfico y la velocidad, características que son fundamentales en escenarios donde suelen presentarse embotellamientos.

Finalmente, la perspectiva dinámica y en tiempo real proporcionada por el quinto patrón revela cambios temporales en la densidad del tráfico, lo cual es esencial para comprender la evolución del flujo vehicular.

En conjunto, estos patrones presentan características asociadas a las constantes fluctuaciones de densidad en el tráfico, algunas de estas están asociadas a situaciones de congestión vehicular, lo cual sugiere que las técnicas de agrupamiento son factibles para identificar patrones cuyas características indican que existe la presencia de problemas en el flujo vehicular. El autor lo demuestra con las diversas perspectivas que ha investigado que reflejan la eficacia de las técnicas de agrupamiento en identificar patrones representativos asociados a la congestión.

Para concluir este capítulo, es importante destacar que uno de los aspectos de mayor relevancia en esta investigación son las fluctuaciones de densidad en el flujo vehicular. Estas fluctuaciones se manifiestan de manera específica en contextos particulares abordados en el ámbito de investigación de esta tesis. La precisión lograda mediante los algoritmos de agrupamiento no solo destaca la presencia de dichas fluctuaciones, sino que también resalta su relevancia en escenarios concretos analizados dentro de los límites de este estudio. La capacidad de identificar y analizar estas fluctuaciones mediante técnicas de minería de datos proporciona una comprensión más profunda de las circunstancias específicas que generan variaciones en la densidad del flujo vehicular. Además, algunas características de estas fluctuaciones podrían estar asociadas a fenómenos como la congestión, constituyendo un pilar fundamental para el análisis detallado de las trayectorias examinadas. Con este análisis realizado, se destaca la utilidad de los algoritmos de agrupamiento en la identificación de fenómenos complejos, además de orientar la investigación hacia las dinámicas y factores involucrados en las fluctuaciones de densidad en el flujo vehicular.

## Capítulo 4

# Variabilidad del Flujo Vehicular

La identificación y comprensión de patrones vehiculares, revelados a través de técnicas de minería de datos y algoritmos de agrupamiento, no solo enriquecen el entendimiento de situaciones específicas que se generan en la infraestructura vial, sino que también ofrecen un fundamento para el análisis detallado de las trayectorias examinadas. Este análisis no solo destaca la utilidad de los algoritmos de agrupamiento en la identificación de patrones complejos, sino que también orienta la investigación hacia las dinámicas y factores involucrados en situaciones específicas de altas densidades vehiculares.

En este sentido, este capítulo presenta una propuesta de método innovador diseñado para abordar los desafíos inherentes a la detección de áreas afectadas por altas densidades vehiculares. Los fundamentos que respaldan la creación de este método se describen, destacando su relevancia en el análisis de la movilidad urbana al lograr identificar de manera efectiva el fenómeno de posible congestión causado por una alta densidad del flujo vehicular. Para ello el método permite reconocer los diferentes niveles de variabilidad en la densidad vehicular de los grupos en diferentes zonas urbanas.

De esta manera, el método propuesto permite en base a las fluctuaciones de la densidad vehicular, obtener los niveles de variabilidad proporcionando una herramienta robusta y precisa en el análisis del flujo vehicular. Esta herramienta facilita una comprensión de los patrones de tráfico y su comportamiento bajo diversas situaciones y escenarios que se presentan en las ciudades.

Por ello, es importante describir que la densidad es una medida que cuantifica la concentración de una determinada magnitud dentro de una unidad específica de espacio o volumen. La variabilidad, por su parte, describe la medida en que los valores de un conjunto de datos se dispersan alrededor de su media, reflejando la heterogeneidad del sistema en estudio.

El análisis de las fluctuaciones de la densidad y la obtención de los diferentes niveles de variabilidad en el flujo vehicular es fundamental para comprender y gestionar la dinámica del tráfico en las carreteras.

En el ámbito del flujo vehicular, la densidad se expresa como el número de vehículos por metro cuadrado ( $\text{vehículos}/m^2$ ), cuantificando la concentración de

tráfico o saturación en una carretera específica en un momento dado. Una alta densidad puede indicar congestión, mientras que una baja densidad puede sugerir fluidez en el tráfico.

Por otra parte, la variabilidad del flujo vehicular se manifiesta a través de las oscilaciones en la densidad y/o en la velocidad de los vehículos a lo largo del tiempo y del espacio. Estas fluctuaciones pueden ser causadas por diversos factores, como la variación en la demanda de tráfico en diferentes momentos del día, las condiciones meteorológicas, y eventos imprevistos como accidentes o trabajos en la carretera, los cuales afectan la fluidez y pueden originar una congestión de tráfico.

Analizar la variabilidad es crucial para identificar patrones y prever posibles problemas en el flujo vehicular, permitiendo a los tomadores de decisiones la implementación de medidas adecuadas para mejorar la eficiencia del tráfico y reducir la congestión.

El análisis de las fluctuaciones en la densidad vehicular permite identificar y abordar diversas problemáticas en la gestión del flujo vehicular. Una de las líneas de análisis más importantes es la congestión vial, donde se estudia la relación entre la densidad y el flujo vehicular para identificar zonas y períodos de alta congestión. Este análisis se enfoca en cómo los diferentes niveles de variabilidad pueden afectar la severidad y frecuencia de la congestión, proporcionando información crítica para el diseño de estrategias de mitigación.

Otra posible línea de análisis esencial es la capacidad y demanda de la vía, que determina la capacidad máxima de las carreteras y la compara con la demanda actual. Este enfoque permite identificar las necesidades de infraestructura y planificar posibles mejoras para evitar saturaciones futuras. Esto es posible mediante el análisis de patrones temporales de tráfico permitiendo identificar fluctuaciones en la densidad y el flujo a lo largo del día, la semana o el año, destacando horas pico y temporadas de alta demanda, lo cual es vital para una planificación eficiente.

El agrupamiento en el análisis del flujo vehicular es una técnica que se destaca principalmente en dos áreas clave: la identificación de patrones de congestión y la detección de anomalías en el comportamiento del tráfico, tal como se indicó en el capítulo anterior. En primer lugar, al aplicar técnicas de agrupamiento a datos de densidad y flujo vehicular, es posible identificar patrones recurrentes de congestión en diferentes momentos del día o en áreas específicas de una ciudad. Estos patrones ayudan a comprender las causas subyacentes de la congestión y facilitan el diseño de estrategias efectivas para su mitigación.

En segundo lugar, el agrupamiento permite detectar anomalías en el flujo vehicular, como picos inesperados de densidad o fluctuaciones bruscas en el tráfico. Estas anomalías pueden indicar problemas en la infraestructura vial o eventos inesperados que requieren atención inmediata, permitiendo una respuesta rápida y eficiente para mantener la seguridad y fluidez del tráfico.

## 4.1. Método de identificación de variabilidad en el flujo del tráfico

Se plantea un método para abordar la problemática de encontrar fluctuaciones de la densidad en el flujo vehicular en entornos urbanos. El método propuesto está basado en un enfoque combinado de dos niveles de agrupamiento, se plantean como una estrategia prometedora para abordar el análisis de la densidad vehicular. La complejidad relacionada a la dinámica del tráfico exige enfoques avanzados que no solo identifiquen patrones en las trayectorias vehiculares, sino que también permita el análisis efectivo de las fluctuaciones de densidad en el tránsito en las zonas correspondientes. Con este propósito, en esta tesis se propone la implementación de un algoritmo de agrupamiento basado en las características espaciales de áreas con mayor densidad vehicular, diseñado para adaptarse a las cambiantes condiciones del tráfico, con el fin de descubrir patrones subyacentes en las trayectorias de los vehículos a lo largo del tiempo. Este enfoque permitirá identificar agrupaciones significativas de comportamientos vehiculares, ofreciendo una comprensión más profunda de la dinámica del tráfico en diferentes momentos del día. El método propuesto integra dos niveles de agrupamiento que analiza las características del desplazamiento, densidad y variabilidad de los vehículos que circulan por las vías para determinar de manera precisa.

Como se observa en la Figura 4.1, el método aquí propuesto tiene dos módulos principales que operan de manera independiente: uno dedicado a identificar zonas en las que el tránsito vehicular presenta características similares y otro enfocada en analizar las características de densidad y variabilidad de las distintas zonas identificadas. Adicional a los módulos anteriores, se dispone de dos sub-módulos complementarios para el funcionamiento del método, el primero se encarga de preprocesar la información referencial de las áreas para la evaluación de la densidad y el segundo se encarga de la visualización de los resultados.

En la configuración del método propuesto, se destaca que los dos módulos principales, operan de manera simultánea, es decir se realizan en paralelo para maximizar la eficiencia en el análisis de trayectorias de origen vehicular. Mientras uno se dedica a la identificación de zonas con características de tránsito similares, el otro se enfoca en evaluar la densidad y variabilidad vehicular en dichas áreas. Además, es crucial destacar que los sub-módulos complementarios cumplen roles específicos para el funcionamiento integral del método. En primer lugar, el sub-módulo encargado de generar áreas preprocesadas debe ejecutarse de manera previa, proporcionando así la base esencial para la evaluación subsiguiente de la densidad vehicular. Por otro lado, el módulo principal de agrupamiento por proximidad depende de la información generada por el sub-módulo complementario de preprocesamiento de áreas de tráfico históricas y de la existencia de instantáneas generadas por el módulo principal de agrupamiento dinámico de ubicaciones, estableciendo una interdependencia necesaria para el flujo eficiente del proceso analítico.

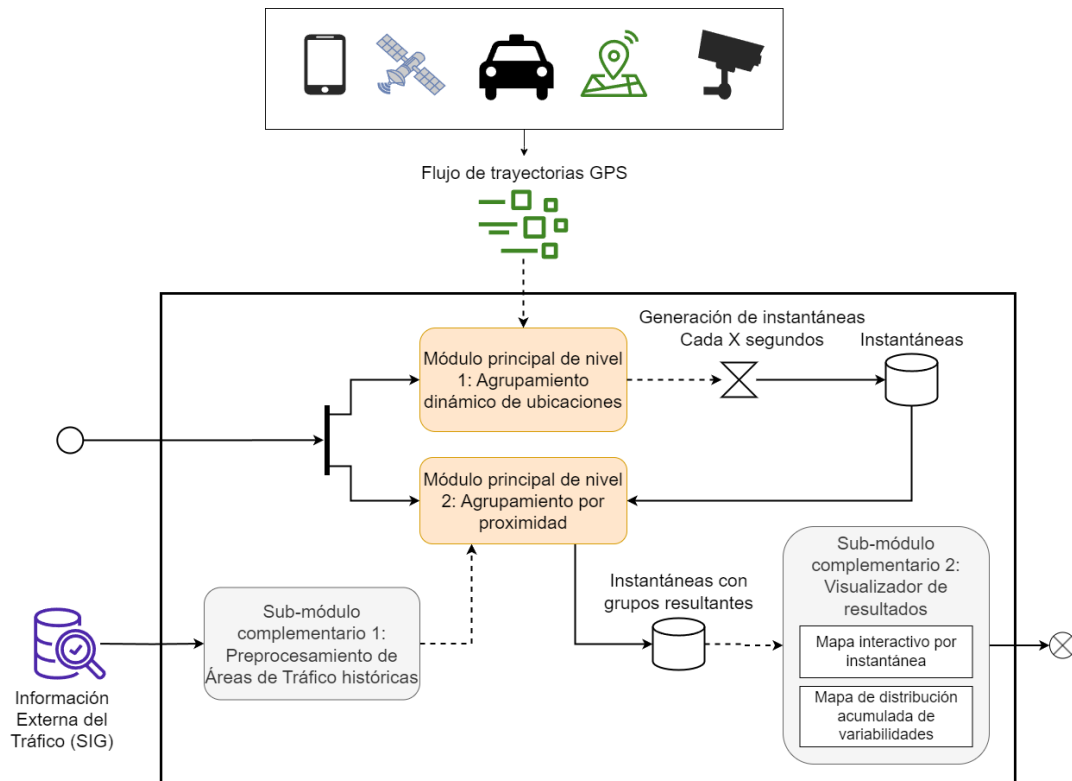


FIGURA 4.1: Diagrama correspondiente al método propuesto

Una descripción generalizada de los elementos que componen cada módulo puede encontrarse en el Anexo A.

#### 4.1.1. Módulo principal de nivel 1: Agrupamiento dinámico de ubicaciones

Para procesar en tiempo real un flujo constante de trayectorias en este estudio, se implementó un proceso de procesamiento de datos ágil y eficiente, que se deriva del método de agrupamiento expuesto por el autor en (Reyes et al. 2023b).

Para este trabajo se simuló un flujo constante que es procesado por un método de agrupamiento espacial para agrupar las ubicaciones GPS con características de movimiento similares. Este método permite procesar eficientemente el flujo constante de trayectorias y utilizar la información resultante para el análisis.

Las partes principales de este método que son la formación de microgrupos y el olvido se relacionan con la manera en que se van a constituir los microgrupos y cómo controlan que las ubicaciones GPS utilizadas sean actuales; especialmente pensando en la temporalidad de los datos.



### Formación de microgrupos

El microgrupo se compone de un conjunto de ubicaciones GPS, las cuales están representadas por un centroide o ubicación promedio. Este centroide, definido como la ubicación geográfica que representa el centro del conjunto de ubicaciones del microgrupo, sirve como una referencia representativa del microgrupo. Esto facilita la comprensión y el análisis de la distribución espacial del mismo.

Para delimitar el área de cobertura del microgrupo, se utiliza una representación rectangular. Esta representación proporciona una manera clara de establecer un área alrededor del centroide, lo que facilita la visualización espacial del microgrupo en el análisis y determina el alcance del mismo. Esta área delimitada se caracteriza como una estructura geométrica rectangular en un espacio bidimensional, y su posición en el espacio está determinada por el centroide del microgrupo al que está asociada. Es importante destacar que el centro de esta área delimitada coincide con el centroide del microgrupo.

Una representación visual de un microgrupo se muestra en la Figura 4.2.

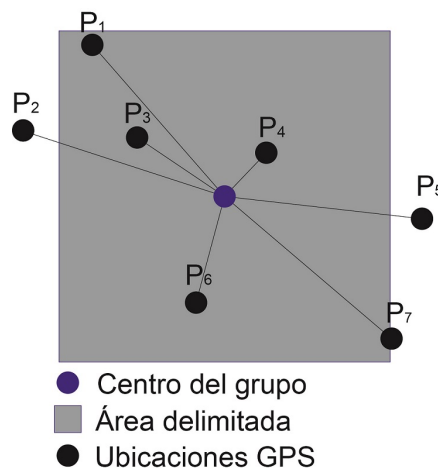


FIGURA 4.2: Elementos que componen un microgrupo.

Cada ubicación GPS extraída del flujo se procesa de manera individual y contienen información sobre las coordenadas geográficas (latitud y longitud), el identificador del vehículo y el tiempo de ingreso. Para realizar el agrupamiento, se utiliza un criterio de similitud basado en la distancia euclidiana y se utilizan los atributos latitud y longitud de las ubicaciones GPS.

Cada ubicación GPS es analizada y se calcula la distancia Euclidiana entre la ubicación y los centroides de todos los microgrupos existentes. Se determina el microgrupo que tenga la menor distancia espacial entre la ubicación analizada y el centroide del microgrupo. Si la coordenada de la ubicación GPS analizada se localiza dentro del área delimitada del microgrupo seleccionado, la ubicación GPS se asigna al microgrupo y se considerará parte integral del microgrupo. El centroide se actualiza cada vez que se integran nuevas ubicaciones GPS a un microgrupo en

particular. Durante este proceso, se irán actualizando los microgrupos que reciban nuevas ubicaciones.

En el caso de que la ubicación GPS analizada no se encuentre ubicada dentro del área delimitada del microgrupo seleccionado mas cercano, se procede a generar un nuevo microgrupo. Una vez asociado una ubicación a un microgrupo, las ubicaciones no puede ser reasignadas a ningún otro microgrupo.

La principal información que contiene un microgrupo se puede observar en la Figura 4.3.

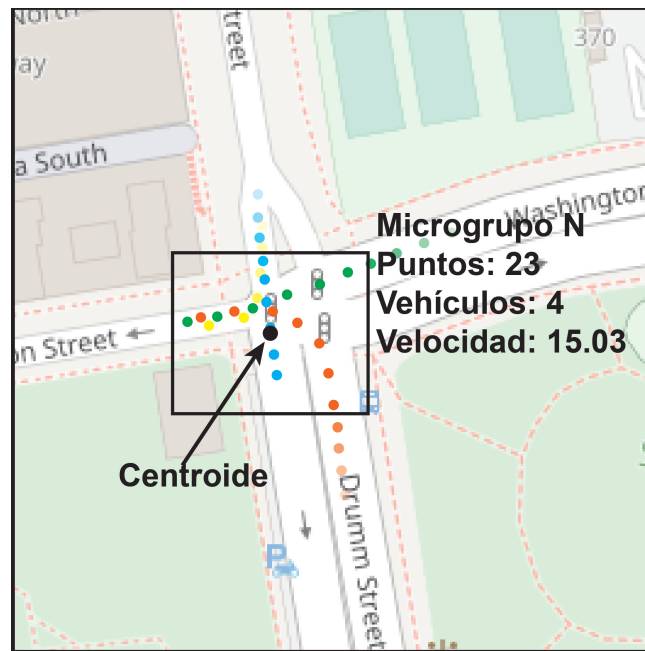


FIGURA 4.3: Información asociada a un microgrupo.

Es importante señalar que, en situaciones donde el centroide del microgrupo se desplaza significativamente, las ubicaciones más antiguas que se alejan del centroide serán eventualmente eliminadas mediante un mecanismo de olvido. Este mecanismo permite descartar las ubicaciones que, debido al desplazamiento del centroide, terminan estando fuera del área del microgrupo. Por lo tanto, se evita que coordenadas que originalmente pertenecían a un microgrupo se mantengan asignadas a este cuando, por proximidad, podrían estar más cerca de un nuevo centroide que se haya desplazado en una dirección similar.

#### **Olvido porcentual de datos basado en el tiempo de ingreso**

Para asegurarse que el sistema esté actualizado y no retenga datos demasiado antiguos, se implementa un mecanismo de olvido porcentual basado en el tiempo de ingreso de la última ubicación GPS.

Este mecanismo de olvido indica el porcentaje de relevancia de los datos por cada unidad de tiempo en segundos transcurrida, se determina mediante la Ecuación 4.1.

$$F = e^{-1*\lambda*\delta t} \quad (4.1)$$

donde,  $e$  corresponde a la función exponencial,  $\lambda$  (lambda) es el parámetro que controla la rapidez de la disminución. Cuanto mayor sea el valor de  $\lambda$ , más rápida será la disminución. Finalmente,  $\delta t$  indica la diferencia, en segundos, entre los tiempos de la ubicación analizado y el tiempo de la última ubicación integrada al microgrupo.

Cada microgrupo tiene un indicador numérico que refleja la cantidad de ubicaciones que debe recordar, este indicador primero disminuye un porcentaje calculado por el mecanismo de olvido a partir de cada segundo de tiempo que ha transcurrido desde la última ubicación que ha ingresado al microgrupo y posteriormente aumenta en 1 cada vez que se añade una nueva ubicación GPS al microgrupo.

Se establece un umbral de tolerancia del 5 %, el cual está relacionado con la mejor configuración para el parámetro de lambda y facilita la elección de este parámetro en términos de tiempo transcurrido. Este umbral se ha definido a partir de simulaciones que determinan aproximadamente qué porcentaje de datos se debe considerar según el tiempo que se desea conservar los datos. Es importante tener en cuenta que este porcentaje puede variar según el conjunto de datos específico y las necesidades del análisis. Por lo tanto, se recomienda una evaluación cuidadosa de los requisitos de tiempo y datos para determinar el umbral de tolerancia adecuado en cada caso.

Este mecanismo de olvido se utiliza para filtrar datos o resultados, concentrando la atención en aquellos que superan el umbral de tolerancia y descartando aquellos que no lo hacen. Al aplicar un tope del 5 % de relevancia, se busca asegurar que solo los valores que tienen un grado significativo de importancia sean recordados en el análisis de los microgrupos.

Este mecanismo determina la cantidad de ubicaciones GPS que permanecerán en el microgrupo durante el agrupamiento. Esto asegurará que el sistema se adapte a cambios en el tráfico y evite acumulación de datos obsoletos. Aquellos microgrupos cuyas ubicaciones GPS han perdido importancia por no recibir una cantidad suficiente de nuevas ubicaciones GPS, se eliminan para mantener un análisis actualizado de la situación del tráfico. Al mismo tiempo, aquellos microgrupos activos que sigan recibiendo nuevas ubicaciones GPS se mantendrán actualizados para el correcto análisis del tráfico.

Una representación visual del porcentaje de olvido y relevancia que se considera durante el mecanismo de olvido se muestra en la Figura 4.4, se puede observar que con una diferencia de tiempo de cero segundos la relevancia es del 100 %, y a medida que aumenta la diferencia de tiempo transcurrido en segundos, el porcentaje de relevancia va disminuyendo.

La zona con el porcentaje relevante se marca con fondo de color verde, y cuando el porcentaje cae por debajo del umbral de tolerancia establecido en 5 % se considera no relevante esta zona se encuentra marcada con fondo color rojo.

El mecanismo de olvido porcentual se aplica cada vez que una ubicación se añade a un microgrupo. Adicionalmente el mecanismo de olvido se aplica de manera periódica a los microgrupos que no han recibido ubicaciones recientes, esto con el fin de asegurarse que estos microgrupos particulares se mantengan actualizados.

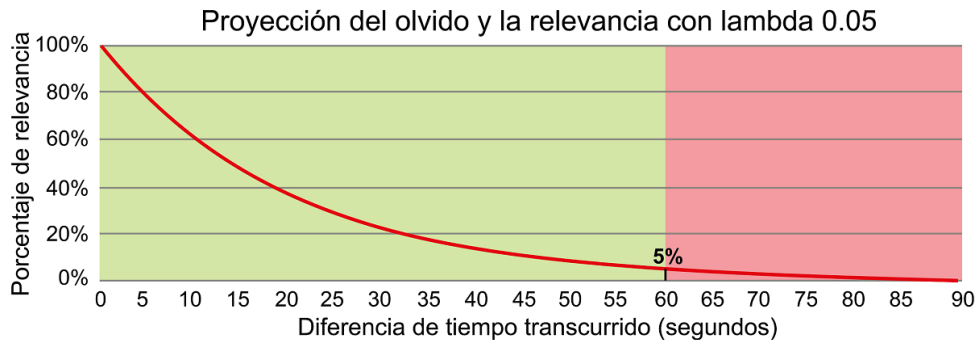


FIGURA 4.4: Porcentaje de olvido y relevancia de los datos por cada unidad de tiempo transcurrido para un valor de lambda establecido en 5%.

Este módulo periódicamente se encarga de realizar capturas denominadas instantáneas del estado de los microgrupos formados a partir de los resultados de agrupamiento hasta el momento que se realiza la instantánea en intervalos regulares.

En cada captura de instantánea, se registran variables relevantes que describen el estado de los microgrupos. Esto incluye información como la ubicación geográfica, la cantidad de vehículos presentes en el área, la velocidad promedio, y otros datos relacionados con las condiciones del tráfico vehicular.

Cada captura de instantáneas se almacena en memoria, con esto se consigue que los tiempos de espera necesarios para almacenar cada instantánea sean los mínimos posibles sin que afecte al proceso de agrupamiento, evitando en gran medida la pérdida de ubicaciones GPS del flujo. Este almacenamiento resulta fundamental para analizar tendencias a lo largo del tiempo y detectar patrones en el comportamiento del tráfico.

Con cada captura que se almacena, el módulo encargado del agrupamiento por proximidad puede iniciar su operación.

El Pseudocódigo 2 muestra el algoritmo que resume el proceso de este módulo principal de agrupamiento dinámico de ubicaciones.

#### 4.1.2. Módulo principal de nivel 2: Agrupamiento por proximidad.

La ejecución de este módulo permite analizar y agrupar por proximidad los microgrupos resultantes del módulo principal 1, lo que ayudará a identificar áreas problemáticas y áreas con alta densidad vehicular. Mientras se realiza el agrupamiento de los datos en tiempo real, se opta por un proceso que se ejecuta en paralelo que realiza la captura de resultados de los agrupamientos de manera periódica y realiza un agrupamiento sobre los microgrupos formados.

**Pseudocódigo 2** Módulo de agrupamiento dinámico de ubicaciones.

**Input:**  $S$ : flujo de datos;  $t_i$ : Tiempo inicial de captura de datos;  $t_e$ : tiempo final de captura de datos;  $lambda$ : valor de lambda

**Output:**  $I_{List}$ : lista de instantáneas,  $CLUSTERING$ : Estado del agrupamiento

```

1:  $MG_{List} \leftarrow \emptyset$  {lista de microgrupos}
2:  $T \leftarrow \emptyset$  {tiempo general del entorno}
3:  $LS_{(x,y)} \leftarrow \emptyset$  {suma lineal de las coordenadas}
4:  $MG = (time, n, LS_{(x,y)}, points_{List}) \leftarrow \emptyset$ 
5: while  $S \neq \emptyset$  do
6:    $CLUSTERING \leftarrow TRUE$ 
7:    $P = (x, y, time, speed, id) \leftarrow$  extraer ( $S$ )
8:    $T \leftarrow P.time$ 
9:    $MG \leftarrow$  identificar microgrupo mas cercano ( $MG_{List}, P_{x,y}$ ) usando distancia Euclidiana
10:  if  $MG_{List} = \emptyset$  or  $MG = \emptyset$  then
11:     $MG \leftarrow$  crear nuevo microgrupo ( $P$ )
12:     $MG.n \leftarrow 1$ 
13:     $MG.LS_{(x,y)} \leftarrow P_{x,y}$ 
14:     $MG.time \leftarrow P.time$ 
15:     $MG_{List} \leftarrow$  añadir ( $MG$ )
16:  else
17:     $O \leftarrow$  calcular olvido ( $lambda, MG.time, P.time$ )
18:     $MG.n \leftarrow MG.n * O$ 
19:     $MG.LS_{(x,y)} \leftarrow MG.LS_{(x,y)} * O$ 
20:     $MG.n \leftarrow MG.n + 1$  {Aumenta el contador de ubicaciones en el microgrupo}

21:     $MG.LS_{(x,y)} \leftarrow MG.LS_{(x,y)} + P_{x,y}$  {Acumula la suma lineal de longitud(x) y latitud(y) de la nueva ubicación GPS}
22:     $MG.points \leftarrow$  filtrar ( $MG.points, MG.n$ ) {Conserva la cantidad de ubicaciones definidas por el contador de ubicaciones}
23:     $MG.time \leftarrow P.time$ 
24:  end if
25:  Determinar actualización periódica de microgrupos ( $MG_{List}, T$ )
26:  Determinar eliminación periódica de microgrupos ( $MG_{List}, T$ )
27:  Determinar captura periódica de instantáneas ( $I_{List}, MG_{List}, T$ ) {Lo realiza el Módulo principal de nivel 2 de acuerdo a las parametrizaciones del usuario}
28: end while
29:  $CLUSTERING \leftarrow FALSE$ 

```

**Sub-módulo complementario 1: Preprocesamiento de áreas de tráfico históricas**

Este sub-módulo se encarga de preprocesar áreas de tráfico históricas con dimensiones similares al tamaño de los microgrupos, con el objetivo de reducir el costo de procesamiento necesario para analizar la información referencial de las densidades habituales en las diversas áreas cubiertas por los futuros microgrupos. Además, se encarga de extraer e integrar la información referencial a partir de datos históricos de tráfico, los cuales son proporcionados por un ente externo afín a la gestión o monitorización del tráfico.

La ejecución de este sub-módulo se realiza antes de los módulos de agrupamiento dinámico de ubicaciones y módulo de agrupamiento por proximidad. La información que genera este sub-módulo puede aplicarse a diferentes situaciones para análisis en la misma ciudad, para ello guarda en un almacenamiento recurrente la información preprocesada de las densidades habituales en las distintas áreas. El diseño de este sub-módulo está basado en el uso e importancia de información histórica de tráfico, similar al procesamiento de datos históricos de la propuesta del autor presentada en (Reyes et al. 2023a).

Primero se realiza la definición de ubicaciones uniformemente distribuidos que sirven de referencia para ubicar cada área, luego se realiza una proyección alrededor de la ubicación, esta será la extensión del área para el análisis de densidades que proporcionará el ente externo.

El proceso implica la integración de datos suministrados por una entidad externa, los cuales contienen el conteo de vehículos y las densidades registradas de las carreteras. Estos datos son analizados para comprender el comportamiento habitual de los vehículos en cada área y determinar las densidades medias históricas correspondientes.

La entidad externa que proporciona esta información a menudo suelen ser Sistemas de Información Geográfica (SIG) de alta precisión, en el caso de no tener disponibilidad de estos sistemas, el autor de este estudio ha planteado un procedimiento externo alternativo de carácter accesible y de ejecución simplificada con el fin de adquirir dicha información, como se detalla en (Reyes et al. 2022a).

Los resultados obtenidos son áreas históricas preprocesadas las cuales se almacenan eficientemente en sistemas especializados de gestión de bases de datos para asegurar su integridad y disponibilidad en análisis futuros. Esta metodología contribuye significativamente a optimizar el rendimiento del método, al generar un recurso inicial que no requiere actualizaciones frecuentes durante el procesamiento de nuevos flujos de datos.

Sin embargo, se requiere información recopilada en las diferentes temporadas del año para evitar que eventos particulares afecten a las mediciones habituales, además, para mantener la información actualizada respecto al tráfico reciente, se recomienda una frecuencia de actualización que puede ser semanal o mensual, ajustada según la disponibilidad y la variabilidad de los datos de tráfico. Esta estrategia asegura que el sistema cuente siempre con datos precisos y relevantes, mejorando la efectividad de los nuevos análisis.

La principal ventaja de este sub-módulo es que permite preprocesar la información de tráfico y mantenerla almacenada en un repositorio accesible para el método. Esto facilita el acceso rápido a la información sin necesidad de consultar constantemente a la entidad externa, aunque siempre se tiene la opción de realizar consultas puntuales en tiempo real para obtener datos actualizados si fuera necesario.

Finalmente, es importante destacar que este sub-módulo puede requerir el uso de software especializado de SIG, indispensable para manejar grandes volúmenes de

datos y realizar análisis complejos de manera eficiente. Esta dependencia de proveedores externos de datos garantiza el suministro continuo y preciso de la información necesaria para el funcionamiento del sub-módulo.

El Pseudocódigo 3 resume el proceso del sub-módulo de preprocesamiento de áreas de tráfico históricas que se ha descrito anteriormente.

---

**Pseudocódigo 3** Sub-módulo de preprocesamiento de áreas de tráfico históricas.

---

**Input:**  $R_{List}$ : lista de redes de carreteras;  $H_{List}$ : conjunto de datos vehiculares externos;  $max_{inst}$ : cantidad de períodos históricos a analizar;  $ADim_{(x,y)}$ : Dimensiones del área de procesamiento;  $Dis$ : Distancia de separación entre centros para áreas preprocesadas

**Output:**  $AREAS_{List}$ : lista de áreas preprocesadas

- 1:  $AREAS_{List} \leftarrow \emptyset$  {lista de áreas preprocesadas}
  - 2:  $maxX \leftarrow$  estimar cantidad de centros en eje Longitud ( $ADim_{(x)}, Dis$ )
  - 3:  $maxY \leftarrow$  estimar cantidad de centros en eje Latitud ( $ADim_{(y)}, Dis$ )
  - 4: **for**  $CELDA_{(x,y)}$  **to**  $x := maxX, y := maxY$  **do**
  - 5:    $CENTRO = (longitude, latitude) \leftarrow$  establecer el centro del área
  - 6:    $L_{List} \leftarrow$  calcular límites mínimos y máximos en base a  $CENTRO$
  - 7:    $P_{List} \leftarrow$  filtrar  $H_{List}$  dentro de  $L_{List}$
  - 8:    $Vehiculos \leftarrow$  determinar cantidad de vehículos ( $P_{List}$ )
  - 9:    $Densidad \leftarrow$  calcular la densidad del área ( $P_{List}, ADim_{(x,y)}$ )
  - 10:    $CELDA_{(x,y)} = (id(x,y), CENTRO, Vehiculos, Densidad)$
  - 11:    $AREAS_{List} \leftarrow$  añadir  $CELDA_{(x,y)}$
  - 12: **end for**
- 

### Agrupamiento por proximidad

Entre las principales funciones que desempeña el módulo principal de nivel 2 se encuentran el monitoreo periódico de instantáneas y el agrupamiento de microgrupos en función de su proximidad.

Este módulo se asegura que la información sobre el estado de los microgrupos que fue capturada por las instantáneas sea la más reciente y que el agrupamiento de nivel 2 se base en datos recientes. Esto resulta crucial para la precisión de los resultados y la comprensión efectiva de la evolución de la densidad del tráfico en las áreas de estudio a lo largo del tiempo.

El agrupamiento de microgrupos se realiza con cada nueva captura, proporcionando así una visión dinámica de la distribución del tráfico. Esta actualización periódica permite detectar cambios en el tiempo y analizar cómo varía el flujo del tráfico en función de diferentes condiciones y eventos.

A diferencia del módulo principal de nivel 1, este módulo no se realiza en tiempo real, esto guarda relación con la frecuencia en que se generan las instantáneas, por lo que configurar períodos de tiempos cortos entre instantáneas puede causar tiempos de espera adicionales debido a una alta recurrencia del proceso de agrupamiento. Este factor debe ser tomado en cuenta si se quiere mostrar resultados muy cercanos al tiempo real.

El proceso de analizar cada instantánea involucra la ejecución de manera iterativa sobre cada microgrupo de los procesos que se describen a continuación:

Primero, se procede a la identificación del área preprocesada más cercana, con un enfoque preciso en determinar cuál área preprocesada tiene la mayor similitud con respecto al área que es cubierta por el microgrupo analizado. Esto se logra mediante el cálculo de la distancia euclidiana entre los centros de las áreas preprocesadas y los centroides correspondientes a los microgrupos analizados. Esta métrica proporciona una medida objetiva de la proximidad de cada microgrupo hacia las áreas preprocesadas, asegurando que la elección del área preprocesada más cercana determine las condiciones referenciales que se usarán para el análisis de la densidad del microgrupo.

Una vez identificada el área preprocesada más cercana, se procede a recuperar la información detallada sobre la información del tráfico histórico en dicha área. Esto incluye datos sobre la capacidad máxima de las carreteras, densidades vehiculares históricas, y otros atributos relevantes que faciliten un análisis de la información histórica de las áreas.

Con el reconocimiento de las áreas con mayor similitud se procede al cálculo de densidad actual de cada microgrupo. La densidad actual de cada microgrupo se calcula mediante la división de la cantidad de vehículos observada entre el área que cubre el microgrupo (vehículos/ $m^2$ ). Se calcula la diferencia de densidades con respecto a la densidad histórica y se obtiene un valor de variabilidad respecto a la densidad histórica para cada microgrupo.

Los microgrupos cuya densidad actual supera la densidad media histórica serán categorizados como variabilidad positiva. Esto implica que en estas áreas el flujo del tráfico vehicular es superior al promedio histórico, indicando una mayor actividad y posible congestión. Por otro lado, los microgrupos cuyo valor de densidad actual es menor o igual a la densidad media histórica serán categorizados como de variabilidad negativa, lo cual sugiere que estas áreas experimentan un flujo vehicular menor o igual al promedio, indicando una menor actividad comparada con los registros históricos.

Utilizando la densidad histórica proporcionada por el área preprocesada se proporciona una descripción cuantitativa de la operación de cada zona cubierta y permite identificar cuáles presentan un mayor flujo vehicular con respecto al tráfico habitual y, por lo tanto, una mayor probabilidad de estar congestionado.

El proceso de agrupamiento por proximidad de los microgrupos se lleva a cabo mediante un algoritmo de agrupamiento simplificado que se fundamenta en la proximidad entre los microgrupos, uniéndolos para formar grupos dentro de una distancia establecida. Este proceso se rige por un parámetro único, denotado como epsilon, que define el radio de proximidad alrededor de cada microgrupo.

Para este método se considerará como microgrupos no visitados aquellos que no han sido considerados en ninguna fase del agrupamiento y serán eventualmente seleccionados para iniciar o ser parte de un grupo. Por otro lado, se considerará



como microgrupos visitados a aquellos que han sido procesados en el agrupamiento y han sido integrados a algún grupo.

La ejecución de este segundo agrupamiento comprende varias etapas. En la fase de inicialización, se identifica la existencia de centros de masas previos que serán las semillas para los primeros grupos, estas se utilizan como ubicaciones iniciales. En caso de ausencia de grupos previos, se comienza con un microgrupo no visitado seleccionado arbitrariamente, como se muestra en la Figura 4.5.

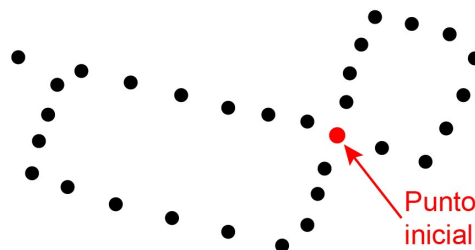


FIGURA 4.5: Selección de la ubicación inicial en el nuevo grupo.

Luego, se procede a la expansión de los grupos, donde se determina el conjunto de microgrupos localizados dentro del radio  $\epsilon$  del centro de masa previo o microgrupo inicial, los microgrupos identificados son considerados como parte del mismo grupo, la Figura 4.6 presenta como se identifican los microgrupos a partir del radio  $\epsilon$ .

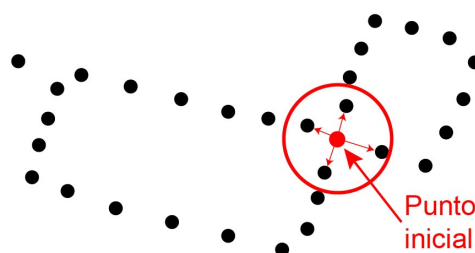


FIGURA 4.6: Área de expansión del grupo.

Los microgrupos que integran un determinado grupo pueden ser considerado microgrupos analizados si se han examinado detalladamente para identificar microgrupos cercanos mediante la expansión, o pueden ser microgrupos no analizados si aún no se ha realizado la expansión.

Este proceso de expansión se repite iterativamente para agregar nuevos microgrupos al grupo, realizando el análisis de cada microgrupo que se ha integrado al grupo para identificar nuevos microgrupos no visitados, un ejemplo de este proceso

iterativo de los integrantes del grupo después de 2 iteraciones se puede observar en la Figura 4.7(a), después de 6 iteraciones en la Figura 4.7(b) y después de 9 iteraciones en la Figura 4.7(c). Ya sea que el proceso de expansión se origine a partir de los centros de masas de las instantáneas previas o de un microgrupo no visitado, este proceso se detiene cuando todos los microgrupos hayan sido visitados y asignados a algún grupo.

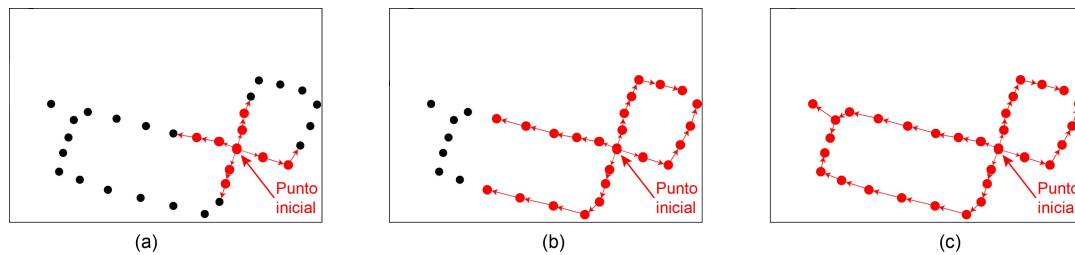


FIGURA 4.7: Expansión iterativa de un grupo.

Los nuevos centros de masa de los grupos formados se basan en atributos específicos, como la cantidad de vehículos en los microgrupos, calculando temporalmente el centro de masa de cada uno proporcionalmente al atributo mencionado. Posteriormente, el centro de masa previamente calculado se reemplaza con la ubicación del microgrupo más cercano, este será el nuevo centro de masa del grupo, con esto se logra una representación adecuada del grupo en términos del atributo seleccionado; para el grupo presentado previamente como ejemplo, el centro de masa representativo del grupo se puede observar en la Figura 4.8.

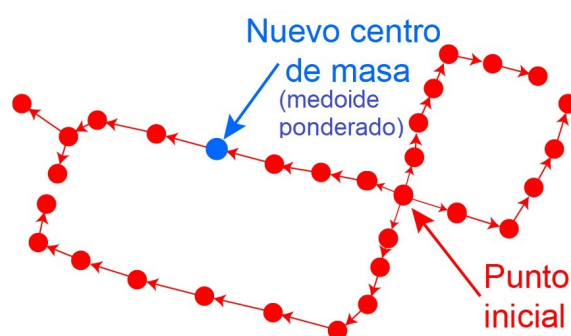


FIGURA 4.8: Ubicación del nuevo centro de masa.

Este enfoque presenta ventajas significativas, incluyendo su simplicidad de implementación al no requerir parámetros adicionales y no gestionar elementos como ruido. Además, puede formar grupos de cualquier forma siempre que los microgrupos estén dentro del radio épsilon unos de otros, y no requiere especificar el número de grupos, ya que estos se forman automáticamente en función de la proximidad.

Una vez formados los grupos y asignados todos los microgrupos, se procede al análisis de los grupos obtenidos y sus características para interpretar los resultados y extraer información relevante sobre los cambios en el flujo de tráfico.

El Pseudocódigo 4 muestra el algoritmo que resume el proceso del módulo principal de nivel 2 encargado del agrupamiento por proximidad. La primera condición de parada o fin del algoritmo viene dada por el estado de ejecución del módulo principal de nivel 1 siempre que siga realizando su función de agrupamiento dinámico de ubicaciones, la segunda condición de parada es la existencia de instantáneas por procesar.

### **Sub-módulo complementario 2: Visualizador de resultados**

En esta investigación, se analiza la información de las trayectorias a intervalos regulares, lo que facilita la detección precisa de las variaciones en el tráfico de vehículos.

Con el objetivo de proporcionar una presentación visual y activa de los resultados de cada proceso de agrupamiento, se utiliza el sub-módulo visual para la representación de los patrones en cada grupo presentado por el autor en (Reyes et al. 2022c), este sub-módulo genera un conjunto de mapas interactivos a partir de las instantáneas procesadas por el método propuesto y además genera un mapa resumido que muestra la distribución acumulada de las variabilidades positivas en la densidad vehicular basada en los resultados del agrupamiento del método propuesto.

#### **Mapa interactivo por instantánea.**

Este primer mapa permite el análisis gráfico y en tiempo real de la información relevante de cada grupo. Cada área con características similares se representa mediante un color de acuerdo al tipo de variabilidad que se ha identificado, esta puede ser de color rojo para los microgrupos con variabilidades positivas y verde para los microgrupos con variabilidades negativas, como se muestra en la ilustración de la Figura 4.9.

Es importante aclarar que el sub-módulo de visualización se despliega con el propósito de presentar los resultados derivados del agrupamiento por proximidad, ya sea de manera inmediata o haciendo uso de datos obtenidos en agrupamientos previos. En esta última situación, el módulo de agrupamiento por proximidad guarda una réplica de las instantáneas con el fin de someterlas a un análisis adicional por parte del usuario en el futuro.

El Pseudocódigo 5 muestra el algoritmo que resume el proceso del sub-módulo complementario.

Cabe destacar que este método está enfocado a la obtención de resultados en tiempo real, por lo que si esta no es la prioridad, el autor ha propuesto un método alternativo en donde el procesamiento se enfoca en flujos de datos periódicos mediante un modelo en microlotes (Reyes et al. 2023b).

#### **Mapa de distribución acumulada de variabilidades**

**Pseudocódigo 4** Módulo de agrupamiento por proximidad.

**Input:**  $I_{List}$ : lista de instantáneas,  $CLUSTERING$ : Estado del agrupamiento;  $AREAS_{List}$ : lista de áreas preprocesadas;  $Epsilon$ : radio de detección de microgrupos

**Output:**  $I_{List}$ : lista de instantáneas

```

1:  $PCM_{List} \leftarrow \emptyset$  {lista de centros de masa previos}
2:  $CM = (x, y)$  {centro de masa}
3:  $MGNA_{List} \leftarrow \emptyset$  {lista de microgrupos no analizados}
4:  $G = (CM_{x,y}, MG_{List}) \leftarrow \emptyset$  {Grupo}
5:  $G_{List} \leftarrow \emptyset$  {lista de grupos}
6: while  $CLUSTERING$  is true or  $\exists I \in I_{List}$  do
7:   if  $\exists I \in I_{List}$  then
8:     for each  $MG = (time, centroid_{x,y}, n, LS_{(x,y)}, points_{List}) \in I = MG_{List}$  do
9:        $A = (centroid_{x,y}, Vehiculos, Densidad) \leftarrow$  buscar área preprocesada mas
       cercana ( $MG.centroid_{x,y}, AREAS_{List}$ )
10:       $MG.Densidad \leftarrow$  calcular densidad de grupo ( $MG.points_{List}$ )
11:       $MG.Variabilidad \leftarrow$  calcular variabilidad de grupo
       ( $MG.Densidad, A.Densidad$ )
12:      if  $MG.Variabilidad > 0$  then
13:         $G.tipo_{variabilidad} \leftarrow$  "Positiva"
14:      else if  $MG.Variabilidad \leq 0$  then
15:         $MG.tipo_{variabilidad} \leftarrow$  "Negativa"
16:      end if
17:    end for
18:    while  $\exists MG \in I = MG_{List}$  do
19:      if  $\exists CM \in PCM_{List}$  then
20:         $CM \leftarrow$  extraer centro de masa ( $PCM_{List}$ )
21:      else
22:         $MG \leftarrow$  extraer microgrupo aleatorio ( $I = MG_{List}$ )
23:         $CM \leftarrow$  asignar microgrupo como centro de masa ( $MG.centroid_{x,y}$ )
24:      end if
25:       $MGNA_{List} \leftarrow$  añadir ( $CM$ )
26:      for each  $MG \in MGNA_{List}$  do
27:         $G.MG_{List} \leftarrow$  añadir ( $MG$ )
28:         $MGNA_{List} \leftarrow$  extraer y añadir microgrupos cercanos
       ( $CM, MG_{List}, Epsilon$ )
29:      end for
30:       $G.CM_{x,y} \leftarrow$  calcular nuevo centro de masa ( $G.MG_{List}$ )
31:       $G_{List} \leftarrow$  añadir ( $G$ )
32:       $PCM_{List} \leftarrow$  añadir ( $G.CM_{x,y}$ )
33:    end while
34:     $I_{List} \leftarrow$  Actualizar ( $I$ )
35:    Ejecutar sub-módulo: Visualizador de resultados ( $I_{List}$ )
36:  end if
37: end while

```

En este segundo mapa, se analizan las distribuciones acumuladas de los grupos con variabilidades positivas en la densidad vehicular en las diferentes ciudades.

Para ello, se ha implementado un sistema de reticulado que divide el área urbana

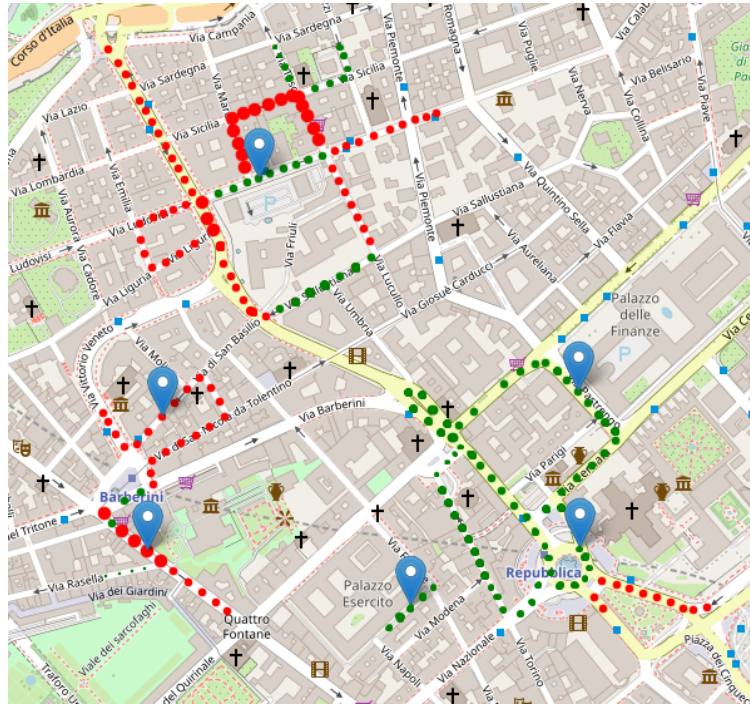


FIGURA 4.9: Grupos proyectados sobre el mapa.

---

**Pseudocódigo 5** Sub-módulo visualizador de resultados.
 

---

**Input:**  $I_{List}$ : lista de instantáneas

**Output:**  $M_i$ : Mapas de las instantáneas

```

1: for each  $I = G_{List} \in I_{List}$  do
2:    $M \leftarrow \emptyset$  {Render de mapa vacío}
3:   for all  $G = (time, centroid_{x,y}, n, LS_{(x,y)}, points_{List}, estado) \in I$  do
4:      $C \leftarrow \emptyset$  {Capa vacía}
5:      $C \leftarrow$  añadir  $G.centroid_{x,y}$ 
6:      $C \leftarrow$  añadir  $G.points_{x,y}$ 
7:      $C \leftarrow$  añadir estadísticos generales
8:      $M \leftarrow$  añadir  $C$ 
9:   end for
10:  Almacenar  $M$ 
11: end for

```

---

en celdas de tamaño uniforme. Este reticulado permite identificar y contabilizar la frecuencia de detección de grupos cuya densidad actual excede la densidad histórica en cada celda, los grupos con una densidad inferior son excluidos en este análisis. El proceso de contabilización se realiza en los intervalos de las instantáneas durante el período total de análisis, donde cada celda puede registrar un máximo de contabilizaciones que es equivalente a la cantidad de instantáneas procesadas por el método, independientemente del número de grupos detectados en cada período.

Para cada celda del reticulado se utiliza un índice de distribución acumulada, que determina la cantidad de períodos expresada de manera porcentual con respecto a la cantidad de instantáneas empleadas en el análisis.

Por cada instantánea se analizan las celdas de manera independiente y se determina si en cada instantánea existe por lo menos un grupo categorizado con variabilidad positiva, en este caso se contabiliza para la celda en la respectiva instantánea con un 1 indicando que se ha encontrado un grupo (o varios) con esta variabilidad en específico, se hace uso de la Ecuación 4.2.

$$n_{celda} = \frac{\sum_{i=1}^m VP_i}{m} \quad (4.2)$$

Donde:

$n$  indica el índice de distribución acumulada de cada celda;  $m$  indica el máximo de instantáneas analizadas;  $i$  representa cada instantánea;  $VP$  indica la cantidad de instantáneas que han presentado variabilidad positiva, donde se contabiliza 1 en el caso de presentar un grupo (o varios) con variabilidad positiva y 0 en caso de no presentar;  $celda$  representa cada una de las celdas del área analizada

Después de haber determinado el índice de distribución acumulada, se puede representar cada celda en el mapa, visualizadas mediante una escala que indica la cantidad (o porcentaje respecto al máximo) de instantáneas en las que se ha determinado que una celda ha presentado grupos con variabilidad positiva.

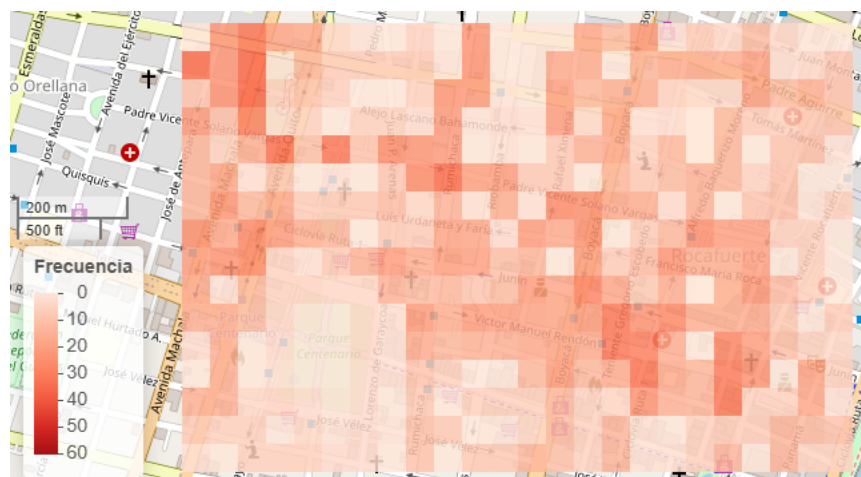


FIGURA 4.10: Distribución acumulada de variabilidades positivas.

Para generar los mapas de distribución, se utiliza una escala de color que varía de tonos claros a oscuros de rojo, indicando desde una baja hasta una alta frecuencia de detección, respectivamente. Este método permite visualizar de manera efectiva las áreas con mayor variabilidad positiva, proporcionando una herramienta crucial para el análisis urbano ya que estas se enfocan en presentar de manera visual las zonas con las fluctuaciones de densidad que estarían ocasionando problemas en el tráfico vehicular. Las condiciones específicas consideradas incluyen la identificación de grupos que presentan una diferencia positiva significativa entre la densidad actual y la densidad histórica, garantizando que cada celda refleje una medición precisa de la actividad dinámica en el área estudiada.

El análisis resultante ofrece una visión detallada de las áreas urbanas con mayor presencia de variabilidades positivas que reflejan una alta frecuencia de densidades superiores a las habituales, facilitando la identificación de zonas estratégicas que pueden requerir intervenciones específicas en términos de planificación urbana, gestión del flujo de personas y seguridad, así como potenciales oportunidades comerciales.

## 4.2. Resultados obtenidos

### 4.2.1. Datos utilizados

En este estudio se emplearon los conjuntos de datos correspondientes a las ciudades de San Francisco, Roma y Guayaquil, estos corresponden a los casos de estudio que se analizaron.

El conjunto de datos de la ciudad de San Francisco se recopiló el 02 de junio de 2008; contiene 290 trayectorias registradas por taxis utilizando dispositivos GPS. Cada registro contiene los siguientes datos: id de trayectoria, latitud, longitud, tiempo, velocidad y dirección. Para este conjunto de trayectorias, el análisis incluyó todas las trayectorias registradas entre las 12:30 p.m. y 13:30 p.m. Como resultado de este proceso de filtrado, se obtuvieron 2 382 registros, que representan 290 trayectorias de todo el conjunto de datos. Estas trayectorias han sido reconstruidas aplicando un método de enrutamiento e interpolación cada 5 metros de los cuales se han obtenido 182 559 ubicaciones en total.

El conjunto de datos de la ciudad de Roma fue recolectado el día 12 de febrero de 2014 y contiene 137 trayectorias realizadas por taxis recolectadas por dispositivos GPS con un intervalo de tiempo promedio entre dos ubicaciones consecutivas de 10 segundos. Cada registro contiene id\_trayectoria, latitud, longitud, tiempo, velocidad, dirección. Para este segundo conjunto de trayectorias, se efectuó el análisis en horario comprendido entre las 18:00 hasta las 19:00 hs. Como resultado de este proceso de filtrado se obtuvieron 33 793 registros que representan 137 trayectorias de todo el conjunto de datos. Estas trayectorias han sido reconstruidas aplicando un método de enrutamiento e interpolación cada 5 metros de los cuales se han obtenido 197 790 ubicaciones en total.

El conjunto de datos de la ciudad de Guayaquil fue recolectado el 28 de octubre de 2017 y corresponde a 218 trayectorias realizadas por estudiantes universitarios que viajan en algún medio de transporte como taxi, motocicleta, metro-vía. Las ubicaciones en este conjunto de datos fueron recolectadas por smartphones con un intervalo de tiempo promedio entre dos ubicaciones consecutivas de 5 segundos. Cada registro contiene id\_trayectoria, latitud, longitud, tiempo, nombre de usuario, email y tipo de transporte. Dado que se trata de un conjunto de trayectorias reducido, se efectuó el análisis en el horario comprendido entre las 16:30 hasta las 18:30 hs por ser considerado el de mayor concentración de registros. Como resultado de este proceso

de filtrado se obtuvieron 30 557 registros que representan 206 trayectorias de todo el conjunto de datos. Estas trayectorias han sido reconstruidas aplicando un método de enrutamiento e interpolación cada 5 metros de los cuales se han obtenido 135 237 ubicaciones en total.

#### 4.2.2. Parametrizaciones iniciales

En este trabajo, se ha delimitado un área de análisis de  $1200 \times 800$  metros cuadrados en un espacio bidimensional. Es importante destacar que las parametrizaciones específicas pueden variar según las necesidades de análisis y las características de la región en estudio. Sin embargo, en este contexto particular, se decidió utilizar el mismo tamaño de área de análisis para las tres ciudades evaluadas. Tanto las áreas preprocesada como las delimitadas representan aproximadamente el 3 % del área de análisis y mantienen un tamaño constante de alrededor de  $35 \times 25$  metros. La captura y análisis de instantáneas se realiza con una frecuencia de 1 minuto cada una. La medida de similitud utilizada es la distancia Euclidiana, y el parámetro lambda ( $\lambda$ ) se estableció en 0,068, lo que significa que las ubicaciones GPS de los microgrupos se consideran relevantes hasta los 45 segundos, las ubicaciones GPS que superan ese tiempo se los consideran poco relevantes y se eliminan. Los microgrupos con poca actividad se actualizan cada 30 segundos, y los microgrupos antiguos se eliminan si no han recibido ubicaciones GPS durante al menos 2 minutos. El agrupamiento por proximidad considera el parámetro épsilon establecido en 35 metros para la expansión de los grupos finales.

#### 4.2.3. Análisis de resultados obtenidos en los casos de estudio

Como estrategia para identificar las fluctuaciones de densidad en el flujo del tráfico, se abordan los siguientes aspectos a analizar:

- Aspecto a analizar 1: Desplazamiento.
- Aspecto a analizar 2: Densidad.
- Aspecto a analizar 3: Variabilidad.

En esta sección se muestran las fluctuaciones en la densidad y desplazamiento de los grupos, con acumulaciones significativas en ciertas intersecciones y una variabilidad que refleja cambios en la estabilidad del tráfico vehicular.

##### **Aspecto a analizar 1: Desplazamiento.**

El análisis de densidad de los microgrupos y el desplazamiento del centro de masa de los grupos en las zonas urbanas son una herramienta que permite entender la dinámica del tráfico vehicular y su impacto en la planificación urbana. Estos análisis



permiten identificar patrones de congestión y fluidez en el tráfico, proporcionando información valiosa para la gestión eficiente de la movilidad en las ciudades.

A través de la observación de los resultados del agrupamiento en sucesivos eventos temporales, se evaluará cómo los microgrupos se distribuyen y densifican en función del tiempo, también se analizará como se desplazan los grupos analizados en las diferentes ciudades.

La Figura 4.11 muestra un grupo seleccionado de manera aleatoria el cual es resultante del agrupamiento realizado en la ciudad de Roma.



FIGURA 4.11: Grupo extraído en la ciudad de Roma.

El desplazamiento del grupo vehicular en Roma muestra un patrón dinámico a lo largo del tiempo. Inicialmente, el centro de masa del grupo se encuentra hacia el sur, reflejando la acumulación inicial de tráfico en esta área. A medida que los ciclos de semaforización avanzan y el tráfico se dispersa, el centro de masa se desplaza progresivamente hacia el sur.

Con la aparición de nuevos flujos vehiculares en el centro del mapa, el centro de masa del grupo se desplaza hacia el área central, indicando una redistribución del tráfico en respuesta a los nuevos patrones de movilidad. Este ajuste refleja la adaptación del tráfico a las condiciones cambiantes de la carretera.

Finalmente, el centro de masa se desplaza nuevamente hacia el sur debido a la incorporación de microgrupos densos en esa zona. Sin embargo, la dirección del desplazamiento sugiere un posible retorno hacia el norte, siguiendo una vía con alta

acumulación de tráfico. Este movimiento continuo ilustra la naturaleza dinámica del tráfico vehicular en entornos urbanos complejos.

La Figura 4.12 muestra un grupo extraído de los resultados del agrupamiento en la ciudad de San Francisco.

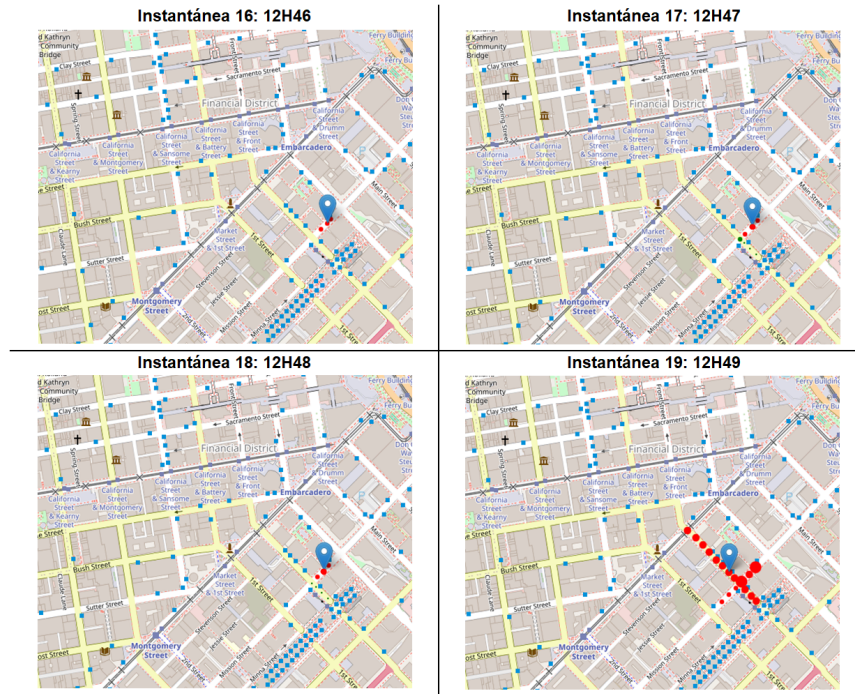


FIGURA 4.12: Grupo extraído en la ciudad de San Francisco.

El estudio del flujo vehicular en San Francisco revela un patrón de desplazamiento distintivo en la zona sureste. Inicialmente, el centro de masa del tráfico muestra una distribución estable de microgrupos sin acumulaciones significativas, lo que indica un flujo vehicular próximo a los valores de densidad media habituales. Sin embargo, a medida que el tiempo avanza, el centro de masa comienza a desplazarse debido a las fluctuaciones en la densidad de los microgrupos, con algunas áreas experimentando disminuciones que sugieren una variabilidad moderada en el tráfico.

Con el tiempo, se observa una dispersión del tráfico, evidenciada por la disminución de las variabilidades negativas en ciertas zonas. A pesar de la presencia de microgrupos con densidades inferiores a la media histórica, el centro de masa se desplaza hacia áreas donde la variabilidad positiva sigue predominando. Esto sugiere un ajuste del tráfico hacia calles con densidades ligeramente superiores a la media.

En etapas más avanzadas, el centro de masa se desplaza notablemente hacia la zona central, donde se intensifica la acumulación vehicular. La alta variabilidad positiva en esta área indica una saturación considerable y una creciente inestabilidad en el flujo de tráfico, con la congestión propagándose a lo largo de las principales vías de conexión.

La Figura 4.13 muestra los resultados del agrupamiento en la ciudad de Guayaquil.

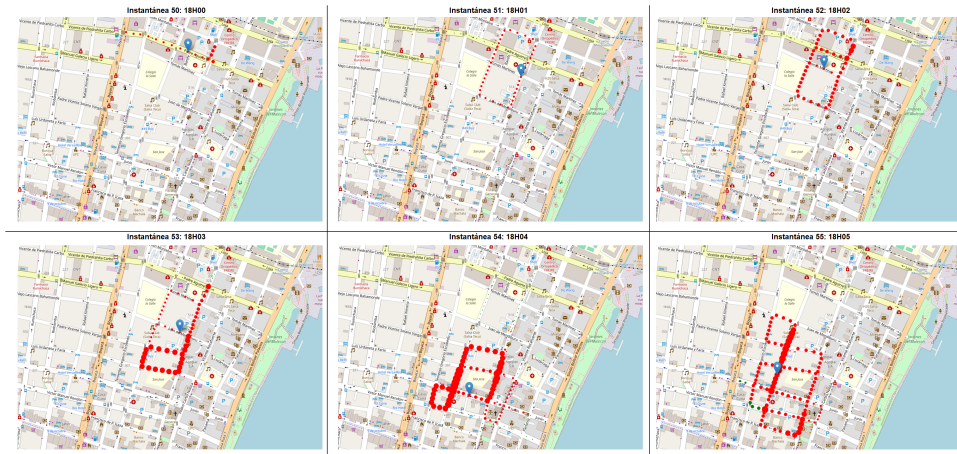


FIGURA 4.13: Grupo extraído en la ciudad de Guayaquil.

El análisis del flujo vehicular en Guayaquil revela un patrón de desplazamiento significativo a lo largo del tiempo. Inicialmente, el centro de masa vehicular muestra una distribución mixta de microgrupos con variabilidades positivas y negativas, indicando una variabilidad en el tráfico que se distribuye alrededor del centro de masa. Durante esta fase, no se observa un desplazamiento claro del centro de masa, lo que sugiere una relativa estabilidad en el patrón de tráfico.

A medida que el tiempo avanza, el centro de masa del grupo vehicular comienza a desplazarse hacia el norte, coincidiendo con el aumento de microgrupos con variabilidad positiva en esa área. Este desplazamiento indica una acumulación creciente de tráfico en el norte, lo que sugiere una tendencia hacia la congestión en esa región. Al mismo tiempo, el centro de masa se aleja de las zonas con variabilidad negativa, que representan áreas con menor densidad vehicular.

En las instantáneas siguientes, la expansión de microgrupos densos se refleja en el desplazamiento del centro de masa hacia las principales vías cercanas a esta concentración. La propagación de la congestión refuerza este movimiento hacia las áreas de mayor densidad, con el centro de masa ajustándose para alinearse con la creciente acumulación vehicular en el área central.

Eventualmente, se observa una ligera dispersión en el tráfico en algunas zonas, lo que provoca un leve movimiento del centro de masa. Este ajuste refleja una redistribución del tráfico, con la aparición de más microgrupos con variabilidades negativas indicando una disminución local en la congestión. En la última instantánea, el centro de masa muestra un equilibrio más marcado, reflejando tanto áreas de alta como de baja densidad, y sugiriendo una posible estabilización del tráfico en ciertas partes de Guayaquil.

### Aspecto a analizar 2: Densidad.

Antes de analizar la variabilidad del tráfico, es fundamental examinar la densidad vehicular, ya que esta proporciona una base para comprender el comportamiento general del flujo vehicular en las áreas urbanas. La densidad vehicular refleja la cantidad de vehículos presentes en un área específica durante un período determinado, lo cual es esencial para identificar determinados patrones y tendencias en el tráfico. Un análisis exhaustivo de la densidad permite identificar puntos críticos de congestión, evaluar la efectividad de la infraestructura vial existente y entender cómo factores como el crecimiento poblacional o las horas pico afectan el tráfico.

En la Tabla 4.1 se presentan los estadísticos descriptivos sobre las densidades vehiculares de los grupos resultantes en los casos de uso.

TABLA 4.1: Resumen estadístico de densidades.

Ciudad	Media	Mediana	Desviación	Mínimo	Máximo
San Francisco	0,09990	0,00248	0,21502	0,00000	0,83726
Roma	0,01378	0,00268	0,02350	0,00000	0,12381
Guayaquil	0,06395	0,00870	0,10210	0,00000	0,44237

San Francisco presenta una media de densidad de vehículos de 0,09990, con una mediana de 0,00248, indicando que la distribución de densidades es asimétrica y presenta una alta variabilidad. La desviación es de 0,21502, lo que refleja una dispersión significativa en los valores de densidad, que oscilan entre 0,00000 y 0,83726. Esta amplia gama y alta desviación sugieren que en San Francisco existen áreas con densidades de vehículos muy bajas y otras con concentraciones extremadamente altas. Esta heterogeneidad puede deberse a la estructura urbana diversa y a la presencia de zonas con diferentes características de infraestructura y demanda vehicular.

Roma tiene una media de densidad de 0,01378 y una mediana de 0,00268, lo que revela una distribución que está centrada en valores bajos. La desviación es de 0,02350, indicando una menor variabilidad en comparación con San Francisco, con densidades que varían entre 0,00000 y 0,12381. La baja desviación y la concentración de datos en el rango inferior reflejan que la distribución es relativamente uniforme y menos concentrada de vehículos en Roma. Esto podría estar relacionado con la planificación urbana y las políticas de transporte que limitan la densidad vehicular en ciertas áreas.

Guayaquil muestra una media de densidad de 0,06395 y una mediana de 0,00870. La desviación de 0,10210 indica una variabilidad moderada en las densidades, con un rango de 0,00000 a 0,44237. La diferencia entre la media y la mediana indica que hay algunas áreas con densidades de vehículos significativamente más altas que el promedio. Este patrón podría reflejar un crecimiento urbano desigual o áreas con concentraciones de tráfico más intensas, lo que contribuye a una distribución de densidades más heterogénea en comparación con Roma, pero menos extrema que en San Francisco.

**Aspecto a analizar 3: Variabilidad.**

El análisis de la variabilidad realizado sobre los microgrupos permite comprender las dinámicas del tráfico urbano. Este enfoque se centra en evaluar las diferencias entre la densidad actual de vehículos y la densidad histórica, lo que se logra es obtener una visión simplificada de las fluctuaciones y patrones de densidades a lo largo del tiempo. Monitorear esta métricas proporciona una visión acerca de la adaptabilidad del algoritmo para asegurar la formación de grupos consistentes y su aplicación en diversos contextos.

En la Figura 4.14 se pueden observar los diferentes niveles de variabilidades de los grupos en el tráfico vehicular. Un grupo con la variabilidad positiva, indicado por la totalidad de sus microgrupos en color rojo, todos sus microgrupos superan sus respectivas densidades medias históricas. Este grupo se encuentra predominantemente en la zona superior izquierda del mapa, señalando áreas donde la densidad actual de vehículos supera a la densidad histórica, lo que sugiere un aumento significativo en el tráfico.

Otro grupo está representado por la variabilidad negativa, denotado por la totalidad de sus microgrupos en color verde, en esta variabilidad todos sus microgrupos no superan sus respectivas densidades medias históricas. Este grupo aparece en la zona superior derecha, destacando zonas donde la densidad actual es menor que la histórica, reflejando una disminución en el flujo vehicular.

Por último, se presenta un grupo con la variabilidad mixta, indicado por microgrupos en color verde y rojo. Este grupo se manifiesta en la zona central derecha, estas zonas sugieren fluctuaciones más complejas donde se alternan períodos de alta y baja densidad vehicular en los microgrupos.

El objetivo de analizar este aspecto es determinar las características presentes en las tendencias de la variabilidad, lo cual puede indicar cambios significativos en el rendimiento del tráfico o en las características de los datos utilizados.

**Variabilidad Positiva.**

En la Figura 4.15 se analizan la variabilidad en la densidad de un grupo a lo largo de varios ciclos, extraído en la ciudad de San Francisco, comparando la densidad actual medida en cada ciclo con la densidad media histórica.

En todos los ciclos analizados, se observa una variabilidad positiva, indicando que la densidad actual en cada ciclo es mayor que la densidad media histórica. Esto sugiere una tendencia de incremento en la densidad actual respecto a las densidades históricas, reflejada también en la figura. La variabilidad gráfica también muestra fluctuaciones en los valores, con picos significativos en los ciclos 52 y 55, que coinciden con los mayores incrementos de densidad actual respecto a la media.

El análisis de la variabilidad de las densidades a lo largo de los ciclos muestra una tendencia clara de incremento positivo en cada ciclo. Las diferencias entre la densidad actual y la media histórica indican que, en general, las características de los datos o el rendimiento del algoritmo de agrupamiento han mejorado o cambiado

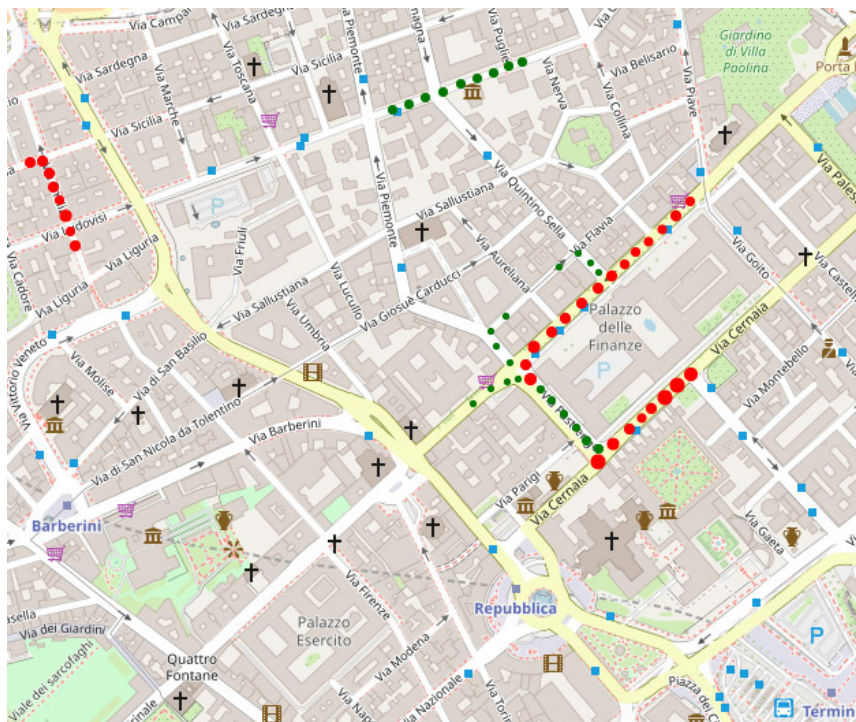


FIGURA 4.14: Ejemplos de los distintos niveles de variabilidad que se pueden obtener.

significativamente. Los picos de variabilidad en los ciclos 52 y 55 sugieren eventos específicos que han afectado la densidad de los grupos, mientras que los incrementos más moderados en otros ciclos podrían indicar una estabilización gradual en el rendimiento del algoritmo.

#### Variabilidad Negativa.

La Figura 4.16 se presenta un grupo que ha presentado variabilidades negativas en la densidad de los grupos generados por el agrupamiento a lo largo de varios ciclos, para este caso se extrajo un grupo de los resultados de la ciudad de Guayaquil.

Los tres ciclos analizados presentan una variabilidad negativa significativa en la densidad de los grupos formados. Los ciclos 12 y 13 muestran una disminución constante del 80% respecto a la densidad media histórica, indicando que el grupo ha presentado una densidad inferior respecto a la densidad habitual.

#### Variabilidad Mixta.

La Figura 4.17 se presenta el ejemplo de un grupo que ha presentado variabilidades mixtas en la densidad de los grupos generados por el agrupamiento a lo largo de varios ciclos, comparando la densidad actual medida en cada ciclo con la densidad media histórica en la ciudad de Guayaquil.

El análisis de la variabilidad de las densidades a lo largo de los ciclos muestra tanto incrementos como decrementos significativos. Los ciclos 16 y 19 presentan una variabilidad positiva, sugiriendo que en el grupo formado por el algoritmo existe

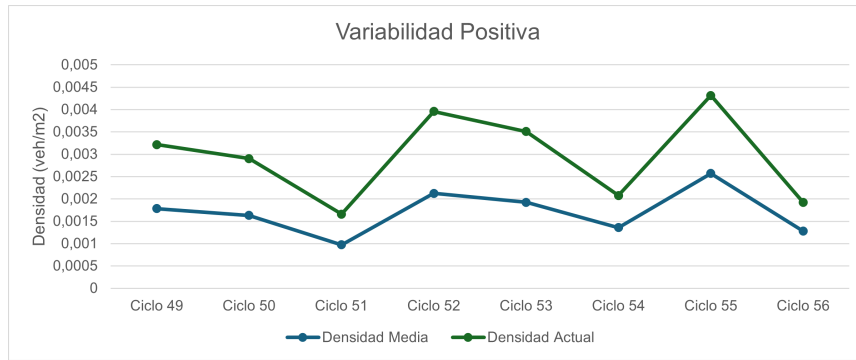


FIGURA 4.15: Ejemplo de variabilidad positiva.

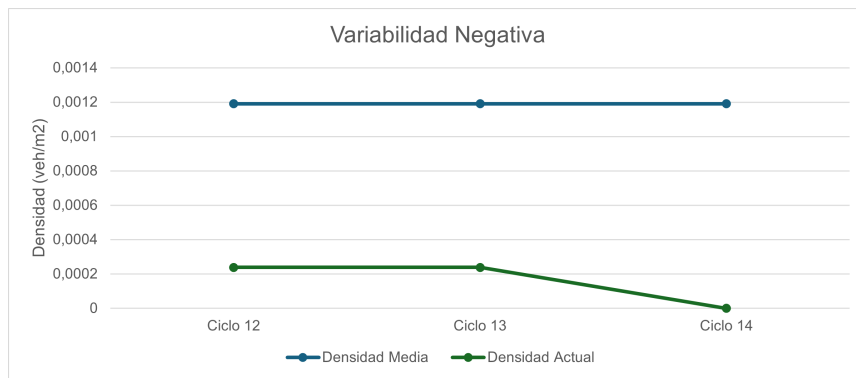


FIGURA 4.16: Ejemplo de variabilidad negativa.

una presencia mayoritaria de microgrupos que superan la densidad histórica, mientras que los ciclos 17 y 18 muestran una variabilidad negativa, indicando posibles ausencia de densidad respecto a la densidad histórica en las características de los datos.

Es importante destacar que los patrones de variabilidad observados en este aspecto se han manifestado en distintas proporciones en todas las ciudades analizadas, por lo que la presencia de cada patrón es independientemente de la ciudad seleccionada para el análisis.

En base a las observaciones realizadas sobre los grupos, se puede identificar las siguientes características recurrentes:

**Desplazamiento.** El centro de masa de los grupos se desplaza conforme cambia la distribución de densidades. Los desplazamientos de los centros de masa están influenciados por la incorporación de nuevos flujos vehiculares y la dispersión de los microgrupos existentes. Hay una tendencia a que los centros de masa se mueva hacia otra ubicación, aunque en ciclos posteriores, se puede llegar a observar un retorno de la ubicación del centro de masa.

**Densidad.** La densidad de los microgrupos varía a lo largo de los ciclos, con zonas que inicialmente tienen alta densidad y luego disminuyen con el paso del tiempo. Las colas de densidad superior a la media histórica se forman y posteriormente

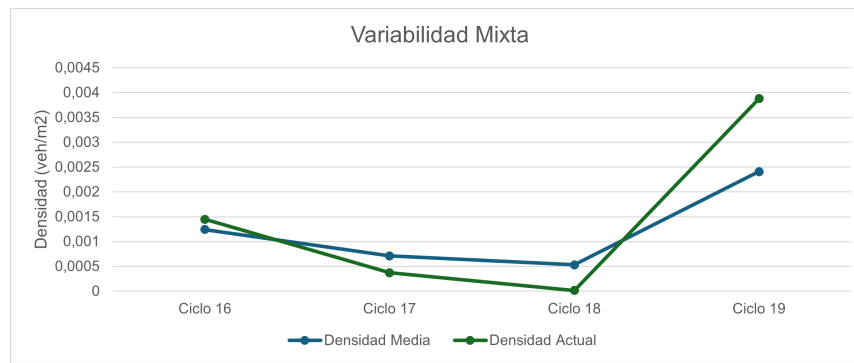


FIGURA 4.17: Ejemplo de variabilidad mixta.

disminuyen, a veces hasta desaparecer. La incorporación de nuevos microgrupos con densidad alta o media afecta la densidad total y la ubicación del centro de masa.

**Variabilidad de la densidad.** La densidad de los microgrupos y la ubicación del centroide muestran una variabilidad notable a lo largo de los ciclos. Las zonas de alta densidad no son estáticas, y su distribución varía, desplazándose entre diferentes partes del mapa. La influencia de elementos externos, como la semaforización, y la llegada de nuevos flujos vehiculares densos contribuyen a la variabilidad observada en los patrones de densidad y desplazamiento.

#### 4.2.4. Distribución acumulada de variabilidades

El mapa presentado en la Figura 4.18 muestra un reticulado sobre la ciudad de Roma que indica la frecuencia de detección de grupos con una variabilidad positiva en la densidad, comparando la densidad actual con la histórica. Cada celda del reticulado tiene un valor que representa el número de períodos de un minuto durante una hora en los que al menos un grupo con variabilidad positiva ha sido identificado en esa ubicación. El máximo de contabilizaciones por celda es de 60.

La distribución de la variabilidad en el mapa revela áreas con mayor frecuencia de detección, destacadas en tonos más oscuros de rojo, lo cual indica zonas con alta actividad o densidad de grupos que varían positivamente respecto a su densidad histórica. Por otro lado, las celdas con menor o nula detección, representadas en tonos más claros o blancos, corresponden a áreas con poca o ninguna actividad de variabilidad positiva durante el período de observación.

En cuanto a los patrones espaciales, se observa una concentración significativa de celdas con alta frecuencia en las inmediaciones de las estaciones de metro. Esto sugiere que estas zonas son puntos de convergencia o de paso frecuente de grupos, posiblemente debido a la afluencia de personas que utilizan el transporte público. Además, las calles principales y áreas comerciales cercanas a estos nodos también muestran un incremento en la actividad, lo cual puede estar asociado a la dinámica económica y social de estas áreas.



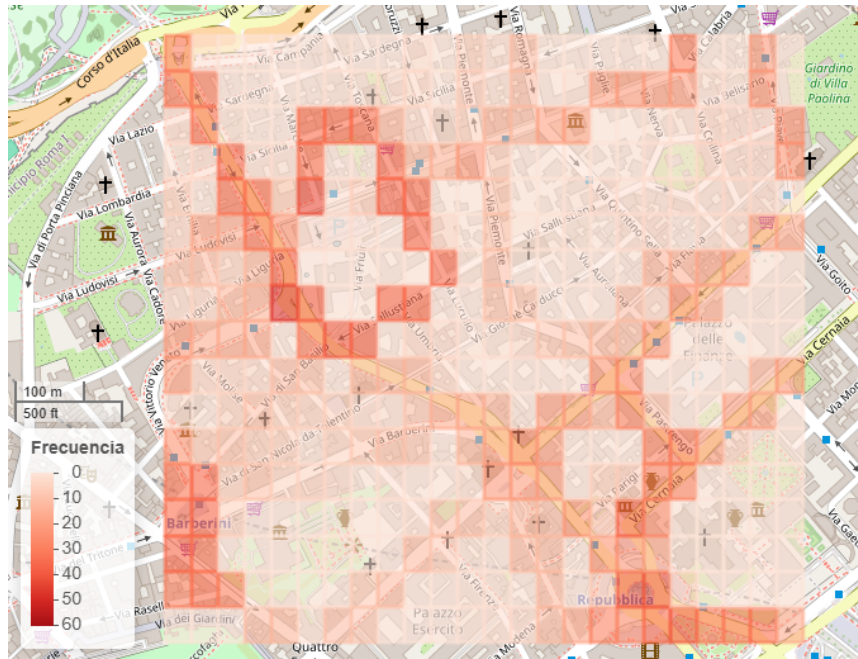


FIGURA 4.18: Distribución acumulada por celdas en la ciudad de Roma.

Desde una perspectiva de implicaciones urbanas, las zonas con mayor variabilidad positiva pueden indicar áreas de interés para la planificación urbana, la seguridad y la gestión del flujo de personas. Por ejemplo, un incremento en la densidad en ciertas horas puede requerir medidas adicionales de control de tráfico peatonal o vehicular. Asimismo, en términos comerciales, estas áreas pueden ser potencialmente atractivas para negocios debido al aumento de la presencia de personas.

La comparación histórica de los datos sugiere que las áreas con variabilidad positiva han experimentado un cambio significativo en comparación con períodos anteriores. Este cambio podría estar relacionado con eventos específicos, modificaciones en la infraestructura urbana o variaciones en el comportamiento de la población. Identificar las causas de estas variaciones es crucial para entender los patrones de movilidad y densidad en la ciudad.

El mapa presentado en la Figura 4.19 correspondiente a la ciudad de Guayaquil muestra, se observa una concentración significativa de celdas con alta frecuencia a lo largo de las principales avenidas. Esto sugiere que estas zonas son puntos clave de tránsito, además, áreas cercanas a parques y zonas verdes, también muestran un incremento en la actividad, lo cual puede estar asociado a la agrupación de personas en espacios recreativos.

Las zonas con mayor variabilidad positiva pueden señalar áreas de interés para la planificación urbana, la seguridad y la gestión del flujo de personas. Por ejemplo, el aumento en la densidad en ciertas horas podría requerir medidas adicionales de control de tráfico peatonal o vehicular. En términos comerciales, las áreas con alta frecuencia de detección pueden ser atractivas para negocios debido al incremento

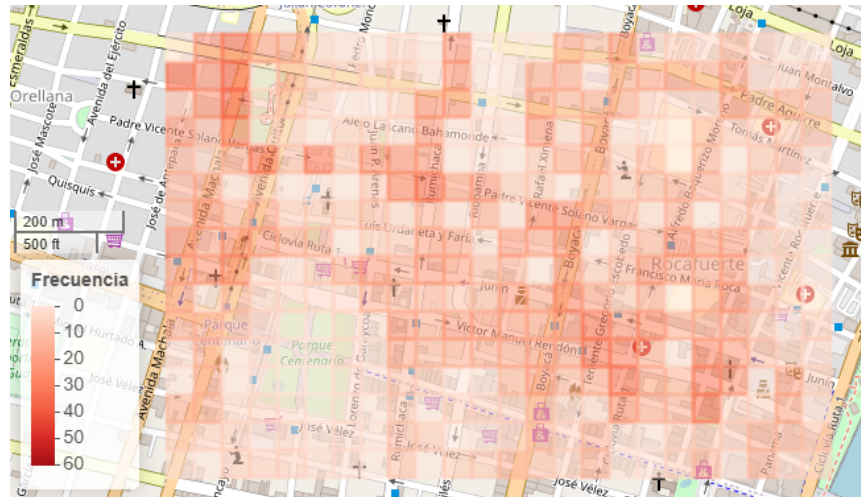


FIGURA 4.19: Distribución acumulada por celdas en la ciudad de Guayaquil.

en la presencia de personas, lo que podría traducirse en mayores oportunidades de ventas y actividad económica.

La detección de variabilidad positiva sugiere que estas áreas han experimentado cambios significativos en comparación con períodos históricos. Esto podría estar relacionado con eventos específicos, cambios en la infraestructura urbana o variaciones en el comportamiento de la población.

El mapa proporcionado de la ciudad de San Francisco mostrado en la Figura 4.20 presenta una concentración significativa de celdas con alta frecuencia en las zonas financieras y comerciales. Esto sugiere que estas zonas son puntos recurrentes de tránsito o actividades rutinarias, posiblemente debido a su importancia económica y comercial. La alta densidad de detección en estas áreas podría estar asociada a la congregación de personas por motivos laborales y turísticos.

El mapa también revela patrones espaciales de actividad alrededor de las principales carreteras y calles comerciales. Estas áreas muestran un incremento en la actividad, lo cual puede estar relacionado con la presencia de oficinas, comercios y atracciones turísticas.

#### 4.2.5. Validación del agrupamiento

En algoritmos de agrupamiento la validación de resultados es una etapa crítica para determinar la eficacia de los grupos formados.

En esta tesis, se presenta un algoritmo de agrupamiento compuesto por dos niveles, en el cual se aplican diversas métricas para evaluar la calidad de los agrupamientos generados. Estas métricas han sido seleccionadas meticulosamente para asegurar una evaluación adecuada del proceso de agrupamiento.

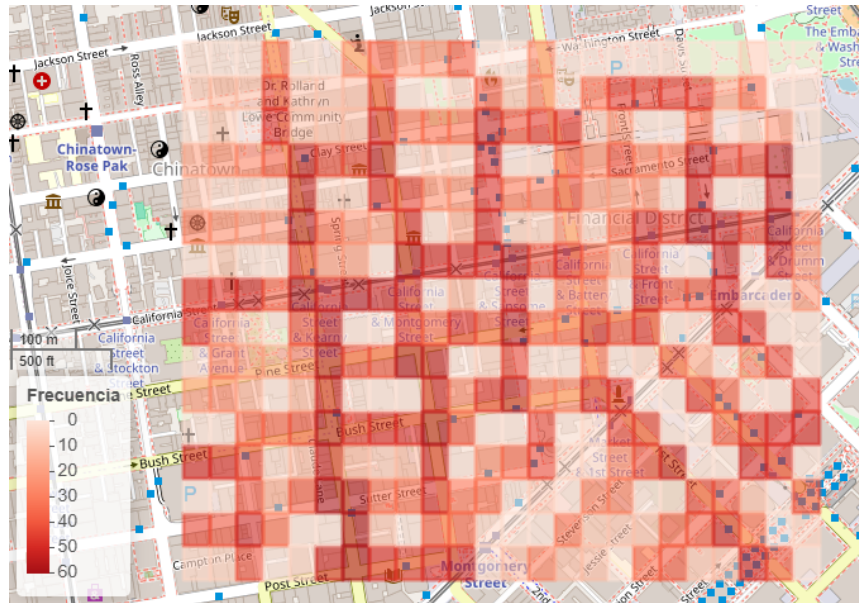


FIGURA 4.20: Distribución acumulada por celdas en la ciudad de San Francisco.

### Indicadores de calidad aplicada a los microgrupos del agrupamiento de nivel 1

En el primer nivel, se emplean la métrica de Silhouette, las cuales proporcionan una medida de la cohesión y separación de los grupos, ambas métricas permiten evaluar la eficacia del proceso de agrupamiento.

**Coefficiente de Silhouette (SC).** Es una métrica esencial utilizada para evaluar la calidad de los resultados generados por algoritmos de agrupamiento (Liu et al. 2024; Kossakov et al. 2024). Se basa en dos conceptos fundamentales: cohesión y separación. El primero mide qué tan cerca está un punto de datos de los otros puntos dentro del mismo grupo, calculando la distancia promedio entre un punto y todos los demás puntos de su grupo, y el segundo mide la distancia promedio entre un punto y los puntos del grupo más cercano al que no pertenece.

Un puntaje de Silhouette cercano a 1 indica un buen resultado de agrupamiento caracterizado por sus elementos que están bien agrupado dentro de su grupo y alejado de otros. Una valoración cercana a 0 indica que los elementos está en el límite de dos grupos, mientras que un valor negativo indica una asignación incorrecta a un grupo diferente. Esta métrica es esencial para determinar la calidad de los grupos formados y es ampliamente utilizada para seleccionar la configuración óptima de agrupamiento.

La Tabla 4.2 presenta los resultados del coeficiente de Silhouette para los microgrupos obtenidos en el nivel 1 del agrupamiento.

Los resultados obtenidos en las tres ciudades indican que se obtienen agrupamientos adecuados, los resultados con respecto a la formación de los grupos son razonables tomando en cuenta el rango de valoración del indicador. Únicamente en el caso de Guayaquil se puede observar una mayor dispersión en su resultado.

TABLA 4.2: Resultados del coeficiente de Silhouette.

Ciudad	Media	Desviación estándar
San Francisco	0,4641	0,0096
Roma	0,4720	0,0186
Guayaquil	0,4280	0,1455

### Indicadores de calidad aplicada a los grupos del agrupamiento de nivel 2

En el segundo nivel, la calidad de los grupos se valida mediante una métrica que se enfoca en evaluar la estructura de densidad de los grupos formados.

**Índice de Validación de Agrupamiento Basado en Densidad (DBCW).** Es una herramienta diseñada para evaluar la calidad de un agrupamiento, especialmente cuando se utilizan algoritmos basados en densidad. El propósito principal del DBCW es medir cuán bien definidos y separados están los grupos generados por un algoritmo de agrupamiento de densidad (Moulavi et al. 2014).

DBCW tiene la capacidad de manejar grupos con diferentes densidades y formas, ya que por un lado se basa en la noción de una distancia, que permite capturar propiedades de densidad, y por otro lado hace uso de un árbol de expansión mínima (MST) por grupo para captar la forma de los mismos.

Este índice se basa en dos componentes esenciales, el primero es la densidad intra-cluster y el segundo es la separación inter-cluster. La densidad intra-cluster se determina calculando la densidad local de los puntos dentro de un grupo, lo cual se logra contando la cantidad de puntos del grupo. Esto proporciona una medida de cuán concentrados están los puntos dentro del grupo. Por otro lado, la separación inter-cluster se evalúa mediante el cálculo de las distancias entre puntos de diferentes grupos, lo que indica la claridad con que los grupos están separados unos de otros. Un buen agrupamiento debe mostrar grupos que están bien separados entre sí.

Para cada grupo  $C_i$  resultante de un agrupamiento, se calcula un índice  $V_c(C_i)$  que determina qué tan concentrados están los elementos dentro del grupo en comparación con la distancia que los separa de otros grupos. Esto se logra mediante la relación entre la distancia mínima de separación con otros grupos y la densidad de separación dentro del propio grupo. Para obtener el  $V_c$  para cada grupo, se emplea la Ecuación 4.3.

$$V_C(C_i) = \frac{\min_{1 \leq j \leq l, j \neq i} (DSCP(C_i, C_j)) - DSC(C_i)}{\max(\min_{1 \leq j \leq l, j \neq i} (DSCP(C_i, C_j)), DSC(C_i))} \quad (4.3)$$

donde,  $DSCP$  define la distancia mínima de alcanzabilidad entre los nodos internos de dos pares de grupos. Para ello utiliza el árbol de expansión mínima, que es esencial para determinar la estructura de los grupos, que toma forma de un esqueleto que indica como están conectados sus elementos, esto se realiza para medir

y evaluar la distancia identificando los bordes entre los grupos al considerar las distancias que conectan los elementos más cercanos.

Por otro lado, el *DSC* corresponde a la menor densidad dentro de un grupo.

Finalmente, para obtener un panorama general de todo el agrupamiento, se calcula el índice  $DBCVC(C)$ , que es el promedio ponderado de los índices de cada grupo, ajustado por el tamaño de cada grupo en relación al total de datos, este cálculo emplea la Ecuación 4.4.

$$DBCVC(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_C(C_i) \quad (4.4)$$

donde,  $l$  corresponde a la cantidad total de grupos que se obtuvieron durante el agrupamiento;  $C_i$  indica la cantidad de elementos que conforman el grupo;  $O$  indica el número total de elementos analizados en el agrupamiento y  $V_C(C_i)$  corresponde al cálculo del *DBCVC* de un grupo específico.

El *DBCVC* combina estos factores para generar un índice que varía en una escala que va desde -1 a 1, donde los valores más altos sugieren una buena formación de grupos, caracterizados por alta densidad interna y separación adecuada entre ellos, mientras que los valores más bajos o negativos pueden señalar que los grupos no están bien definidos o se solapan (Thieu 2024; Van Thieu y Mirjalili 2023).

Los valores del *DBCVC* para los casos de estudio se presentan en la Tabla 4.3, se observan los valores medios y desviaciones para las tres ciudades. Estos son valores adecuados para este indicador y se observa una dispersión moderada, esto indica que los grupos se han formado de una manera apropiada.

TABLA 4.3: Resultados del índice *DBCVC*.

Ciudad	Media	Desviación estándar
San Francisco	0,669980	0,006714
Roma	0,779924	0,020378
Guayaquil	0,449972	0,022218

#### 4.2.6. Análisis de los tiempos de ejecución

Luego de haber realizado la ejecución con instantáneas de 60 segundos en la que se utilizó los tres conjuntos de datos San Francisco, Roma y Guayaquil, para evaluar el rendimiento del método propuesto, se ha generado el siguiente cuadro con los tiempos de ejecución. Este cuadro proporciona una visión detallada y objetiva de cómo se comporta el método en condiciones prácticas.

En el proceso de ejecución en paralelo, es relevante destacar que los tiempos del módulo de agrupamiento se miden de manera independiente, mientras que los tiempos del módulo de agrupamiento por proximidad y del sub-módulo de visualización se evalúan de forma conjunta, y por separado al módulo de agrupamiento

por proximidad, los resultados de los tiempos de ejecución se muestran en la Tabla 4.4, en unidad de medida en minuto y son obtenidos a partir de la ejecución de una hora de datos a intervalos de 60 segundos en las ciudades de San Francisco, Roma y Guayaquil. Esta estrategia de medición individualizada permite un análisis más preciso de cada módulo y su contribución al tiempo total de ejecución. Además, en las ejecuciones en paralelo es común observar que se toma como referencia el tiempo máximo de los procesos a medir.

TABLA 4.4: Tiempos de ejecución (minutos) del método propuesto.

Módulo	San Francisco	Roma	Guayaquil
Primer agrupamiento	28:23	10:46	18:32
Segundo agrupamiento	01:41	00:47	00:56
Tiempo máximo	28:23	10:46	18:32

Los resultados muestran diferencias significativas en los tiempos de ejecución de los módulos de nivel 1 y 2 en las ciudades de San Francisco, Roma y Guayaquil. A pesar de que se utilizaron áreas de dimensiones muy parecidas, existen varias razones que pueden explicar estas disparidades.

En el caso del módulo del primer agrupamiento, el escenario con instantáneas de 60 segundos mostraron que la ciudad de San Francisco muestra el tiempo más prolongado, con 28 minutos y 23 segundos. Esto podría deberse a la complejidad de los datos en esa ciudad o a una mayor cantidad de datos que requieren procesamiento. Por otro lado, Roma y Guayaquil presentan tiempos más cortos, 10 minutos con 46 segundos y 18 minutos con 32 segundos, respectivamente. Los valores de eficiencia en el proceso de agrupamiento en esas ciudades está directamente relacionado a la cantidad de trayectorias.

En los escenarios con San Francisco se procesó 290 trayectorias y se ha obtenido el mayor tiempo de procesamiento, Guayaquil con 218 trayectorias ha obtenido un tiempo menor que San Francisco, y finalmente Roma obtuvo el menor tiempo de procesamiento al procesarse 137 trayectorias, lo que reafirma que existe una relación directa entre el tiempo de procesamiento y la cantidad de trayectorias agrupadas.

El almacenamiento de cada instantánea toma menos de 0,02 segundos, durante este tiempo el agrupamiento se detiene momentáneamente, sin embargo, este tiempo es relativamente pequeño lo que no afecta al tiempo de procesamiento del agrupamiento ya que si se lo compara con el tiempo entre instantánea el cual es de 60 segundos, este tiempo representa el 0,03 %.

En cuanto al módulo del segundo agrupamiento, los escenarios realizados con instantáneas de 60 segundos muestran que el conjunto de datos de Roma se destaca con el tiempo más bajo con 47 segundos, seguido por Guayaquil con 56 segundos. San Francisco, en este caso, muestra el tiempo más largo, 1 minuto y 41 segundos. Estos tiempos podrían relacionarse con la disponibilidad de recursos de procesamiento en cada ubicación o la distribución y centralización de zonas densas de los

microgrupos resultantes en el primer agrupamiento.

La diversidad en la disposición de las carreteras disponibles en una ciudad puede influir directamente en el rendimiento de este tipo de soluciones, se plantean algunos ejemplos que ilustran algunos casos a los que se enfrenta la dinámica del tráfico, la Figura 4.21(a) muestra el caso más sencillo y que requerirá menor cantidad de recursos para analizar, el caso de la Figura 4.21(b) es un escenario con una distribución urbana muy común entre las ciudades que requerirá mayor tiempo hasta obtener la valoración del tráfico, y la Figura 4.21(c) muestra el caso de una red vial compleja que esta compuesta por múltiples carreteras que se interceptan.



FIGURA 4.21: Ejemplo de diferentes complejidades de redes. Una sola carretera (a), cuatro carreteras (b) y seis carreteras (c).

Desde la perspectiva de la ejecución en paralelo, estos resultados resaltan la importancia de considerar el rendimiento de cada módulo por separado. La ejecución en paralelo permite distribuir la carga de trabajo de manera eficiente, pero los tiempos de ejecución varían según la capacidad de procesamiento de cada módulo y la forma en que interactúan entre sí.

### 4.3. Discusiones

#### Acerca de los grupos resultantes en el agrupamiento

En la presente sección, se analiza y compara el comportamiento de los grupos resultantes del proceso de agrupamiento en tres ciudades: San Francisco, Roma y Guayaquil. El análisis realizado sobre las densidades y vehículos de los grupos resultantes en San Francisco, Roma y Guayaquil revela diferencias notables en la distribución y concentración de estas variables, la tabla que resume los estadísticos descriptivos respecto a la densidad se pueden visualizar en la Tabla 4.5.

En San Francisco se presenta una alta densidad vehicular y una considerable desviación en la concentración de vehículos, indicando diferencias significativas entre áreas de la ciudad. Esta desviación también se refleja en la cantidad de vehículos totales, señalando una distribución dispersa en diversas zonas. En cuanto a los microgrupos, San Francisco muestra una notable desviación en su formación, lo que sugiere una estructura urbana compleja y diversa.

TABLA 4.5: Densidades de los grupos obtenidos.

Ciudad	Variable	Media	Desviación	Mínimo	Máximo
San Francisco	Microgrupos	60,5	125,1	1	453
	Vehículos Totales	80,4	173,1	0	674
	Densidad Total	0,09990	0,21502	0	0,8376
Roma	Microgrupos	12,3	18,2	1	89
	Vehículos Totales	11,6	19,7	0	104
	Densidad Total	0,01378	0,02350	0	0,12381
Guayaquil	Microgrupos	46,5	64,0	1	281
	Vehículos Totales	51,5	82,2	0	356
	Densidad Total	0,06395	0,10210	0	0,44237

Por su parte, Roma exhibe menores densidades vehiculares y una menor cantidad de microgrupos y vehículos totales, con una desviación menos pronunciada. Esto indica una distribución más homogénea de estas variables en la ciudad.

Guayaquil, similar a Roma, presenta una densidad vehicular moderada y una menor cantidad de microgrupos y vehículos totales en comparación con San Francisco. La desviación en Guayaquil, aunque presente, es también menos destacada, indicando una distribución más uniforme de vehículos y microgrupos en su infraestructura urbana.

Estos resultados se ven influenciados por la infraestructura urbana que tiene cada ciudad, esta tiene un impacto notable en la distribución y desviación de los resultados, San Francisco destaca por su alta densidad y dispersión vehicular, mientras que Roma y Guayaquil presentan patrones más homogéneos.

La Tabla 4.6 presenta un resumen de los resultados de los desplazamientos en metros de los centros de masa de los grupos en las ciudades de San Francisco, Guayaquil y Roma, complementando el análisis de densidades previamente efectuado. Los estadísticos descriptivos proporcionan una comprensión detallada de la distribución y dispersión de estos desplazamientos en cada ciudad.

TABLA 4.6: Desplazamientos de los centros de masa de los grupos (en metros).

Ciudad	Media	Mediana	Desviación	Mínimo	Máximo
San Francisco	85,25	12,40	164,09	0	924,03
Roma	36,05	0,94	69,78	0	506,14
Guayaquil	57,92	11,12	87,53	0	432,51

El análisis de los desplazamientos muestra características distintivas para cada ciudad. En San Francisco, se observa una considerable desviación en los desplazamientos, con una tendencia central alta y una distribución asimétrica que sugiere la presencia de desplazamientos extremadamente grandes.



En Guayaquil, los desplazamientos también son relativamente altos, aunque con una menor desviación que en San Francisco, reflejando una mezcla de movimientos cortos y largos.

Roma presenta una tendencia diferente, con la mayoría de los desplazamientos siendo pequeños y una distribución que incluye algunos valores extremos pero menos frecuentes. Esto indica dinámicas urbanas particulares, donde la dispersión de los grupos y la densidad urbana influyen significativamente en la magnitud y desviación de los desplazamientos en cada ciudad.

En general, San Francisco y Guayaquil han presentado una mayor desviación en los desplazamientos, reflejando movimientos más amplios de los centros de masa de los grupos. En contraste, Roma muestra una distribución de desplazamientos predominantemente más pequeña y menos variable.

La Figura 4.22 proporciona un análisis de las variabilidades de los grupos y presenta los estadísticos descriptivos correspondientes. Esta información sobre las variabilidades de los datos en cada ciudad es valiosa para comprender las fluctuaciones en los flujos de tráfico de las diferentes localidades.

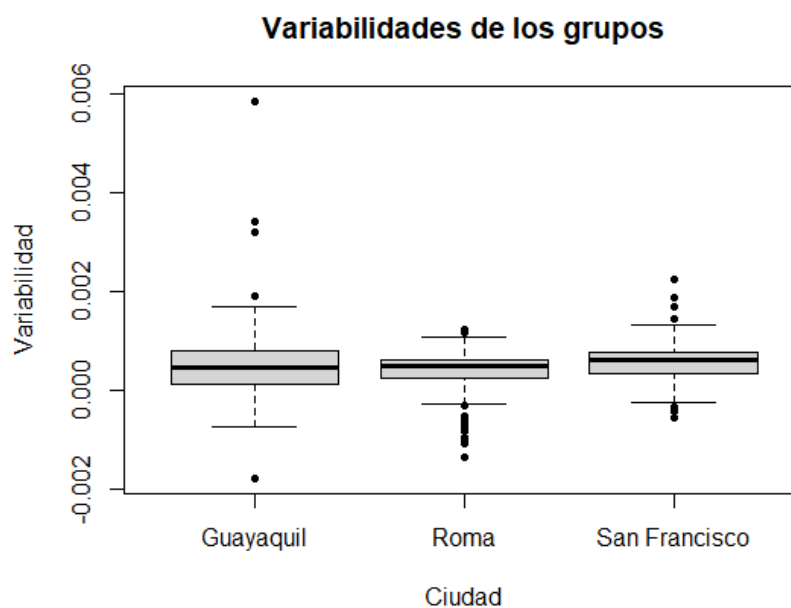


FIGURA 4.22: Variabilidad de los flujos de tráfico en diferentes entornos urbanos.

El análisis de la variabilidad de los grupos revela distintas características en cada ciudad.

Se observa que en los tres casos de estudio existen valores fuera de rango y en Guayaquil esos valores se observan mucho más dispersos, lo que indica que esta ciudad presenta determinadas zonas con una alta posibilidad de congestión. La mayor dispersión en Guayaquil puede estar asociada a un entorno urbano con una mayor

cantidad de vehículos y microgrupos en movimiento, lo que requiere enfoques más adaptativos.

En Roma existen ciertos valores fuera de rango, poco dispersos, que indica que se presentan zonas con menores fluctuaciones de variabilidad, con una tendencia hacia una menor posibilidad de congestión.

San Francisco presenta una dispersión similar a la de Roma, pero su tendencia es contraria presentando mayor posibilidad de congestión, sus valores atípicos no se encuentran tan dispersos a comparación de Guayaquil. Esta distribución aunque presenta fluctuaciones, se mantiene dentro de límites moderados.

### **Acerca de la distribución acumulada de variabilidades**

Los mapas de distribución acumulada de las variabilidades positivas en la densidad vehicular, presentados mediante un reticulado, ofrecen una herramienta valiosa para diversos ámbitos de la gestión urbana y la planificación estratégica. Los resultados obtenidos permiten identificar con precisión las áreas de mayores densidades, proporcionando información crucial que puede ser utilizada para una serie de aplicaciones prácticas.

En primer lugar, la utilidad de estos mapas se manifiesta en la planificación urbana. Al identificar las zonas con alta variabilidad positiva en la densidad, se puede focalizar los esfuerzos para mejorar la infraestructura y los servicios en estas áreas. Por ejemplo, en las zonas con alta concentración de personas, es posible que se requieran más recursos para el manejo del tráfico peatonal y vehicular, la instalación de señalización adecuada, y la mejora del transporte público para evitar congestiones y mejorar la movilidad urbana.

Estos mapas permiten identificar puntos críticos donde la densidad vehicular puede alcanzar niveles que requieran una vigilancia adicional o la implementación de medidas de seguridad específicas. En eventos masivos o durante horas pico, la identificación de estas áreas puede ayudar a prevenir incidentes y a gestionar mejor las multitudes, garantizando la seguridad de los ciudadanos.

Además, los resultados obtenidos permiten una mejor comprensión del comportamiento y las dinámicas de la población urbana que se desplaza principalmente en vehículos. La capacidad de comparar la densidad actual con la histórica y detectar cambios significativos ofrece una visión profunda sobre cómo ciertos eventos, cambios en la infraestructura, o variaciones estacionales afectan la movilidad y la concentración de personas. Esta información es esencial para la planificación a largo plazo y para la toma de decisiones basadas en datos precisos y actualizados.

Las posibilidades que ofrecen estos mapas son amplias. Pueden integrarse con otros SIG y bases de datos para realizar análisis más complejos, como la relación entre la variabilidad de la densidad y factores socioeconómicos, ambientales, y de salud pública. Además, su uso puede extenderse a estudios de impacto de políticas públicas, como la implementación de nuevas líneas de transporte público o la reorganización de espacios públicos.

### **Acerca de los tiempos de ejecución resultantes**

El método propuesto es utilizado para identificar las fluctuaciones en la densidad vehicular, los resultados están basados en un período de una hora de datos en cada ciudad, ofrece una aproximación valiosa para comprender y abordar los problemas de tráfico en áreas urbanas. Este método se basa en la recopilación y análisis de datos en tiempo real, lo que permite una evaluación dinámica de las condiciones de tráfico en diferentes áreas. Los tiempos de ejecución resultantes son consecuencia de la disponibilidad de recursos de procesamiento en cada ubicación y centralización de zonas densas de los microgrupos. En una ejecución en paralelo los resultados con respecto al tiempo de ejecución obtenidos tanto en nivel 1 como en nivel 2 del agrupamiento resaltan la importancia de considerar el rendimiento de cada módulo por separado porque permite distribuir la carga de trabajo de manera eficiente.

A continuación, se discuten algunas de las ventajas y limitaciones de este método, teniendo en cuenta los resultados obtenidos en las ciudades San Francisco, Roma y Guayaquil.

Este método presenta notables ventajas, como la precisión espacial y temporal, ya que el análisis de datos en una hora específica en cada ciudad proporciona una visión precisa y en tiempo real de las condiciones de tráfico, permitiendo la identificación de zonas de densidades vehiculares altas con gran detalle. La detección en tiempo real resulta esencial, ya que posibilita tomar medidas inmediatas, como ajustar rutas o gestionar el tráfico, lo que a su vez contribuye a la reducción de la congestión y mejora la movilidad urbana. Además, este método es altamente flexible y adaptable a distintos escenarios urbanos, lo que lo convierte en una herramienta versátil para abordar los desafíos del tráfico en diferentes ciudades.

Sin embargo, es relevante considerar sus limitaciones. La representatividad de los datos es una preocupación, dado que la elección de una hora específica puede no reflejar completamente las condiciones de tráfico a lo largo del día, especialmente en momentos de mayor afluencia o eventos especiales. La dependencia de datos en tiempo real puede ser un obstáculo en áreas con infraestructura de recopilación de datos menos avanzada. Por último, la ubicación de las estaciones de monitoreo de tráfico puede influir en la representatividad de los resultados, generando ausencia de datos en áreas específicas o falta de representación en zonas menos transitadas.

## **4.4. Conclusiones**

Los resultados obtenidos indican que el método propuesto basado en un agrupamiento dinámico y un agrupamiento de microgrupos es efectivo y preciso en la identificación de zonas con presencia de altas densidades vehiculares. La capacidad de agrupar dinámicamente los datos de trayectorias de origen vehicular en grupos y realizar un análisis específico para cada uno, permite una mejor identificación de

patrones y similitudes en el flujo vehicular, lo que facilita la detección temprana y precisa de áreas con problemas de saturación en el tránsito.

La aplicación del método de agrupamiento dinámico y el análisis de las fluctuaciones en la densidad vehicular se presenta como una estrategia con gran potencial. Estos métodos se han adaptado a la evolución constante del tráfico urbano, capturando patrones cambiantes de movilidad. La importancia del factor de olvido radica en su capacidad para mantener el agrupamiento actualizado, seleccionando adecuadamente las ubicaciones GPS más recientes y eliminando las más antiguas. Este método permite que los grupos reflejen de manera precisa la dinámica actual del tráfico y, al mismo tiempo, permiten detectar densidades altas emergentes de manera temprana.

El primer aspecto analizado se centró en el análisis del desplazamiento y la densidad de los microgrupos vehiculares en zonas urbanas. Se observó que la dinámica de la densidad vehicular varía significativamente en función del tiempo y la disposición de los semáforos. Estos hallazgos señalan la influencia de la infraestructura y la regulación del tráfico en la distribución y desplazamiento de los grupos vehiculares.

El segundo aspecto analizado se centró en la densidad del tráfico, este demostró que varía significativamente entre los casos de uso analizados, reflejando tanto las políticas de transporte como las características urbanas específicas de cada localidad, este destacó las zonas con alta concentración de densidad y los elementos que han provocado los aumentos de densidades en determinados momentos.

El tercer aspecto analizado, evaluó la variabilidad en el flujo del tráfico, analizando las diferencias entre la densidad actual de vehículos y la densidad media histórica. Se identificaron tres niveles de variabilidad: positiva, negativa y mixta. La variabilidad positiva, caracterizada por una densidad actual superior a la histórica, predominó en determinadas zonas de cada ciudad, sugiriendo un incremento significativo del tráfico. En contraste, la variabilidad negativa, con densidad actual inferior a la histórica, se observó en las zonas donde hay menor presencia de flujos vehiculares. Por último, la variabilidad mixta, que alterna períodos de alta y baja densidad vehicular, se presenta de manera irregular reflejando fluctuaciones complejas en el flujo vehicular.

El método propuesto basado en agrupamiento dinámicos se ha destacado por su capacidad para adaptarse a cambios en el tráfico y el análisis de las fluctuaciones de densidad ha proporcionado una visión completa y actualizada del comportamiento vehicular en áreas urbanas. Estos resultados respaldan la eficacia de las técnicas basadas en agrupamientos como una herramienta valiosa para mejorar la gestión del tráfico e identificar problemas relacionados con la congestión en las ciudades.

## Capítulo 5

# Conclusiones y trabajos futuros

### 5.1. Conclusiones

La investigación abordó de manera integral la problemática de las fluctuaciones de densidad en el tráfico urbano, reconociendo la importancia de encontrar soluciones efectivas para su gestión en entornos dinámicos y complejos. Se contextualizó la relevancia del problema, destacando los desafíos que enfrentan las autoridades de transporte y planificación urbana en la optimización de la movilidad en las ciudades. En este sentido, se procedió a realizar la exploración y análisis de diversos algoritmos de agrupamiento dinámico, fundamentales para el procesamiento y análisis de datos de tráfico.

El estudio destacó la necesidad de identificar métodos que no solo sean eficaces, sino también precisos en la identificación de patrones de tráfico, dado el constante cambio y evolución de las condiciones de las vías urbanas. Además se presentó mecanismo de olvido utilizado en algunas soluciones, así como su importancia en el procesamiento de flujos de datos vehiculares dinámicos.

Una vez establecido el contexto y la relevancia de los algoritmos de agrupamiento dinámico, se profundizó en el desarrollo del modelo propuesto. Se resaltó la importancia de la identificación de información geoespacial y de tráfico a partir de trayectorias vehiculares y de sistemas de monitorización como componentes esencial de la investigación. Este proceso proporcionó la base necesaria para la creación de un modelo robusto capaz de identificar correctamente las fluctuaciones de densidad del tráfico en entornos urbanos.

El modelo desarrollado se caracterizó por su capacidad para adaptarse de manera dinámica a las variaciones temporales del tráfico vehicular, representando así un avance significativo en la gestión de la movilidad urbana.

Los resultados presentados en esta investigación destacan la efectividad del método propuesto para el análisis de densidad y variabilidad del tráfico vehicular en diversas ciudades. Se observaron patrones claros en los aspectos de desplazamiento y densidades en grupos vehiculares, así como diferentes niveles de variaciones significativas en la densidad vehicular que fueron cruciales para entender las dinámicas de tráfico en cada caso de estudio. La regulación semafórica, la diversidad del flujo de tráfico y los tiempos de circulación vehicular fueron aspectos clave que

influyeron en los resultados, mostrando cómo las características específicas de cada entorno urbano afectan el comportamiento del tráfico.

El análisis de los casos de estudio con las ciudades de San Francisco, Roma y Guayaquil ha proporcionado una comprensión general y diversa de las dinámicas del tráfico vehicular en distintos contextos urbanos. Cada ciudad presentó características únicas y desafíos específicos que permitieron evaluar la eficacia y adaptabilidad del método de agrupamiento propuesto. La implementación de estos casos de estudio ha sido fundamental para validar el método en escenarios reales y ha proporcionado información crucial que puede servir para el análisis y desarrollo de políticas urbanas más eficientes y adaptativas.

Finalmente, la validación exitosa del modelo en entornos controlados confirma su aplicabilidad práctica y su relevancia para abordar eficazmente los desafíos asociados al análisis de densidad en el tráfico vehicular. En conjunto, estos hallazgos aportan una contribución significativa al campo de la gestión de la movilidad urbana, resaltando su potencial para informar futuras iniciativas y estrategias enfocadas en mejorar las condiciones de los entornos urbanos congestionados.

## 5.2. Trabajos futuros

Como trabajos futuros, se plantea una serie de mejoras y ampliaciones para el método propuesto. En primer lugar, se busca aumentar su adaptabilidad en entornos urbanos complejos, enfocándose en la optimización para intersecciones y la diversidad de carreteras, así como abordar la interacción entre vehículos y peatones.

Además, se proponen casos de estudio en áreas urbanas extendidas, considerando condiciones climáticas severas y escenarios complejos, para evaluar la escalabilidad y robustez del método en contextos más amplios. Para fortalecer la comprensión del problema de congestión, se plantea explorar la influencia de la topología de la red vial en la dinámica del tráfico, analizando la interconexión entre diferentes nodos y su impacto en la congestión.

Otra línea de trabajo futuro implica la inclusión de otros modos de transporte, como vehículos privados, bicicletas y peatones, con análisis específicos por tipo de vehículo para comprender mejor las variaciones en los patrones de conducción. Asimismo, se considera esencial integrar datos multimodales, como transporte público, puntos de interés y semáforos, para obtener una visión más completa de la movilidad urbana.

En el ámbito de la inteligencia artificial, se propone implementar modelos predictivos respaldados por datos históricos y en tiempo real para anticipar y prevenir patrones de congestión. Se plantea un modelo de clasificación para la identificación de áreas congestionadas y no congestionadas, considerando variables adicionales como velocidad del flujo vehicular y capacidad de la carretera.

Otra mejora que se plantea añadir es la inclusión de una métrica para la identificar sectores con congestión vehicular por medio de un indicador que considere

las características de las carreteras como la saturación de estas y pueda procesar un conjunto de datos completo de trayectorias vehiculares.

Finalmente, se sugiere validar el nuevo método en otros contextos espaciales y temporales para verificar su aplicabilidad y generabilidad, lo que permitiría mejorar la gestión del tráfico en diversas condiciones y entornos urbanos.





## Apéndice A

# Diagramas de procesos por componentes

En este apéndice, se presentan los diagrama de procesos de los módulos desarrollados en esta tesis para abordar el agrupamiento, la identificación y visualización de grupos relacionados con las fluctuaciones de densidad del tráfico en entornos urbanos.

### **A.1. Módulo principal de nivel 1: Agrupamiento dinámico de ubicaciones**

Este algoritmo, diseñado para gestionar un flujo continuo de datos, se enfoca en la creación y gestión dinámica de microgrupos. Inicia con la preparación de una lista de grupos vacía y una variable de tiempo general. Operando en un bucle mientras hay datos en el flujo, el algoritmo extrae información de cada ubicación, actualiza el tiempo general, y asigna o crea grupos según la cercanía de las coordenadas. Se ajustan y filtran datos en microgrupos existentes, y se captura una instantánea del estado actual para su almacenamiento. El bucle continúa hasta que no hay más datos en el flujo, concluyendo el proceso de agrupamiento de primer nivel.

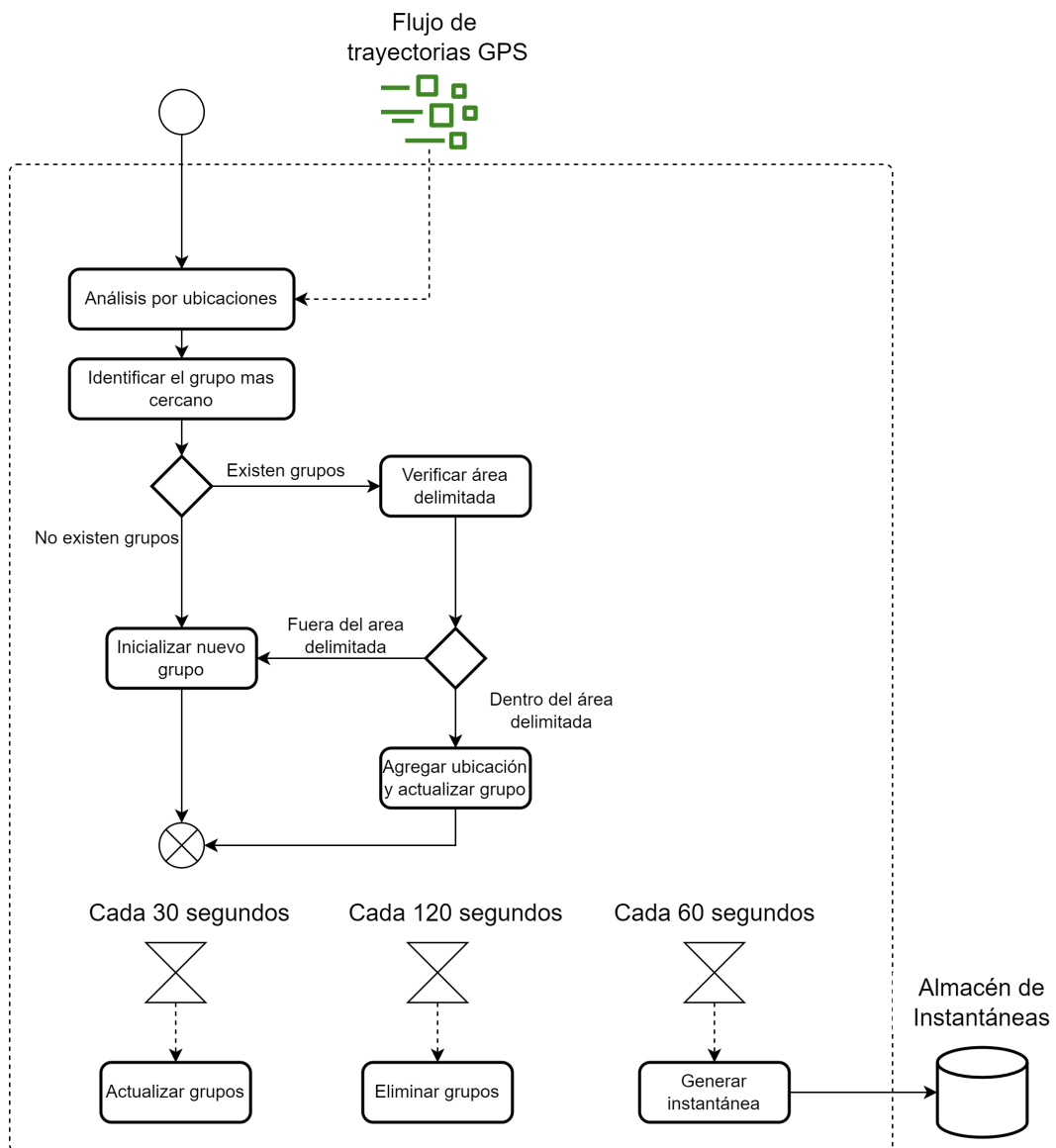


FIGURA A.1: Pasos del módulo principal de nivel 1.

## A.2. Sub-módulo complementario 1: Preprocesamiento de áreas de tráfico históricas

Este sub-módulo fue diseñado para crear áreas preprocesadas localizadas en ubicaciones específicas en una ciudad. Comienza examinando cada área de la ciudad, establece un punto central para áreas específicas y define límites geográficos. Luego, para cada área se extrae la información de datos históricos para comprender patrones pasados. Estos datos históricos proporcionados por una entidad externa proporciona información de los vehículos que han circulado por las carreteras de las áreas previamente definidas.

La información de cada área se resume, incluyendo información correspondiente a la identificación del área, límites y datos de la densidad histórica media de cada área identificada. Este proceso se repite para cada área de la ciudad, creando un conjunto de áreas preprocesadas con información detallada sobre el tráfico y la movilidad. Esta información será relevante para su uso posterior.

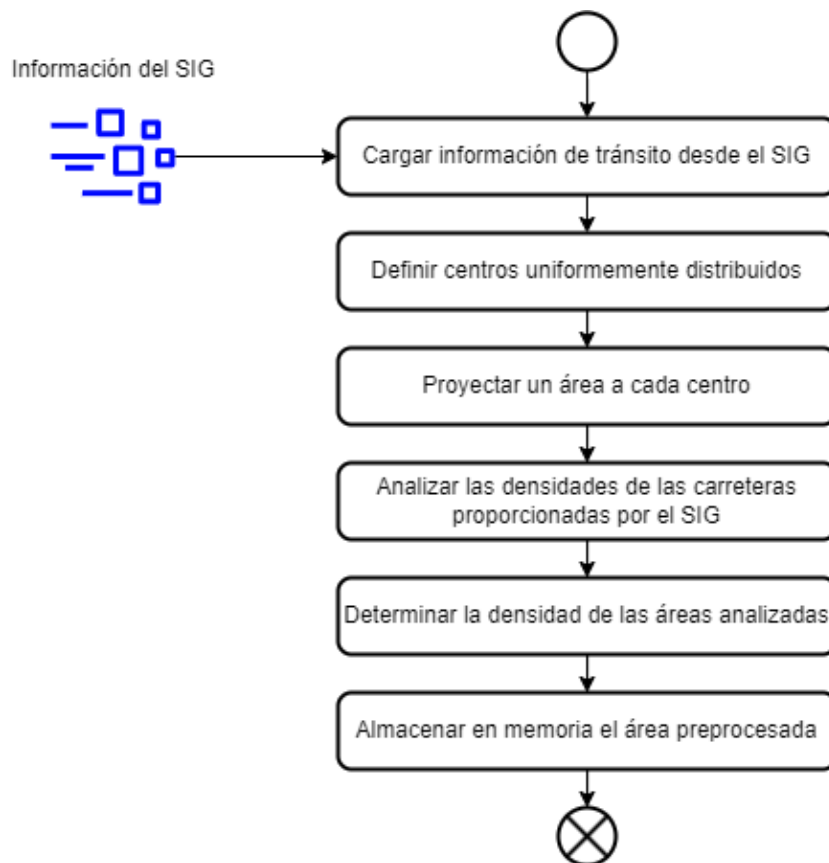


FIGURA A.2: Pasos del sub-módulo complementario 1

### **A.3. Módulo principal de nivel 2: Agrupamiento por proximidad.**

Este algoritmo en constante ejecución se enfoca en evaluar la densidad del tráfico en grupos durante distintos momentos llamados “instantáneas”. A través de un proceso continuo, se verifica la existencia de estas instantáneas, se identifica el área más cercana, se determina la densidad y variabilidad de los microgrupos en cada una de ellas. Luego, los microgrupos son agrupados de acuerdo al criterio de proximidad para formar los grupos finales. Los resultados se actualizan y visualizan periódicamente. Este análisis dinámico proporciona una comprensión en tiempo real de la congestión del tráfico en distintas partes de la ciudad, siendo útil para la toma de decisiones en movilidad urbana. El proceso continúa hasta que no existan más instantáneas para evaluar.

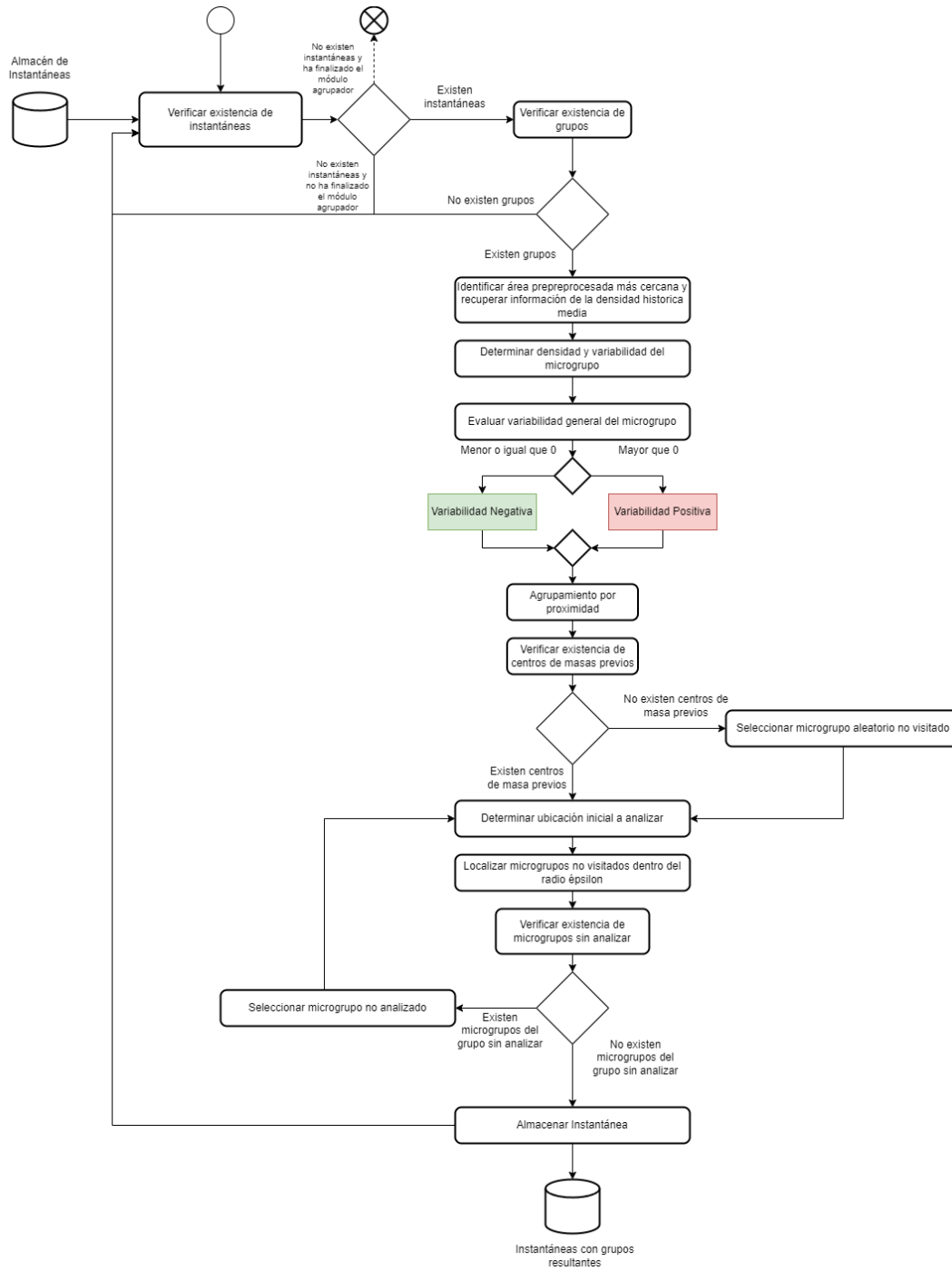


FIGURA A.3: Pasos del módulo principal 2

#### A.4. Sub-módulo complementario 2: Visualizador de resultados

Este sub-módulo está diseñado para representar gráficamente la información de grupos en un mapa de la ciudad, utilizando una lista de instantáneas. A través de un proceso que explora cada momento registrado, se prepara un mapa en blanco para mostrar la distribución y características de cada grupo en ese instante específico. Cada grupo se representa con su posición central, los puntos individuales que lo componen, y estadísticas generales. Estos mapas se almacenan para proporcionar una visualización clara y detallada de la evolución de los grupos a lo largo del tiempo. Esta representación gráfica puede ser crucial para comprender patrones y dinámicas en la congestión del tráfico o los movimientos en la ciudad.

En este mismo submódulo también se determina el índice de densidad alta para un conjunto de zonas dispuestas sobre una rejilla, se contabilizan las zonas con variabilidades altas que son las mas propensas a presentar un estado próximo a la congestión.

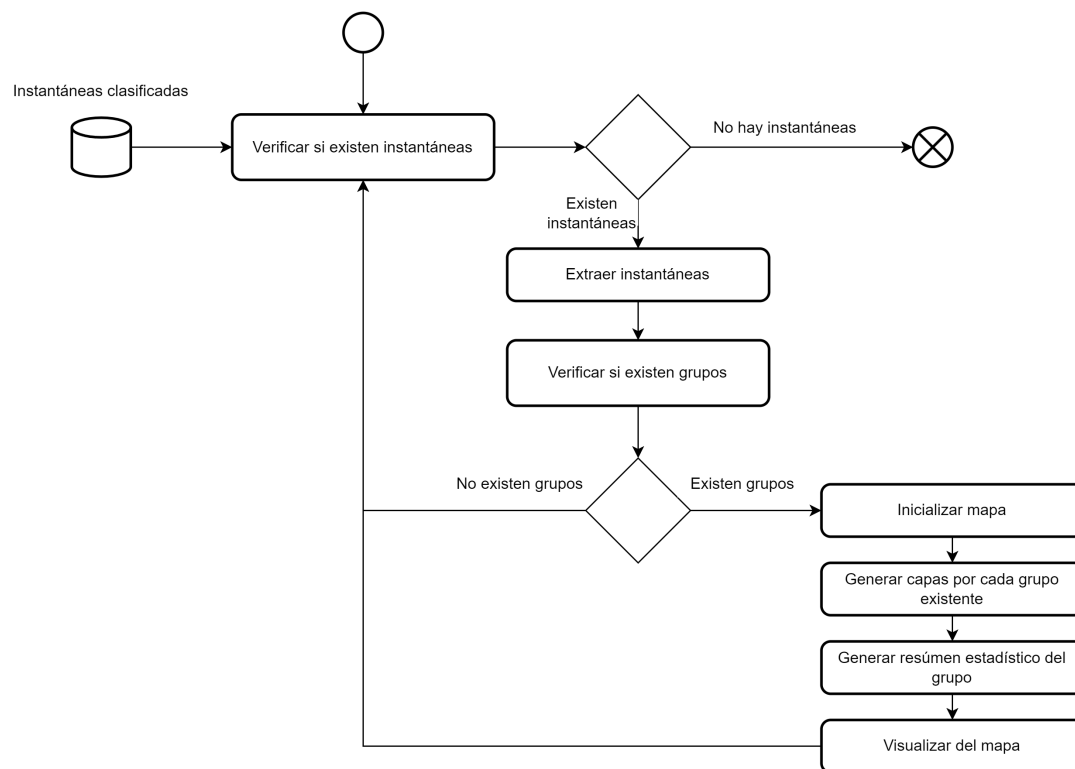


FIGURA A.4: Pasos del sub-módulo complementario 2

## Apéndice B

# Evaluación de una rejilla para la identificación de patrones de congestión del tráfico

### B.1. Introducción

La gestión eficiente del tráfico es un desafío constante en las zonas urbanas, donde el flujo de vehículos puede ser intenso y congestionado. La comprensión y el análisis precisos del flujo vehicular son esenciales para mejorar la planificación del transporte y la toma de decisiones sobre infraestructura vial. En este contexto, este artículo presenta una metodología para analizar el flujo vehicular basada en el análisis de datos de trayectoria de vehículos. El estudio del flujo vehicular ha sido abordado desde diferentes enfoques, y uno de los aspectos clave reside en la adecuada representación de los datos de la trayectoria. En este sentido, la metodología propuesta se basa en la división de un área geográfica de interés en celdas uniformes, donde cada celda contiene información detallada sobre la velocidad y el número de vehículos. Esta representación simplifica el análisis y visualización de los datos, permitiendo una comprensión más precisa del flujo vehicular en el área de estudio. Se utiliza un algoritmo de agrupamiento para realizar el análisis del flujo de tráfico. Se basa en la metodología propuesta por Reyes y otros (Reyes et al. 2021b). Este algoritmo se centra en la agrupación eficiente y precisa de celdas que contienen trayectorias de vehículos. Utilizando una métrica de distancia que considera la diferencia mínima en velocidad y número de vehículos entre los grupos y celdas que se analizan, el algoritmo agrupa las celdas de manera óptima. Esto permite identificar patrones y similitudes en el flujo vehicular de cada grupo, facilitando la detección de áreas con potenciales problemas de congestión. La evaluación de la congestión del flujo vehicular se realiza utilizando un conjunto de datos de trayectorias históricas. Las características de las celdas pertenecientes a los microgrupos resultantes del agrupamiento se analizan y comparan con valores de referencia obtenidos de celdas adyacentes. Mediante una evaluación lógica se determina si las celdas cumplen con las condiciones de congestión establecidas. Esta evaluación proporciona una visión clara de las áreas con problemas de flujo de tránsito, brindando información

relevante para las decisiones de gestión del tránsito.

## B.2. Método propuesto

Este artículo presenta una metodología de evaluación de celdas para la identificación de patrones de congestión, la metodología propuesta en este artículo se basa en la metodología presentada por Reyes y otros (Reyes et al. 2022c) utilizando un algoritmo de agrupamiento dinámico, representación de datos por celdas y visualización de resultados de agrupamiento. Sin embargo, se incorporan elementos adicionales, como una dimensión adicional para la agrupación, el uso de un conjunto de datos complementario y la evaluación de las características de congestión en los grupos. En primer lugar se realiza la definición de las celdas a analizar. El segundo paso implica el uso de un algoritmo de agrupamiento dinámico propuesto para identificar patrones de áreas que podrían presentar congestión. En el tercer paso, los patrones identificados se analizan y evalúan con un método de evaluación de características de congestión. Luego, en el cuarto paso, se procede a generar una visualización adecuada a los resultados obtenidos. Cada uno de estos pasos se describe en detalle a continuación.

### B.2.1. Definición de celdas

Dentro de la estructura de los registros a procesar se encuentran los atributos más importantes como la longitud, latitud, tiempo, velocidad y el identificador de trayectoria. El primer paso crucial en la metodología es asegurar una representación adecuada de los datos que componen las trayectorias. Para lograrlo, partimos de definir el área de interés, delimitando así la región geográfica a la que pertenecen las trayectorias a analizar. Una vez establecida la zona, procedemos a subdividirla en celdas uniformes o zonas más pequeñas. El tamaño de cada celda se determina según la precisión requerida para el análisis del flujo vehicular, permitiéndonos capturar los detalles más relevantes de los patrones de movimiento. Se han realizado varios experimentos con el objetivo de determinar el tamaño de celda apropiado para el análisis de flujo vehicular. Estos experimentos previos a la parametrización ideal han sido fundamentales para seleccionar la configuración óptima que garantice una representación adecuada de los patrones de movimiento en el área de interés. En este trabajo se utilizaron celdas de 30×30 metros y las características esenciales de estas celdas por las cuales se agruparán serán la velocidad de las celdas con rangos entre 20 km/h y el número de vehículos con rangos entre 5 vehículos. Este aspecto adquiere una relevancia fundamental, ya que implica la comprensión de la información correspondiente a cada celda durante un intervalo de tiempo específico. La metodología propuesta en este estudio consiste en analizar integralmente los eventos en cada celda en lugar de abordar las trayectorias de vehículos individuales.



Este enfoque simplifica tanto el análisis como la visualización de los datos. Específicamente, en este trabajo, los datos relacionados con el flujo vehicular, representados en cada celda, fueron sometidos a análisis utilizando un enfoque de microlotes en períodos de tiempo uniformes. La duración de cada período es un parámetro esencial del algoritmo, que debe establecerse de forma predefinida. Cada período se denomina ciclo y representa una evolución de los grupos formados, ya que los resultados de la agrupación se actualizan incorporando cada bloque de datos. Esta metodología permite capturar la dinámica de manera progresiva, a medida que se integran nuevos datos, generando una comprensión más precisa y actualizada del flujo vehicular. En cada ciclo de evolución, se almacena un flujo continuo de datos en un buffer temporal y se realizan los cálculos pertinentes para cada celda, permitiendo obtener información característica y relevante.

### **B.2.2. Agrupación dinámica de celdas**

En el segundo paso de la metodología, se utiliza un algoritmo para procesar celdas que contienen trayectorias. Este algoritmo ha sido seleccionado por su capacidad para analizar de manera eficiente y precisa las características específicas de las trayectorias en cada celda. Como se mencionó anteriormente, la primera etapa del algoritmo DyClee se utiliza para construir microgrupos. Sin embargo, en este trabajo se ha realizado una adaptación de la metodología propuesta original, donde en lugar de utilizar directamente las ubicaciones GPS y su densidad, se consideran para cada tramo las velocidades de los tramos de trayectoria y el número de vehículos presentes en cada celda. En la etapa de agrupación de las celdas en cada ciclo se utiliza una métrica de distancia que considera la mínima diferencia de velocidad y número de vehículos entre los microgrupos y la celda analizada. Si la diferencia de velocidad y de vehículos es suficientemente pequeña, la celda se incorpora al microgrupo existente y se actualiza la información correspondiente. Si no hay ningún grupo cercano, se crea un nuevo microgrupo con la celda analizada. De esta forma se consigue una agrupación adecuada de celdas en función de similitudes en velocidad y número de vehículos. En la segunda etapa del algoritmo se realiza un análisis de las densidades de los microgrupos clasificándolos en dos categorías: densos y poco densos. El número de microgrupos formados está relacionado con el valor del parámetro "tamaño relativo", que modifica directamente el tamaño de las "hipercajas". Esto se refleja en una mayor o menor amplitud del rango de asignación de celdas a los microgrupos, según su velocidad y número de vehículos. Durante estos procesos, la información relativa al desempeño de las etapas se registra por separado, permitiendo un análisis detallado de su funcionamiento.

### **B.2.3. Evaluación de la congestión**

En este estudio, es importante tener un conjunto de datos de trayectorias históricas que difieran en fecha pero que coincidan en tiempo y ubicación con el utilizado

en el proceso de agrupación. Este conjunto de datos históricos corresponde a un flujo de datos real, proporciona información complementaria y enriquece el análisis del flujo de vehículos en tránsito. Esto asegura una perspectiva más completa y precisa del comportamiento de los vehículos en las celdas de la zona de interés. Para la representación del flujo vehicular se utiliza una proyección de una rejilla similar a la utilizada en el apartado B.2.1 del artículo, dicha rejilla se aplica al conjunto de datos históricos. Luego, se realiza un análisis específico para determinar un valor de referencia de velocidad y número de vehículos. Este valor de referencia se obtiene considerando celdas adyacentes y utilizando hasta cinco niveles de distancia, estos valores de referencia históricos se calculan proporcionalmente donde cada nivel aporta una proporción específica. Como se muestra en la Figura B.1, las celdas de la periferia se determinarán con valores de proporción más bajos. Siendo el nivel 1, la celda analizada, representa el 30 % del valor histórico total, disminuyendo en porcentaje según el nivel, hasta el nivel 5, que aporta el 10 % del total. De esta forma, se utiliza la información de las celdas cercanas para establecer una base comparativa más precisa.

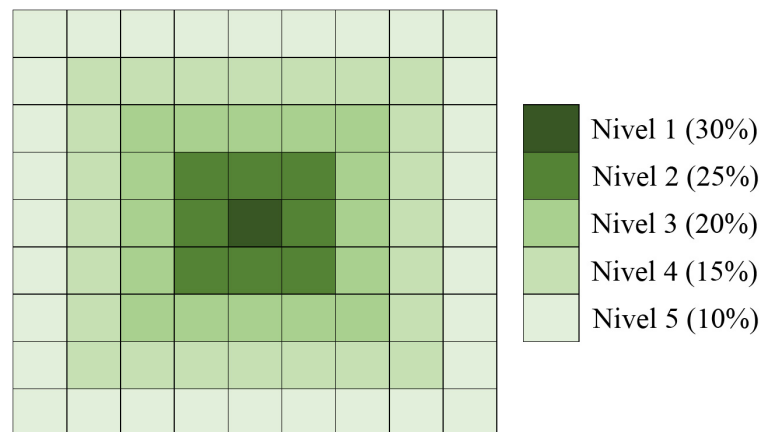


FIGURA B.1: Proporciones según el nivel de las celdas

Posteriormente, se comparan las características de las celdas de los grupos resultantes del agrupamiento con los valores de referencia obtenidos previamente. Esta comparación permite evaluar si las celdas analizadas cumplen con las condiciones de congestión establecidas. Las condiciones para la detección de congestión se basan en la evaluación propuesta por He y otros (He et al. 2023). Se aplica una evaluación lógica para determinar si las celdas analizadas cumplen con las siguientes condiciones de congestión:

- El número de puntos de ruta en la celda es mayor o igual que el número de puntos de ruta establecidos como referencia.
- La velocidad media de los vehículos dentro de la celda es inferior a la velocidad de referencia establecida.

Esta evaluación lógica permite identificar las celdas que presentan signos de congestión y contribuye a la detección temprana y precisa de zonas con potenciales problemas de circulación. Para proporcionar una visión general de las evaluaciones realizadas, el índice de congestión se calcula utilizando la fórmula propuesta por He y otros (He et al. 2023). Éste define el índice como el número total de ciclos en los que cada celda ha cumplido con las condiciones de congestión establecidas indicando que han sido identificadas como áreas de congestión.

#### B.2.4. Visualización de agrupaciones

En este estudio la información de la trayectoria se analiza en intervalos de 3 minutos, este tiempo permite identificar cambios en el flujo vehicular de una manera más precisa y detallada. Para proporcionar una representación visual interactiva de los resultados de cada grupo, se ha desarrollado un mapa interactivo. Este mapa permite analizar de forma gráfica y dinámica la información relevante de cada grupo. Cada grupo está representado en el mapa con un color diferente para cada patrón según la cantidad de vehículos y velocidades que el algoritmo de agrupamiento ha identificado, como se muestra en la Figura C.6.

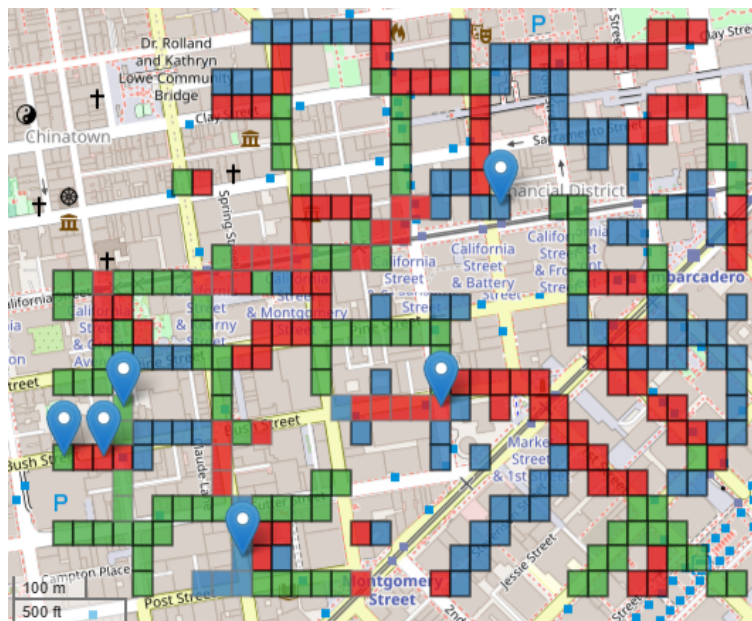


FIGURA B.2: Grupos proyectados sobre el mapa

Este estudio registra el resultado del proceso de agrupación realizado, el cual permite reconstruir todos los mapas desde el inicio del análisis de flujo vehicular. Esto proporciona una descripción general rápida del estado del tráfico.

### B.3. Resultados obtenidos

Para realizar el análisis, el conjunto de datos se dividió en ciclos que abarcaron 3 minutos de datos, lo que generó un total de 20 ciclos para la ciudad de San Francisco. Estos bloques fueron analizados consecutivamente, considerando que el período de 3 minutos es adecuado en relación al volumen de datos disponibles. A continuación, se procedió a definir los valores de los parámetros “tamaño relativo” e “hipercaja” para utilizar el algoritmo DyClee. Se fijó un valor de “tamaño relativo” de 0,2, con límites de velocidad entre 0 y 120 km/h, y límites de cantidad de vehículos entre 1 y 25 para cada celda.

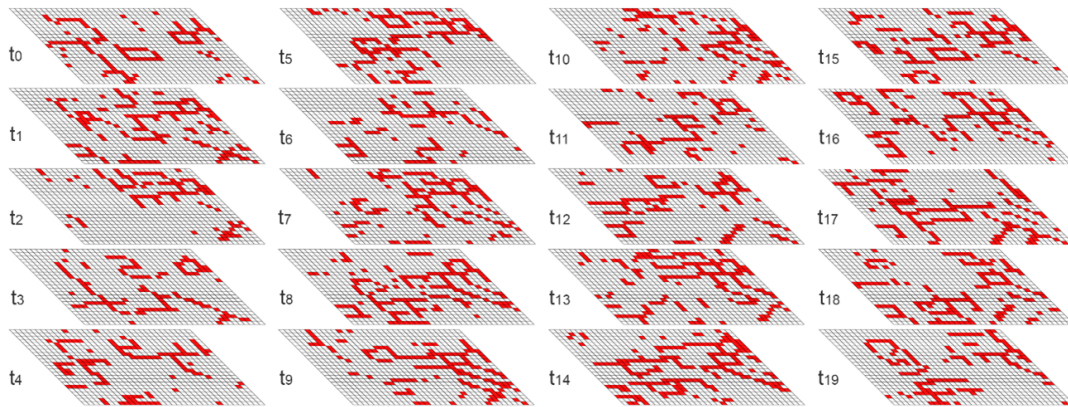


FIGURA B.3: Evaluación de celdas congestionadas usando promedios históricos

La Figura B.3 muestra la proyección de las celdas sobre un plano equivalente al mapa del área de procesamiento analizado como resultado de las evaluaciones según los valores referenciales de cada ubicación realizadas en los 20 ciclos del conjunto de datos. Cada celda congestionada se muestra en color rojo, lo que indica que estas celdas han cumplido las dos condiciones establecidas en la Sección B.2.3 que indican que las celdas han sido evaluadas como congestionadas. Cada mapa muestra las localidades donde se cumplen las condiciones de congestión, se puede observar que estas varían en el tiempo y son diferentes dependiendo de la localidad. Estas variaciones pueden variar desde concentraciones en zonas específicas como la Figura B.3( $t_2$ ); zonas con congestión dispersa como la Figura B.3( $t_{10}$ ); o zonas con congestión en áreas periféricas como la Figura B.3( $t_{18}$ ).

Estos patrones de variación en las condiciones de congestión pueden atribuirse a diversos factores, como la hora del día, eventos especiales o incluso las condiciones climáticas. También se puede observar que en determinados momentos se identifican grupos de celdas congestionadas que forman patrones espaciales específicos. Por ejemplo, en algunos ciclos se observa que se forman “puntos” de congestión en áreas específicas, como carreteras interconectadas que están muy transitadas, mientras que en otros ciclos se puede identificar una congestión más difusa y dispersa,

TABLA B.1: Tiempos de ejecución

Descripción	Agrupamiento	Evaluación	Total
Tiempo total	2 458,218	13,333	2 471,551
Tiempo promedio por ciclo	122,911	0,667	123,578
Desviación	$\pm 39,410$	$\pm 0,124$	$\pm 39,534$

lo que indica que la congestión puede estar ocurriendo debido a situaciones particulares. Estos patrones espaciales pueden estar relacionados con la infraestructura vial, la distribución de puntos de interés o incluso la densidad de población en diferentes áreas. Es importante señalar que no todas las celdas evaluadas presentan congestión en todos los ciclos. Algunas celdas pueden alternar entre condiciones congestionadas y no congestionadas a lo largo del tiempo, lo que indica la dinámica y variabilidad relacionadas con el flujo vehicular. Los experimentos y análisis de la metodología propuesta se realizaron utilizando una computadora con las siguientes especificaciones técnicas: procesador AMD A9-9425 3.1 GHz; 8 GB de RAM DDR4; y sistema operativo Windows 10 Home.

Los resultados obtenidos sobre el desempeño del algoritmo se muestran en la Tabla B.1, esta tabla contiene un resumen de los tiempos de ejecución en segundos de la metodología propuesta. El tiempo de agrupamiento utilizó 2458,218 segundos y la evaluación de las celdas tomó 13,333 segundos, lo que suma un tiempo total de 2471,551 segundos. Analizando el tiempo promedio por ciclo, se observa que el agrupamiento requiere en promedio 122,911 segundos con una desviación de  $\pm 39,410$  segundos, mientras que la evaluación solo necesita 0,667 segundos con una desviación de  $\pm 0,124$  segundos. Estos datos indican que el proceso de agrupación requiere más tiempo en comparación con la evaluación. En promedio, cada ciclo del conjunto de datos tarda aproximadamente 123,578 segundos en completarse. Estos resultados reflejan la efectividad de la metodología en términos de tiempos de ejecución.

Para proporcionar una visión general de las evaluaciones realizadas y representar cuantitativamente la congestión vehicular en el área de procesamiento analizada, se calcula el índice de congestión que proporciona una medida única para cada celda y se basa en la acumulación de las celdas identificadas como congestionadas a lo largo del 20 ciclos del conjunto de datos. El resultado de este cálculo proporciona un valor numérico que representa el nivel de congestión en cada celda. Cuanto mayor sea el valor del índice de congestión, mayor será la frecuencia de congestión vehicular en ese lugar específico. Este enfoque proporciona una perspectiva global de la congestión en el área de procesamiento y facilita la identificación de áreas críticas que requieren atención y posibles soluciones. Para la visualización de estos índices se traza una distribución de la congestión mostrando una representación espacial de las áreas con mayor o menor frecuencia de congestión, lo que ayuda a comprender la distribución geográfica de este fenómeno. Esta distribución se muestra en la Figura B.4. Para comprender la relación entre densidad y clasificación de la congestión,

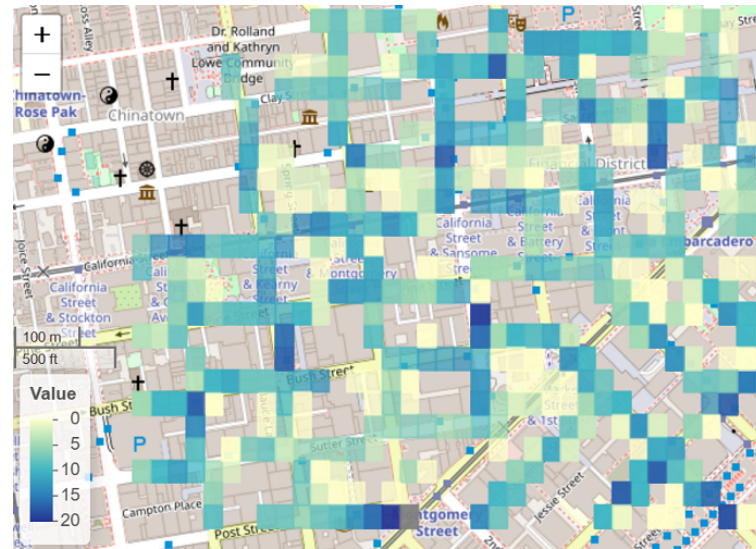


FIGURA B.4: Distribución de la congestión por promedios históricos

TABLA B.2: Matriz de contingencia basada en la coincidencia de resultados de la evaluación del flujo

Categorías	Congestionadas	No congestionadas
Poco densos	60,50 %	57,60 %
Densos	22,31 %	60,68 %

se ha utilizado una matriz de contingencia que se muestra en la Tabla B.2. El análisis de esta matriz de contingencia revela tendencias en la precisión de las clasificaciones de congestión en relación con la densidad de los conglomerados resultantes.

Estos valores representan el porcentaje de celdas clasificadas correctamente en cada combinación de categorías. En la categoría de grupos dispersos, resultantes de la agrupación, el 60,50 % de las celdas congestionadas coinciden con las celdas clasificadas como congestionadas por el flujo real de celdas, y el 57,60 % de las celdas no congestionadas coinciden con las celdas clasificadas como no congestionadas por el flujo real. Asimismo, en la categoría de grupo denso, el 22,31 % de las celdas congestionadas coinciden con celdas clasificadas como congestionadas del flujo real de celdas, y el 60,68 % de las celdas no congestionadas coinciden con celdas clasificadas como no congestionadas del flujo real. Se observa que la precisión de las clasificaciones varía según la categoría de densidad resultante del agrupamiento. Esto indica que el resultado de la clasificación funciona mejor para identificar la congestión en grupos con celdas de baja densidad. Al comprender la relación entre la densidad del tráfico y las clasificaciones de la congestión, se pueden desarrollar estrategias de clasificación más efectivas para una identificación precisa de la congestión vehicular en entornos de tráfico variables. Para mejorar las tasas de coincidencia de detección de congestión utilizando la metodología propuesta, es crucial considerar varios factores. En primer lugar, se debe garantizar la exactitud de los datos de tráfico obteniendo información actualizada y de alta calidad de fuentes confiables. Además, los

parámetros del algoritmo de agrupamiento y los umbrales de congestión deben revisarse y ajustarse para adaptarse a las características específicas de la ciudad o área de estudio. La inclusión de factores externos, como eventos especiales o cambios climáticos, también es relevante para tener una comprensión completa de la congestión y puede mejorar significativamente la precisión de la detección de la congestión de vehículos. La metodología desarrollada en este estudio presenta similitudes con la propuesta por He y otros (He et al. 2023) para procesos de detección de eventos de congestión de tráfico. Aunque existen diferencias en los detalles de implementación y objetivos específicos, ambas metodologías comparten una base conceptual común para abordar la identificación de la congestión vehicular. Esta similitud permite la posibilidad de realizar comparaciones parciales entre ambas metodologías y enriquecer el conocimiento general sobre la detección de congestiones de tráfico en entornos dinámicos. Es importante reconocer que la metodología propuesta también enfrenta ciertas limitaciones y desafíos potenciales que deben tenerse en cuenta al interpretar los resultados. Dado que las condiciones de congestión pueden variar ampliamente según la hora del día, eventos especiales o incluso las condiciones climáticas, esta metodología puede generar resultados diferentes en diferentes momentos. Un análisis cuidadoso de los resultados en diferentes contextos temporales es esencial para comprender la estabilidad de los patrones identificados. Además, aunque la metodología ha demostrado ser eficaz en la ciudad de San Francisco, los resultados deberían validarse en otras zonas urbanas con características diferentes. Cada ciudad puede tener infraestructuras viales únicas, patrones de tráfico distintos y dinámicas de congestión particulares. Esto puede requerir ajustes y adaptaciones específicas para obtener resultados precisos y significativos. Otra posible limitación es la necesidad de datos de alta calidad. El éxito de la metodología depende en gran medida de la disponibilidad de datos precisos sobre el tráfico vehicular. Si los datos son insuficientes o de baja calidad, esto podría afectar la precisión y confiabilidad de las evaluaciones de la congestión.

## B.4. Conclusiones

La distribución de la congestión vehicular en el área de estudio ha sido claramente identificada y visualizada, lo que permite una mejor comprensión de los patrones espaciales y temporales de la congestión vehicular. Se ha demostrado la eficacia de la metodología propuesta para evaluar y monitorear de manera precisa y sistemática la congestión vehicular. Los resultados obtenidos han proporcionado información valiosa sobre las zonas y momentos de mayor congestión, que puede contribuir a la toma de decisiones informadas en la planificación y gestión del tráfico urbano. Se ha observado una variabilidad significativa en las condiciones de congestión a lo largo del tiempo, lo que destaca la importancia de considerar la estacionalidad y la dinámica del flujo vehicular en las estrategias de gestión del tráfico. La visualización de los resultados ha permitido identificar patrones específicos de congestión en

áreas específicas, que pueden guiar acciones encaminadas a mejorar la infraestructura vial o implementar medidas específicas de control del tráfico en esas ubicaciones. La metodología ha demostrado ser una herramienta eficaz para la identificación y seguimiento de zonas congestionadas, que puede contribuir a la optimización de los recursos y la eficiencia en la gestión del transporte urbano. Los resultados obtenidos apoyan la necesidad de implementar estrategias de control de tráfico y planificación urbana que tengan en cuenta la variabilidad y dinámica de la congestión vehicular, con el objetivo de mejorar la fluidez y reducir los tiempos de viaje en el área de estudio. Como líneas de trabajo futuro se propone mejorar el modelo de clasificación para mejorar la precisión en la clasificación de celdas congestionadas y no congestionadas. Otra propuesta es considerar variables adicionales como la velocidad del flujo de vehículos, la capacidad de la vía o el tipo de vehículos para obtener una comprensión más completa de la congestión. Evaluar la congestión de vehículos a lo largo del tiempo para desarrollar modelos predictivos que identifiquen patrones y tendencias a largo plazo. Además, se propone la validación de la nueva metodología en otros contextos espaciales y temporales para verificar su aplicabilidad y generalización. Esto permitiría utilizar la nueva metodología para mejorar la gestión del tráfico en una variedad de entornos.



## Apéndice C

# Metodología para la Identificación de Congestionamientos Vehiculares Basada en Agrupación Dinámica

### C.1. Introducción

La movilidad sostenible ha surgido en respuesta a los desafíos ambientales y sociales asociados con el crecimiento urbano y el aumento del tráfico vehicular. Este paradigma busca transformar los modos de viaje, promoviendo alternativas que reduzcan las emisiones de gases de efecto invernadero y minimicen el impacto en los ecosistemas. Los esfuerzos de investigación se han centrado en una variedad de frentes, desde el desarrollo de tecnologías de vehículos más eficientes y limpios hasta la planificación urbana orientada a la movilidad. La investigación sobre movilidad sostenible se ha convertido en un campo interdisciplinario y en constante evolución, impulsado por la urgente necesidad de encontrar soluciones viables y sostenibles a las crecientes demandas de transporte en las ciudades.

Las áreas de investigación más cruciales, que revelan un panorama diverso y complejo, han surgido como una convergencia crucial con los Sistemas de Transporte Inteligentes (ITS), marcando una transición hacia soluciones más avanzadas y efectivas. Estos sistemas, respaldados por tecnologías innovadoras como el análisis de datos en tiempo real y la inteligencia artificial, ofrecen oportunidades sin precedentes para mejorar la gestión del tráfico, facilitar la planificación urbana y fomentar la adopción de modos de transporte sostenibles. La investigación y análisis de Sistemas Inteligentes de Transporte en áreas urbanas se han vuelto imprescindibles en la actualidad debido a la complejidad de este problema y su profundo impacto en la sociedad. El crecimiento constante de la población en las zonas urbanas, así como el aumento del tráfico vehicular, son factores obvios que requieren una atención cuidadosa (Soumia Goumiri y Djahel 2023).

Los Sistemas de Transporte Inteligentes surgen como una respuesta innovadora y tecnológica para abordar estos desafíos de manera eficiente y sostenible. En su búsqueda de soluciones efectivas a los desafíos de la movilidad urbana, emplean una

variedad de técnicas de aprendizaje automático para obtener aplicaciones prácticas y ofrecer enfoques analíticos en el campo del transporte.

Los sistemas de transporte inteligentes utilizan algoritmos de aprendizaje automático para detectar patrones en el comportamiento de los vehículos, como la congestión habitual en determinadas zonas o las rutas preferidas de los conductores. Esta información es esencial para la predicción de la congestión, la planificación óptima de rutas y la adaptación en tiempo real de las estrategias de gestión del tráfico.

La aplicabilidad de estos enfoques es amplia y abarca desde la gestión del tráfico en tiempo real hasta la planificación a largo plazo de la infraestructura de transporte. Al comprender mejor los patrones de tráfico y los comportamientos de los conductores, los sistemas inteligentes pueden ofrecer soluciones más efectivas, como la optimización de los semáforos, la gestión de rutas de transporte público y la implementación de políticas de movilidad sostenible.

Para el análisis de los datos de trayectoria de los vehículos se pueden observar diversas aproximaciones, siendo las más destacadas la aplicación de un algoritmo de agrupamiento dinámico y la evaluación de la congestión del tráfico mediante un indicador que se adapta a diferentes zonas. La combinación de estas perspectivas se presenta como un área de investigación potencial y sugerente, que ofrece una comprensión integral para abordar el flujo vehicular. La combinación de evaluación de la congestión y la aplicación de algoritmos de agrupamiento puede ser un área de estudio valiosa para mejorar la eficiencia y la planificación en entornos urbanos.

Este artículo propone una metodología para analizar el flujo vehicular agrupando datos de trayectoria de vehículos con puntos GPS. Esta metodología permite una representación precisa de los datos, especialmente útil cuando los puntos son escasos. Utiliza grupos para detectar áreas de patrones de congestión. La actualización constante de los grupos garantiza datos actualizados y una gestión real de la congestión. También utiliza un indicador de congestión para medir la saturación del tráfico, lo que permite una visión dinámica de la situación del tráfico en diferentes áreas.

## **C.2. Materiales y métodos**

Este artículo presenta una metodología para la identificación de zonas de congestión basada principalmente en agrupamiento dinámico. La metodología utilizada consta de cinco pasos ilustrados en la Figura C.1. En el primer paso, se carga la información de datos de la red de carreteras. En el segundo paso, se procesa un flujo de datos de trayectorias. En el tercer paso, se utiliza un algoritmo de agrupamiento dinámico basado en la distancia para identificar áreas con patrones similares. En el cuarto paso, las áreas se evalúan con un indicador de congestión para su posterior clasificación. En el quinto paso se procede a generar una visualización adecuada de los grupos resultantes ya clasificados. Cada uno de estos pasos se describe en detalle a continuación.

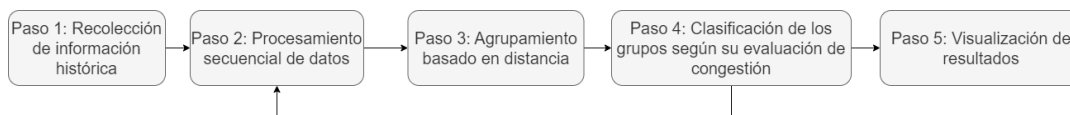


FIGURA C.1: Pasos de la metodología propuesta.

### C.2.1. Paso 1: Cargar Información de la Red Vial

La estimación de datos de referencia es importante para asegurar la confiabilidad de los resultados y para fortalecer la validez y consistencia de los análisis. realizado. El objetivo principal de este paso es cargar en memoria la información relevante sobre la infraestructura vial de la zona a analizar. Esto incluye datos detallados sobre el trazado de la carretera, ubicación geográfica, capacidad, número de carriles y límites de velocidad. Estos datos no sólo se utilizarán para el análisis, sino que también permitirán una comparación efectiva entre diferentes tramos de carretera, estableciendo así un punto de referencia sólido para detectar y evaluar con precisión la congestión vial.

La solidez de los resultados depende de la calidad y la exhaustividad de los datos recopilados en este paso. La precisión en la representación de la red de carreteras y la minuciosidad en la recopilación de datos son esenciales para garantizar resultados sólidos y confiables en los pasos posteriores de este estudio.

Se recomienda implementar un proceso de actualización mensual para garantizar la actualidad y precisión de los datos de la red de carreteras en entornos urbanos cambiantes. Para un cambio de sentido de una vía que modifica la dinámica del flujo vehicular, por ejemplo, dependiendo de la infraestructura tecnológica en la fuente que almacena la información sobre las redes viales, la actualización de cambios en las fuentes podría demorar varios días.

### C.2.2. Paso 2: Procesamiento de datos

En este paso, se establece un método para recibir y procesar puntos GPS en tiempo real o desde un repositorio accesible. Estos datos provienen de vehículos equipados con GPS o aplicaciones móviles en vehículos. El procesamiento se realiza en microlotes en intervalos de tiempo uniformemente distribuidos, llamados "ciclos", que representan momentos en la evolución de los flujos de datos.

En cada ciclo, los datos se acumulan en un búfer temporal y los resultados del método de agrupación se actualizan a medida que se agregan nuevos datos. Para abordar la falta de información en los datos de trayectorias, se implementa un método de enrutamiento y un método de interpolación que operan simultáneamente durante la recopilación de nuevas ubicaciones GPS.

La decisión de realizar enrutamiento e interpolación junto con la acumulación de datos se justifica por su capacidad para generar un búfer de datos enriquecido y densamente poblado. Esto garantiza que a medida que avanza el flujo de datos,

funcione con información de ruta completa y precisa para cada vehículo, mejorando la eficiencia del análisis de agrupación posterior.

Una de las funciones subyacentes de este paso es mitigar los desafíos asociados con la presencia de datos GPS afectados por el ruido y la información incompleta. Routing se propone como una estrategia eficaz para tratar datos ruidosos, contribuyendo a la limpieza y mejora de la calidad de los datos de trayectoria de los vehículos. La interpolación se presenta como una herramienta clave para estimar los valores faltantes, lo que permite una representación detallada de la información de ubicación del vehículo observada. Estas estrategias se implementan con el propósito de optimizar la precisión y coherencia de los datos de la red de carreteras utilizados en el proceso de agrupación. El método de enrutamiento utiliza el servicio Open Source Routing Machine (OSRM), basado en jerarquías de contracción y optimización del tiempo de viaje. Este servicio calcula la ruta más corta y genera puntos espaciales clave en función de la geometría de las redes viarias, proyectando una ruta artificial sobre las vías por las que podría haber circulado el vehículo.

El método de interpolación basado en distancia consiste en estimar continuamente valores de datos a lo largo de una ruta con un intervalo constante de 5 m. Estas técnicas se aplican para completar rutas en los casos en los que no existen registros para determinados tramos, proporcionando una representación más detallada y uniforme de los datos, especialmente útil cuando los puntos de recogida están dispersos o incompletos.

El método de interpolación basado en distancia consiste en estimar continuamente valores de datos a lo largo de una ruta con un intervalo constante de 5 m. Estas técnicas se aplican para completar rutas en los casos en los que no existen registros para determinados tramos, proporcionando una representación más detallada y uniforme de los datos, especialmente útil cuando los puntos de recogida están dispersos o incompletos.

Las rutas y la interpolación generan un conjunto de datos más denso y continuo, lo que permite un análisis preciso y una visualización perfecta del comportamiento del vehículo a lo largo de la ruta. Esta información interpolada es valiosa para identificar y analizar tendencias de vehículos individuales, como se ilustra en la Figura C.2.

### C.2.3. Paso 3: agrupación basada en la distancia

Un grupo se compone principalmente de un centroide, un hipercuadro y puntos GPS vinculados. El centroide, que es el punto geográfico que representa el centro del conglomerado, se utiliza como referencia representativa del conglomerado en el análisis. El hiperbox es una estructura rectangular; se posiciona de acuerdo con el centroide del grupo. Esta forma rectangular delimita un área alrededor del centroide, lo que simplifica la representación espacial y determina su área de influencia. La representación visual de un grupo se presenta en la Figura C.3.

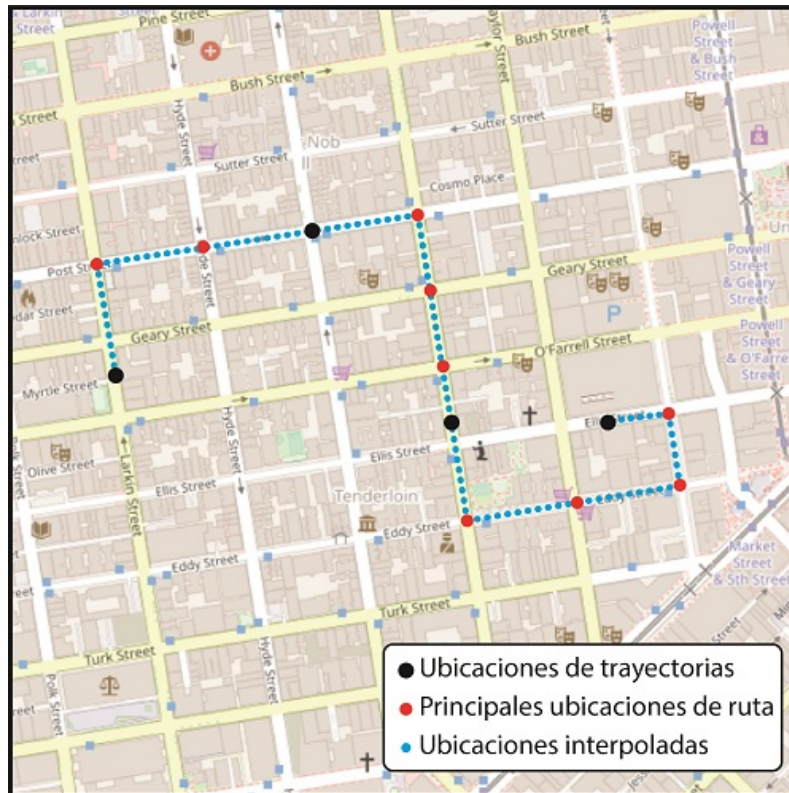


FIGURA C.2: Enrutamiento e interpolación de una trayectoria.

Cada punto GPS procesa información de ubicación geográfica, identificación del vehículo y hora de ingreso. La agrupación se realiza mediante similitud basada en la distancia euclidiana, considerando los atributos de latitud y longitud de los puntos GPS. Cada punto GPS se analiza calculando la distancia euclidiana con los centroides de los grupos existentes. Cada punto se asigna al grupo con la distancia espacial más pequeña y dentro del área del hipercuadro.

En el caso de que no esté en el área del hipercuadro, se crea un nuevo grupo. El centroide se actualiza cuando se integran nuevos puntos GPS en un grupo y se crean nuevos grupos si no hay grupos cercanos.

Para garantizar que los grupos se actualicen y evitar la retención de datos antiguos, se utiliza un método de olvido basado en el tiempo de entrada del último punto GPS para determinar la pérdida de relevancia a medida que transcurre el tiempo y se calcula mediante la Ecuación ??.

$$F = e^{-1 \times \lambda \times \delta t} \quad (C.1)$$

donde,  $e$  representa la función exponencial,  $\lambda$  controla la velocidad de pérdida de relevancia y  $\delta t$  es la diferencia entre el tiempo del punto analizado y el tiempo del último punto integrado al grupo.

Se establece un umbral de relevancia del 5 % para determinar cuándo un valor deja de ser relevante para el análisis. Este método determina el número de puntos

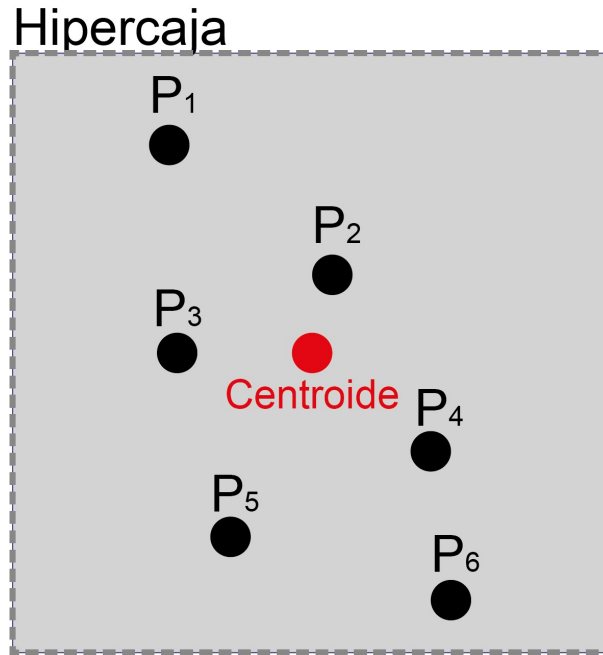


FIGURA C.3: Elementos que conforman un grupo.

GPS que permanecerán en el grupo durante el periodo de retención, adaptando así el grupo a los cambios de tráfico y evitando la acumulación de datos obsoletos. Los grupos que pierden relevancia por falta de nuevos puntos GPS se eliminan, mientras que los grupos activos que siguen recibiendo datos se mantienen actualizados.

#### C.2.4. Paso 4: Clasificación de los grupos según su evaluación de congestión

Una vez asignados a un grupo todos los puntos del buffer temporal, cada grupo formado se evalúa mediante un indicador que analiza el estado del tráfico en función de los resultados del agrupamiento en ese momento. El búfer temporal se puede utilizar además para procesar un nuevo flujo de datos con su respectiva agrupación. Este paso nos permite analizar y clasificar cada grupo individualmente según su nivel de congestión, lo que nos ayudará a identificar áreas problemáticas y áreas con mejor flujo de tráfico.

Cada grupo es examinado individualmente para comprender su comportamiento y características particulares; estas características, reflejadas a través de mediciones estadísticas implícitas en cada tramo vial que se encuentra dentro del área del grupo, han estado presentes desde su formación inicial y se actualizan cada vez que el grupo presenta cambios; a partir de la información de los segmentos podemos obtener la cantidad de puntos GPS, la velocidad promedio de los vehículos (unidad medida en kilómetros por hora, km/h) y la cantidad de vehículos, entre otros. Se proyecta espacialmente un hipercuadro para cada grupo para delimitar su respectiva área de análisis. Los datos cartográficos se utilizan para identificar las carreteras

y los segmentos de carreteras contenidos dentro del área delimitada de cada grupo. Se aplica un recorte espacial a las carreteras para que quepan dentro del área definida por el hipercuadro para cada grupo; esto permite aislar los tramos de carretera relevantes que influyen en cada grupo específico. Un ejemplo de este recorte se puede ver en la Figura C.4. Además, se utilizará información asociada a la geometría y metadatos de cada vía.

Para la clasificación del estado de congestión de los grupos se utiliza un Indicador de Coeficiente de Tráfico como indicador de congestión. El indicador de congestión utilizado indica un valor que refleja el nivel de congestión en una ubicación o vía (**nguyen\_ficient\_2018**), midiendo cuantitativamente la congestión en función de la densidad de vehículos y sus velocidades. Un valor alto de este indicador de congestión indica una congestión significativa, mientras que un valor bajo del indicador de congestión sugiere un tráfico fluido.

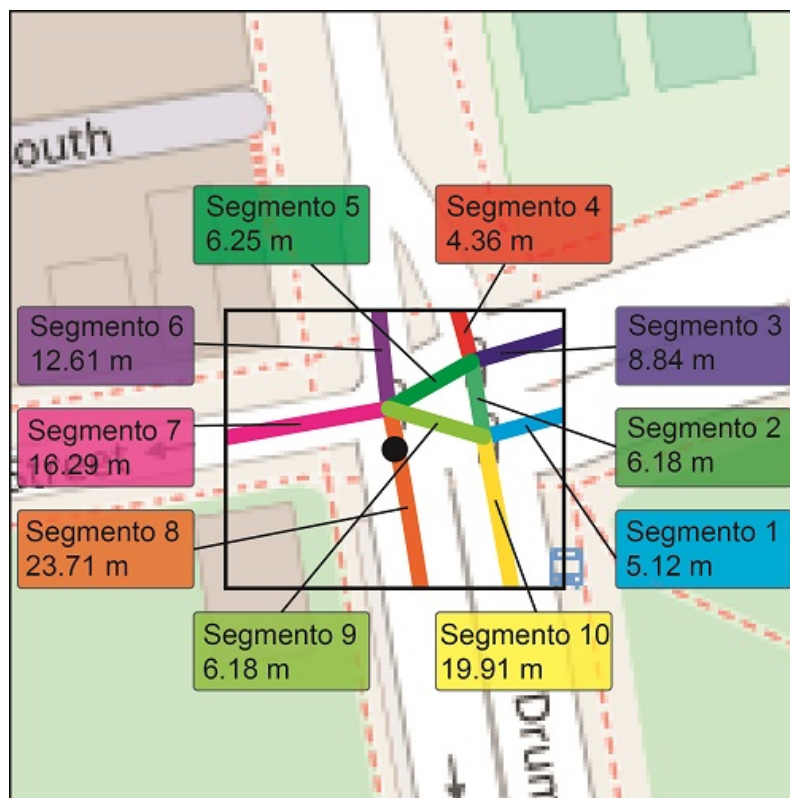


FIGURA C.4: Segmentos de carretera analizadas en el área de un grupo.

El Indicador de Coeficiente de Tráfico, utilizado para evaluar la congestión, se basa de forma teórica y experimental en un artículo anterior (**nguyen\_ficient\_2018**). La validación de este indicador se realizó mediante experimentación sometiendo el modelo a diversas condiciones y escenarios. Algunas limitaciones que se observaron son la dependencia de la disponibilidad y exactitud de la información de las

redes viales; además, durante los primeros ciclos de ejecución dentro de nuestra metodología, se observó una disminución en la precisión del indicador, fenómeno que atribuimos a la adaptación del modelo a las particularidades específicas de los datos en cada ejecución. El cálculo de este indicador de congestión se realiza mediante tramos de vía. Dado que un hipercuadro de un grupo puede comprender varios segmentos, el indicador de congestión se calcula individualmente para cada segmento. Luego, se genera un valor unificado en función de la longitud de los segmentos con al menos un vehículo para que cada grupo tenga su propio valor generado por el indicador de congestión.

El valor del indicador de congestión se calcula mediante la relación entre el Índice de Densidad y el Índice de Velocidad.

El índice de densidad representa la cantidad de vehículos en un segmento de carretera en un momento específico. Se calcula dividiendo el número de vehículos observados en la zona por el número máximo registrado previamente en su respectivo tramo vial. Este monto máximo se basa en datos históricos o estudios de tráfico realizados previamente. En esta investigación se establece un procedimiento sistemático para la determinación dinámica del valor máximo de densidad de tráfico. El procedimiento comienza identificando los tramos viales en el área de estudio, como se muestra en la Figura C.5, y calculando la densidad de tránsito de cada uno. Estas densidades se convierten en valores de densidad por unidad de longitud ( $D/L$ ), considerando las diferentes longitudes de los segmentos.



FIGURA C.5: Tramos identificados en una región de la red vial. Los segmentos con diferentes colores representan el flujo vehicular y los segmentos negros no muestran ningún flujo vehicular.

Luego, se calcula la proporción de cada segmento en función de su longitud con



respecto al número total de segmentos recorridos. Los valores ponderados de densidad por unidad de longitud se obtienen multiplicando el  $D/L$  con las proporciones de cada segmento y sumándolos para obtener una medida representativa del conglomerado considerado.

Luego se utilizan y promedian las densidades de todos los conglomerados en el ciclo para obtener un valor general para ese ciclo. Este valor se suma a un registro histórico que se actualiza en cada ciclo. Con este valor histórico se puede estimar la densidad máxima de tráfico de manera generalizada para diferentes longitudes de tramos de vía multiplicando el valor histórico por la longitud de la vía bajo análisis. La inclusión de múltiples longitudes de segmentos de carretera en el registro histórico garantiza precisión y confiabilidad al determinar la densidad máxima de tráfico, independientemente de la longitud del segmento bajo análisis.

Cuando el Índice de Densidad se acerca o llega a 1, indica que el número de vehículos en esa zona se acerca o ha superado la capacidad máxima observada. Esto sugiere una alta probabilidad de congestión.

El Índice de Velocidad refleja la velocidad media de los vehículos en las vías analizadas. Se calcula dividiendo la velocidad media del vehículo por el límite de velocidad establecido por las normas de tráfico locales. Estas normas se determinan de acuerdo a la normativa de cada ciudad, con el fin de asegurar un adecuado flujo de tránsito.

Se calcula dividiendo la velocidad promedio de los vehículos observados en cada segmento por la velocidad máxima permitida en el segmento de carretera respectivo. Cuando el índice de velocidad es cercano o igual a 1, los vehículos viajan a la velocidad máxima permitida, lo que indica un flujo de tráfico fluido y baja congestión. Por el contrario, una velocidad más baja indica un flujo de tráfico más lento, lo que podría indicar la presencia de congestión.

### C.2.5. Paso 5: Visualización de resultados

En este estudio, la información de la trayectoria se examina a intervalos regulares, lo que permite la detección precisa de cambios en el flujo vehicular.

Para brindar una representación visual e interactiva de los resultados de cada grupo, se desarrolla un mapa interactivo que se puede generar en cualquier ciclo. Este mapa permite el análisis dinámico y gráfico de la información relevante para cada grupo. Cada área con características similares está representada con un color diferente en el mapa, como se ilustra en Figura C.6.

Para su aplicabilidad en los sistemas de gestión de tráfico, proponemos una estrategia gradual que priorice la interoperabilidad con los sistemas existentes. Proponemos el desarrollo de una interfaz que facilite una integración perfecta, permitiendo una transición fluida a la implementación. Además, sugerimos una evaluación

exhaustiva de la usabilidad mediante pruebas piloto en entornos urbanos específicos, abordando desafíos como la adaptación a patrones de tráfico complejos, la gestión eficiente de grandes volúmenes de datos en tiempo real y la interacción efectiva con los operadores de tráfico. Destacamos la importancia de una estrecha colaboración con las autoridades de tránsito y los profesionales en el campo para obtener comentarios valiosos sobre la integración de la herramienta en las prácticas cotidianas.

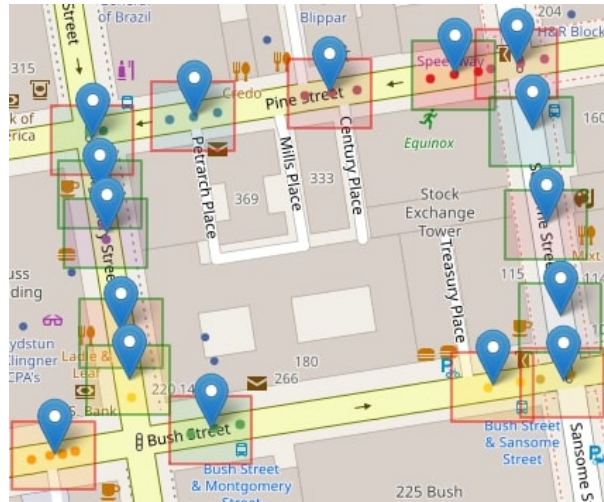


FIGURA C.6: Grupos proyectados en una parte del mapa.

### C.3. Resultados obtenidos

En esta investigación, se compararon los resultados de la metodología de agrupamiento dinámico con los resultados del indicador de congestión aplicado a celdas estáticas para analizar el flujo vehicular utilizando el Indicador de Coeficiente de Tráfico para medir la congestión, como se muestra en la Figura C.7.

En la metodología de agrupamiento dinámico, los datos del vehículo se agrupan en patrones similares y la ubicación del grupo determinó el área de comparación, mientras que en las celdas estáticas, el área se dividió en celdas uniformes. Ambos casos se evalúan utilizando el indicador de congestión.

Se aplicó una tolerancia a los valores de los indicadores en grupos dinámicos para tener en cuenta la variabilidad natural de los datos y evitar la identificación errónea de posibles estados de congestión. La tolerancia determina qué tan grande es el margen de error permitido al ajustar el indicador de congestión, influyendo directamente en el número de coincidencias observadas en la clasificación de congestión. Luego se compararon los resultados de la clasificación congestionada entre celdas y grupos, registrando coincidencias válidas cuando al menos una celda coincidía con la misma clasificación de congestión que el grupo. Para determinar si la clasificación de conglomerados se realizó correctamente, los resultados del modelo

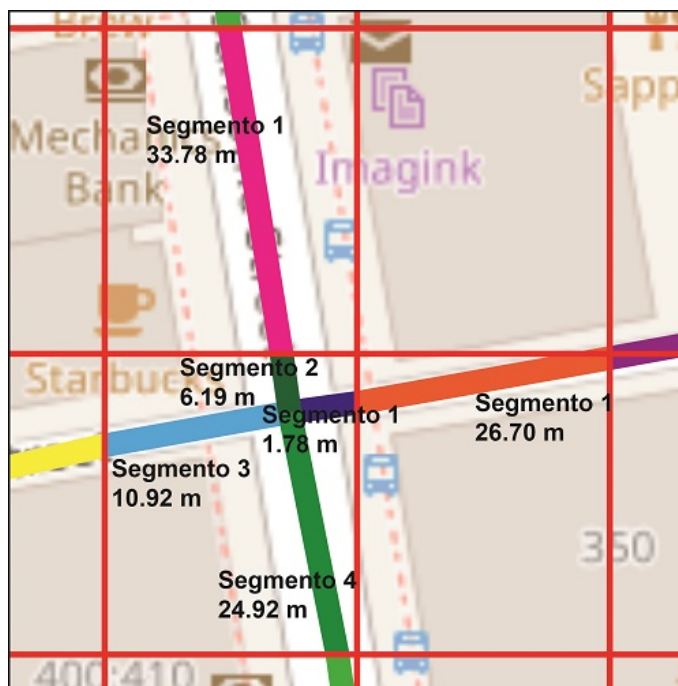


FIGURA C.7: Grupo proyectado en la rejilla.

que utiliza matrices de confusión se comparan con los resultados de la ciudad de San Francisco. Las siguientes matrices de confusión presentan una representación de las capacidades del modelo de clasificación de grupos dinámicos en relación con el modelo de cuadrícula estática. Los valores reflejan el número de predicciones precisas y erróneas en las categorías congestionadas y no congestionadas. Estos resultados proporcionan información detallada sobre cómo el modelo interpreta y predice el estado de congestión en los grupos, lo cual es esencial para evaluar su desempeño en situaciones de tráfico reales.

La matriz de confusión para la ciudad de San Francisco, con un valor de olvido de 45 s y sin tolerancia, se presenta en la Tabla C.1. En los casos de congestión se clasificaron correctamente 6 433 y de los casos no congestionados se clasificaron correctamente 10 932. Sin embargo, se cometieron 1 273 errores al clasificar erróneamente situaciones no congestionadas como congestión, y 6 307 errores al clasificar situaciones congestionadas como no congestionadas.

TABLA C.1: Matriz de confusión para la ciudad de San Francisco utilizando 45 segundos de olvido y sin tolerancia.

	Celdas congestionadas	Celdas no congestionadas
Grupos congestionados	6 433	1 273
Grupos no congestionados	6 307	10 932

La matriz de confusión para la ciudad de San Francisco, con un valor de olvido de 45 s y una tolerancia de 0,2, se presenta en la Tabla C.2. En el caso de los grupos congestionados, 6 881 casos se clasificaron correctamente como congestión, y en el caso de los grupos no congestionados, 12 387 casos se clasificaron correctamente como no congestionados. Sin embargo, se cometieron 825 errores al clasificar erróneamente situaciones no congestionadas como congestión, y 4 852 errores al clasificar situaciones congestionadas como no congestionadas.

TABLA C.2: Matriz de confusión para la ciudad de San Francisco usando 45 segundos de olvido y tolerancia de 0,2.

	<b>Celdas congestionadas</b>	<b>Celdas no congestionadas</b>
Grupos congestionados	6 881	825
Grupos no congestionados	4 852	12 387

La matriz de confusión para la ciudad de San Francisco, con un valor de olvido de 60 s y sin tolerancia, se presenta en la Tabla C.3. Para los conglomerados congestionados, 6 708 casos se clasificaron correctamente como congestión, y para los no congestionados, 10 776 casos se clasificaron correctamente como no congestionados. Sin embargo, se cometieron 1 390 errores al clasificar erróneamente situaciones no congestionadas como congestión, y 6 293 errores al clasificar situaciones congestionadas como no congestionadas.

TABLA C.3: Matriz de confusión para la ciudad de San Francisco utilizando 60 segundos de olvido y sin tolerancia.

	<b>Celdas congestionadas</b>	<b>Celdas no congestionadas</b>
Grupos congestionados	6 708	1 390
Grupos no congestionados	6 293	10 776

La matriz de confusión para la ciudad de San Francisco, con un valor de olvido de 60 s y una tolerancia de 0,2, se presenta en la Tabla C.4. Para los conglomerados congestionados, 7 177 casos se clasificaron correctamente como congestión, y para los no congestionados, 12 232 casos se clasificaron correctamente como no congestionados. Sin embargo, se cometieron 921 errores al clasificar erróneamente situaciones no congestionadas como congestión, y 4 837 errores al clasificar situaciones congestionadas como no congestionadas.

TABLA C.4: Matriz de confusión para la ciudad de San Francisco utilizando 60 segundos de olvido y tolerancia de 0,2.

	Celdas congestionadas	Celdas no congestionadas
Grupos congestionados	7 177	921
Grupos no congestionados	4 837	12 232

Estos resultados indican que, en comparación con parametrizaciones anteriores con un valor de olvido de 45 s, el modelo presenta cantidades ligeramente mejores en la identificación de situaciones de congestión. Al evaluar la tasa de verdaderos positivos, es decir, la capacidad de identificar correctamente el estado de congestión del tráfico, los grupos obtuvieron un alto número de coincidencias en comparación con las celdas de la cuadrícula congestionada en la ciudad de San Francisco.

En esta sección examinamos los resultados obtenidos de la comparación entre la metodología de agrupamiento dinámico y el indicador de congestión aplicado a celdas estáticas en la ciudad de San Francisco. La métrica de evaluación está representada por las tasas de precisión que proporcionan una comprensión profunda del desempeño de ambos resultados.

La precisión se obtiene usando la Ecuación eq:precision:

$$\text{Precisión} = \frac{TP}{TP + FP} \quad (\text{C.2})$$

donde  $TP$  representa el número de verdaderos positivos, es decir, casos que el modelo clasifica correctamente como congestionados, y  $FP$  representa el número de falsos positivos, es decir, casos que el modelo clasifica incorrectamente como congestionados cuando en realidad no están congestionados. .

En esta investigación, se llevó a cabo un análisis detallado de la clasificación de los conglomerados congestionados en el estudio de congestión del tráfico, ya que esta clasificación juega un papel central en la gestión del tráfico urbano y en la mejora de la movilidad en las ciudades. Los grupos congestionados representan escenarios de tráfico problemáticos que pueden tener un impacto significativo en la movilidad urbana y la calidad de vida de los ciudadanos. Identificar y clasificar con precisión estas situaciones es esencial para tomar decisiones informadas sobre la gestión del tráfico y aplicar estrategias eficaces de alivio de la congestión. Los resultados de precisión para los grupos categorizados como congestionados se muestran en la Tabla C.5. La Tabla de resultados proporciona información importante sobre la precisión de la comparación entre los grupos congestionados del método de grupo dinámico y las celdas congestionadas del método de cuadrícula estática para identificar la congestión del tráfico. Se han evaluado dos parámetros clave, el olvido y la tolerancia, para comprender su impacto en la precisión de los resultados.

TABLA C.5: Resultados de precisión en situaciones de congestión.

<b>Olvido</b>	<b>Tolerancia</b>	<b>Precisión</b>
45	0	83,48 %
45	0,2	89,29 %
60	0	82,84 %
60	0,2	88,63 %

La metodología de identificación de áreas congestionadas se caracteriza por la variación de los valores de olvido entre 45 y 60 s, evidenciándose una relación inversa entre este parámetro y la precisión. Al elegir el valor del olvido, es importante equilibrar la captura de la dinámica reciente y la relevancia de los acontecimientos pasados. Sin embargo, explorar la tolerancia, especialmente en niveles como 0,2, sugiere que permitir un margen de error puede mejorar la precisión.

La identificación de combinaciones óptimas, como un olvido de 45 s con una tolerancia de 0,2, destaca la existencia de combinaciones de parámetros adecuadas para una detección eficaz basada en los índices de precisión de zonas congestionadas en entornos dinámicos.

Los resultados resaltan la influencia crucial de los parámetros olvido y tolerancia en la precisión de la predicción de la congestión del tráfico. Está claro que la elección adecuada de estos valores es un factor determinante para lograr una precisión óptima en la clasificación de las situaciones de congestión. La configuración de estos parámetros debe estar alineada con precisión con los requisitos de aplicación específicos y los objetivos de predicción.

Sin embargo, es importante señalar que esta mejora en la precisión al reducir el valor de olvido y aumentar la tolerancia también puede tener implicaciones para otros aspectos del análisis. Un valor de olvido más bajo significa que se está considerando una ventana de tiempo más estrecha, lo que puede resultar en la pérdida de información relevante a largo plazo. Además, una mayor tolerancia implica un margen de error más amplio, lo que podría permitir la inclusión de datos ruidosos que afecten a la precisión en determinadas situaciones.

Por lo tanto, encontrar el equilibrio adecuado entre estos parámetros es un desafío clave en la aplicación práctica de estos métodos. La elección de valores óptimos para el olvido y la tolerancia dependerá de las necesidades específicas de la tarea de predicción de la congestión y de la importancia de mantener la precisión en comparación con otros factores, como la retención de información histórica y la gestión del ruido en los datos.

Al analizar estos resultados, es fundamental resaltar la eficiencia del algoritmo de agrupamiento en la detección de congestión de vehículos en comparación con el método basado en celdas estáticas en regiones fijas. Los altos niveles de precisión fortalecen la capacidad del algoritmo para identificar patrones de congestión en los datos y anticipar situaciones futuras.

Además, observando la comparación realizada en las pruebas del modelo, es evidente que la falta de adaptabilidad del indicador de congestión aplicado a las celdas estáticas para ajustarse a la evolución de los datos y a los cambios en la distribución de los grupos puede afectar la calidad de los resultados. Si los datos de congestión y los flujos de tráfico no se identifican adecuadamente debido a esta falta de adaptabilidad, la detección de áreas congestionadas puede carecer de confiabilidad.

Por otro lado, una metodología de agrupación dinámica que tenga en cuenta las variaciones en los flujos de datos y se adapte a los cambios en la distribución de los grupos proporciona una representación precisa de la dinámica del tráfico. Como se muestra en la Tabla C.6, los datos estadísticos para un conglomerado específico indican que las velocidades del ciclo 4, aunque pertenecen al mismo conglomerado, han sufrido cambios inusuales. Esto se debe a que el grupo incorpora información de diferentes vehículos matriculados, lo que permite una mejor adaptación a la evolución del tráfico.

Al recalibrar continuamente la posición del centroide y ajustar el hipercuadro en función de los datos en evolución, esta metodología captura de manera efectiva las variaciones en las densidades y formas de los segmentos de la carretera. Esto se puede ver en la Tabla C.7, que muestra la información utilizada para analizar los tramos de la vía del ciclo 5. En esta tabla se puede observar que de cuatro segmentos identificados, sólo tres tienen vehículos registrados. Como cada segmento se analiza de forma independiente, es posible que un vehículo que viaja a través de varios segmentos se cuente como un solo vehículo en el contexto de otro segmento. La representación visual de esta tabla está asociada a la Figura ??c. Esto permite realizar agrupaciones basadas en datos actualizados y evaluaciones realistas.

TABLA C.6: Ejemplo de la evolución de un grupo.

Ciclos	Puntos	Vehículos	Velocidad
1	4	2	17,94
2	3	2	17,91
3	2	1	16,43
4	3	1	3,47
5	9	3	16,96
6	6	3	14,32

TABLA C.7: Ejemplo del dinamismo de los segmentos de carreteras en un grupo del ciclo 5.

Segmento	Vehículos	Longitud (Metros)	Densidad	Indicador de congestión
1	3	13,73	0,218	2,14
2	2	14,29	0,139	1,39
3	1	7,87	0,126	1,89
4	0	2,14	0,000	N/A
Total	6	35,89		

#### C.4. Conclusiones

Los resultados obtenidos resaltan la efectividad de la metodología de agrupamiento dinámico en comparación con el método estático basado en celdas para clasificar condiciones de congestión. Al permitir la agrupación dinámica de datos de trayectoria de vehículos y realizar análisis específicos para cada grupo, esta metodología facilita la detección temprana y precisa de áreas con problemas de tráfico congestionado.

La aplicación de métodos de agrupamiento dinámico se presenta como una estrategia muy prometedora. Estos métodos tienen la capacidad de adaptarse a los cambios constantes del tráfico urbano, capturando patrones de movilidad en constante evolución. La relevancia del factor olvido radica en su capacidad para mantener actualizados los grupos, considerando tanto ubicaciones recientes como antiguas. Esto garantiza que los grupos reflejen con precisión la dinámica del tráfico actual, lo que permite identificar tempranamente la congestión emergente.

Ajustar cuidadosamente los parámetros de olvido y tolerancia es fundamental para obtener resultados precisos en la comparación entre los métodos de agrupamiento dinámico y de cuadrícula estática en la predicción de la congestión del tráfico. Estos hallazgos son esenciales para mejorar la eficacia de los modelos de clasificación en aplicaciones de gestión del tráfico.

La metodología basada en agrupamiento dinámico destaca por su adaptabilidad a los cambios de tráfico, proporcionando una visión completa y actualizada del comportamiento vehicular en el entorno urbano. Estos resultados respaldan la eficacia de los grupos como una herramienta valiosa para mejorar la gestión del tráfico y reducir los problemas de congestión en las ciudades.

Nuestro estudio reconoce varias limitaciones importantes. La generalización de los hallazgos a diferentes entornos urbanos resalta la necesidad de validar la metodología en varias ciudades y regiones urbanas. Las pruebas se centraron en comparaciones estáticas. Es importante ampliar las comparaciones a enfoques dinámicos para una evaluación más completa de los desafíos clave. Además, el rendimiento



podría ser más lento de lo esperado debido a la implementación del microlotes, por lo que es necesario investigar y optimizar la eficiencia del sistema, y el indicador de congestión, aunque dinámico, puede presentar errores iniciales y requerir estabilización. Este es un proceso que se puede mejorar.

Es necesario abordar el rendimiento en tiempo real, optimizando la implementación de microlotes o explorando diferentes enfoques que permitan una respuesta más rápida, ligada a una mejora del modelo con foco en la eficiencia y el cálculo del indicador de congestión de forma dinámica.

En cuanto a futuras investigaciones, es fundamental explorar en profundidad las posibles razones detrás de la disminución observada en la precisión en la clasificación congestionada. Además, se propone realizar nuevos casos de estudio en áreas más grandes, en condiciones de congestión extrema y en escenarios de tráfico de alta complejidad. Estos casos de estudio proporcionarán información adicional sobre los límites y la solidez de la metodología propuesta.



# Bibliografía

- Abulude, Francis, Akinyinka Akinnusotu y Adewale Adeyemi (2015). «GLOBAL POSITIONING SYSTEM AND IT'S WIDE APPLICATIONS». En: *Continental J. Information Technology*. DOI: [10.5707/cjit.2015.9.1.22.32](https://doi.org/10.5707/cjit.2015.9.1.22.32).
- Anil Meera, Ajith y Martijn Wisse (2021). «Dynamic Expectation Maximization Algorithm for Estimation of Linear Systems with Colored Noise». En: *Entropy* 23.10, pág. 1306. ISSN: 1099-4300. DOI: [10.3390/e23101306](https://doi.org/10.3390/e23101306).
- Arrais De Freitas, Nicksson, Ticiana Coelho Da Silva, José Fernandes De Macêdo, Leopoldo Melo Junior y Matheus Cordeiro (2021). «Using Deep Learning for Trajectory Classification:» en: *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. Online Streaming: SCITEPRESS - Science and Technology Publications, págs. 664-671. ISBN: 978-989-758-484-8. DOI: [10.5220/0010227906640671](https://doi.org/10.5220/0010227906640671).
- Azimi, Mehdi y Yunlong Zhang (2010). «Categorizing Freeway Flow Conditions by Using Clustering Methods». En: *Transportation Research Record: Journal of the Transportation Research Board* 2173.1, págs. 105-114. ISSN: 0361-1981, 2169-4052. DOI: [10.3141/2173-13](https://doi.org/10.3141/2173-13).
- Bahmani, Bahman, Benjamin Moseley, Andrea Vattani, Ravi Kumar y Sergei Vassilvitskii (2012). «Scalable k-means++». En: *Proc. VLDB Endow.* 5.7, 622–633. ISSN: 2150-8097. DOI: [10.14778/2180912.2180915](https://doi.org/10.14778/2180912.2180915). URL: <https://doi.org/10.14778/2180912.2180915>.
- Barbosa Roa, Nathalie, Louise Travé-Massuyès y Victor H. Grisales-Palacio (2019). «DyClee: Dynamic clustering for tracking evolving environments». En: *Pattern Recognition* 94, págs. 162-186. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2019.05.024>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320319301992>.
- Bathae, Nadia, Alireza Mohseni, Sejoon Park, J David Porter, David S Kim, Nadia Bathae, Alireza Mohseni, Sejoon Park, J David Porter y S David (2018). «A Cluster Analysis Approach For Differentiating Transportation Modes Using Bluetooth Sensor Data». En: *Journal of Intelligent Transportation Systems* 22.4, págs. 353-364. DOI: [10.1080/15472450.2018.1457444](https://doi.org/10.1080/15472450.2018.1457444).
- Besse, Philippe C., Brendan Guillouet, Jean-Michel Loubes y Francois Royer (2016). «Review and Perspective for Distance-Based Clustering of Vehicle Trajectories». En: *Trans. Intell. Transport. Sys.* 17.11, 3306–3317. ISSN: 1524-9050. DOI: [10.1109/TITS.2016.2547641](https://doi.org/10.1109/TITS.2016.2547641).

- Bian, Jiang, Dayong Tian, Yuanyan Tang y Dacheng Tao (2018). «A Survey on Trajectory Clustering Analysis». En: DOI: [10.48550/ARXIV.1802.06971](https://doi.org/10.48550/ARXIV.1802.06971). arXiv: [1802.06971](https://arxiv.org/abs/1802.06971) [cs].
- Byoung-Kee Yi, H.V. Jagadish y C. Faloutsos (1998). «Efficient Retrieval of Similar Time Sequences under Time Warping». En: *Proceedings 14th International Conference on Data Engineering*. Orlando, FL, USA: IEEE Comput. Soc, págs. 201-208. ISBN: 978-0-8186-8289-6. DOI: [10.1109/ICDE.1998.655778](https://doi.org/10.1109/ICDE.1998.655778).
- Chen, Jinyang, Rangding Wang, Liangxu Liu y Jiatao Song (2011). «Clustering of Trajectories Based on Hausdorff Distance». En: *2011 International Conference on Electronics, Communications and Control (ICECC)*. Ningbo, China: IEEE, págs. 1940-1944. ISBN: 978-1-4577-0320-1. DOI: [10.1109/ICECC.2011.6066483](https://doi.org/10.1109/ICECC.2011.6066483).
- Chen, Wen, Minhe Ji y Jianmei Wang (2014). «T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation». En: *International Journal of Online and Biomedical Engineering (iJOE)* 10.6, págs. 19-24. ISSN: 2626-8493. DOI: [10.3991/ijoe.v10i6.3881](https://doi.org/10.3991/ijoe.v10i6.3881).
- Choong, Mei Yeen, Renee Ka Yin Chin, Kiam Beng Yeo y Kenneth T. K. Teo (2016). «Trajectory pattern mining via clustering based on similarity function for transportation surveillance». En: *International Journal of Simulation-Systems, Science & Technology* 17.34, págs. 19-1. DOI: [10.5013/IJSSST.a.17.34.19](https://doi.org/10.5013/IJSSST.a.17.34.19).
- Cobo, M.J., A.G. López-Herrera, E. Herrera-Viedma y F. Herrera (2011). «An Approach for Detecting, Quantifying, and Visualizing the Evolution of a Research Field: A Practical Application to the Fuzzy Sets Theory Field». En: *Journal of Informetrics* 5.1, págs. 146-166. ISSN: 17511577. DOI: [10.1016/j.joi.2010.10.002](https://doi.org/10.1016/j.joi.2010.10.002).
- Dziadczyk, Emil, Wojciech Zabierowski y A. Napieralski (2007). *Satellite Navigation System GPS*, pág. 506. DOI: [10.1109/CADSM.2007.4297633](https://doi.org/10.1109/CADSM.2007.4297633).
- Eiter, Thomas y Heikki Mannila (1994). «Computing Discrete Fréchet Distance». En.
- Endo, Yuki, Hiroyuki Toda, Kyosuke Nishida y Jotaro Ikedo (2016). «Classifying Spatial Trajectories Using Representation Learning». En: *International Journal of Data Science and Analytics* 2.3-4, págs. 107-117. ISSN: 2364-415X, 2364-4168. DOI: [10.1007/s41060-016-0014-1](https://doi.org/10.1007/s41060-016-0014-1).
- Erdelić, Tomislav, Tonči Carić, Martina Erdelić, Leo Tišljarić, Ana Turković y Niko Jelušić (2021). «Estimating congestion zones and travel time indexes based on the floating car data». En: *Computers, Environment and Urban Systems* 87, pág. 101604. ISSN: 0198-9715. DOI: <https://doi.org/10.1016/j.compenvurbsys.2021.101604>. URL: <https://www.sciencedirect.com/science/article/pii/S0198971521000119>.
- Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu et al. (1996). «A density-based algorithm for discovering clusters in large spatial databases with noise.» En: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Vol. 96. 34. AAAI Press, págs. 226-231.

- Etemad, Mohammad, Zahra Etemad, Amilcar Soares, Vania Bogorny, Stan Matwin y Luis Torgo (2020). «Wise Sliding Window Segmentation: A Classification-Aided Approach for Trajectory Segmentation». En: *arXiv:2003.10248 [cs, stat]*. eprint: [2003.10248](https://arxiv.org/abs/2003.10248) (cs, stat).
- Etemad, Mohammad, Amílcar Soares Júnior, Jordan Rose, Arazoo Hoseyni y Stan S Matwin (2019). «A Trajectory Segmentation Algorithm Based on Interpolation-based Change Detection Strategies». En: DOI: [10.13140/RG.2.2.34157.03049](https://doi.org/10.13140/RG.2.2.34157.03049).
- Feng, Huifang, Fengshan Bai y Youji Xu (2019). «Identification of critical roads in urban transportation network based on GPS trajectory data». En: *Physica A: Statistical Mechanics and its Applications* 535, pág. 122337. ISSN: 03784371. DOI: [10.1016/j.physa.2019.122337](https://doi.org/10.1016/j.physa.2019.122337).
- Ferreira, Nivan, James T. Klosowski, Carlos Scheidegger y Claudio Silva (2012). *Vector Field k-Means: Clustering Trajectories by Fitting Multiple Vector Fields*. arXiv: [1208.5801](https://arxiv.org/abs/1208.5801) [cs.LG].
- Gao, Yongsheng y Maylor K H Leung (2002). «Line segment Hausdorff distance on face matching». En: *Pattern Recognition* 35.2, págs. 361-371. ISSN: 0031-3203. DOI: [10.1016/S0031-3203\(01\)00049-8](https://doi.org/10.1016/S0031-3203(01)00049-8).
- Han, Jiawei, Micheline Kamber y Anthony K.H. Tung (2001). «Spatial Clustering Methods in Data Mining: A Survey». En: *Geographic data mining and knowledge discovery*, págs. 188-217.
- He, Yufeng, Barbara Hofer, Yehua Sheng, Yue Yin y Hui Lin (2023). «Processes and Events in the Center: A Taxi Trajectory-Based Approach to Detecting Traffic Congestion and Analyzing Its Causes». En: *International Journal of Digital Earth* 16.1, págs. 509-531. ISSN: 1753-8947. DOI: [10.1080/17538947.2023.2182374](https://doi.org/10.1080/17538947.2023.2182374).
- Hinneburg, Alexander y Daniel A. Keim (1999). «Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering». En: *Proceedings of the 25th International Conference on Very Large Data Bases*. Vldb '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., págs. 506-517. ISBN: 1-55860-615-7.
- Hong, Zihan, Ying Chen y Hani S. Mahmassani (2018). «Recognizing Network Trip Patterns Using a Spatio-Temporal Vehicle Trajectory Clustering Algorithm». En: *IEEE Transactions on Intelligent Transportation Systems* 19.8, 2548–2557. ISSN: 1524-9050, 1558-0016. DOI: [10.1109/TITS.2017.2754401](https://doi.org/10.1109/TITS.2017.2754401).
- Hu, Hyejung, Gunwoo Lee, Jae Hun Kim y Hyunju Shin (2020). «Estimating Micro-Level On-Road Vehicle Emissions Using the K-Means Clustering Method with GPS Big Data». En: *Electronics*.
- Hu, Xianbiao, Yifei Yuan, Xiaoyu Zhu, Hong Yang y Kun Xie (2018). «Behavioral responses to pre-planned road capacity reduction based on smartphone GPS trajectory data : a functional data analysis approach». En: *Journal of Intelligent Transportation Systems* 0.0, págs. 1-11. ISSN: 1547-2450. DOI: [10.1080/15472450.2018.1488133](https://doi.org/10.1080/15472450.2018.1488133).

- Jain, Anil K. (2009). «Data Clustering: 50 Years Beyond K-Means.» En: *Pattern Recognition Letters* 31.8, págs. 651-666. ISSN: 0167-8655. DOI: [10.1016/j.patrec.2009.09.011](https://doi.org/10.1016/j.patrec.2009.09.011). URL: <https://www.sciencedirect.com/science/article/pii/S0167865509002323>.
- Kamble, Shridevi Jeevan y Manjunath R Kounte (2020). «Machine Learning Approach on Traffic Congestion Monitoring System in Internet of Vehicles». En: *Procedia Computer Science* 171, págs. 2235-2241. ISSN: 18770509. DOI: [10.1016/j.procs.2020.04.241](https://doi.org/10.1016/j.procs.2020.04.241).
- Kim, Jiwon y Hani S Mahmassani (2015). «Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories». En: *Transportation Research Procedia* 9, págs. 164-184. DOI: [10.1016/j.trpro.2015.07.010](https://doi.org/10.1016/j.trpro.2015.07.010). URL: <https://www.sciencedirect.com/science/article/pii/S2352146515001702>.
- Kossakov, Marlen, Assel Mukasheva, Gani Balbayev, Syrym Seidazimov, Dinargul Mukammejanova y Madina Sydybayeva (2024). «Quantitative Comparison of Machine Learning Clustering Methods for Tuberculosis Data Analysis». En: *CIEES 2023*. MDPI, pág. 20. DOI: [10.3390/engproc2024060020](https://doi.org/10.3390/engproc2024060020).
- Lee, Jae-Gil, Jiawei Han y Kyu-Young Whang (2007). «Trajectory clustering: a partition-and-group framework». En: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. New York, NY, USA: Association for Computing Machinery, págs. 593-604. ISBN: 9781595936868. DOI: [10.1145/1247480.1247546](https://doi.org/10.1145/1247480.1247546). URL: <https://doi.org/10.1145/1247480.1247546>.
- Leite Da Silva, Camila, Lucas May Petry y Vania Bogorny (2019). «A Survey and Comparison of Trajectory Classification Methods». En: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. Salvador, Brazil: IEEE, págs. 788-793. ISBN: 978-1-72814-253-1. DOI: [10.1109/BRACIS.2019.00141](https://doi.org/10.1109/BRACIS.2019.00141).
- Liu, Chang, Shize Zhang, Lufang Cao y Bin Lin (2023). «The Identification of Ship Trajectories Using Multi-Attribute Compression and Similarity Metrics». En: *Journal of Marine Science and Engineering* 11.10, pág. 2005. ISSN: 2077-1312. DOI: [10.3390/jmse11102005](https://doi.org/10.3390/jmse11102005).
- Liu, Jincheng, Di Wu, Hadi Mohammed y Razak Seidu (2024). «A Novel Method for Anomaly Detection and Signal Calibration in Water Quality Monitoring of an Urban Water Supply System». En: *Water* 16.9, pág. 1238. ISSN: 2073-4441. DOI: [10.3390/w16091238](https://doi.org/10.3390/w16091238).
- Liu, Liang Xu, Jia Tao Song, Bo Guan, Zhao Xiao Wu y Ke Jia He (2012). «Tra-DBScan: A Algorithm of Clustering Trajectories». En: *Frontiers of Manufacturing and Design Science II*. Vol. 121. Applied Mechanics and Materials. Trans Tech Publications Ltd, págs. 4875-4879. DOI: [10.4028/www.scientific.net/AMM.121-126.4875](https://doi.org/10.4028/www.scientific.net/AMM.121-126.4875).
- Lou, Jingfeng y Aiguo Cheng (2020). «Detecting Pattern Changes in Individual Travel Behavior from Vehicle GPS/GNSS Data». En: *Sensors* 20.8. ISSN: 1424-8220.

- DOI: 10.3390/s20082295. URL: <https://www.mdpi.com/1424-8220/20/8/2295>.
- Luo, Ting, Xinwei Zheng, Guangluan Xu, Kun Fu y Wenjuan Ren (2017). «An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories». En: *ISPRS International Journal of Geo-Information* 6.3. ISSN: 2220-9964. DOI: 10.3390/ijgi6030063. URL: <https://www.mdpi.com/2220-9964/6/3/63>.
- Madhulatha, T Soni (2012). «An overview on clustering methods». En: *arXiv preprint arXiv:1205.1117*.
- Magdy, N, M A Sakr, T Mostafa y K El-Bahnasy (2015). «Review on trajectory similarity measures». En: *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, págs. 613-619. ISBN: VO -. DOI: 10.1109/IntelCIS.2015.7397286.
- Mahmood Hussien, Nadia, Yasmin Mohialden, Basim Kudaer e Itidal Mohammed (2023). «Review of an Accurate System Utilizing GPS Technology». En: *Journal La Multiapp* 3, págs. 266-269. DOI: 10.37899/journallamultiapp.v3i6.746.
- Mao, Yingchi, Haishi Zhong, Hai Qi, Ping Ping y Xiaofang Li (2017). «An Adaptive Trajectory Clustering Method Based on Grid and Density in Mobile Pattern Analysis». En: *Sensors* 17.9, pág. 2013. ISSN: 1424-8220. DOI: 10.3390/s17092013. URL: <http://www.mdpi.com/1424-8220/17/9/2013>.
- Markovi, Nikola, Przemysław Sekuła, Zachary Vander Laan, Gennady Andrienko y Natalia Andrienko (2018). «Applications of Trajectory Data from the Perspective of a Road Transportation Agency : Literature Review and Maryland Case Study». En: *IEEE Transactions on Intelligent Transportation Systems*. arXiv: arXiv:1708.0719.
- Mazimpaka, Jean Damascène y Sabine Timpf (2016). «Trajectory Data Mining: A Review of Methods and Applications». En: *Journal of Spatial Information Science* 13, págs. 61-99. ISSN: 1948-660X. DOI: 10.5311/JOSIS.2016.13.263.
- Molina, Roberto y Waldo Hasperué (2018). «D3CAS: un Algoritmo de Clustering para el Procesamiento de Flujos de Datos en Spark». En: *XXIV Congreso Argentino de Ciencias de la Computación*. La Plata, págs. 452-461. ISBN: 978-950-658-472-6.
- Moulavi, Davoud, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek y Jörg Sander (2014). «Density-Based Clustering Validation». En: *Proceedings of the 2014 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, págs. 839-847. ISBN: 978-1-61197-344-0. DOI: 10.1137/1.9781611973440.96.
- Mu, Chunxiao, Yanchen Hou, Jindong Zhao, Shouke Wei y Yuxuan Wu (2023). «Stream-DBSCAN: A Streaming Distributed Clustering Model for Water Quality Monitoring». En: *Applied Sciences* 13.9, pág. 5408. ISSN: 2076-3417. DOI: 10.3390/app13095408.
- Rempe, Felix, Gerhard Huber y Klaus Bogenberger (2016a). «Spatio-Temporal Congestion Patterns in Urban Traffic Networks». En: *Transportation Research Procedia* 15, págs. 513-524. ISSN: 23521465. DOI: 10.1016/j.trpro.2016.06.043.

- Rempe, Felix, Gerhard Huber y Klaus Bogenberger (2016b). «Spatio-Temporal Congestion Patterns in Urban Traffic Networks». En: *Transportation Research Procedia* 15, págs. 513-524. ISSN: 23521465. DOI: [10.1016/j.trpro.2016.06.043](https://doi.org/10.1016/j.trpro.2016.06.043).
- Reyes, Gary, Christopher Crespo, Oscar León-Granizo, Wellington Bazán y Richard Horta (2022a). «Propuesta de método de extracción de ubicaciones georreferenciales de una red de carreteras para el análisis de trayectorias GPS». En: *Investigación, Tecnología e Innovación* 14.16, págs. 1-15. ISSN: 2661-6548. DOI: [10.53591/iti.v14i16.1465](https://doi.org/10.53591/iti.v14i16.1465).
- Reyes, Gary, Laura Lanzarini, César Estrebow y Aurelio Bariviera (2022b). «Data Stream Processing Method for Clustering of Trajectories». En: *Technologies and Innovation*. Ed. por Rafael Valencia-García, Martha Bucaram-Leverone, Javier Del Cioppo-Morstadt, Néstor Vera-Lucio y Emma Jácome-Murillo. Vol. 1658. Cham: Springer International Publishing, págs. 151-163. ISBN: 978-3-031-19960-8 978-3-031-19961-5. DOI: [10.1007/978-3-031-19961-5\\_11](https://doi.org/10.1007/978-3-031-19961-5_11).
- Reyes, Gary, Laura Lanzarini, César Estrebow, Aurelio Bariviera y Victor Maquilón (2023a). «Evaluation of a Grid for the Identification of Traffic Congestion Patterns». En: *Technologies and Innovation*. Ed. por Rafael Valencia-García, Martha Bucaram-Leverone, Javier Del Cioppo-Morstadt, Néstor Vera-Lucio y Pablo Humberto Centanaro-Quiroz. Communications in Computer and Information Science. Cham: Springer Nature Switzerland, págs. 277-290. ISBN: 978-3-031-45682-4. DOI: [10.1007/978-3-031-45682-4\\_20](https://doi.org/10.1007/978-3-031-45682-4_20).
- Reyes, Gary, Laura Lanzarini, César Estrebow y Aurelio F. Bariviera (2024). *Un Análisis Bibliométrico de La Producción Científica Acerca Del Agrupamiento de Trayectorias GPS*. DOI: [10.48550/ARXIV.2404.17761](https://doi.org/10.48550/ARXIV.2404.17761).
- Reyes, Gary, Laura Lanzarini, Cesar Estrebow y Aurelio Fernandez Bariviera (2022c). «Dynamic Grouping of Vehicle Trajectories». En: *Journal of Computer Science and Technology* 22.2, e11. ISSN: 1666-6038, 1666-6046. DOI: [10.24215/16666038.22.e11](https://doi.org/10.24215/16666038.22.e11).
- Reyes, Gary, Laura Lanzarini, Waldo Hasperué y Aurelio F. Bariviera (2020). «GPS Trajectory Clustering Method for Decision Making on Intelligent Transportation Systems». En: *Journal of Intelligent & Fuzzy Systems* 38.5. Ed. por José M. Merigó, Salvador Linares-Mustaros y Joan Carles Ferrer-Comalat, págs. 5529-5535. ISSN: 10641246, 18758967. DOI: [10.3233/JIFS-179644](https://doi.org/10.3233/JIFS-179644).
- (2021a). «Proposal for a Pivot-Based Vehicle Trajectory Clustering Method». En: *Transportation Research Record: Journal of the Transportation Research Board* 2676, pág. 036119812110584. DOI: [10.1177/03611981211058429](https://doi.org/10.1177/03611981211058429).
- Reyes, Gary, Laura Cristina Lanzarini, César Armando Estrebow y Victor Maquilón (2021b). «Vehicular Flow Analysis Using Clusters». En: *XXVII Congreso Argentino de Ciencias de La Computación (CACIC)*. Vol. XXVII. Universidad Nacional de Salta, 261-270. ISBN: 978-987-633-574-4.
- Reyes, Gary, Roberto Tolozano-Benites, Laura Lanzarini, César Estrebow, Aurelio F. Bariviera y Julio Barzola-Monteses (2023b). «Methodology for the Identification



- of Vehicle Congestion Based on Dynamic Clustering». En: *Sustainability* 15.24, pág. 16575. ISSN: 2071-1050. DOI: [10.3390/su152416575](https://doi.org/10.3390/su152416575).
- Reyes-Zambrano, Gary, Laura Lanzarini, Waldo Hasperué y Aurelio F Bariviera (2020). «GPS trajectory clustering method for decision making on intelligent transportation systems». En: *Journal of Intelligent & Fuzzy Systems* 38.5, págs. 5529-5535. ISSN: 1875-8967. DOI: [10.3233/JIFS-179644](https://doi.org/10.3233/JIFS-179644).
- Rockafellar, R Tyrrell y Roger J-B Wets (2009). *Variational analysis*. Vol. 317. Berlin Heidelberg: Springer-Verlag. DOI: [10.1007/978-3-642-02431-3](https://doi.org/10.1007/978-3-642-02431-3).
- Saxena, Abhineet, Lalit Goyal y Mamta Mittal (2015). «Comparative Analysis of Clustering Methods». En: *International Journal of Computer Applications* 118, págs. 30-35. DOI: [10.5120/20873-3452](https://doi.org/10.5120/20873-3452).
- Schweppe, Fred C. (1967). «On the Bhattacharyya Distance and the Divergence between Gaussian Processes». En: *Information and Control* 11.4, págs. 373-395. ISSN: 00199958. DOI: [10.1016/S0019-9958\(67\)90610-9](https://doi.org/10.1016/S0019-9958(67)90610-9).
- Shahraki, Amin, Mahmoud Abbasi, Amir Taherkordi y Anca Delia Jurcut (2022). «A Comparative Study on Online Machine Learning Techniques for Network Traffic Streams Analysis». En: *Computer Networks* 207, pág. 108836. ISSN: 1389-1286. DOI: [10.1016/j.comnet.2022.108836](https://doi.org/10.1016/j.comnet.2022.108836).
- Shang, Qiang, Yang Yu y Tian Xie (2022). «A Hybrid Method for Traffic State Classification Using K-Medoids Clustering and Self-Tuning Spectral Clustering». En: *Sustainability* 14.17, pág. 11068. ISSN: 2071-1050. DOI: [10.3390/su141711068](https://doi.org/10.3390/su141711068).
- Soumia Goumiri, Saïd Yahiaoui y Soufiene Djahel (2023). «Smart Mobility in Smart Cities: Emerging Challenges, Recent Advances and Future Directions». En: *Journal of Intelligent Transportation Systems* 0.0, págs. 1-37. DOI: [10.1080/15472450.2023.2245750](https://doi.org/10.1080/15472450.2023.2245750). eprint: <https://doi.org/10.1080/15472450.2023.2245750>.
- Su, Xing, Wenjie Sun, Chenting Song, Zhi Cai y Limin Guo (2023). «A Latent-Factor-Model-Based Approach for Traffic Data Imputation with Road Network Information». En: *ISPRS International Journal of Geo-Information* 12.378. ISSN: 2220-9964. DOI: [10.3390/ijgi12090378](https://doi.org/10.3390/ijgi12090378).
- Sun, Shuming, Juan Chen y Jian Sun (2019). «Traffic congestion prediction based on GPS trajectory data». en. En: *International Journal of Distributed Sensor Networks* 15.5. Publisher: SAGE Publications. ISSN: 1550-1329. DOI: [10.1177/1550147719847440](https://doi.org/10.1177/1550147719847440). URL: <https://doi.org/10.1177/1550147719847440>.
- Tang, Juanyu, Yi Zhu, Yizhe Huang, Zhong-ren Peng y Zhanyong Wang (2018). «Identification and interpretation of spatial – temporal mismatch between taxi demand and supply using global positioning system data». En: *Journal of Intelligent Transportation Systems* 0.0, págs. 1-13. ISSN: 1547-2450. DOI: [10.1080/15472450.2018.1518137](https://doi.org/10.1080/15472450.2018.1518137).
- Thieu, Nguyen Van (2024). «PerMetrics: A Framework of Performance Metrics for Machine Learning Models». En: *Journal of Open Source Software* 9.95, pág. 6143.

- DOI: [10.21105/joss.06143](https://doi.org/10.21105/joss.06143). URL: <https://joss.theoj.org/papers/10.21105/joss.06143>.
- Upreti, Swagata y Manish Kumar (2008). *Perspectives of Global Positioning System (GPS) Applications*.
- Van Thieu, Nguyen y Seyedali Mirjalili (2023). «MEALPY: An open-source library for latest meta-heuristic algorithms in Python». En: *Journal of Systems Architecture*. DOI: [10.1016/j.sysarc.2023.102871](https://doi.org/10.1016/j.sysarc.2023.102871).
- Varghese, Bindiya M, A Unnikrishnan y Jacob K Poullose (2013). «Spatial clustering algorithms – An overview». En: *Asian Journal of Computer Science and Information Technology* 3.1, págs. 1-8.
- Villegas, Fulgencio (2021). «Relatividad y el Sistema de Posicionamiento Global (GPS)». En: *Revista de Investigación de Física* 23.1, págs. 44-47. ISSN: 1728-2977, 1605-7724. DOI: [10.15381/rif.v23i1.20289](https://doi.org/10.15381/rif.v23i1.20289).
- Wang, Chao, Jitong Li, Yunhua He, Ke Xiao y Chunqiang Hu (2020). «Segmented Trajectory Clustering-Based Destination Prediction in IoVs». En: *IEEE Access* 8, págs. 98999-99009. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2020.2998063](https://doi.org/10.1109/ACCESS.2020.2998063).
- Wang, Xin, Zhaowei Qu, Xianmin Song, Haitao Li y Zhaotian Pan (2021). «Using Taxi GPS Trajectory Data to Optimize the Spatial Layout of Urban Taxi Stands». En: 2675, 301–312. ISSN: 0361-1981, 2169-4052. DOI: [10.1177/0361198120970537](https://doi.org/10.1177/0361198120970537).
- Wei, Long-xiang, Xiao-hai He, Qi-zhi Teng y Ming-liang Gao (2014). «Trajectory Classification Based on Hausdorff Distance and Longest Common SubSequence: Trajectory Classification Based on Hausdorff Distance and Longest Common SubSequence». En: *Journal of Electronics & Information Technology* 35.4, págs. 784-790. ISSN: 1009-5896. DOI: [10.3724/SP.J.1146.2012.01078](https://doi.org/10.3724/SP.J.1146.2012.01078).
- Wu, Haixiao, Zhongming Wu, Junfeng Lu y Li Sun (2024). «Anti-Rollover Trajectory Planning Method for Heavy Vehicles in Human–Machine Cooperative Driving». En: *World Electric Vehicle Journal* 15.8, pág. 328. ISSN: 2032-6653. DOI: [10.3390/wevj15080328](https://doi.org/10.3390/wevj15080328).
- Xiong, Hui, Pengjian Shang y Songhan Bian (2017). «Detecting intrinsic dynamics of traffic flow with recurrence analysis and empirical mode decomposition». En: *Physica A: Statistical Mechanics and its Applications* 474, 70–84. ISSN: 03784371. DOI: [10.1016/j.physa.2017.01.060](https://doi.org/10.1016/j.physa.2017.01.060).
- Yang, Xin-She (2019). «Data Mining Techniques». En: *Introduction to Algorithms for Data Mining and Machine Learning*. Elsevier, págs. 109-128. ISBN: 978-0-12-817216-2. DOI: [10.1016/B978-0-12-817216-2.00013-2](https://doi.org/10.1016/B978-0-12-817216-2.00013-2).
- Yang, Xue, Kathleen Stewart, Luliang Tang, Zhong Xie y Qingquan Li (2018). «A Review of GPS Trajectories Classification Based on Transportation Mode». En: *Sensors* 18.11, pág. 3741. ISSN: 1424-8220. DOI: [10.3390/s18113741](https://doi.org/10.3390/s18113741).
- Yao, Jianjun, Chen Qian, Yikun Zhang y Geyang Yu (2023). «Multi-Objective Redundancy Optimization of Continuous-Point Robot Milling Path in Shipbuilding». En: *Computer Modeling in Engineering & Sciences* 134.2, págs. 1283-1303. ISSN: 1526-1506. DOI: [10.32604/cmescs.2022.021328](https://doi.org/10.32604/cmescs.2022.021328).

- Yu, Qingying, Yonglong Luo, Chuanming Chen y Shigang Chen (2019a). «Trajectory similarity clustering based on multi-feature distance measurement». En: *Applied Intelligence* 49.6, págs. 2315-2338. DOI: [10.1007/s10489-018-1385-x](https://doi.org/10.1007/s10489-018-1385-x).
- Yu, Qingying, Yonglong Luo, Chuanming Chen y Xiaoyao Zheng (2019b). «Road Congestion Detection Based on Trajectory Stay-Place Clustering». En: *ISPRS International Journal of Geo-Information* 8.264. ISSN: 2220-9964. DOI: [10.3390/ijgi8060264](https://doi.org/10.3390/ijgi8060264).
- Yuan, Guan, Penghui Sun, Jie Zhao, Daxing Li y Canwei Wang (2017). «A review of moving object trajectory clustering algorithms». En: *Artificial Intelligence Review* 47.1, 123-144. ISSN: 0269-2821, 1573-7462. DOI: [10.1007/s10462-016-9477-7](https://doi.org/10.1007/s10462-016-9477-7). URL: <http://link.springer.com/10.1007/s10462-016-9477-7>.
- Yuan, Guan, Mingjun ZHU, Shaojie QIAO, Zhixiao WANG y Lei Zhang (2018). «Sparse High-Noise GPS Trajectory Data Compression and Recovery Based on Compressed Sensing». En: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E101.A, págs. 811-821. DOI: [10.1587/transfun.E101.A.811](https://doi.org/10.1587/transfun.E101.A.811).
- Yuan, Nicholas Jing, Yu Zheng, Liuhang Zhang y Xing Xie (2013). «T-Finder: A Recommender System for Finding Passengers and Vacant Taxis». En: *IEEE Transactions on Knowledge and Data Engineering* 25.10, págs. 2390-2403. ISSN: 1041-4347. DOI: [10.1109/TKDE.2012.153](https://doi.org/10.1109/TKDE.2012.153).
- Zhang, Ruobing, Jiayi Guo, Jianming Hu y Xin Pei (2019). «Deep Trajectory Similarity Model: A Fast Method for Trajectory Similarity Computation». En: *International Conference on Transportation and Development 2019*. American Society of Civil Engineers, págs. 13-23. DOI: [10.1061/9780784482582.002](https://doi.org/10.1061/9780784482582.002). eprint: <https://ascelibrary.org/doi/pdf/10.1061/9780784482582.002>.
- Zhang, Yingya, Ning Ye, Ruchuan Wang y Reza Malekian (2016). «A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis». En: *ISPRS International Journal of Geo-Information* 5.5. ISSN: 2220-9964. DOI: [10.3390/ijgi5050071](https://doi.org/10.3390/ijgi5050071). URL: <https://www.mdpi.com/2220-9964/5/5/71>.
- Zhang, Yu, Kanat Tangwongsan y Srikanta Tirthapura (2017). «Streaming K-Means Clustering with Fast Queries». En: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. San Diego, CA, USA: IEEE, págs. 449-460. ISBN: 978-1-5090-6543-1. DOI: [10.1109/ICDE.2017.102](https://doi.org/10.1109/ICDE.2017.102).
- Zheng, Liyuan y Weiming Liu (2024). «A Comprehensive Investigation of Lane-Changing Risk Recognition Framework of Multi-Vehicle Type Considering Key Features Based on Vehicles' Trajectory Data». En: *Electronics* 13.6, pág. 1097. ISSN: 2079-9292. DOI: [10.3390/electronics13061097](https://doi.org/10.3390/electronics13061097).
- Zheng, Yu (2015). «Trajectory Data Mining: An Overview». En: *ACM Transactions on Intelligent Systems and Technology* 6.3, 29:1-29:41. ISSN: 2157-6904. DOI: [10.1145/2743025](https://doi.org/10.1145/2743025).
- Zheng, Yu y Xiaofang Zhou, eds. (2011). *Computing with Spatial Trajectories*. New York, NY: Springer New York. ISBN: 978-1-4614-1628-9 978-1-4614-1629-6. DOI: [10.1007/978-1-4614-1629-6](https://doi.org/10.1007/978-1-4614-1629-6).

- Zhou, Ruiyu, Hong Chen, Hengrui Chen, Enze Liu y Shangjing Jiang (2021). «Research on Traffic Situation Analysis for Urban Road Network Through Spatio-temporal Data Mining: A Case Study of Xi'an, China». En: *IEEE Access* 9, págs. 75553-75567. DOI: [10.1109/ACCESS.2021.3082188](https://doi.org/10.1109/ACCESS.2021.3082188).
- Zhou, Xiangbing, Fang Miao, Hongjiang Ma, Hua Zhang y Huaming Gong (2018). «A Trajectory Regression Clustering Technique Combining a Novel Fuzzy C-Means Clustering Algorithm with the Least Squares Method». En: *ISPRS International Journal of Geo-Information* 7.5. ISSN: 2220-9964. DOI: [10.3390/ijgi7050164](https://doi.org/10.3390/ijgi7050164). URL: <https://www.mdpi.com/2220-9964/7/5/164>.
- Zhou, Xiaofang, Kai Zheng, Hoyoung Jueng, Jiajie Xu y Shazia Sadiq (2015). «Making Sense of Spatial Trajectories». En: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. Melbourne Australia: ACM, págs. 671-672. ISBN: 978-1-4503-3794-6. DOI: [10.1145/2806416.2806418](https://doi.org/10.1145/2806416.2806418).