Estudio de los robos en Pereira mediante la detección y dinámica de patrones en redes espacio temporales

Miguel Ángel Quintero Martinez

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2023

Estudio de los robos en Pereira mediante la detección y dinámica de patrones en redes espacio temporales

Miguel Ángel Quintero Martinez

Trabajo de grado presentado como requisito parcial para optar al título de:
Magíster en ciencias - Estadística

Directora:
Ph.D. Martha Patricia Bohórquez

Codirector:
Ph.D. Rafael Ricardo Rentería

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2023

Y lo asustó la sospecha tardía de que es la vida, más que la muerte, la que no tiene límites.
Gabriel Garcia Márquez

Agradecimientos

Quisiera agradecerle a mis directores, Martha Bohorquez y Rafael Renteria por su valiosa orientación y conocimientos que han sido fundamentales para la elaboración de este proyecto. Su indicaciones y sugerencias definieron la metodología que finalmente este trabajo tendría.

A mi mamá, a mi hermana y mi cuñado, por el constante apoyo, amor y cuidado que me han brindado durante todo el proceso de la maestría. Gracias a su paciencia y comprensión en los momentos más difíciles, fueron posibles todos los logros alcanzados.

Finalmente quisiera agradecer a todas esas personas y amigos que me han acompañado y apoyado durante todo este proceso. En especial quiero agradecer a Santiago, Steven, Gabriel, Nata, Atenea, Lau y Cami. El apoyo recibido por ellos durante los momentos más difíciles en el transcurso de este proceso, junto con su preocupación, constante ánimo, risas y compañía, me llenaron de motivación y fuerzas en los momentos en los que más lo necesité para lograr este objetivo. Sin ellos este trabajo no hubiese sido posible.

Resumen

Estudio de los robos en Pereira mediante la detección y dinámica de patrones en redes espacio temporales

Este trabajo presenta un análisis espacio-temporal del crimen en la ciudad de Pereira, Colombia, mediante un modelamiento por medio de redes complejas. Se construye una red de eventos criminales considerando cada crimen como un nodo y las relaciones espaciales y temporales entre ellos como aristas. Los datos utilizados para este estudio consisten en hurtos reportados al departamento de policía local entre 2018 y 2021. Se estudia la estructura de la red mediante la identificación de patrones recurrentes, conocidos como motifs, que describen comportamientos emergentes del crimen. Para mejorar la eficiencia computacional para la detección de motifs, se propone una metodología que utiliza la estructura de estos subgrafos que optimiza el conteo dentro del algoritmo. Se utilizan algoritmos de clusterización en los datos para identificar las áreas que tienen patrones espacio-temporales similares de delincuencia. Se observó que una distribución log-normal se ajustaba adecuadamente a la distribución del grado de la red de eventos, permitiendo definir la distancia en la que dos sucesos pueden estar relacionados en el espacio. Los resultados muestran que la metodología propuesta es eficaz en la identificación de motifs que capturan patrones espacio-temporales del hurto de personas. Este estudio demuestra la utilidad del análisis de redes en la modelización del crimen y aporta ideas que pueden servir de base a las estrategias de prevención de la delincuencia.

Palabras clave: *Redes Complejas, Redes de Eventos, Motifs, Datos Espacio-Temporales, Clusterización, Centralidad.*

Abstract

Study of thefts in Pereira through the detection and dynamics of patterns in spatial-temporal networks

This work presents a spatio-temporal analysis of crime in the city of Pereira, Colombia, using a complex network modeling approach. A network of criminal events is constructed considering each crime as a node and the spatial and temporal relationships between them as edges. The data used for this study consists of thefts reported to the local police department between 2018 and 2021. The network structure is studied by identifying recurring patterns, known as motifs, which describe emergent crime behaviors. To improve computational efficiency for motif detection a methodology is proposed that uses the structure of these subgraphs to optimize the count within the algorithm. Clustering algorithms are used on the data to identify areas that have similar spatio-temporal patterns of crime. A log-normal distribution was found to adequately fit the event network degree distribution, allowing to define the appropriate distance in which two events can be related in space. The results show that the proposed methodology is effective in identifying motifs that capture spatio-temporal patterns of crime. This study demonstrates the usefulness of network analysis in crime modeling and provides insights that can inform crime prevention strategies.

Keywords: *Complex Networks, Event Networks, Motifs, Spatio-temporal Data, Clustering, Centrality.*

Índice general

Índice de figuras **xiii**

Índice de tablas **xv**

1 Introducción **1**

2 Marco teórico **5**

2.1 Datos espaciales y crimen 5

2.1.1 Conceptos básicos 5

2.1.2 Sistema coordenado 6

2.1.3 Criminología ambiental 10

2.2 Herramienta de visualización del crimen espacial 11

2.3 Redes complejas 12

2.3.1 Conceptos básicos 14

2.3.2 Medidas en redes 16

2.3.3 Algunos ejemplos de redes complejas 18

2.3.4 Redes espacio-temporales: redes de eventos 21

2.3.5 Subgrafos y motifs 23

2.4 Clusterización de datos 24

2.4.1 *k*-means 25

2.4.2 DBSCAN 27

2.5 Método ERGEN para identificación de subgrafos 30

2.5.1 Propuesta: formula de Stirling 31

3 Metodología **35**

3.1 Metodología del caso de estudio 35

3.1.1 Análisis descriptivo y exploratorio 37
3.1.2 Clusterización 37
3.1.3 Distancia entre eventos 38
3.1.4 Construcción de la red de eventos 38
3.1.5 Identificación de subgrafos 39
3.1.6 Análisis de las redes de eventos 39
3.2 Implementación computacional 39

4 Análisis de resultados **45**
4.1 Descripción de los datos 45
4.2 Contextualización de la ciudad de estudio: Pereira 47
4.3 Análisis exploratorio de los datos 51
4.4 Clusterización 57
4.4.1 *k*-means 58
4.4.2 DBSCAN 59
4.4.3 Elección del método 60
4.5 Red de eventos y subgrafos 63
4.5.1 Selección de la distancia entre eventos 63
4.5.2 Construcción de la red y conteo de subgrafos: Contribución computacional 67
4.5.3 Resultados y generalización de conteo de motifs 71
4.5.4 Medidas de centralidad en las redes de eventos 73

5 Conclusiones y trabajo futuro **77**
5.1 Conclusiones 77
5.2 Trabajo futuro 81

Bibliografía **83**

Índice de figuras

2.1 Dimensiones de los datos en un SIG [1] 12
2.2 Ejemplo de diferentes redes complejas 13
2.3 Ejemplo de subgrafos 23
2.4 Motifs de estudio en el trabajo propuesto 24
2.5 Ejemplo de *k*-means 27
2.6 Ejemplo de DBSCAN con MinPts $= 2$ y ε dado por el radio de los círculos observados 28
2.7 Ejemplo de DBSCAN para diferentes valores de MinPts y ε sobre datos normales multivariados 29

3.1 Metodología a implementar en el caso de estudio 36

4.1 Número de crímenes por: día, semana, mes. En rojo las semanas y meses en donde no se tienen datos 47
4.2 Mapa de Pereira y Dosquebradas 48
4.3 Mapa temático de la criminalidad en Pereira por Barrios 49
4.4 Marcación espacial de los crímenes en la ciudad de Pereira 50
4.5 Mapas temáticos de los crímenes en Pereira para distintas fechas 51
4.6 Pirámide poblacional de los crímenes en Pereira 52
4.7 (a) Número de casos por hora del delito. (b) Número de casos por día de la semana y parte del día 53
4.8 Porcentaje del total de casos en el mes por parte del día 54
4.9 Distribución del porcentaje de casos por tipo de modalidad y parte del día . 55
4.10 Mapas temáticos por modalidades de crimen 56
4.11 *k*-means con distinto número de clusters 58
4.12 Método del *elbow* para *k*-means 59

4.13 Método del *elbow* para DBSCAN 60
4.14 Resultados de los clusters obtenidos por DBSCAN 61
4.15 Resultado final por DBSCAN aplicado a todos los datos 62
4.16 Distribución del grado de las redes generadas vs distribución log normal . . 64
4.17 Ejemplo de diferentes valores de radio espacial sobre Pereira 67
4.18 (a) Grafo construido para la fecha 2019-05-01. (b) Grafo construido para la fecha 2020-06-01 68
4.19 (a) Conteo de motifs por medio de la metodología propuesta por Forero A., et al.[2] (b) Conteo de motifs mediante la metodología propuesta. 70
4.20 (a) Motfis hasta diciembre de 2019.(b) Motifs desde julio de 2020 70
4.21 Grafo construido para la fecha 2019-05-01 con atributos para (a) centralidad de intermediación (b) Centralidad de cercanía 74

Índice de tablas

4.1 Número de casos para las distintas modalidades de crímenes reportados en Pereira y Dosquebradas. 56
4.2 Valor p para las distintas redes y diferentes distancias 64
4.3 p-valor medio para cada una de las distancias 65
4.4 Muestra de los conteos para las distintas permutaciones dentro de cada mes 69
4.5 Valor-p de la prueba de Dickey-Fuller aumentada 71
4.6 Matriz de correlación de las series de tiempo estandarizadas. 72
4.7 Asortatividad de las redes de eventos generadas para una distancia de $D = 500$. 75

Capítulo 1

Introducción

La seguridad ciudadana es un tema de gran interés para el desarrollo y bienestar de las comunidades, por lo cual estudiar la criminalidad se vuelve necesario ya que es un problema recurrente en muchas ciudades. Por este motivo, durante los últimos años se han desarrollado distintos estudios para abordar esta problemática desde perspectivas tanto sociales como económicas. Dentro de las metodologías que se han utilizado para el estudio y mitigación de esta problemática, aparece la ciencia de redes, que permite modelar esquemas de sistemas complejos por medio de grafos. En este trabajo de grado se aborda un caso de estudio sobre la criminalidad en la ciudad de Pereira, Colombia, utilizando redes espacio-temporales. Se busca identificar factores y patrones espacio-temporales que contribuyan a su incidencia. Para lograr esto, se realiza el modelamiento por medio de redes de eventos [3] de tal forma que se pueda relacionar tanto la dimensión espacial como la temporal de los crímenes, junto con un un análisis descriptivo y exploratorio que permita dar un contexto general del crímen en la ciudad. Además, se realiza una clusterización de los datos que se sustenta en la teoría de la Criminología Ambiental. Con el generador empírico aleatorio de eventos de red (*Empirical Random Network Generator* o ERGEN) [2], se identificarán patrones de comportamiento del crimen en la ciudad, que junto con medidas de centralidad de las redes de eventos, permitirá hacer recomendaciones a las autoridades locales para la toma de decisiones y la implementación de políticas públicas efectivas para mitigar este problema.

Resulta interesante realizar un modelamiento de la criminalidad por medio de redes complejas, ya que para los propósitos de este trabajo, se puede identificar que la distribución espacial de eventos criminógenos corresponderán a los nodos de la red. Al ser eventos espacio-temporales, las reglas lógicas que se definan para la conexión entre los nodos (definición del radio espacial y temporal) deben tener en cuenta estas dimensiones. Dependiendo

del tipo de reglas que se definan, se pueden generar distintos tipos de redes con diferentes características. En la literatura es posible encontrar propuestas metodológicas para agregar la dimensión temporal a las redes. Por ejemplo, en las redes multicapa [4], en donde cada conjunto de eventos que ocurren dentro de un determinado tiempo se configuran dentro de una capa diferente de la red, agregando de esta forma el factor temporal de los datos. Otra metodología corresponde al modelamiento por medio de redes de eventos y la búsqueda de patrones que caractericen el problema de estudio [3]. Esta última ha demostrado ser eficaz en caracterizar los sistemas de estudio, ya que la búsqueda de patrones en el espacio-tiempo permite entender las dinámicas generadas por los agentes en la red [5].

Uno de los principales desafíos en el estudio de redes espacio-temporales es la identificación de patrones como subgrafos con topologías específicas. Se han desarrollado un gran número de trabajos enfocados en la identificación de dichos patrones que permitan caracterizar la dinámica del sistema [3, 6]. Con lo cual la creación de algoritmos altamente eficientes computacionalmente para dicha detección es uno de los mayores retos en esta área [2, 7, 8].

La propuesta metodológica en este trabajo consiste en realizar un análisis descriptivo y exploratorio de la base de datos, teniendo en cuenta el aspecto espacial de los mismos. Este análisis tiene como objetivo entender las características principales del crimen en la ciudad de estudio. Luego se realiza una clusterización de los datos para determinar si existen diferencias en la criminalidad en la ciudad, ya que los crímenes tienden a clusterizarse debido al contexto social y económico de la zonas de ocurrencia. Una vez identificados los clusters, se modela el problema como una red de eventos. Para esto, es necesario definir la distancia para la cual dos nodos estarán conectados. Por lo tanto, se hace un estudio de la topología de la red para distintas distancias y mediante pruebas de bondad de ajuste se busca cuál de estas se ajusta mejor a resultados en la literatura. Finalmente, se realiza un análisis estadístico de los subgrafos generados por la red con métodos computacionales para su identificación y comportamiento, que junto con medidas de centralidad de las redes generadas, permitirá crear propuestas de valor para la toma de decisiones en busca de mejorar la seguridad de la población.

En este trabajo, dicha metodología se aplicará sobre los datos para el crimen en la ciudad de Pereira y el municipio de Dosquebradas. Adicionalmente, en el trabajo se realiza una contribución para mejorar la robustez y eficiencia computacional de los algoritmos para la identificación de subgrafos, siendo un aporte de valor sobre los métodos ya existentes. Esta metodología tiene especial importancia ya que no solamente se enmarca en el estudio de

la criminalidad de Pereira. Además de ser reproducible para el crimen de otras ciudades, también puede funcionar para cualquier tipo de sistema que pueda ser modelado como una red espacio-temporal. Por ejemplo, podrá ser reproducible para problemas en epidemiología, dinámicas poblaciones, genética, física, entre otras áreas, donde la caracterización de la red permita entender más a fondo el sistema de estudio.

Este trabajo se organiza de la siguiente manera: en el segundo capítulo se presenta el marco teórico, en donde se revisa la teoría necesaria y los conceptos básicos para trabajar con datos espaciales, la identificación de subgrafos, el modelamiento de redes y la clusterización de los datos; en el tercer capítulo se presenta la metodología del proyecto, sustentada en los métodos y la teoría establecida en el marco teórico; en el cuarto capítulo se presentan los resultados al aplicar la metodología propuesta sobre los datos; finalmente, en el quinto capítulo se presentan las conclusiones del modelamiento anteriormente realizado, junto con el trabajo futuro.

Este trabajo se desarrolla principalmente en Python, ya que en gran parte este caso de estudio requiere de un componente computacional. Todos los resultados acá mostrados, incluidos los análisis descriptivos y exploratorios junto con los códigos para el modelamiento del problema como una red de eventos y la identificación de patrones, se podrán consultar en un repositorio que se menciona en el tercer capítulo del proyecto. Todos los requerimientos necesarios para poder duplicar el proyecto estarán listados en dicho repositorio.

Capítulo 2

Marco teórico

En este capítulo se presenta el marco teórico necesario para el caso de estudio de este proyecto. Se introducen los conceptos estadísticos necesarios para el manejo de datos espacio-temporales, junto con los principales conceptos de las redes complejas y como éstos se relacionan con datos espacio-temporales mediante el concepto de redes de eventos. Adicionalmente, se menciona el concepto de subgrafo y se plantea una metodología para su identificación computacional. Por último, se aborda la teoría de dos métodos de clasificación no supervisada para la clusterización de datos espaciales.

2.1 Datos espaciales y crimen

En la historia el crimen se ha estudiado principalmente por medio de la criminalística y la criminología, en donde la primera se encarga de estudiar la escena en donde ocurre el hecho, y la segunda se encarga de estudiar los aspectos sociales y psicológicos asociados al crimen. No obstante, un crimen es un suceso con una ubicación espacial puntual georreferenciada, lo cual ha permitido durante las últimas décadas usar estos datos espaciales para examinar patrones y procesos que ayuden a identificar zonas con un mayor riesgo en un determinado lugar. Para esto, es importante mencionar algunos conceptos básicos que permiten el análisis del crimen por medio de datos geoespaciales, junto con herramientas que permitan dar a estos datos un contexto social.

2.1.1 Conceptos básicos

En primer lugar, se define un proceso espacio-temporal como un proceso estocástico:

$$\{Z(\boldsymbol{s},t) : (\boldsymbol{s},t) \in D_{\boldsymbol{s}} \times D_t\} \tag{2.1}$$

Donde el conjunto $D_{\boldsymbol{s}} \subset \mathbb{R}^d$ es el conjunto índice de la ubicación espacial $\boldsymbol{s}$ y el conjunto $D_t \subset \mathbb{R}$ es el conjunto índice del tiempo t. Teniendo en cuenta que el estudio de los crímenes en este trabajo se realiza por medio de su ubicación espacial fija, este corresponderá a un proceso espacial puntual, en donde ambos conjuntos D_s y D_t serán discretos.

Un proceso espacial puntual corresponde a un conjunto de ubicaciones espaciales observadas llamadas eventos, en donde el evento $\boldsymbol{s}_i = (x,y), i = 1,\ldots,k$ y $\boldsymbol{s}_i \in D_{\boldsymbol{s}} \subset \mathbb{R}^2$ indica la ubicación del i-ésimo evento en una ventana de observación.

Cuando se habla de la ubicación espacial del evento $\boldsymbol{s}_i$, se debe tener en cuenta que el sistema coordenado que será utilizado permite identificar la **ubicación** de estos. Esto ayudará a darle un contexto a los datos de estudio, además de tener que ser universal para que cualquier persona sea capaz de reproducir las ubicaciones espaciales de la misma manera en que el estudio se realiza. En la sección 2.1.2 se presentan diferentes tipos de sistemas coordenados y además se menciona el que será utilizado en este trabajo.

Si bien los procesos espaciales puntuales corresponden a una de las grandes áreas de estudio de la estadística espacial, en este trabajo se realizará un análisis desde una perspectiva de las redes complejas cuyo marco teórico se abarca en la sección 2.3, por lo cual no se abarcarán métodos propios de la estadística espacial para este tipo de datos.

2.1.2 Sistema coordenado

Con el objetivo de poder realizar un análisis espacial, es importante tener en cuenta la forma en que se ubican los distintos eventos en un mapa. Para esto, se utiliza un sistema coordenado de referencia, lo cual permite identificar cualquier lugar de la tierra mediante vectores desde un punto origen. En general, un sistema coordenado de referencia se puede dividir en: un sistema de coordenadas geográficas (*geographic coordinate system* o GCS), que mide las posiciones en la tierra directamente como longitudes λ y latitudes ϕ teniendo en cuenta un sistema coordenado ortogonal esférico o elíptico; y un sistema plano coordenado, el cual es una representación de las ubicaciones de la tierra mediante coordenadas cartesianas sobre una superficie plana.

Un sistema de coordenadas ortogonal corresponde a un conjunto de coordenadas $\boldsymbol{q} = (q_1, q_2, \ldots, q_d)$ en donde cada uno de los vectores tangentes $\boldsymbol{e}_i$ a las curvas coordenadas son ortogonales entre sí. Utilizar un sistema de coordenadas ortogonal permite hacer transformaciones con coordenadas cartesianas mediante ecuaciones paramétricas, es decir:

$$q_i = f_i(x, y, z)$$

Y equivalentemente:

$$x = x(q_1, \ldots, q_d) \quad y = y(q_1, \ldots, q_d) \quad z = z(q_1, \ldots, q_d) \tag{2.2}$$

Esto permite que muchos problemas puedan ser resueltos mediante separación de variables y que, bajo cualquier sistema coordenado utilizado que denote el mismo espacio físico, el valor o atributo que tiene ese punto debe ser siempre el mismo. Adicionalmente, en teoría de transformaciones, se puede demostrar cómo este tipo de coordenadas tienen la propiedad de invarianza para campos escalares y vectoriales. Los vectores $\boldsymbol{e}_i$ se obtienen mediante la variación de un vector posición de un punto con respecto a una sola coordenada. Es decir, teniendo en cuenta que en el espacio euclídeo se puede denotar la posición mediante coordenadas cartesianas como:

$$\boldsymbol{r} = x\hat{\boldsymbol{i}} + y\hat{\boldsymbol{j}} + z\hat{\boldsymbol{k}}$$

Y teniendo en cuenta que estas coordenadas son función del nuevo sistema coordenado ortogonal, entonces:

$$\boldsymbol{e}_i = \frac{\partial \boldsymbol{r}}{\partial q_i} \tag{2.3}$$

En donde el vector obtenido no necesariamente es unitario, por lo cual se utilizan factores de escala h_i de tal forma que el vector unitario asociado a la coordenada q_i será:

$$\hat{\boldsymbol{e}}_i = \frac{1}{h_i}\frac{\partial \boldsymbol{r}}{\partial q_i}$$

Usando (2.3):

$$\hat{\boldsymbol{e}}_i = \frac{\boldsymbol{e}_i}{h_i} = \frac{\boldsymbol{e}_i}{|\boldsymbol{e}_i|}$$

Y al tratarse de un sistema de coordenadas ortogonal, se tiene la siguiente propiedad:

$$\hat{\boldsymbol{e}}_i \cdot \hat{\boldsymbol{e}}_j = \delta_{ij}$$

Con δ_{ij} la función delta de Kronecker que se define como $\delta_{ij} = 0$ si $i \neq j$ y $\delta_{ij} = 1$ si $i = j$. Al tratarse de sistemas que proyectan ubicaciones en la tierra mediante algún sistema coordenado ortogonal, esto implica que no se trata de ubicaciones exactas, ya que la tierra no tiene una forma parametrizable, por lo cual siempre existirán incertidumbres en la medición que serán más grandes o pequeñas dependiendo del área de medición y el sistema utilizado, ya que corresponde a la proyección de la tierra mediante un sistema muchas veces simétrico. Por este motivo, utilizar un sistema plano coordenado mediante coordenadas cartesianas no está mal si el área de extensión de estudio no es muy extensa, ya que las diferencias con un sistema elíptico o esférico no serán importantes, siempre y cuando los datos de estudio así lo permitan.

En este trabajo se utilizará el sistema WGS 84, el cual es un sistema de coordenadas geográficas que identifica las ubicaciones en la tierra mediante longitudes y latitudes medidas en grados mediante un sistema coordenado que modela a la tierra mediante un esferoide oblato, que se obtiene a través de las coordenadas elípticas por rotación alrededor del eje y. Las coordenadas con las que se identifican una posición en el espacio corresponden a (ϕ, λ, ξ), en donde la latitud $\phi : -\pi/2 \longrightarrow \pi/2$, la longitud $\lambda : 0 \longrightarrow 2\pi$, y la altitud $\xi : 0 \longrightarrow \infty$, y con reglas de transformación a un sistema coordenado cartesiano:

$$x = a\cosh\xi\cos\phi\cos\lambda \qquad y = a\cosh\xi\cos\phi\sin\lambda \qquad z = a\sinh\xi\sin\phi \tag{2.4}$$

En donde a corresponde al punto focal del elipsoide. Para el caso particular de la tierra, se tiene que la coordenada ξ será constante. En el caso en que sea necesario realizar una medición mediante una distancia euclidiana, hay que considerar que el rango de la latitud varía entre los polos de la tierra y la longitud para el meridiano, por lo cual una distancia euclidiana no es recomendable para este tipo de coordenadas sin realizar algún tipo de transformación a coordenadas cartesianas. No obstante, se podría considerar su uso en ubicaciones espaciales cercanas y cerca al Ecuador, o en casos en que no es necesario una gran precisión. Sin embargo, cerca a los polos una distancia euclidiana no será de utilidad. Por este motivo, para ubicaciones espaciales con latitudes y longitudes, se puede utilizar la distancia de Haversine que facilita el cálculo de una distancia sin necesidad de realizar transformación a coordenadas cartesianas.

La distancia de Haversine está determinada por la siguiente ecuación:

$$\mathrm{Hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) \tag{2.5}$$

donde θ corresponde a $\frac{d}{r}$, donde d es la distancia entre dos puntos dentro de un radio de una esfera n-dimensional, y r es el radio de dicha esfera. Para el caso particular de longitudes y latitudes, la ecuación puede ser escrita como:

$$\mathrm{Hav}(\theta) = \mathrm{Hav}(\phi_1 - \phi_2) + [1 - \mathrm{Hav}(\phi_1 - \phi_2) - \mathrm{Hav}(\phi_1 + \phi_2)]\,\mathrm{Hav}(\lambda_1 - \lambda_2) \tag{2.6}$$

donde ϕ_i y λ_i, corresponden a las latitudes y longitudes, respectivamente, de los puntos 1 y 2. Finalmente, se resuelve esta ecuación para encontrar la distancia d entre los dos puntos mediante la inversa de Haversine, es decir:

$$d = 2r \arcsin\left[\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos\phi_1 \cos\phi_2 \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right] \tag{2.7}$$

Es importante considerar que la distancia de Haversine es una fórmula matemática que se utiliza para calcular la distancia entre dos puntos en una superficie esférica. Es especialmente

útil para aplicaciones en las que se requieren cálculos precisos de distancias cortas y en las que las diferencias de altitud son mínimas, por lo que es importante seleccionar cuidadosamente la elipsoide que se va a utilizar en cada caso. Sin embargo, no es adecuada para situaciones en las que se requieren cálculos de distancias más largas ya que, como se mencionó anteriormente, WGS84 asume que la tierra es una esferoide oblato. En estos casos, para tener una mayor precisión y menor incertidumbre, se recomienda utilizar la formula de Vincenty [9].

2.1.3 Criminología ambiental

En la criminología ambiental se busca estudiar los distintos aspectos y configuraciones en el que el crímen ocurre [10], contrario a la criminología tradicional en donde se busca explicar la causa raíz del crimen y por qué las personas se convierten en criminales. Patrones puntuales con ubicación espacial $\boldsymbol{s}_i$, configurados dentro de un lugar por medio de un sistema coordenado universal, permiten que el estudio por medio de la criminología ambiental adquiera un sentido importante, ya que se estudia el crímen no solo por su configuración espacial sino por los **atributos** del mismo. Como mencionan Felson y Boba [11], algunas de estas configuraciones contienen mayores oportunidades para que un crimen ocurra y la criminología ambiental permite entender como estas oportunidades son creadas y luego se clusterizan. En particular, para esta última idea en la sección 2.4 se explora la teoría para la clusterización de datos que permitan observar dicho comportamiento, no solo desde el análisis social y psicológico, sino también desde el análisis de datos y como esto contribuye en el análisis del crimen.

Un resultado interesante de criminología ambiental que estudia como dos crímenes están relacionados en un lugar y una distancia determinada, viene de la Prevención de la Delincuencia Mediante el Diseño Ambiental (*Crime prevention through environmental design* o CPTED), la cual es una teoría que está basada en la Yuxtaposición Geográfica (*Geographic Juxtaposition* o GJ), que corresponde al entorno próximo al lugar del delito. Cozens et al. [12] señala que el concepto de Yuxtaposición Geográfica sirve para comprender una gama mucho más amplia de crímenes que se ven influenciados por algún actor dependiendo de la lejanía al lugar del delito. Cozens et al. [12] determina los impactos que tiene esta distancia en CPTED. Para esto, definen las siguientes formas de GJ: Micro GJ, el cual corresponde a factores GJ que actúan en el lugar del delito; proximal GJ, que corresponde a factores que actúan desde lugares próximos o contiguos al lugar del delito; Meso GJ, que corresponde a factores originados en zonas desde la más próxima hasta las más alejadas del lugar del delito; y Macro GJ, que corresponde a factores que actúan como influencias remotas sobre el delito sin importar su ubicación. En este sentido, la construcción de la red y posterior

aparición de subgrafos como se mencionará a continuación, permiten establecer mecanismos de prevención de acuerdo a las formas de GJ.

En el trabajo realizado por Sanabria, Bohorquez y Renteria [2] se explora cómo el CPTED permite explicar cómo los crímenes que ocurren forman subgrafos y motifs, lo cual se abarca en la sección 2.3.5, como una interacción entre las dimensiones espacio-temporales de los datos, y los aspectos sociales y geográficos del lugar de estudio. En particular, este trabajo será una guía metodológica importante para el desarrollo de este caso de estudio.

2.2 Herramienta de visualización del crimen espacial

Para poder trabajar visualmente con datos con una ubicación espacial, y además poder asociarlos con el contexto social, económico y ambiental en el que se encuentran los eventos de estudio y realizar su respectivo análisis, normalmente se utiliza Sistema de Información Geográfica (SIG) o más comúnmente GIS por sus siglas en inglés. En un sentido muy amplio, GIS corresponde a un conjunto de herramientas que permiten realizar visualizaciones interactivas para analizar información georreferenciada mediante la edición de datos y mapas. No obstante, en este trabajo solo se mencionarán los conceptos necesarios para poder identificar los eventos del caso de estudio en un mapa, y cómo estos mapas pueden dar información de valor para la toma de decisiones.

Para representar un evento geográfico en el espacio con un alto nivel de precisión, es necesario mencionar los siguientes elementos geométricos:

- **Puntos:** Un punto corresponde a las parejas de coordenadas espaciales en un sistema coordenado previamente definido. Este punto representa el evento, que para este caso de estudio corresponderá a los crímenes. Como se mencionó anteriormente, desde la estadística espacial, estos puntos forman patrones puntuales, lo cual será el objeto de análisis.

- **Lineas:** Una línea corresponde a puntos conectados que pueden representar carreteras, calles, ríos, etc.

- **Polígonos:** Un polígono es un objeto bidimensional que representa un área geográfica en particular. Como se verá más adelante, estos polígonos representarán comunas o barrios de la ciudad de Pereira. Estos polígonos permiten obtener información de interés, ya que es posible obtener el valor del perímetro o el área.

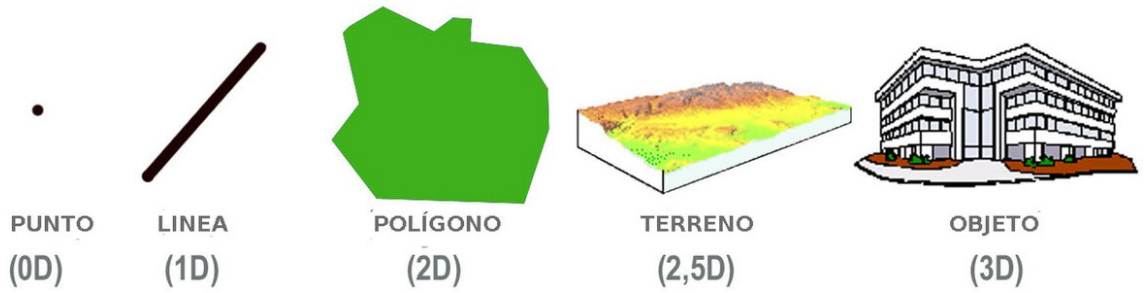


Figura 2.1 Dimensiones de los datos en un SIG [1]

En la figura 2.1, se presenta un ejemplo de cómo se visualizan estos elementos, junto con dos elementos adicionales que no se abarcarán en este trabajo.

Un mapa que presente estos objetos espaciales como puntos o polígonos, permiten una primera y rápida identificación de patrones y outliers para el caso de estudio. No obstante, existen otras herramientas como los mapas temáticos, los cuales permiten representar la distribución espacial de algunas estadísticas o atributos de los crímenes. Por ejemplo, en el capítulo 4 se presenta un mapa temático del número de casos de los crímenes por barrio. Adicionalmente, se utilizará un mapa de referencia que permite dar un contexto social y económico más amplio a los datos, ya que este mapa muestra los objetos espaciales que componen el lugar de estudios, como lo son las calles, los ríos, las montañas, etc. En el capítulo 3 de metodología se explorará a fondo la manera en que estas representaciones pueden ser realizadas en Python para el caso de estudio.

2.3 Redes complejas

De manera general, una red compleja es un sistema formado por una colección de elementos conectados entre sí, representados mediante vértices o nodos (puntos) y enlaces (líneas). En la figura 2.2 se presentan ejemplos de lo que puede definirse como una red compleja. Desde un punto de vista matemático, las redes están representadas por grafos y la teoría de grafos es la herramienta matemática utilizada para estudiar y analizar estas estructuras. Conceptualmente, los vértices o nodos representan los elementos individuales de la red, mientras que los enlaces o aristas representan las relaciones o conexiones entre ellos. Si bien según el área de estudio estos puntos y líneas pueden tener nombres diferentes, durante el desarrollo de este trabajo se utilizará indistintamente los términos vértices o nodos para hacer referencia a los elementos de las redes de estudio.

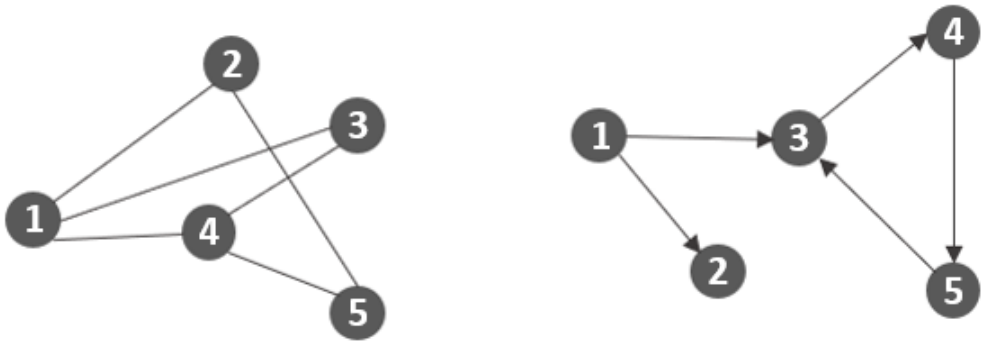


Figura 2.2 Ejemplo de diferentes redes complejas

Las redes complejas han cobrado especial importancia gracias a su gran capacidad de adaptación en diferentes temáticas. Sus aplicaciones van desde la física [13], epidemiología, formación de opinión [14], o cualquier sistema que pueda ser modelado a través de nodos y enlaces [15–20]. Si bien sus orígenes se remontan a principios del siglo XIX, en donde las principales metodologías y resultados se pueden encontrar en los trabajos recopilatorios realizados por Freeman[21] y Wasserman[22], no ha sido sino hasta los últimos 20 años en donde el estudio de la complejidad ha cobrado especial importancia, centrándose principalmente en los comportamientos emergentes en las redes complejas, lo cual presenta efectos de interés. Actualmente existen una gran diversidad de libros en donde se recopila tanto la parte histórica, como los fundamentos conceptuales, teóricos y aplicaciones de las redes [15, 20, 18, 23, 16, 17]. En particular, esta sección se construirá principalmente de los trabajos realizados por Newman [15], Estrada [18] y Latora [20].

En la siguiente sección se abarcarán los conceptos básicos necesarios para realizar un estudio de las redes complejas desde el punto de vista matemático. Es importante mencionar que la notación utilizada en redes complejas no tiene un consenso y puede variar entre autores.

Nota: *Esta sección se construye sobre el marco teórico presentado en la tesis de pregrado en física de Miguel Ángel Quintero titulada "Ising Model applied to Complex Networks: Opinion formation in dialectic processes" [24], que a su vez está basada principalmente en los autores anteriormente mencionados. Adicionalmente, se complementan y se reorganizan los conceptos que son necesarios para este proyecto en particular.*

2.3.1 Conceptos básicos

Una red compleja o grafo G consiste en un conjunto de nodos $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ y un conjunto de parejas ordenadas $\mathscr{L} = \{(v_i, v_j)\} \subset \mathcal{V} \times \mathcal{V}$, donde cada pareja ordenada (v_i, v_j) se llama conexión dirigida del nodo v_i al nodo v_j. Se dice que la red es no dirigida si para cada pareja $(v_i, v_j) \in \mathscr{L}$ también existe la pareja $(v_j, v_i) \in \mathscr{L}$, de lo contrario, la red se denomina dirigida. En la figura 2.2, se presenta un ejemplo de una red dirigida (derecha) y una no dirigida (izquierda). Todos los nodos conectados a un nodo v_i serán llamados los vecinos del nodo v_i.

Para representar un grafo, se define la matriz de adyacencia A, la cual corresponde a la matriz de elementos a_{ij} que determina la forma en que los nodos están conectados entre sí. Para N nodos en la red, el tamaño de esta matriz será de $N \times N$. Matemáticamente, A se define de la siguiente manera:

$$a_{ij} = \begin{cases} 1 & \text{si hay una conexión entre el nodo } i \text{ y el nodo } j \\ 0 & \text{en caso contrario} \end{cases}$$

Utilizando los ejemplos de la figura 2.2, la matriz de adyacencia del grafo no dirigido (izquierda) correpsonde a:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

En donde se puede observar que la matriz es simétrica, lo cual es una propiedad de todos los grafos no dirigidos. Para el caso del grafo dirigido (derecha) se tiene que la matriz de adyacencia corresponde a:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

En donde contrario al caso anterior, la matriz obtenida no es simétrica. Adicionalmente, en ambos casos la diagonal de las matrices es cero, debido a que en este tipo de grafos no se presenta una conexión de un nodo con el mismo; no obstante, pueden existir grafos con esta característica.

Otro concepto corresponde a grafos ponderados o con pesos. En estos casos los enlaces de la red tendrán un valor. En la red de adyacencia esto se puede representar mediante los valores a_{ij} los cuales pueden ser diferentes a 1. Este concepto es importante porque nos permite observar que los vínculos, e incluso los nodos, pueden tener atributos que en este caso corresponden a un peso, pero también pueden tener características cualitativas. Existen otros tipos de grafos, como hypergrafos, redes bipartitas, redes cíclicas y acíclicas, entre otros. No obstante, para este trabajo no se abarcarán este tipo de redes.

Gran parte de los estudios realizados en el campo de las redes complejas corresponde al estudio de la estructura que estas poseen, es decir, a la forma en que los nodos están conectados entre ellos, lo cual recibe el nombre de **topología de la red**. Varios trabajos se han realizado en esta área, por ejemplo los estudios realizados por Erdös y Rényi en 1959[25] sobre redes aleatorias, los modelos de Watts y Strogatz con respecto a las redes de mundo pequeño en 1998 [26] o el trabajo de Redner, Albert y Barabasi en redes libre de escala [27]. La topología de la red es importante porque se ha evidenciado que dependiendo del área de estudio, la red generada se puede enmarcar en alguna topología en particular, como las anteriormente explicadas.

Teniendo en cuenta esto, una de las principales definiciones corresponde a la conectividad o el grado de la red. La conectividad de un nodo corresponde al número de vecinos que este tiene, por lo cual el grado de una red es el valor medio sobre todas las conectividades:

$$\langle \mu \rangle = \frac{1}{N} \sum_{i=1}^{N} \mu_i \tag{2.8}$$

donde μ_i es la conectividad o el grado del nodo i. Otra definición corresponde a la accesibilidad entre dos nodos diferentes en un grafo independiente de que éstos no sean vecinos. Dos nodos en una red pueden acceder entre ellos mediante algún camino. Un camino $W(v_i, v_j)$ de un nodo v_i a un nodo v_j es una secuencia de nodos y enlaces empezando por el nodo v_i y

terminando en el nodo v_j. La distancia o distancia geodésica en un camino corresponde al número de enlaces en esa secuencia. Un camino simple o *path* es un camino en donde los nodos son visitados una única vez.

2.3.2 Medidas en redes

Existen diferentes tipos de medidas en redes complejas cuyo uso dependerá del tipo de estudio que se esté realizando. En esta sección se presenta un pequeño resumen de algunas medidas de centralidad y otro tipo de medidas para redes.

Centralidad de una red

La centralidad de una red permite conocer cómo y cuán importante son los nodos que la componen. La importancia dependerá de la definición dada en el estudio, ya que, a modo de ejemplo, ésta puede corresponder al número de enlaces que un nodo tiene, o puede corresponder a los nodos que están conectados a otros nodos con muchos enlaces. En esta sección se revisarán las medidas más utilizadas o históricamente relevantes, ya que existen cientos de este tipo de medidas y que su uso dependerá siempre del caso de estudio.

La primera medida corresponde a **Centralidad de Grado**. Esta corresponde al grado de un nodo, y está definida por:

$$C_{deg}(v) = \frac{k_v}{|N| - 1} \tag{2.9}$$

Donde N es el conjunto de nodos de la red, $|N|$ es el número de nodos de la red y k_v es el grado del nodo v.

La segunda medida es la centralidad de cercanía. Esta corresponde al promedio de la distancias desde un nodo a todos los demás por medio del camino más corto. Es decir, un nodo tendrá un valor alto de centralidad de cercanía si este se encuentra conectado a nodos con una alta conectividad. Esta medida está dada por la siguiente expresión:

$$C_{close}(v) = \frac{|N| - 1}{\sum_{u \in N - v} d(v,u)} \tag{2.10}$$

donde $d(v,u)$ corresponde a la distancia entre los nodos v y u, y $N-v$ es el conjunto de nodos en la red sin el nodo sobre el cual se hace la medición.

Otra medida corresponde a la centralidad de intermediación. Esta medida cuantifica cuántas veces un nodo se encuentra entre los caminos más cortos del grafo. Por lo tanto, ésta se encuentra definida por la ecuación:

$$C_{btw}(v) = \sum_{s,t \in N} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}} \tag{2.11}$$

donde $\sigma_{s,t}$ corresponde al número de geodésicas entre el nodo s y t, y $\sigma_{s,t}(v)$ es el número de geodésicas entre el nodo s y t que cruzan por el nodo v.

Existen otras medidas de centralidad que históricamente han tenido una gran importancia como lo son la centralidad de vector propio y *Page Rank*; no obstante, para este trabajo no se revisarán los conceptos de estas medidas. Entre estas medidas, *Page Rank* ha tenido la mayor importancia, ya que es utilizada y patentada por *Google* desde 1999 [28] y fue la primera versión utilizada para su buscador web.

Otras medidas

Existen otras medidas que también son importantes para comprender la estructura de la red compleja de estudio. Al igual que las medidas de centralidad, existen cientos de este tipo de medidas. En este trabajo solo se citan algunas de estas de forma ilustrativa e introductoria, y debido a su relevancia histórica.

La primera corresponde al **Coeficiente de Jaccard**, que da el número de vecinos comunes entre dos nodos normalizado por el número total de vecinos, el cual está dado por la siguiente expresión:

$$C_{jacc_coeff}(v,u) = \frac{|\Gamma(v) \cap \Gamma(u)|}{|\Gamma(v) \cup \Gamma(u)|} \tag{2.12}$$

donde $\Gamma(v)$ corresponde al conjunto de los vecinos del nodo v. Otra medida igualmente interesante es ***Resource Allocation***. Esta medida indica la fracción de un "recurso" que un

nodo puede enviar a otro a través de sus vecinos comunes. Esta medida viene dada por la ecuación:

$$C_{res_alloc}(v,s) = \sum_{u \in \Gamma(v) \cap \Gamma(s)} \frac{1}{|\Gamma(u)|} \tag{2.13}$$

Por último, una medida que tiene en cuenta la estructura de las comunidades en una red para la predicción de enlaces es la de Soundarajan-Hopcroft. Ésta viene dada por:

$$C_{SH}(v,u) = |\Gamma(v) \cap \Gamma(u)| + \sum_{u \in \Gamma(v) \cap \Gamma(u)} f(s) \tag{2.14}$$

donde $f(s)$ es 1 si s está en la misma comunidad que v y u, y 0 en caso contrario.

Adicionalmente se tiene la asortatividad, la cual es una medida que indica la preferencia de un nodo a conectarse con otros nodos que presenten características similares, por lo general el valor de su grado. Un valor de la asortatividad cercano a -1 indica que la red es disortativa, un valor de 0 que es no asortativa, y un valor de 1 que es asortativa.

2.3.3 Algunos ejemplos de redes complejas

Existen muchos tipos de redes que se han definido para distintos problemas de interés y que cumplen con características diferentes. No obstante, algunas de estas tienen especial importancia dado que tienen muchas similitudes con muchos sistemas que se modelan como una red, algunos ejemplos son las redes aleatorias o las redes de mundo pequeño. En esta sección se presentan los resultados más importantes que describen las redes aleatorias y las redes libres de escala. En particular, se observan los modelos de Erdös-Rényi [25] y el modelo de Barabási-Albert [27]. En el caso de las redes de mundo pequeño, si bien estas han tenido una especial importancia histórica, no se abarcarán en este trabajo.

Redes aleatorias: El modelo de Erdös-Rényi

Matemáticamente, el modelo de Erdös-Rényi, o también llamado grafos aleatorios binomiales, denotado por $G^{ER}_{N,p}$, consiste en un grafo donde cada par de nodos se conectan entre ellos con una probabilidad p. Entonces, la probabilidad asociada a un grafo es una distribución Bernoulli de la forma:

$$P_G = p^d(1-p)^{m-d} \tag{2.15}$$

donde d es el número de enlaces dentro de la red y m es el número máximo de enlaces posibles dentro de la red, cuyo valor es $m = N(N-1)/2$. La probabilidad de tener cualquier red con la condición de que tenga exactamente d enlaces, a través del modelo de Erdös-Rényi, corresponde a:

$$p_d = \binom{m}{d} p^d(1-p)^{m-d} \tag{2.16}$$

Que corresponde a una distribución binomial. La ecuación (2.16) depende sólo de d a diferencia de la ecuación (2.15), que es la probabilidad de tener un grafo con un orden, nodos y d enlaces dados.

En las redes, suele ser más importante hacer una descripción a partir de la conectividad de los nodos, es decir, de la topología de la red; por lo tanto, para las redes siempre existe una distribución del grado. Para este modelo, la distribución de grado viene dada por:

$$p_k = \binom{N-1}{k} p^k(1-p)^{N-1-k} \tag{2.17}$$

donde k es el grado del nodo. Además, el valor esperado es:

$$\langle k \rangle = p(N-1) \tag{2.18}$$

Las ecuaciones (2.15), (2.16) y (2.17) corresponden a los conceptos básicos para definir una red aleatoria binomial o de Erdös-Rényi. En particular, esta red presenta un comportamiento particular para ciertos valores medios del grado de un nodo. Este comportamiento es llamado una transición de fase, y define los parámetros para los cuales se presenta un componente gigante dentro de la red, esto es, un conjunto de nodos con una conectividad muy alta dentro de la red.

Redes libre de escala: El modelo de Barabási-Albert

Bajo la evidencia de que muchos sistemas muestran muy pocos nodos con un grado muy alto y muchos nodos con un grado muy pequeño, Albert Lászlo Barabási y Réka Albert en 1999 [27] desarrollaron un modelo que muestra estas características. Esto se puede observar a través de la distribución del grado de la red, el cual está dado por una ley de potencias de la forma:

$$p_k = ck^{-\gamma}$$

donde k corresponde al grado del nodo, c una constante de proporcionalidad y γ el exponente. Este modelo define a las redes libre de escala o *scale-free*, en donde el valor de γ en la ecuación anterior dependerá del sistema de estudio.

Este modelo puede desarrollarse mediante el siguiente algoritmo. Sean n_t y l_t el número de nodos y enlaces en una red en el tiempo t. Para cada paso de tiempo se tiene que:

1. Un nuevo nodo se agrega a la red y éste viene con m enlaces.
2. Los m enlaces de este nodo se conectan con alguno de los nodos ya existentes del paso de tiempo anterior.

Así, la probabilidad de que un nuevo nodo u se conecte al nodo v en t está dado por:

$$\Pi_{u\to v} = \frac{k_{v,t-1}}{\sum_{l=1}^{n-1} k_{l,t-1}} \tag{2.19}$$

donde $k_{v,t}$ es el grado del nodo v en el tiempo t. La ecuación (2.19) define el apego preferencial observado en muchos sistemas reales. Teniendo en cuenta que la dependencia temporal está implícita en este sistema, la ecuación anterior puede ser escrita más fácilmente como:

$$p_{k_v} = \frac{k_v}{\sum_{k_u} k_u} \tag{2.20}$$

Es decir, la probabilidad de que el nodo v tenga grado k_v. Adicionalmente, como se men-

cionó anteriormente, es importante conocer la distribución del grado, que para este modelo corresponde a:

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \simeq 2m(m+1)k^{-3} \tag{2.21}$$

Esto es, la distribución del grado para el modelo de Barabási-Albert sigue una ley de potencias [18]. Se dice que este tipo de modelos son libres de escala, ya que independiente de la escala del problema (por ejemplo, el número de nodos), estos siempre tienen el mismo comportamiento, es decir, son independientes de la escala en la que se observan.

2.3.4 Redes espacio-temporales: redes de eventos

En las secciones anteriores se dieron los conceptos para la construcción de una red compleja, no obstante para datos espacio-temporales y su respectivo modelamiento por medio de una red compleja, es necesario definir los mecanismos en que la ubicación espacial y la dimensión temporal permiten definir cómo se realizará una conexión entre nodos. En la literatura es posible encontrar propuestas metodológicas para agregar la dimensión temporal a las redes. Por ejemplo, en las redes multicapa [4], en donde cada conjunto de eventos para un espacio temporal discreto ocurren en una capa diferente, agregando de esta forma el factor temporal de los datos. No obstante, Davies y Marchione [3] proponen un método que relaciona tanto al espacio como al tiempo para la construcción de una red que llaman red de eventos.

De acuerdo a Davies y Marchione [3] para la construcción de la red de eventos, sean N eventos con una ubicación espacial $\boldsymbol{s}_i$ y un tiempo de ocurrencia t_i. Se define la distancia entre dos eventos v_i y v_j como:

$$d_{ij} = d(\boldsymbol{s}_i, \boldsymbol{s}_j) \geq 0 \tag{2.22}$$

De manera similar, se define $t_{ij} = t_j - t_i$ como la separación temporal entre eventos. Adicionalmente, sean D un radio espacial y T un radio temporal que definen cuándo dos eventos son cercanos en cada dimensión. Se define la red de eventos como el grafo G_{dt}^{DT}, en donde dos vértices estarán conectados por un enlace si dos eventos ocurren a una distancia D entre

cada uno, y además, si ocurren en una ventana temporal T, de tal forma que el conjunto de los enlaces corresponde a:

$$\mathscr{L}_{dt}^{DT} = \left\{(v_i, v_j) | d(\boldsymbol{s}_i, \boldsymbol{s}_j) \leq D \wedge (t_j - t_i) \leq T\right\} \tag{2.23}$$

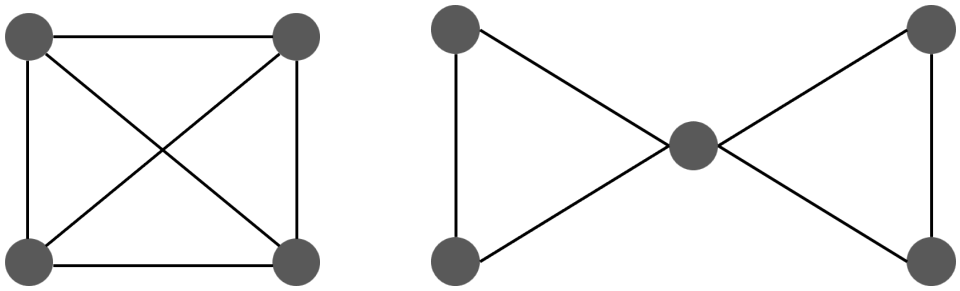
Conceptualmente, una red de eventos corresponde a un conjunto de nodos que representan los eventos de estudio con una ubicación espacial particular y con una dimensión temporal que representa el momento de ocurrencia del evento, es decir, representan los patrones puntuales de estudio. Esta red define cómo dos eventos se relacionan entre sí al definir una distancia $d_{ij} \leq D$ dentro de una ventana temporal T. La red resultante corresponderá a una red dirigida, en donde un evento pasado v_i que ocurre en t se conecta con un evento v_j en $t+1$ mediante un enlace dirigido. Por lo tanto, es necesario definir los parámetros D y T para la construcción de la red.

En el trabajo realizado por Forero, Bohorquez y Renteria [2] se propone realizar un estudio para M radios espaciales $D \in \{D_1, D_2, \ldots, D_M\}$ separados una distancia T constante (ya que no se ha considerado un radio temporal), de tal manera que se se obtiene una serie de tiempo de redes de eventos para cada D tal que:

$$G_{t,D} = \{(v_{dt}^{DT}, \mathscr{L}_{dt}^{DT})\} \tag{2.24}$$

De esta manera, se utiliza una detección de puntos de cambio para determinar el valor ideal de D mediante el concepto de *scan statistics* que se ajusta al caso de estudio. No obstante, la teoría asociada no se presenta en este trabajo ya que, a diferencia del trabajo realizado Forero, Bohorquez y Renteria [2], en este caso de estudio se propone que la distancia se defina de acuerdo a la similitud que presente la red construida con la topología de redes ya conocidas, como pueden ser los ejemplos presentados en la sección 2.3.3.

Para esto, se comparan los resultados con el trabajo realizado por Broide y Clauset [29]. En ese trabajo, los autores ponen a prueba la idea de que las redes que se forman por eventos del mundo real siguen una topología libre de escala, lo cual se presenta en la sección 2.3.3 en el modelo de Barabási-Albert. Para esto, utilizan alrededor de 1000 redes sociales, biológicas, tecnológicas, de transporte e información. En el caso de las redes sociales que son el objeto



(a) Clique con 4 vértices (b) Butterfly

Figura 2.3 Ejemplo de subgrafos

de estudio de este trabajo, se encontró que en la mayoría de los casos la distribución del grado de las redes se pueden ajustar a una log-normal igual o incluso mejor que una ley de potencias.

Por lo tanto, para la construcción de una red de eventos, la distancia se debe escoger de acuerdo al grado de la red de eventos que mejor se ajuste a la distribución del grado de una red con topología conocida. En principio se podría considerar hacer el ajuste a una red libre de escala ya que han sido consideradas las redes que mejor gobiernan la dinámica de distintos eventos del mundo real. No obstante, se pondrá a prueba si lo establecido por Boride y Clauset [29] se cumple en este tipo de redes sociales, es decir, se ajustan más comúnmente a redes donde su distribución del grado sigue una distribución log-normal, y no a redes libre de escala. Por este motivo, se realizarán pruebas de bondad de ajuste sobre redes de eventos construidas para diferentes valores de distancia y así se validará estadísticamente con cuál distancia el comportamiento de la red se corresponde con lo reportado en la literatura y para garantizar un nivel de confianza en los resultados obtenidos.

2.3.5 Subgrafos y motifs

Otro concepto importante corresponde al de subgrafo R, el cual está compuesto por los nodos $v_R = \{v_1, v_2, \ldots, v_R\} \subset G$, en donde $\mathscr{L}_R \subset \mathscr{L}$, es decir, los enlaces son un subconjunto de los enlaces entre parejas ordenadas del grafo G. Algunos ejemplos de subgrafos corresponden a cliques, los cuales son subgrafos donde cada vértice está conectado a todos los demás vértices del subgrafo; motifs, los cuales son subgrafos que se repiten dentro de una red, y por lo tanto, su topología depende del problema de estudio; butterflies, los cuales corresponden a la unión de 2 cliques de 3 vértices cada uno, unidos por un único vértice en común. En la figura 2.3 se presenta un ejemplo de subgrafo de clique y butterfly, respectivamente.

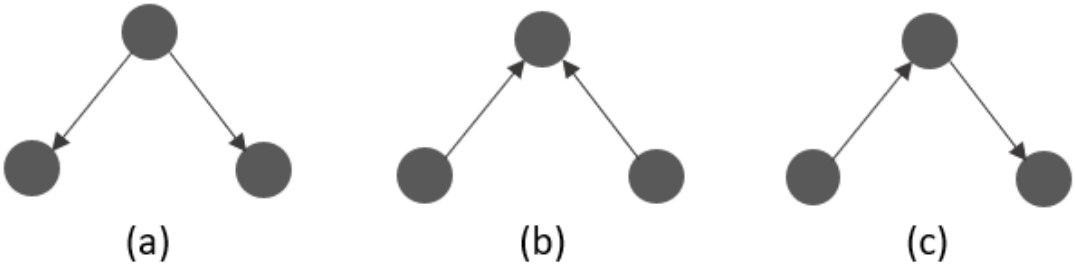


Figura 2.4 Motifs de estudio en el trabajo propuesto

Davies y Marchione [3], y Sanabria, Bohorquez y Renteria [2], buscan caracterizar una red a partir de motifs. Los motifs son subgrafos que se repiten así mismos múltiples veces dentro de una red compleja, es decir, no tienen una forma en particular y depende de su significancia estadística dentro de cada red para identificar si el subgrafo corresponde o no a un motif. En la sección 2.5 se exploran algunos posibles métodos para identificar este tipo de subgrafos. Este tipo de subgrafos tienen una gran importancia ya que permiten caracterizar el comportamiento o la estructura de una red compleja mucho más grande.

En particular Sanabria, Bohorquez y Renteria [2] realizan el estudio del crimen mediante 3 motifs dentro de las redes de eventos. En la gráfica 2.4, se presentan los 3 tipos de motifs de estudio:

- La figura 2.4.a corresponde al motif *out-2-star*, el cual corresponde a un evento que ocurre en un tiempo t_i, y está conectado con dos eventos que ocurren en un $t_j > t_i$ a una distancia d.
- La figura 2.4.b corresponde al motif *in-2-star*, el cual corresponde a dos eventos que ocurren en un tiempo t_i y t_j, y están relacionados con un evento en una distancia d que ocurre en un $t_k > t_i$ y $t_k > t_j$.
- La figura 2.4.c corresponde al motif *2-path*, el cual corresponde a un evento que ocurre en un tiempo t_i y está conectado con un segundo evento que ocurre en un $t_j > t_i$ a una distancia d, y a su vez este segundo evento se relaciona con un tercer evento en $t_k > t_j$.

2.4 Clusterización de datos

En la sección 2.1.3 se presenta como Felson y Boba [11] mencionan que los patrones puntuales del crímen se pueden configurar de tal forma que estos se clusterizan. Para esto, existen

diversos métodos de clasificación no supervisada que permiten la detección de clusters dentro de un conjunto de datos. En particular, para un conjunto de datos con ubicación espacial es posible utilizar diferentes metodologías teniendo en cuenta la naturaleza de los datos de estudio.

En esta sección se exploran los aspectos teóricos de dos metodologías espaciales ampliamente conocidas: *k*-means y *Density-Based Spatial Clustering of Applications with Noise* o DBSCAN. Sí bien existen muchos más métodos de clusterización de datos, para el marco de este trabajo y por la naturaleza espacio-temporal de los datos que se estudiarán, solo se presentan estos métodos.

2.4.1 *k*-means

k-means es un algoritmo de agrupamiento que se utiliza para dividir un conjunto de datos en *k* grupos (o clusters) con características similares. El algoritmo funciona asignando cada punto de datos a un grupo según una medida de distancia entre el punto y el centro del cluster. El objetivo final del algoritmo es minimizar la suma de las distancias cuadradas entre cada punto de datos y el centro del cluster al que pertenece.

El algoritmo fue desarrollado por el estadístico estadounidense Lloyd en 1957, no obstante no fue publicado sino hasta 1982 [30]. A medida que el algoritmo se ha vuelto más popular, se han desarrollado variantes y mejoras, como el algoritmo de agrupamiento jerárquico y el algoritmo de agrupamiento basado en densidad, no obstante dichos métodos no se abarcarán en este documento.

Supóngase que se tiene un conjunto de observaciones $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, donde cada $\boldsymbol{x}_i$ corresponde a un vector *m*-dimensional. En este sentido, el algoritmo buscará hacer una partición de las *n* observaciones en *k* (con $k \leq n$) conjuntos $\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_k)$, de tal manera que la suma de cuadrados dentro de cada cluster se minimice. Es decir:

$$\underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||^2 = \underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i=1}^{k} |\boldsymbol{S}_i| \operatorname{Var}(S_i) \tag{2.25}$$

donde $\boldsymbol{\mu}_i$ corresponde a la media de las observaciones dentro del cluster S_i. En este sentido, el número *k* de clusters para la implementación del método es un parámetro no conocido para poder realizar la clasificación no supervisada. Por lo tanto es importante utilizar otros métodos

de diagnósticos para determinar el número de clusters óptimo. A modo de ejemplo, si se tiene un modelo 2-dimensional y utilizando la distancia euclidiana, el algoritmo correspondería a:

$$\underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||^2 = \underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x,y \in S_i} \left[(x - \bar{x}_i)^2 + (y - \bar{y}_i)^2 \right]$$

Computacionalmente existen múltiples algoritmos que permiten la implementación de *k-means*. Para ilustrar la manera en que estos funcionan, se presenta uno de los más conocidos, el cual corresponde al llamado algoritmo de Lloyd. Este algoritmo funciona de la siguiente manera:

1. Se inicializa el algoritmo con k clusters (con k conocido). Estos clusters corresponderán a k observaciones, consideradas los valores medios del cluster o centroides en el espacio n-dimensional.

2. Cada observación se asignará al centroide más cercano de acuerdo a la medida de distancia definida.

3. Se vuelve a calcular el valor de los centroides de acuerdo a las observaciones que hacen parte de ese cluster.

4. Nuevamente cada observación se asignará al centroide más cercano.

En la gráfica 2.5 se ilustra cómo el algoritmo de k-means se aplica a un conjunto de datos distribuidos de forma normal y multivariada. La ilustración muestra cómo los datos son inicialmente asignados a los centroides más cercanos y cómo, a medida que el algoritmo se ejecuta en iteraciones sucesivas, los centroides se mueven y mejoran la clasificación de los datos. Es importante destacar que los datos en este ejemplo están muy bien separados, lo que permite una rápida convergencia del algoritmo.

Si bien el ejemplo presentado ilustra que el algoritmo de k-means es eficiente en la clasificación de estos datos artificiales, en otras situaciones puede no ser la mejor opción. En términos computacionales, la elección de una medida de distancia diferente a la euclidiana puede causar que el algoritmo no alcance la convergencia. Además, si se utiliza la distancia euclidiana, cuando hay datos superpuestos o cercanos entre sí, los clusters se delimitarán a través de regiones lineales, ya que este método genera diagramas de Voronoi.

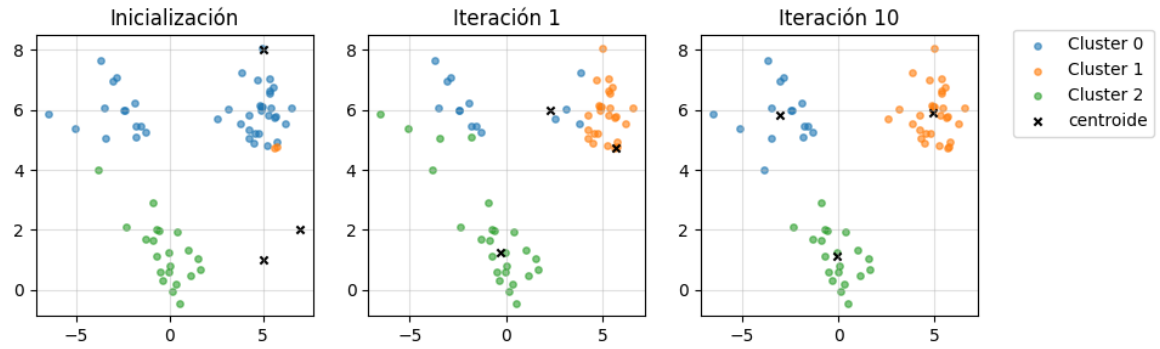


Figura 2.5 Ejemplo de *k*-means

Para la selección del número de clusters k óptimo, existen diferentes metodologías pero con limitaciones. En particular, en este proyecto se usa el método del *elbow* o codo, la cual es una metodología heurística y práctica para determinar el número de clusters. Este método selecciona el valor óptimo de k mediante la ecuación (2.25). El valor de k corresponderá al valor del *elbow* (punto en que la curva cambia significativamente su pendiente) de la curva de la suma de cuadrados dentro de los clusters para diferentes valores de k.

2.4.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo de agrupamiento no supervisado que busca agrupar puntos que se encuentran dentro de una región densa en el espacio. A diferencia de *k*-means, DBSCAN no requiere especificar el número de clusters a priori. El algoritmo funciona asignando un radio ε y un número mínimo de puntos MinPts para cada punto de datos. Si hay al menos MinPts puntos dentro de un radio ε del punto, se considera que dicho punto está dentro de una región densa. Los puntos dentro de una región densa son agrupados juntos en un cluster, en contraste a los puntos que no se encuentran dentro de una región densa, que serán considerados ruido.

El algoritmo DBSCAN fue propuesto en 1996 por Ester, Kriegel, Sander y Xu [31]. DBSCAN es ahora ampliamente utilizado en una variedad de aplicaciones, incluyendo la minería de datos, la inteligencia artificial, la visión por computadora y estadística espacial.

En la gráfica 2.6 se presenta el funcionamiento del método anteriormente explicado. En este caso se tomó MinPts $= 2$ y un valor constante de ε, el cual es el valor del radio de los círculos alrededor de cada punto. Esto significa que un cluster se forma cuando un punto se conecta con 2 puntos dentro del radio ε y los puntos dentro de esta región se siguen

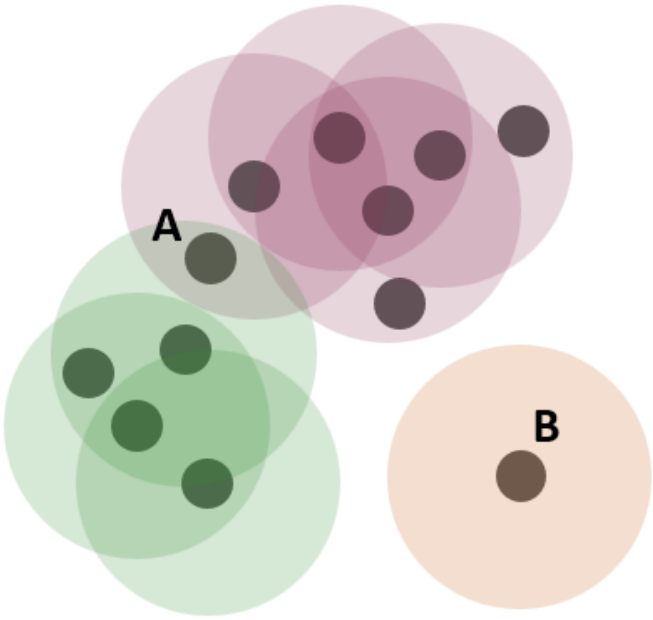


Figura 2.6 Ejemplo de DBSCAN con MinPts $= 2$ y ε dado por el radio de los círculos observados

conectando con la misma cantidad de puntos. Esto se puede observar en dos regiones dentro de la gráfica, la primera con 7 puntos, y la segunda con 5 puntos. En el caso del punto A, este puede pertenecer a cualquiera de los dos clusters en los que se encuentra, que en algunas implementaciones de este método puede llegar a ser considerado como ruido. Por otro lado, se observa el punto aislado B, el cual no conecta con ningún punto dentro del radio ε, por lo cual es tratado igualmente como ruido.
Para la implementación de este método, a diferencia de *k*-means, el algoritmo no hace ninguna estimación de la varianza de los clusters, no obstante si hace uso de la definición de una distancia, que puede ser euclidiana o no. En este caso, se hace una revisión de cada uno de los puntos en el espacio. En general, el algoritmo funciona de la siguiente manera:

1. Se inicializa el algoritmo estableciendo los dos parámetros MinPts y ε.
2. Se selecciona un punto arbitrario y se observa su alrededor a una distancia ε, y si contiene por lo menos MinPts puntos, se inicia un cluster. De lo contrario, este punto se marca como ruido.
3. Tanto el punto que se selecciona, como los puntos dentro de la región pertenecen al mismo cluster.
4. Se selecciona un punto no visitado dentro del cluster y se observa su región. Si cumple nuevamente con los parámetros, estos puntos también harán parte del cluster.

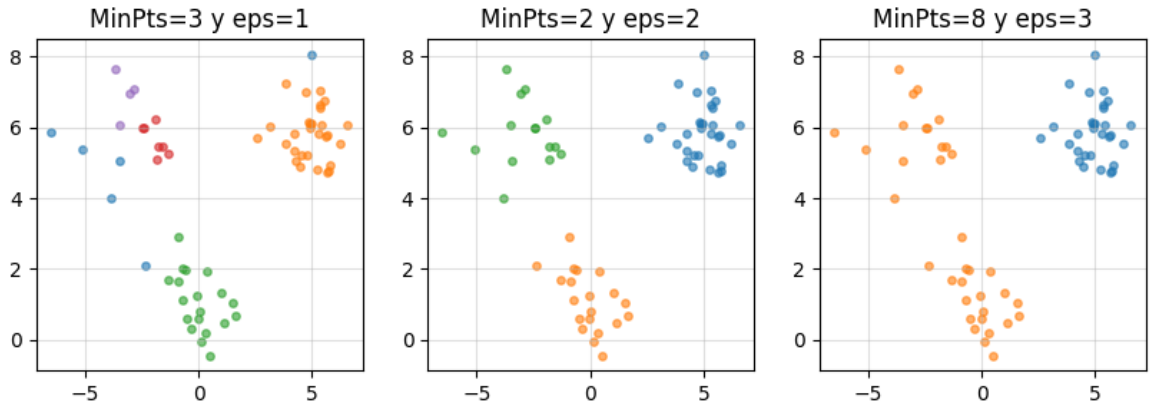


Figura 2.7 Ejemplo de DBSCAN para diferentes valores de MinPts y ε sobre datos normales multivariados

5. Si los puntos visitados dentro de la región ya no cumplen con alguno de los parámetros, se procede con el paso 2.

En la gráfica 2.7, se presenta un ejemplo de este método para el mismo conjunto de datos utilizado en *k*-means. En este caso se puede observar como DBSCAN clusteriza para diferentes valores de los parámetros escogidos. Para el caso en que MinPts $= 2$ y $\varepsilon = 2$, se observa la mejor clusterización de los datos, en donde identifica correctamente los 3 clusters. No obstante, esto no significa que sean la única combinación de parámetros que de esta misma configuración.

En el caso particular de ubicaciones espaciales con latitudes y longitudes, este algoritmo permite el uso de la distancia de Haversine mostrada en la ecuación (2.5) sin ninguna perdida de generalidad o convergencia.

DBSCAN tiene varias ventajas en comparación con otros algoritmos de clusterización, como la capacidad de detectar clusters de cualquier forma geométrica o la capacidad de detectar datos atípicos y no requerir que se especifique el número de clusters. Sin embargo, también tiene algunas desventajas, como la falta de determinismo en la asignación de puntos a clusters cuando estos se encuentran en una región compartida y la dificultad para distinguir clusters con diferentes densidades debido a que los parámetros MinPts y ε son fijos en el método, por lo cual siempre va a tener el mismo sentido de densidad independiente de la ubicación.

Para escoger el valor de MinPts y ε, se utiliza nuevamente el método heurístico del *elbow*. En este caso, se observa el número de puntos que se encuentran a una distancia determinada

con respecto a la distancia de los k vecinos más cercanos, y de esta forma se selecciona el valor de ε. El valor de MinPts dependerá del conocimiento de la densidad de la población y del análisis exploratorio que se realice sobre los datos.

2.5 Método ERGEN para identificación de subgrafos

Como se mencionó en la sección 2.3.5, es necesario utilizar algún tipo de algoritmo que permita identificar los motifs que componen una red compleja. Existen múltiples métodos para esto [7, 8], empezando por un conteo exacto de los subgrafos, como métodos computacionalmente más eficientes basados en muestras de la red compleja. Los métodos que utilizan un conteo exacto son ideales para redes pequeñas, ya que estos no son computacionalmente eficientes en la medida en que una red empieza a crecer, por lo cual es importante utilizar diferentes métodos en los casos en que un conteo exacto no sea viable.

Este caso se estudio se concentra principalmente en el método propuesto por Forero, Bohorquez y Renteria [2], con el método conocido como *Empirical Random Network Generator*. El algoritmo ERGEN consiste en los siguientes pasos:

1. Primero, para cada radio $d \in D$, con D el conjunto de todas las distancias a usar en la construcción de redes de eventos, y para cada ventana de tiempo $t' = 1, \ldots, T$, se tendrá un conjunto de nodos $v_{t',d}$, ordenado por fecha y hora.

2. En cada una de las redes de eventos, para el conjunto $v_{t',d}$ se considerarán todas las formas en que dichos eventos pueden ocurrir. Esto es, se intercambiará el orden de los puntos dentro del conjunto independiente de su hora y fecha de suceso.

3. Se toman B muestras de permutación de $v_{t',d}$. Esto es, solo se realizarán B permutaciones y no todas las permutaciones posibles, ya que esto último es poco práctico.

4. Se construirán B redes para cada radio d y período de tiempo t'. Luego, la serie de tiempo $\{G_{t',d}\}_{b=1}^{B}$ representa una muestra de tamaño B que corresponde a todas las posibles combinaciones de puntos en cada periodo de tiempo.

Se consideran solo las combinaciones posibles dentro de un mismo t'. De esta manera, se realiza el conteo de los motifs de interés para este proyecto para cada $\{G_{t'}\}_b$, generando una matriz $\boldsymbol{Y}_{t'}$ de tamaño $3 \times (B+1)$, donde la i-ésima fila de la matriz será $Y_{i,t'} = (\#out - 2 - stars, \#in - 2 - stars, \#2 - path)_i$, y la primera fila representará el número de motifs para el

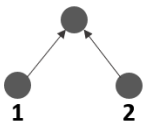
arreglo original de los eventos ordenados. De acuerdo a lo mencionado por los autores, para evitar los problemas de enmascaramiento del conteo de motifs por causa de las diferencias del número de nodos y enlaces dentro de cada período de tiempo, se propone escalar los datos mediante la siguiente ecuación:

$$Y'_{t'} = \boldsymbol{S}_{t'}^{-\frac{1}{2}} \left(\boldsymbol{Y}_{t'} - \bar{\boldsymbol{Y}}_{t'} \right) \tag{2.26}$$

donde $S_{t'}$ corresponde a la matriz de covarianzas de $\boldsymbol{Y}_{t'}$ y $\bar{\boldsymbol{Y}}_{t'}$, que corresponde al valor medio asociado, por lo cual corresponde a un vector de tamaño 3×1. De esta manera es posible realizar un análisis de series de tiempo sobre los resultados para establecer la significancia de dichos motifs y la relación entre ellos.

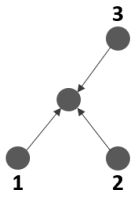
2.5.1 Propuesta: formula de Stirling

En este trabajo se utiliza la formula de Stirling [32] para mejorar la robustez y eficiencia computacional del método ERGEN para el conteo de los motifs. Esta propuesta es importante ya que es una aproximación más cercana al conteo exacto de los subgrafos, en contraste con el método por muestras utilizada por Forero, Bohorquez y Renteria [2], y con una mejora en la eficiencia computacional. Para su aplicación, se generaliza el número de motifs por nodo por medio de una combinatoria como se observa a continuación. En primer lugar, se ve el caso para el motif *in-2-stars*. Las parejas ordenadas de un motif *in-2-stars* harán referencia a los dos nodos que conectan con el nodo *central* para formar el subgrafo. Por lo tanto:

1. 

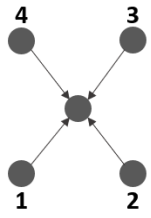
   - Parejas ordenadas: (1,2)
   - Conteo de motifs:

$$\binom{2}{2} = 1$$

2. 

   - Parejas ordenadas: (1,2), (1,3), (2,3)
   - Conteo de motifs:

$$\binom{3}{2} = \frac{3!}{1!2!} = 3$$

3. 

   - Parejas ordenadas: (1,2), (1,3), (1,4), (2,3), (2,4), (3,4)
   - Conteo de motifs:

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

Entonces por inducción, este conteo se puede escribir como:

$$\binom{n}{2} = \frac{n!}{(n-2)!2!} \tag{2.27}$$

donde $n \geq 2$ corresponde al número de nodos que están conectados con el nodo central y cumplen con las condiciones para formar un *in-2-stars*. Ahora, sobre la ecuación (2.27) aplicamos logaritmo natural de forma que:

$$\ln\left(\frac{n!}{(n-2)!2!}\right) = \ln(n!) - \ln(n-2)! - \ln(2) \tag{2.28}$$

Aplicando la formula de Stirling, se tiene que:

$$\ln\left(\frac{n!}{(n-2)!2!}\right) \approx n\ln(n) - n - (n-2)\ln(n-2) + (n-2) - \ln(2) \tag{2.29}$$

Finalmente, al aplicar exponencial a ambos lados, se obtendrá el número de motifs *in-2-stars* por nodo. Este mismo procedimiento también se aplica para el motif *out-2-stars*. Sin embargo, para el motif *2-path*, esta generalización no es posible debido a la dependencia de la forma de la conexión y el orden de los nodos. En el trabajo realizado por Forero, Bohorquez y Renteria [2] se utilizan datos del crimen en Medellín, los autores utilizan muestras de los datos para realizar el conteo, lo cual es común en este tipo de metodologías debido a la complejidad computacional de contar exactamente los datos en caso de combinatorias o permutaciones. La fórmula de Stirling permite aproximar el valor real de estudio en casos con un alto número de nodos, especialmente en áreas altamente densas y con una distancia d que abarque dicha densidad espacial. Por lo cual, se utilizará este método en conjunto con ERGEN para el análisis de los motifs dentro de las redes de eventos.

Capítulo 3

Metodología

En este capítulo se presenta la metodología que será aplicada en este proyecto. Se justifican los métodos utilizados de acuerdo a la teoría presentada en el capítulo anterior. Adicionalmente, se presenta brevemente una descripción de la implementación computacional de cada uno de los pasos a seguir en el caso de estudio.

3.1 Metodología del caso de estudio

En este trabajo final de maestría en profundización en estadística, se presenta un caso de estudio del crímen en Pereira mediante la construcción y el análisis de redes espacio-temporales, las cuales corresponderán a una red social (crímenes cometidos a una persona en una ubicación espacial especifica), y estarán conectados mediante su distancia geográfica y una lógica temporal, de acuerdo a lo presentado en la sección 2.3.4. Para esto, se utilizará Python como lenguaje de programación para todos los análisis requeridos.

Con el objetivo de poder generar conclusiones valiosas para la toma de decisiones y la mitigación del problema del crímen en la ciudad por parte de las autoridades respectivas, la metodología empleada se presenta en las siguientes secciones. A modo de resumen de la metodología propuesta, en la figura 3.1 se presenta un diagrama de los diferentes pasos que se seguirán para el desarrollo del caso de estudio en base al marco teórico presentado en el capítulo 2. Si bien en las posteriores secciones se abarcará cada uno de estos puntos, en resumen estos corresponden a:

1. Análisis descriptivo y exploratorio de la base de datos y las diferentes variables asociadas a cada registro, tales como variables demograficas de las personas que

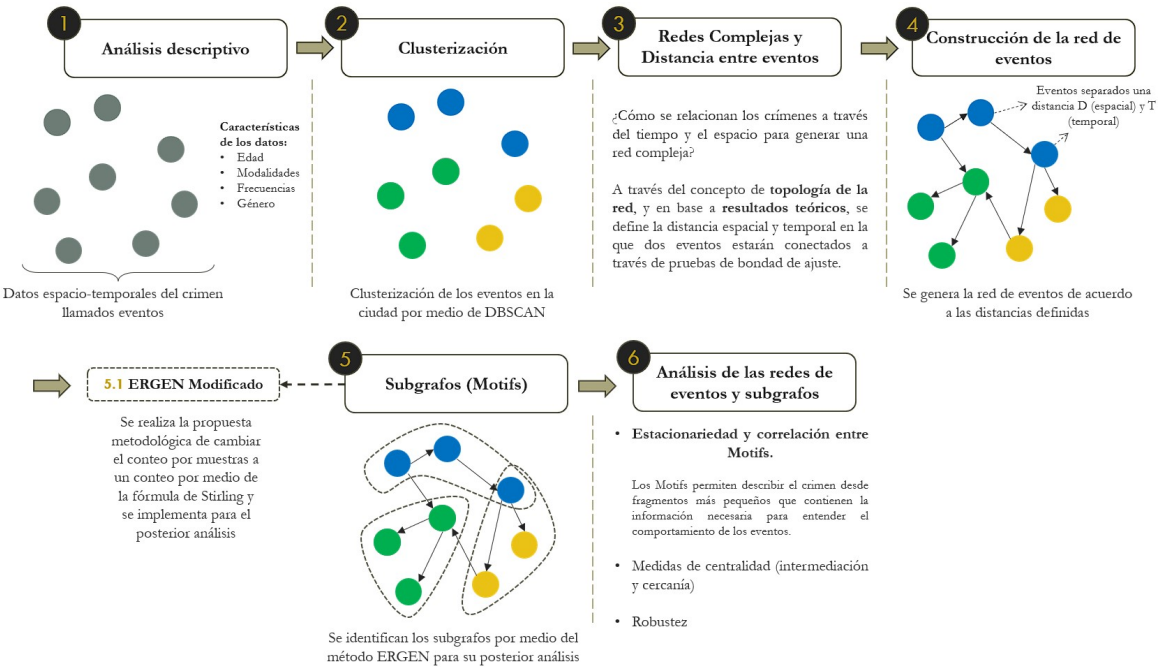


Figura 3.1 Metodología a implementar en el caso de estudio

realizaron la respectiva denuncia y variables descriptivas del crimen tales como la modalidad.

2. Clusterización por del algoritmo DBSCAN, utilizando la distancia de Haversine.

3. Estudio y definición de las distancias espaciales y temporales a utilizar para relacionar dos eventos de forma lógica y correctamente justificada. Para esto se observa la topología (esto es, la forma en que están conectados los nodos entre ellos) de la red compleja generada y su distribución del grado asociada, comparando los resultados con los reportados en otros estudios, por lo cual es necesario realizar pruebas de bondad de ajuste.

4. Se realiza la construcción de la red compleja o red de eventos de acuerdo a las reglas definidas en el paso anterior.

5. Se realiza un análisis de subgrafos, en particular de motifs, por medio de la metodología ERGEN. Para esto se realiza primero la implementación de la formula de Stirling en el conteo, mejorando la eficiencia computacional del algoritmo.

6. Se realiza el análisis de los Motifs por medio de series de tiempo y sus correlaciones, lo cual permite entender los comportamientos emergentes de la red. Adicionalmente se realiza un análisis de la red por medio de medidas de centralidad y robustez.

Finalmente a través de estos análisis realizados por medio de los Motifs, y su capacidad para describir la red, se realizan las recomendaciones a las autoridades locales para combatir la criminalidad en la ciudad.

3.1.1 Análisis descriptivo y exploratorio

En primer lugar, se revisará la base de datos de criminalidad de Pereira y sus municipios aledaños para entender las características principales del crimen en la ciudad. Se realizarán mapas temáticos de la ciudad para observar tendencias en el número de casos por comuna, gráficos de patrones puntuales para entender la distribución de los casos dentro de la ciudad, gráficos de barras y tendencias para las distintas variables categóricas tanto demograficas como características del delito y así detectar elementos explicativos de los crímenes en la ciudad. Además, se investigará la relación que puedan tener algunas de estas variables entre ellas. Es importante destacar que esto implica revisar la base de datos, identificar los tipos de datos y limpiar los mismos.

Esto es importante en el caso de estudio, ya que de acuerdo a lo mencionado en la sección de 2.1.3, es importante conocer el contexto social y económico del lugar del crímen, lo cual es posible mediante las metodologías presentadas en la sección 2.2. Adicionalmente, para poder hacer el análisis de la temporalidad más adecuado, se crearon las siguiente franjas horarias para tener menor cantidad de categorías, las cuales son:

- Entre las 05 : 00 y las 08 : 59 *Mañana temprana*.
- Entre las 09 : 00 y las 12 : 59 *Mañana*.
- Entre las 13 : 00 y las 15 : 59 *Tarde Temprana*.
- Entre las 16 : 00 y las 17 : 59 *Tarde*.
- Entre las 18 : 00 y las 20 : 59 *Noche Temprana*.
- Entre las 21 : 00 y las 04 : 59 *Noche*.

3.1.2 Clusterización

Como se menciona en la sección 2.1.3, Felson y Boba [11] mencionan que los crímenes pueden clusterizarse de acuerdo a sus características ambientales. Teniendo en cuenta este concepto, como siguiente paso en la metodología se propone realizar una clusterización de

los datos teniendo en cuenta que los barrios socialmente presentan estas características de clusterización de los crímenes. Para esto, se realizará la comparación de los dos métodos presentados en la sección 2.4, los cuales corresponde a k-means y DBSCAN, con el objetivo de analizar e identificar si existen zonas de mayor incidencia delictiva en la ciudad. Esto permite observar cuál es la zona con mayor número y densidad de casos. Estos algoritmos son especialmente útiles para la clusterización de datos espaciales, ya que es posible utilizar medidas de distancia acordes a este tipo de datos.

En este caso, se utilizarán dos métricas de distancia, la euclidiana y la distancia de Haversine, de acuerdo a lo presentado en el capítulo anterior. Esta clusterización permitirá que en los próximos pasos de la metodología, los eventos solo puedan conectarse entre ellos sí ocurren dentro del mismo cluster. De acuerdo a los resultados obtenidos, se utilizará el método que mejor describa la naturaleza de los datos.

3.1.3 Distancia entre eventos

Como se mencionó en la introducción a esta sección, se construirá una red espacio-temporal mediante el concepto de redes de eventos presentado en la sección 2.3.4. Para esto, es importante establecer las reglas lógicas que permitirán la construcción de dicha red, es decir, establecer cuál será el valor de la distancia espacial y temporal. Para esto, se utilizará el concepto de topología de la red explicado en la sección 2.3. Como se menciona en la sección 2.3.4, en esta metodología se utilizan principalmente los resultados obtenidos en el trabajo realizado por Broide y Clauset [29]. Es decir, la distancia se escogerá de acuerdo a la distribución del grado de la red de eventos que mejor se ajuste a una red con topología conocida, justificado de manera estadística mediante pruebas de bondad de ajuste.

Para evaluar si dos eventos están a una distancia determinada, es importante considerar que este trabajo utiliza el sistema coordenado WGS84.

3.1.4 Construcción de la red de eventos

Una vez definida la distancia, se construye la red de eventos. Esta se construirá de acuerdo a lo establecido en la sección 2.3.4, en donde la distancia corresponderá al valor con mayor significancia estadística y la distancia temporal se tomará constante y será de un mes, de acuerdo a la metodología sugerida por Forero, Bohorquez y Rentería [2]. De esta manera se obtiene una serie de tiempo de redes de eventos como se establece en la ecuación (2.24).

Además, siguiendo los resultados de la clusterización, se establece que los eventos solo se pueden conectar entre ellos cuando pertenecen a un mismo cluster.

3.1.5 Identificación de subgrafos

Se realiza un análisis de los subgrafos generados en las redes de eventos para cada periodo de tiempo. Para esto se utiliza la metodología ERGEN explicada en la sección 2.5; no obstante, en este trabajo se hace un aporte en el método al modificar la forma en que se hace el conteo de los motifs. Dicha propuesta es presentada en la sección 2.5.1. La identificación y conteo de los diferentes motifs se realiza mediante este método y se realizará una estandarización de los resultados por medio de la ecuación (2.26).

Con estos resultados se lleva a cabo un análisis estadístico, el cual incluye la evaluación de la estacionalidad de las series y la correlación entre ellas. Este análisis permite identificar patrones en los robos, determinar cuáles son más comunes y cómo se relacionan entre si en la ciudad.

3.1.6 Análisis de las redes de eventos

Una vez se ha realizado el estudio de los subgrafos y los motifs obtenidos, se procederá a realizar un análisis de las redes de eventos obtenidas por medio de medidas de centralidad y la asortatividad que puedan ser útiles para entender el comportamiento general del crímen en la ciudad. Lo anterior, de acuerdo a lo mencionado en la sección 2.3.2. Para esto se utiliza principalmente la librería *networkx*, ya que esta permite realizar el cálculo de medidas como la centralidad de intermediación y la centralidad de cercanía fácilmente sobre los grafos generados.

3.2 Implementación computacional

Para poder llevar acabo la metodología anterior, es importante tener en cuenta los métodos computacionales que se utilizarán en cada paso. En este caso, el lenguaje de programación utilizado corresponde a Python. Para un mayor detalle de todos los resultados y métodos descritos, los códigos, bases de datos y otra información de interés se pueden encontrar en el repositorio asociado a este trabajo solicitando el respectivo acceso[1].

[1]Link del repositorio para solicitar acceso: https://github.com/miguelnq/2027283-un_tf_maestria

Principalmente se utilizarán las siguientes librerías para el análisis descriptivo y exploratorio de los datos:

```
# Manejo de bases de datos
import pandas as pd
import numpy as np

# Mapas, patrones puntuales
import geopy
import geopy.distance
import folium
import json
from shapely.geometry import shape, Point

# Gráficos de barras, tendencias
import matplotlib.pyplot as plt
import plotly.graph_objects as go
import plotly.express as px
```

Para los mapas temáticos y los mapas con los patrones puntuales, se utilizará la librería *folium* que permite crear mapas interactivos en html, en donde se pueden ir agregando diferentes capas que pueden activarse y desactivarse dependiendo de los datos que se deseen ver. En este caso, para poder centralizar los datos en la ciudad de Pereira, es necesario crear un mapa base que esté centrado en las coordenadas $(4.804, -75.709)$. De esta forma, se crea el mapa base de la siguiente manera:

```
m = folium.Map([4.8042838,-75.7096671] , zoom_start=13.2)
base_map = folium.FeatureGroup(name='Basemap', overlay=True,
control=False)
folium.TileLayer(tiles=tiles, attr=att).add_to(base_map)
base_map.add_to(m)
```

Una vez se tiene un mapa base, se pueden graficar los patrones puntuales mediante la función *CircleMarker* de *folium* sobre cada una de las coordenadas de los crímenes cometidos en la base de datos. En este caso, los gráficos se realizarán teniendo en cuenta el sistema coordenado WGS84 explorado en la sección 2.1.2.

Para el caso de los mapas temáticos, se utiliza la ubicación geométrica y los atributos

de los diferentes barrios de Pereira en formato Geojson, por lo cual se utiliza la librería *json*. Los datos de la fuente de datos en algunas ocasiones presentan formatos diferentes en los nombres de los barrios con respecto a los nombres en los atributos del Geojson. Por lo tanto, mediante *shapely* se hace la marcación del barrio del Geojson cuando un punto (ubicación del crímen en WGS84) ocurre dentro de la geometría establecida dentro del Geojson, como se muestra a continuación:

```
with open ('../Datos/Barrios_del_municipio_de_Pereira.geojson', 'r')
as jsonFile:
    mapdata = json.load(jsonFile)

def inside_point(row):
    point = Point(row['LONGITUD'], row['LATITUD'])
    for feature in mapdata['features']:
        polygon = shape(feature['geometry'])
        if polygon.contains(point):
            barrio = feature['properties']['NOMBRE']
            return barrio

point = Point(df['LONGITUD'][0], df['LATITUD'][0])
for feature in mapdata['features']:
    polygon = shape(feature['geometry'])
if polygon.contains(point):
    df['BARRIO_GEOJSON'] = df.apply(lambda row: inside_point(row),
    axis=1)
```

No se incluyen los datos de Dosquebradas ya que no fue posible conseguir las geometrías de los barrios que lo dividen.

Para evaluar si dos eventos están a una distancia determinada en metros, se utiliza la librería *geopy*. Geopy contiene una función que calcula la distancia geodésica en WGS84 entre dos puntos (o algún otro sistema coordenado elipsoidal), y la longitud de la distancia se puede obtener en diferentes unidades, en este caso en metros.

```
import geopy

distance = geopy.distance.geodesic(coords_1, coords_2).m
```

De esta manera es posible construir la red de eventos, en donde dos eventos están conectados si ocurren en una distancia determinada mediante la librería *networkx*. Para la construcción de la red es importante tener en cuenta que esta es dirigida, por lo cual es importante determinar los nodos *in* y *out* y que cumplan con el criterio de la distancia. De forma muy general, dicha validación se realiza de la siguiente manera:

```
import networkx as nx

months_list = df['FECHA_MES'].unique()
clusters_list = df['CLUSTERS'].unique()

# Parametros iniciales:

# Radio de distancia entre cada uno de los
# robos para hacer la respectiva conexión.
D = 500 #r el momento

#---------- Creación de los grafos ------------#

# Inicializo ts_graph
ts_graph = {}

for month in months_list:
    # Lista vacia de nodos in, out
    nodes = []

    # Se crea un subdataset con el mes
    df_month = df[df['FECHA_MES']==month].sort_values(['FECHA_HECHO'
                                                       ,'HORA'])
    for i in range(df_month.shape[0]):

        #Inicializo j
        j = i+1

        # Se hace un shift de Latitud y Longitud para calcular
        #la distancia geodesica
        df_month[['LATITUD_shift','LONGITUD_shift','CLUSTERS_shift'
```

```
                ,'index_shift']] = df_month[['LATITUD','LONGITUD',
                                             'CLUSTERS','index']
                                             ].shift(-j)

            # Calculo de la geo_distance_graph
            df_month.apply(lambda x: geo_distance_graph(x,D),axis=1)

        # Se guarda mes + lista de conexiones
        ts_graph[month] = nodes
```

Teniendo en cuenta que este paso corresponde a una sola distancia y para la evaluación de la distancia de la sección anterior, es necesario repetir este paso para todos los valores de distancia a evaluar. La función *geo_distance_graph* corresponde a:

```
def geo_distance_graph(x,D: float):

    # Coordenadas
    if x['CLUSTERS']==x['CLUSTERS_shift']:
        # Coordenadas nodo entrada
        coords_1 = (x['LATITUD'], x['LONGITUD'])
        # Coordenadas nodo salida
        coords_2 = (x['LATITUD_shift'], x['LONGITUD_shift'])

        # Al utilizar shift es mejor utilizar try para
        #evitar los casos en las coordenadas van a quedar con
        # valores nan.

        try:
            # Distancía geodesica de geopy
            distance = geopy.distance.geodesic(coords_1, coords_2).m

        except ValueError:
            # Valor dummy para el caso en el fallo
            distance = D+1

        if distance<=D:
            # Lista de nodos (in, out)
```

```
nodes.append([x['index'],x['index_shift']])
```

Por lo cual en caso de que se desee cambiar la forma en que se mide la distancia entre eventos, basta con cambiar esta función. Finalmente, para las pruebas de bondad de ajuste se utiliza la librería *scipy*, que permite realizar diferentes pruebas estadísticas dependiendo del problema a resolver.

Capítulo 4

Análisis de resultados

En este capítulo, se presentan los datos y el análisis de los resultados detallado del crimen en Pereira mediante la metodología presentada en la sección anterior. Primero, se presentan los datos que serán utilizados para el estudio y se proporciona un contexto socioeconómico y geográfico de la ciudad de Pereira y sus municipios cercanos para comprender mejor el panorama del crimen en la región. Se lleva a cabo un análisis descriptivo y exploratorio de la base de datos para identificar patrones y factores que contribuyen a la ocurrencia del crimen. Adicionalmente, el análisis incluye la clusterización de los datos, la construcción de redes de eventos y el conteo de subgrafos relevantes. Es importante mencionar que este estudio tiene el potencial de expandirse, ya que es importante entender cómo las características socioeconómicas contribuyen a los resultados y cómo estos pueden ayudar en la toma de decisiones para mejorar la seguridad ciudadana.

Como se mencionó en la sección 3.1, para la reproducción de todos los resultados obtenidos, los códigos, bases de datos y otra información de interés se pueden encontrar en el repositorio solicitando el respectivo acceso[1].

4.1 Descripción de los datos

La base de datos proviene de los delitos reportados a la Policía Nacional y utilizados para los respectivos análisis y reportes de la misma entidad. La base contiene 6149 registros que comprenden entre el 25 de marzo de 2018 hasta el 30 de junio de 2021. Cada registro corresponde a un crimen cometido y cuenta con su ubicación espacial en latitud y longitud,

[1] Link del repositorio para solicitar acceso: https://github.com/miguelnq/2027283-un_tf_maestria

junto con una serie de variables que permiten caracterizar el tipo de crimen. Estas variables corresponden a:

- **Conducta:** Hurto de personas.
- **Municipio:** Pereira, Dosquebradas y La Virginia.
- **Comunas:** Comuna donde ocurre el crimen.
- **Barrios y Dirección:** Desagrega comúnas.
- **Sitio:** Descripción del sitio donde ocurre el crimen.
- **Hora, Día, Mes y Año:** Hora y fecha en que ocurre el delito.
- **Modalidad:** Descripción de la modalidad de hurto.
- **Emplea Armas:** índica si se usaron armas en el hurto.
- **Zona:** Rural o Urbana.
- **Hora Hecho y Fecha Hecho:** Desagrega hora y agrupa fecha en formato aaaa-mm-dd.
- **Genero:** Genero de la víctima.
- **Agrupa edad:** Variable categórica que indica el rango de edad de la víctima.
- **Edad:** Edad de la victima.
- **Latitud y Longitud:** índica la ubicación espacial.

Es decir, se tiene 1 columna con valor constante, 6 columnas con diferentes niveles de desagregación de ubicación del delito, 2 columnas con categorías del delito y porte de armas, 3 columnas con información demográfica de la víctima, 6 columnas que indican el día o fecha del hecho y la ubicación espacial del crimen (Longitud y Latitud). En particular, en este trabajo la metodología se realizará principalmente con las variables de fecha, modalidad, comuna, municipio, genero, edad, hora, día y mes. Adicionalmente, solo se trabajará con los municipios de Pereira y La Virginia, teniendo en cuenta la cercanía entre estos dos y los pocos delitos que se presentan en La Virginia.

Con respecto a la temporalidad de los datos, en la figura 4.1 se pueden observar diferentes agregaciones de los números de casos reportados para las distintas fechas. Se resaltan en rojo los datos agregados por semanas y meses para las fechas entre la primera semana de enero

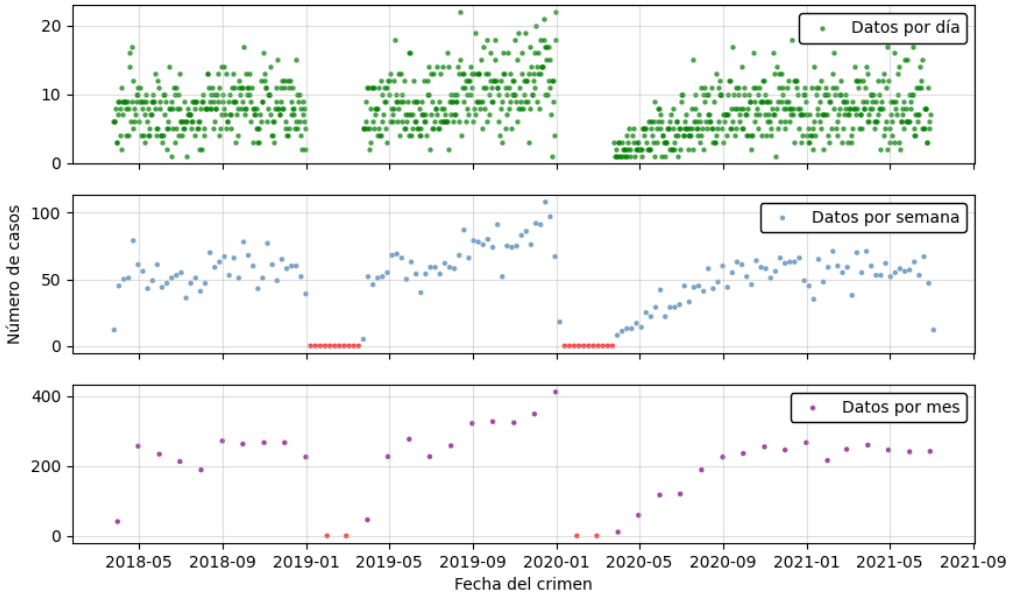


Figura 4.1 Número de crímenes por: día, semana, mes. En rojo las semanas y meses en donde no se tienen datos

del 2019 y la segunda semana del mes de marzo del 2019, junto con el periodo de la segunda semana de enero de 2020 y la última semana de marzo de 2020. Para estos dos periodos no se tienen registros de crímenes, lo cual corresponde a procesos de principio de año en los que se hacen reestructuraciones y también con el principio de la pandemia por Covid-19 en 2020. Para el desarrollo del caso de estudio se tendrán en cuenta estos meses para no cometer errores en los análisis realizados.

En las siguientes secciones se presentará un análisis estadístico exploratorio más detallado de los datos como parte de la metodología propuesta en este trabajo.

4.2 Contextualización de la ciudad de estudio: Pereira

Pereira es una ciudad colombiana, capital del departamento de Risaralda. La ciudad se encuentra ubicada en el Valle de los Rios Cauca y Quito, en el centro-oeste de Colombia. La población de Pereira en 2018 según el Censo Nacional de Población y Vivienda 2018 (CNPV) [33] fue de 409.670 personas (personas efectivamente censadas), con una proyección en base al mismo censo para 2022 de 485.373 personas. El porcentaje de hombres en la ciudad es de aproximadamente el 47.1 % y el porcentaje de mujeres es de aproximadamente el 52.9 %,

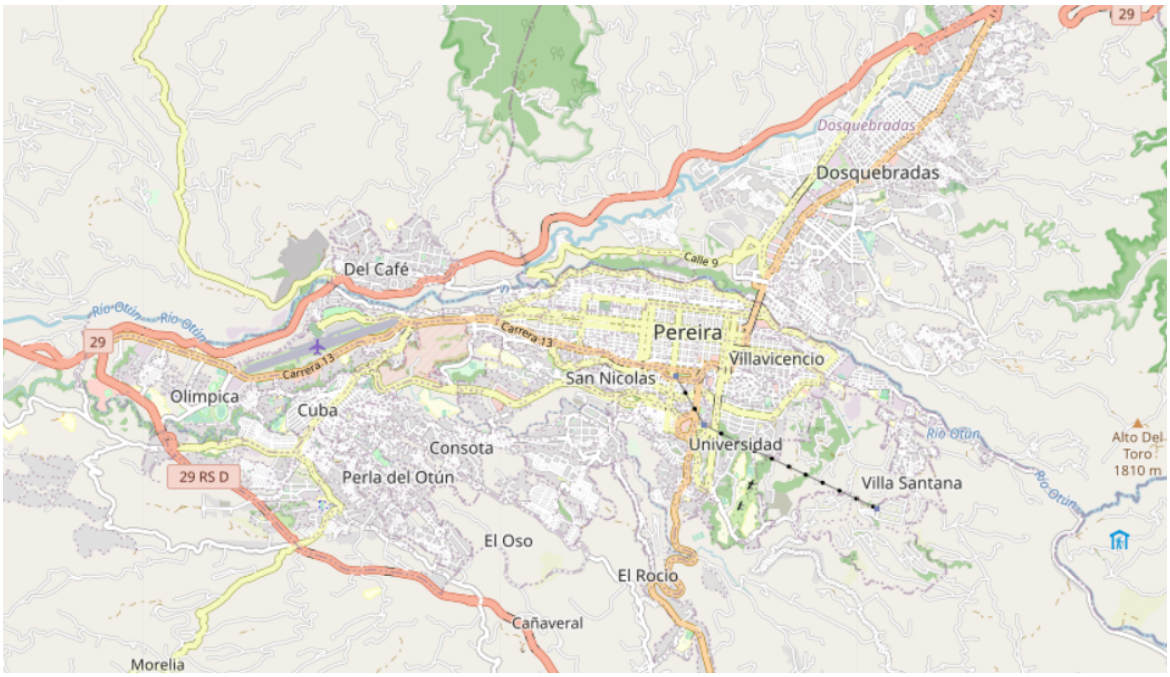


Figura 4.2 Mapa de Pereira y Dosquebradas

teniendo una razón de masculinidad de 89.1 por cada 100 mujeres.

Adicionalmente, según el CNPV 2018, se observan las siguientes características con respecto a la población:

- **Población entre 0 y 14 años:** 22.6 %.
- **Población entre 15 y 64 años:** 68.3 %.
- **Población mayor a 65 años:** 9.1 %.

Esto significa que aproximadamente el 68 % de la población son personas en edad adulta (mayores de 18 años). Estos resultados poblacionales son importantes ya que se contrastarán con las edades de los individuos en la base de datos de crímenes. Adicionalmente, según el CNPV 2018 [33], Risaralda tiene una población de 839.558 personas. Por lo tanto Pereira tiene aproximadamente el 49 % de la población del departamento.

Pereira se encuentra rodeada de varios municipios y ciudades importantes. Al norte de Pereira se encuentra la ciudad de Dosquebradas, considerada la segunda ciudad principal del departamento de Risaralda, y que según el CNPV 2018 [33] tenía 194.229 personas y se estiman

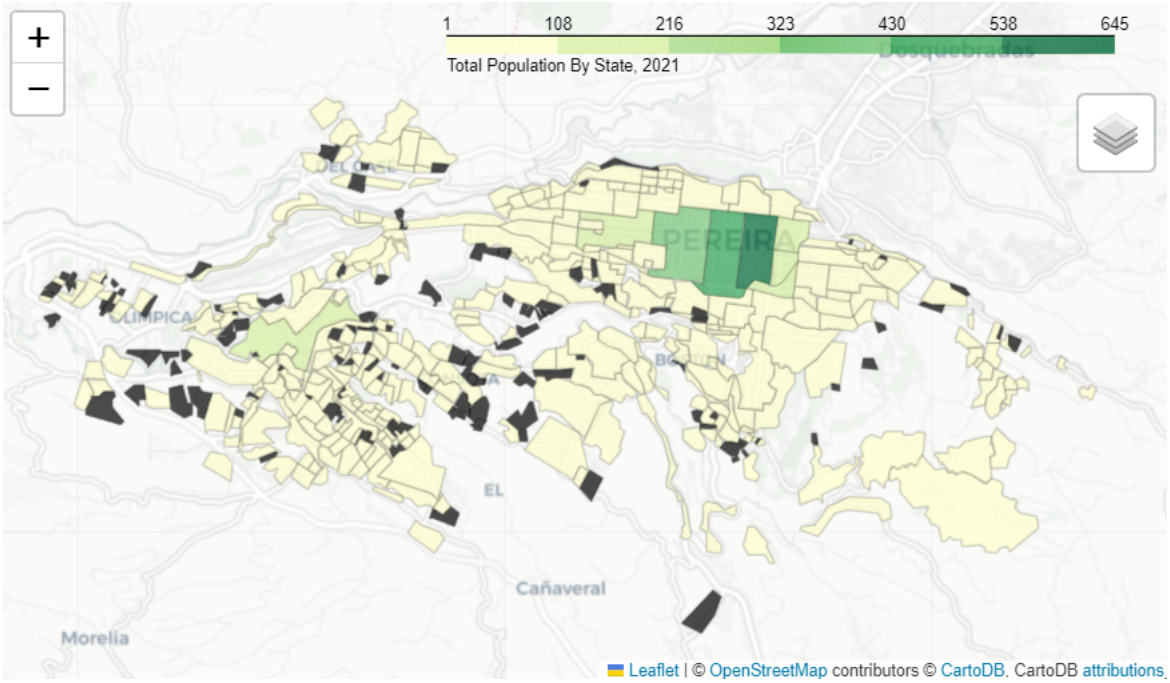


Figura 4.3 Mapa temático de la criminalidad en Pereira por Barrios

227.696 a 2022, con una proporción similar entre hombres y mujeres a Pereira. Al este se encuentra la ciudad de La Virginia, con 25.829 personas al 2018 [33]. Estos tres municipios forman el Área metropolitana de Centro Occidente, con Pereira como el municipio principal y Dosquebradas y La Virginia como los municipios periféricos. Esta Área metropolitana de Centro Occidente tiene un total de 705.431 personas, es decir, aproximadamente el 84 % de toda la población de Risaralda. Entre otros municipios cercanos, se encuentra también Santa Rosa de Cabal y La Celia, al sur y oeste, respectivamente; no obstante, estos municipios no son considerados dentro del área metropolitana.

En la figura 4.2 se puede observar la cercanía geográfica entre la ciudad de Pereira y el municipio de Dosquebradas. Esta proximidad es relevante para dar contexto al estudio del crimen que se realizará, ya que se tendrán en cuenta los datos de ambas áreas. Es posible que estos datos puedan proporcionar información valiosa sobre la relación entre la criminalidad en Pereira y Dosquebradas, teniendo en cuenta que muchos crímenes parecen ocurrir en la frontera entre ambos municipios, lo cual ya ha sido estudiado por Jimenez [34].

En cuanto a territorio, Pereira tiene una superficie de 702 km^2 y Dosquebradas de 70.8 km^2, que en contraste con Bogotá con una superficie de 1636.35 km^2 y una población de aproximadamente 8 millones de habitantes, se podría considerar a Pereira y al Área Metropoli-

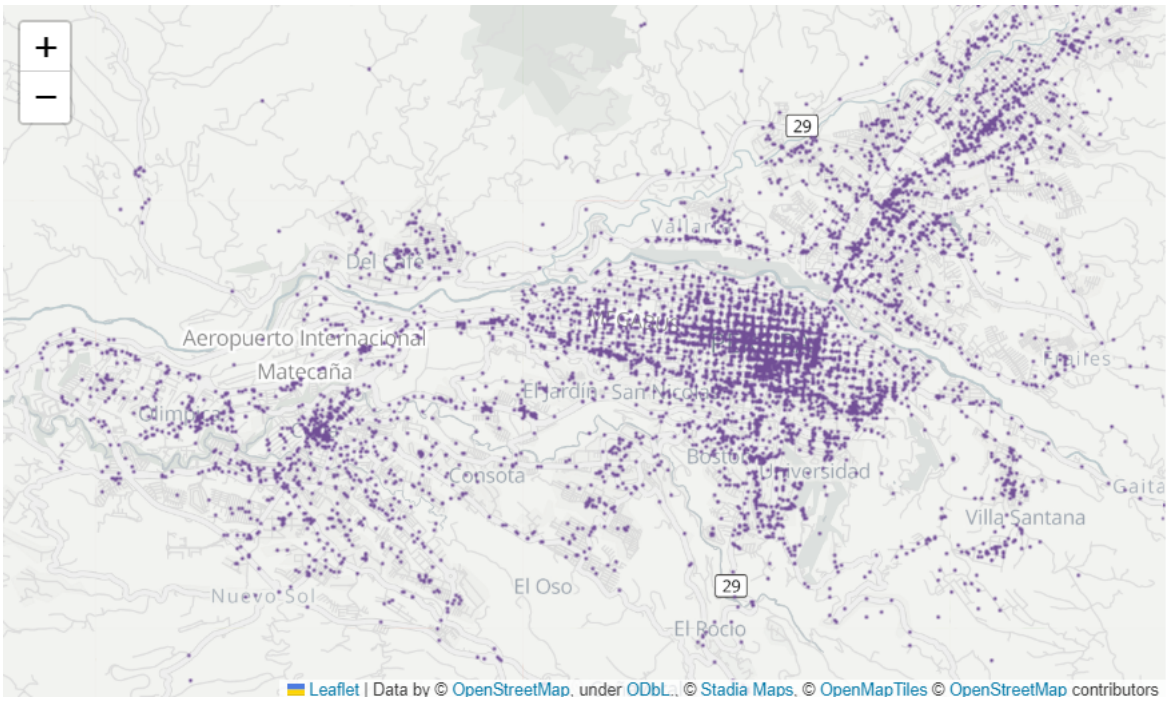


Figura 4.4 Marcación espacial de los crímenes en la ciudad de Pereira

tana como una ciudad pequeña y con una baja densidad de habitantes por km^2, siendo de 841.25 habitantes/km^2.

Pereira es una ciudad que ha enfrentado problemas de violencia y delincuencia, en donde los delitos más comunes incluyen robos, hurtos, extorsiones y homicidios. En el Plan Integral de Seguridad y Convivencia Ciudadana (PISCC) para el 2020-2023 [35], se pueden encontrar las principales conclusiones y lineas de acción para combatir las distintas modalidades de delitos que ocurren en la ciudad. En este plan se menciona cómo Pereira ha sufrido problemas relacionados con el narcotráfico y las bandas criminales, junto con factores culturales como la cultura de la ilegalidad, oportunismo, indiferencia, entre otros, que hacen parte de toda la descripción de la criminología ambiental presentada en la sección 2.1.3, que contribuye de gran forma a la delincuencia y la repetición de la misma. Si bien el PISCC es un plan con lineas de acción tanto sociales como políticas, este trabajo puede servir como complemento para planes de seguridad futuros y fortalecer la toma de decisiones y generación de políticas públicas, ya que se tiene en cuenta un aspecto más cuantitativo de la problemática.

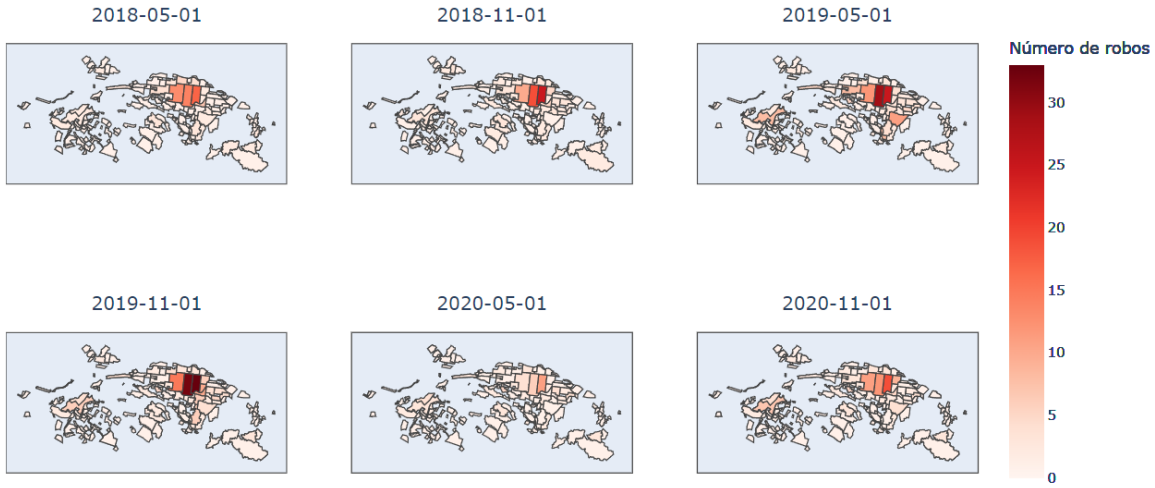


Figura 4.5 Mapas temáticos de los crímenes en Pereira para distintas fechas

4.3 Análisis exploratorio de los datos

Como se menciona en la sección 4.1, la base de datos proviene de los delitos reportados a la Policía Nacional y se tienen un total de 6149 registros entre el 25 de marzo de 2018 y el 30 de junio de 2021. En la figura 4.3 se presenta el mapa temático para la ciudad de Pereira para el número total de casos por barrio en la base de datos. Como se mencionó anteriormente, no se incluyen los datos de Dosquebradas ya que no fue posible conseguir las geometrías de los barrios que lo dividen. Se puede observar que los crímenes se concentran principalmente en tres barrios centrales de Pereira, los cuales corresponden a Galería Central, Plaza de Bolívar y Lago Uribe en la comuna Centro de la ciudad, donde cada uno tiene el 12.51 %, 10.01 % y 6.89 % del total de los datos, respectivamente. En el caso del número de crímenes por Comuna, se obtiene que Centro con el 25.85 %, Universidad con 10.26 % y Rio Otún con 5.14 % son las comunas con el mayor número de casos.

En la figura 4.4 se presentan los patrones puntuales de los crímenes como otro tipo de representación de los datos en contraste al mapa temático observado. En este caso se puede observar los delitos cometidos en Dosquebradas, en donde se hace evidente la cercanía que existe con los delitos cometidos con Pereira, y en especial con el centro de la ciudad que presenta la mayor cantidad de casos. No obstante, no se observa una zona de mayor densidad en Dosquebradas, contrario al caso de Pereira.

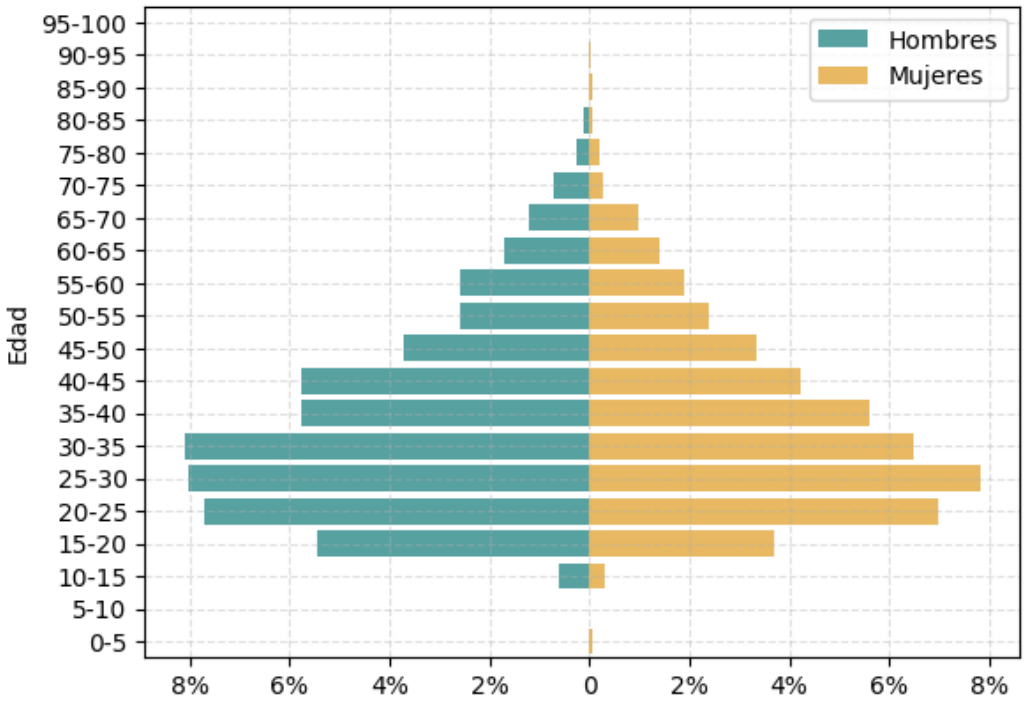
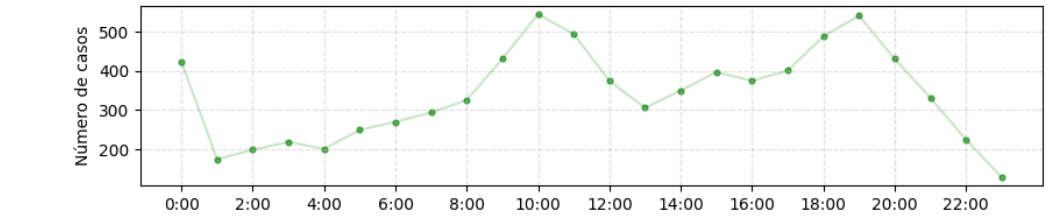


Figura 4.6 Pirámide poblacional de los crímenes en Pereira

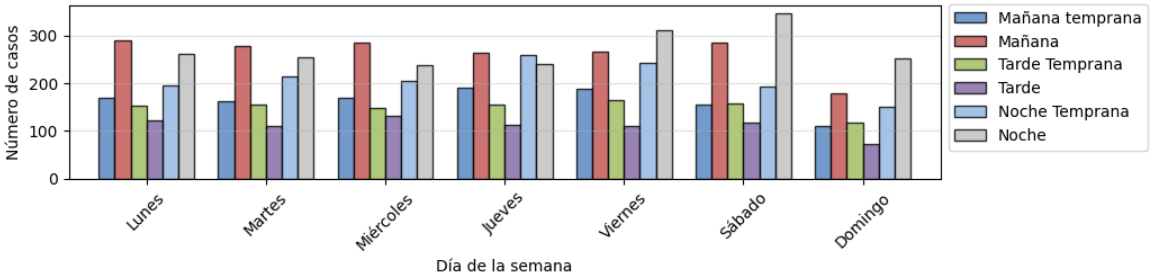
En la figura 4.5 se presentan los mapas temáticos de los delitos para distintos meses[2] para entender la temporalidad de los datos. Se observa que el crimen se comporta como patrones puntuales en la ciudad, donde hay zonas con una mayor densidad de crimen (Centro) y otros barrios con posible datos atípicos o una mayor ocurrencia. Este patrón permite entender que este sistema está presentando relaciones espaciales que pueden ser modeladas a través de una red compleja, y que además se están formando clusters en la ciudad de acuerdo a lo comentado en la sección 2.1.3, donde se hace evidente que existen relaciones sociales y económicas que de acuerdo a la Criminología Ambiental contribuyen a la repetición de la problemática. Si bien en algunos meses parece existir una menor densidad de crímenes, al observar todos los meses se puede encontrar que las zonas con mayor número de delitos son las mismas en todo momento.

Como se mencionaba anteriormente, la base de datos también contiene información demográfica de las víctimas de los delitos. En la figura 4.6, se presenta la pirámide poblacional por grupos quinquenales, la cual es una herramienta valiosa para entender la estructura de

[2]Se recomienda visualizar los datos a través del nootebok en el repositorio compartido ya que se construye un mapa de la ciudad donde cada capa corresponde a cada uno de los crímenes en los distintos meses para un mejor entendimiento.



(a)



(b)

Figura 4.7 (a) Número de casos por hora del delito. (b) Número de casos por día de la semana y parte del día

la población en una determinada área geográfica. En este caso, el gráfico muestra que los hombres presentan un mayor porcentaje de robos en comparación con las mujeres. Esta tendencia se mantiene en todos los grupos de edad, lo que sugiere que los hombres tienen una mayor probabilidad de ser víctimas de un crimen en comparación con las mujeres. Este resultado contrasta con la población de Pereira, que tiene más mujeres que hombres, lo cual sugiere que hay factores específicos en esta área que están contribuyendo a esta distribución desigual de género. No obstante, desde la base de datos no se tiene información que pueda contribuir al por qué de este resultado, por lo cual sería necesario realizar un estudio más a profundidad desde la criminología ambiental para tener en cuenta aspectos sociales, económicos y culturales; sin embargo, estos factores no son objeto de estudio en este proyecto.

Adicionalmente, se hizo un análisis durante el día para ver cuales son las horas en las que más se cometen crímenes. en la figura 4.7a se presenta la gráfica del número de casos por hora para la totalidad de los crímenes. Se observan específicamente 3 picos en los datos: el primero, para la media noche, lo cual se puede deber a que algunas denuncias no registran la hora exacta y puede ponerse por defecto 00 : 00; en segundo lugar, las 10 : 00 ; y por último, las 19 : 00.

En base a las categorías del día establecidas en la sección 3.1.1, se realizó un análisis de

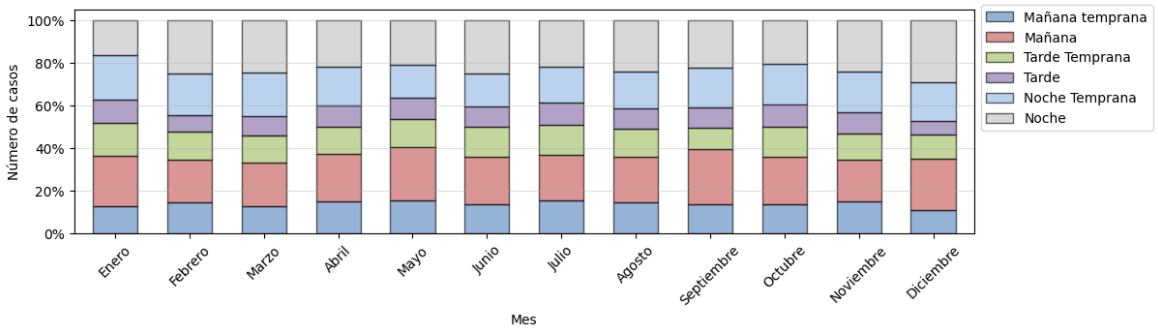


Figura 4.8 Porcentaje del total de casos en el mes por parte del día

los casos de crimen por día de la semana como se observa en la figura 4.7b. Los resultados indican que los viernes, sábados y domingos son los días de la semana en los cuales los crímenes se cometen con mayor frecuencia durante la noche. Por otro lado, de lunes a jueves se observa una mayor cantidad de crímenes durante las horas de *Mañana Temprana*. Y de manera general, las horas más seguras son en la *Tarde Temprana* y en la *Tarde*, lo cual puede estar relacionado con las horas de actividad laboral. A pesar de estas tendencias, se realizó una exploración más a profundidad de los crímenes en la ciudad por medio de estas categorías del día y encontrar si existía algún patrón en la ubicación espacial; no obstante, no se encontró algún patrón al respecto y de hecho siempre se observa que el centro de la ciudad es el área con mayor cantidad de crímenes, dichos resultados se pueden explorar en el repositorio del proyecto. Se podría concluir que la distribución de los crímenes en la ciudad no está necesariamente relacionada con el día o la parte del día, y más bien puede estar influenciada por factores como la densidad de población y la presencia de actividades económicas, que cambian durante los fines de semana y los días laborales.

Este mismo análisis se realiza para los meses del año tal y como se muestra en la figura 4.8. Teniendo en cuenta que para los tres primeros meses del año en algunos casos no se cuenta con datos, en la figura se muestran los resultados como un porcentaje del total de casos de ese mes. En este caso se observa que para el mes de enero la cantidad de crímenes en las noches baja en comparación a los demás meses, y por el contrario, diciembre presenta el mayor número de robos en la noche. No obstante, para los demás meses y partes del día, no se observan diferencias significativas en el porcentaje de casos. Si bien se pueden evidenciar algunas tendencias, sería necesario hacer un análisis estadístico a más profundidad para poder concluir que dichas diferencias son o no son significativas. No obstante, este análisis no entra en el marco de este proyecto.

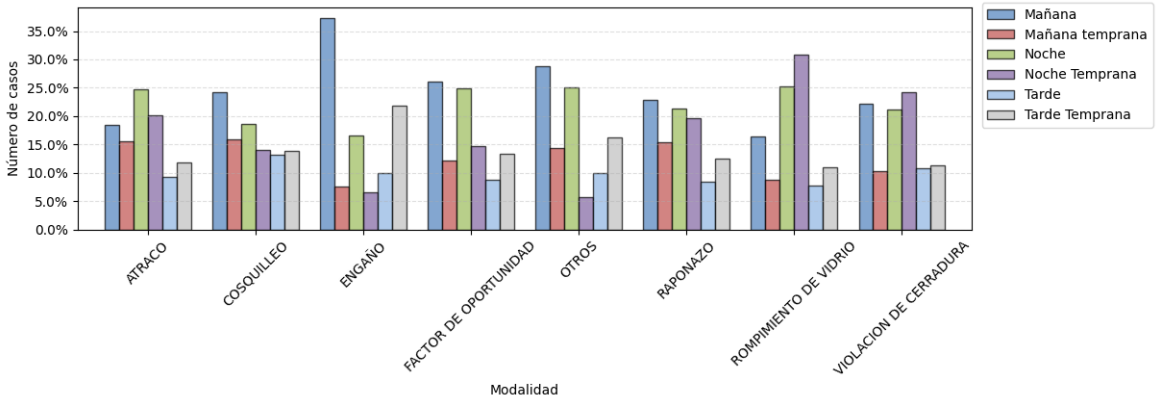


Figura 4.9 Distribución del porcentaje de casos por tipo de modalidad y parte del día

En el caso de las modalidades de robo, se tienen un total de 27 categorías diferentes, en donde solo 6 acumulan la mayoría de casos como se presenta en la tabla 4.1. En la figura 4.9 se presenta el análisis por partes del día para cada una de estas modalidades de robo. En este caso se observan diferencias significativas entre algunas de estas categorías, donde Engaño ocurre mayormente en la *Mañana* y Rompimiento de Vidrios o Violación de Cerradura en la *Tarde*. Si bien se podría pensar en realizar un análisis de redes espacio-temporales independiente para cada tipo de modalidad, es importante tener en cuenta que estas no son necesariamente independientes entre si. Posiblemente algunas de estas modalidades están relacionados con algunas otras, por lo cual en este proyecto se mantendrán las primeras 6 categorías de crímenes con más casos. En el caso de las demás categorías, al tener muy pocos datos, se pueden llegar a falsas relaciones entre robos solo por la cercanía entre estos casos, que son de muy baja ocurrencia y posiblemente con una metodología, trasfondo y objetivo diferente.

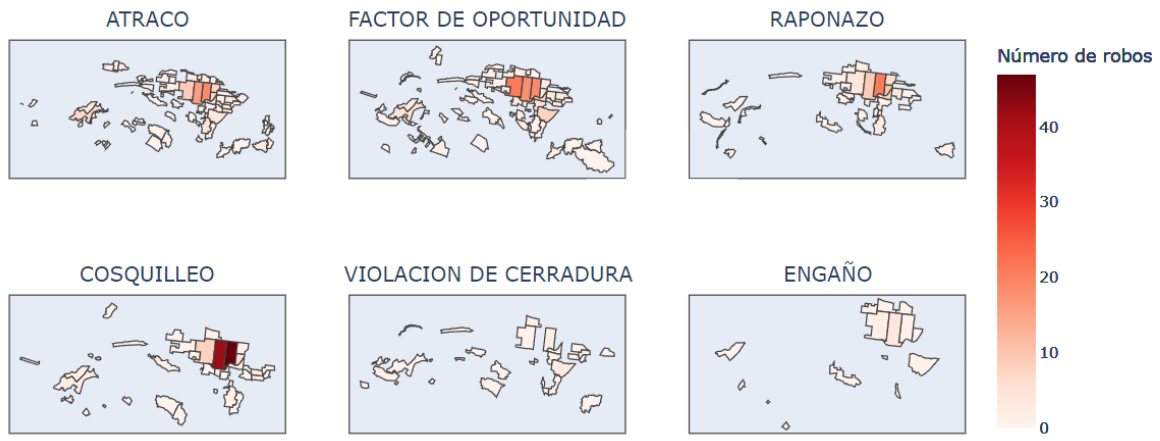


Figura 4.10 Mapas temáticos por modalidades de crimen

Modalidad	Número de Casos	% de Casos
Atraco	3109	38.61
Factor de Oportunidad	1991	24.72
Raponazo	1055	13.10
Cosquilleo	766	9.51
Violación de cerradura	514	6.38
Engaño	284	3.52
Rompimiento de Vidrio	182	2.26
Otros (21 categorías)	152	1.89

Cuadro 4.1 Número de casos para las distintas modalidades de crímenes reportados en Pereira y Dosquebradas.

Finalmente, para analizar la distribución espacial de los crímenes en Pereira, en la figura 4.10 se presentan los mapas temáticos de las diferentes modalidades de crímenes durante el último trimestre del año 2019. Se seleccionó este período en particular para evitar la superposición de patrones y permitir un análisis más detallado. Es importante mencionar que los casos en que los barrios no presenten casos, estos no se presentan en el mapa temático. Aunque algunas modalidades de robos parecen estar más dispersas, no se pueden extraer conclusiones significativas solo a partir de esta visualización. Sin embargo, se puede observar que los casos continúan concentrándose principalmente en el centro de la ciudad. En particular para cosquilleo, se observa un gran número de casos en esta zona de la ciudad, por lo cual parece ser una de las modalidades más predominantes en esta zona.

En resumen, desde el análisis exploratorio es importante identificar algunos patrones que las autoridades pueden utilizar para la mitigación del crimen. En primer lugar, es evidente que la zona centro de la ciudad presenta el mayor número de casos en todo momento. Si bien se puede tratar de un tema social y económico, es importante que las autoridades fortalezcan su presencia en estas zonas. En especial durante la *Mañana* y la *Noche Temprana* de lunes a jueves, y en las noches durante los fines de semana, ya que el crimen tiende a aumentar durante estos horarios.

Además, la distribución espacial del crimen no está necesariamente relacionada con la hora del día, pero la frecuencia de cada modalidad sí varía según el horario, por lo que es importante adaptar el control y la vigilancia en consecuencia por las autoridades. Por ejemplo, se observa que engaño y rompimiento de vidrios tienen tendencias más fuertes en en la *Mañana* y en la *Noche Temprana*, respectivamente. Adicionalmente, también se observa que Cosquilleo es la modalidad con más número de casos en la zona centro de la ciudad, por lo cual las autoridades deben buscar maneras de mitigar esta problemática, que no es común en el resto de Pereira.

4.4 Clusterización

En el análisis exploratorio previo, se observó que la mayoría de los crímenes se concentran en el centro de la ciudad. Al considerar la construcción de una red de eventos para este problema, resulta llamativo que los eventos ocurridos en el centro de la ciudad puedan relacionarse con eventos en zonas periféricas. Aunque es posible que existan dichas relaciones, los análisis de hora, día de la semana y mes, junto con las distribuciones espaciales de las modalidades de robo, sugieren una clara concentración de crímenes en el centro de la ciudad, probablemente debido a motivos económicos, sociales y horarios laborales con mayor ocurrencia de personas.

Por lo tanto, de acuerdo a lo presentado en la sección 3.1.2 y lo mencionado en la sección 2.1.3, se realiza una clusterización de los datos con el objetivo de que al construir la red de eventos solo los eventos que ocurren dentro del mismo cluster se puedan conectar entre si. En esta sección se aplicarán los métodos presentados en 2.4 y se argumentará cuál es la más adecuada para continuar con el proyecto propuesto.

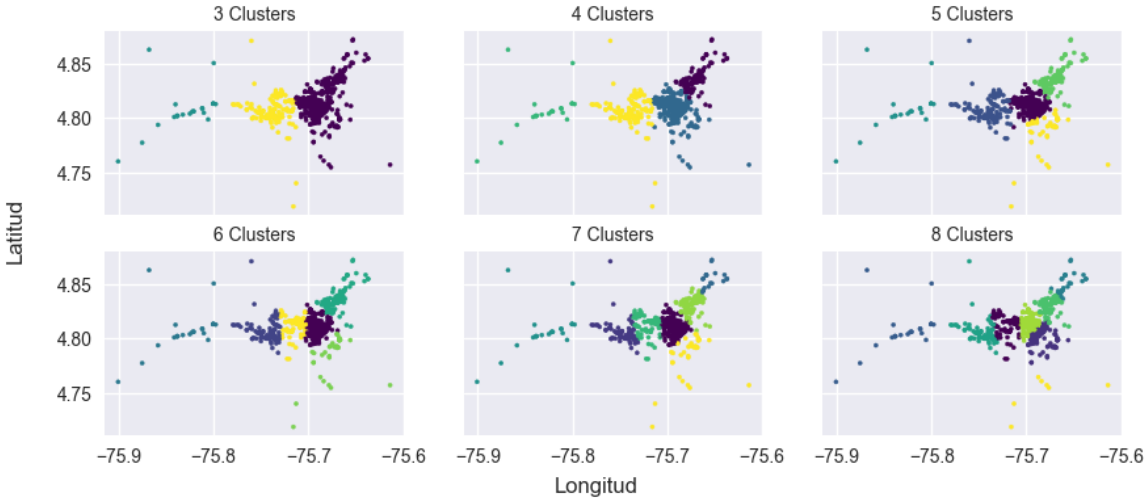


Figura 4.11 *k*-means con distinto número de clusters

4.4.1 *k*-means

Como se mencionó en la sección 2.4.1, la metodología de *k*-means es una técnica de clusterización que agrupa los datos en función de la distancia entre ellos y un punto central. Este enfoque cuenta con un parámetro k llamado número de clusters que debe ser seleccionado cuidadosamente, tal y como se presenta en la ecuación 2.23. Para determinar el número óptimo de clusters, se utiliza el método del *elbow* que permite evaluar la eficacia de la agrupación de los datos de acuerdo a lo presentado en la sección 2.4.1.

En la figura 4.11, se presentan los resultados del análisis de clusterización de los crímenes cometidos en los meses de mayo, junio y julio de 2019 utilizando el algoritmo *k*-means. Para determinar el número adecuado de clusters, se utiliza el método del codo o *elbow* presentado en el capítulo anterior. En la figura 4.12 se presentan los resultados por este método en donde se índica que el número óptimo de clusters para este caso corresponde a 6. Se puede observar que en todos los casos, el centro de la ciudad corresponde a uno de los clusters, y los barrios aledaños, incluyendo el municipio de Dosquebradas, corresponden a clusters independientes. Aunque este resultado sugiere de manera cualitativa que la metodología funciona correctamente, se debe tener en cuenta que el método *k*-means tiene una tendencia a dividir los datos de forma lineal (teniendo en cuenta que se ha utilizado al distancia euclídea).

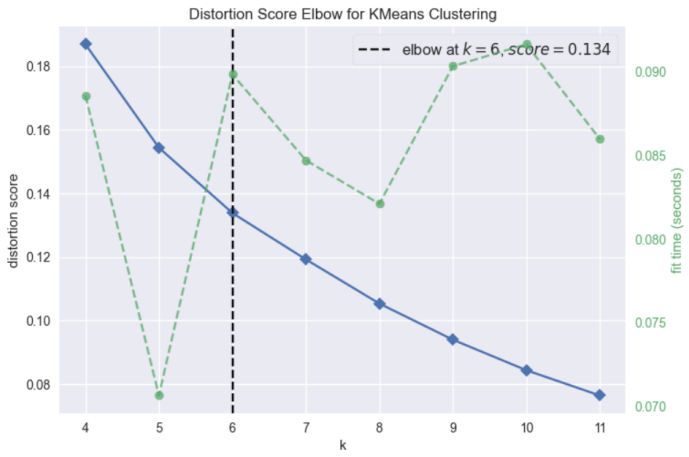


Figura 4.12 Método del *elbow* para *k*-means

4.4.2 DBSCAN

Con el fin de identificar estas relaciones no lineales propias de datos espaciales, se propone utilizar DBSCAN, una metodología de clusterización que se basa en la distancia entre los puntos y el número de puntos en una zona determinada como se presentó en la sección 2.4.2. Para elegir los mejores parámetros para DBSCAN, se utilizó el método del *elbow* nuevamente para encontrar el valor de la distancia ε. Para el número mínimo de puntos dentro del radio ε se escogió 5, de esta manera se mantiene que el algoritmo tenga en cuenta que algunas zonas de la ciudad son de alta densidad, y se evita que toda la ciudad sea considerada un único cluster o que por el contrario se tenga un número muy alto de clusters. Adicionalmente, se utiliza como métrica la distancia de Haversine, contrario a la distancia euclidiana utilizada en *k*-means. Como se mencionó en la sección 2.1.2, al utilizar coordenadas en el sistema WGS84, la distancia de Haversine es mucho más adecuada para estos casos, recordando que esta sigue la ecuación (2.5).

Teniendo en cuenta la metodología para la selección de los parámetros explicada en la sección 2.4.2, en la figura 4.13 se presenta el valor de la distancia entre cada punto (en radianes). En este caso, se puede observar que el valor óptimo corresponde a aproximadamente 0.00005 rad. En la figura 4.14 se pueden observar la clusterización de los datos por medio de DBSCAN con distancia $\varepsilon = 0.00005$ y el número mínimo de puntos MinPts $= 5$. La escogencia de estos parámetros permite capturar la densidad espacial de los datos observados. En este caso los crímenes marcados con color morado corresponden a datos atípicos. Al observar la distribución espacial de estos datos, se observa que la mayoría se encuentran en zonas periféricas o incluso en carreteras lejanas a la ciudad. Al realizar una inspección de estos

Figura 4.13 Método del *elbow* para DBSCAN

datos, se observa que algunos casos marcados en el municipio de La Virginia también se ubican en las mismas zonas, por lo cual pueden corresponder a errores en las marcaciones de los datos. Se observa un gran componente que corresponde al cluster del centro de la ciudad de la ciudad de Pereira y los crímenes en Dosquebradas. Este resultado puede indicar que los crímenes de este último municipio están ocurriendo en zonas muy cercanas al centro de la ciudad de Pereira, y por lo cual el crimen de este municipio está fuertemente influenciado por la zona central de la ciudad.

Finalmente, en la figura 4.15 se presenta la aplicación de este método para todos los datos espaciales y no solo el trimestre seleccionado. En este caso, se han descartado los datos considerados atípicos y se pueden observar claramente los 5 clusters obtenidos.

4.4.3 Elección del método

En la clusterización de los datos espaciales, DBSCAN se destaca como una metodología más adecuada que *k*-means. El principal motivo es que DBSCAN utiliza la distancia de Haversine, la cual es más precisa para medir distancias geográficas que la distancia Euclidiana utilizada por *k*-means, tal y como se menciona en la sección 2.1.2. No obstante, es importante tener en cuenta que la distancia de Haversine no es adecuada en caso de que se quieran manejar distancias altas entre dos eventos, ya que al tratarse de una distancia sobre una superficie esférica, no considera posibles cambios en la altitud, y que el sistema coordenado WGS84 modela a la tiera como un esferoide oblato.

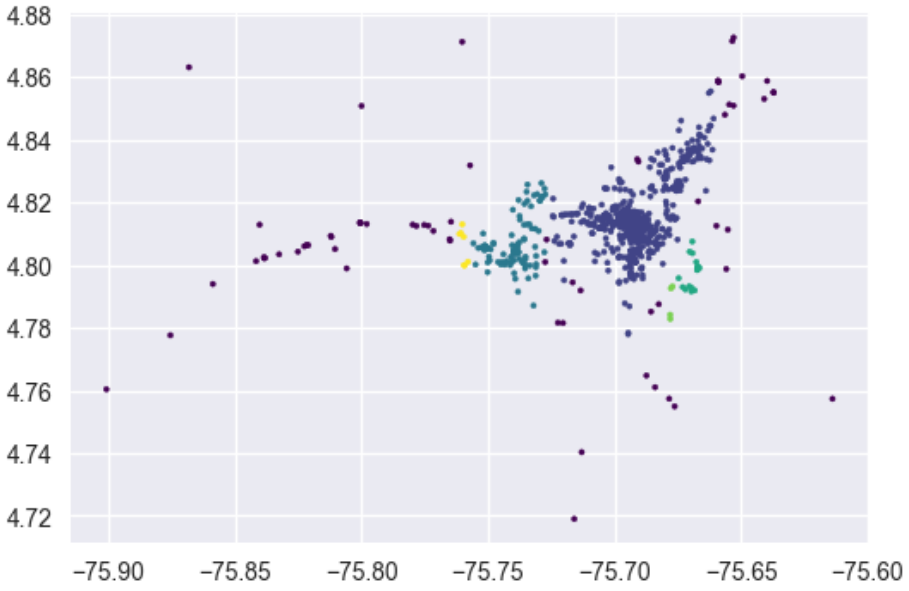


Figura 4.14 Resultados de los clusters obtenidos por DBSCAN

Adicionalmente, DBSCAN es más robusto que *k*-means, lo que permite descartar datos atípicos, es decir, aquellos que se encuentran alejados del resto de los puntos y que, por lo tanto, no tienen relación con el resto de los datos. Como se observó en la figura 4.15, en donde se han descartado los datos atípicos del crimen. En contraste a los resultados de *k*-means, a los cuales no se les puede dar un sentido socio-económico real.

Dada la metodología de DBSCAN, un resultado que llama la atención es la separación del cluster que compone el centro de la ciudad con el cluster del occidente de Pereira. Esto permite deducir que los crímenes que ocurren en la frontera entre ambos clusters no cumplen con las características de densidad propuestas, ya que por la forma en que está construido el algoritmo un cambio de cluster implica que no se cumple con el parámetro de *MinPts*, y por lo tanto, la densidad en ambos casos empieza a disminuir sobre sus respectivos bordes. Es importante que las autoridades tengan en cuenta esta observación, ya que si bien la zona centro presenta el mayor número de casos, también parece existir otra concentración del crimen en la zona oeste de la ciudad, con una menor frecuencia, pero igualmente importante para su mitigación.

El resultado anterior por el contrario no se observa en la conexión entre los municipios de Pereira y Dosquebradas, ya que tanto la zona centro como todo el municipio de Dosque-

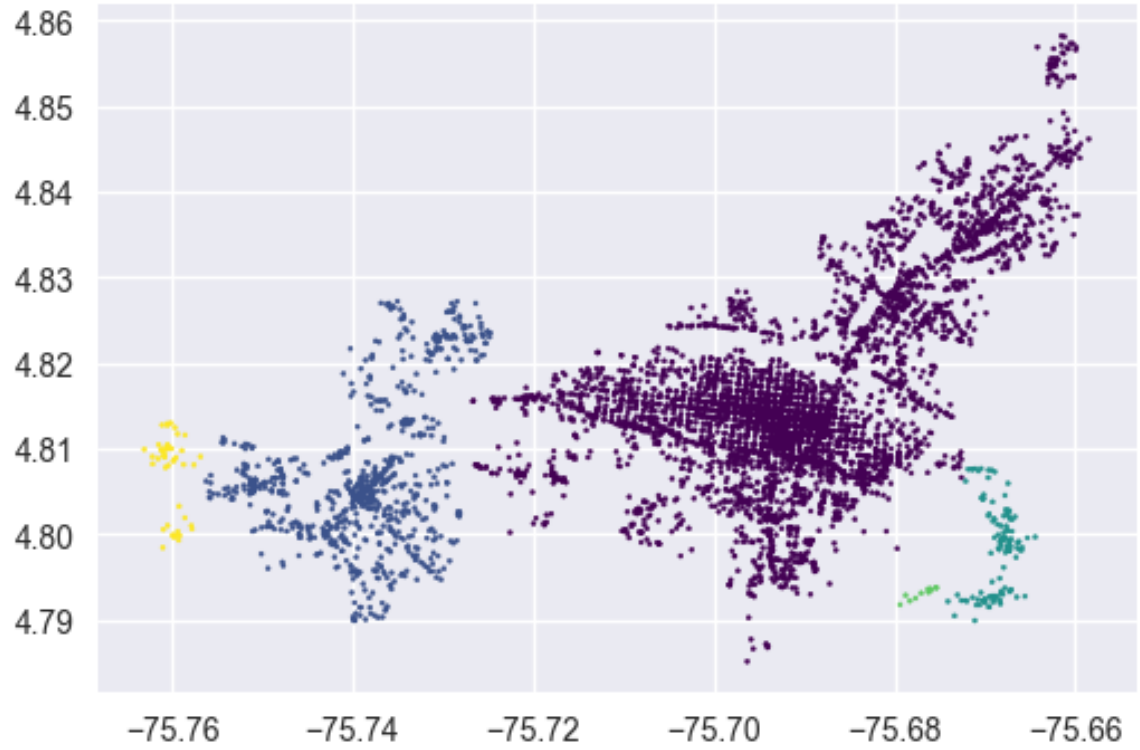


Figura 4.15 Resultado final por DBSCAN aplicado a todos los datos

bradas forman un único cluster. Esto puede significar que los crímenes en ambos municipios están altamente relacionados y que puede existir un efecto de contagio. Como medida de mitigación, las autoridades de Dosquebradas podrían trabajar en la frontera con el centro de Pereira para disminuir la densidad de crímenes en esa área y evaluar si esto tiene un impacto positivo en la reducción de la delincuencia en el municipio.

Dado que DBSCAN parece estar acorde al comportamiento observado para el crímen en la ciudad y los principios de la crimonología ambiental acorde a la realidad del municipio, para la construcción de la red de eventos se continuará con este método, descartando los datos atípicos encontrados. En trabajos futuros se podría explorar la posibilidad de utilizar k-means como alternativa u otro tipo de métodos de clusterización, ya que es importante observar como distintos algoritmos pueden arrojar diferentes resultados y por lo tanto, diferentes perspectivas sobre el crimen.

4.5 Red de eventos y subgrafos

Una vez se ha realizado la clusterización de los datos, es posible realizar la construcción de la red de eventos. Para esto, como se presentó en la sección 3.1.3, es necesario primero definir la distancia máxima en que dos eventos serán considerados cercanos para estar conectados dentro de la red. Una vez definida la distancia, se procederá con la metodología presentada en la sección 3.1.4 para la construcción de la red, y finalmente, la identificación de los subgrafos de acuerdo a lo presentado en la sección 3.1.5.

4.5.1 Selección de la distancia entre eventos

Como se menciona en la sección 2.3.4, para la construcción de la red de eventos es necesario definir el radio espacial D y el radio temporal T que definen cuando dos eventos son cercanos y estarán conectados, tal que el conjunto de los enlaces corresponda al presentado en la ecuación 2.23. Para esto, de acuerdo a lo mencionado en la sección 3.1.4, se utilizará el concepto de topología de la red para definir la distancia D. Para el caso del radio temporal T, se establecerá como un mes, es decir, los eventos dentro de cada mes de los datos estarán conectados.

Como se mencionó anteriormente, el trabajo de Broide y Clauset [29] pone a prueba que las redes que se forman en el mundo real siguen una topología libre de escala, es decir, la distribución del grado de las redes se pueden ajustar a una ley de potencias. Los resultados muestran que para el caso de las redes sociales, se encontró que estas se ajustan de mejor manera a una red con distribución de grado log-normal. Por este motivo, se realizará la bondad de ajuste de la distribución del grado de las redes de eventos obtenidas para las distancias 250m, 290m, 300m, 325m, 350m, 400m, 450m y 500m y ver si se ajusta a una distribución log-normal. Esto quiere decir que, siguiendo la teoría de redes de eventos de la sección 2.3.4, se construye una red para todas las combinaciones entre las distancias y los meses en la base de datos, es decir, se tendrán 8×36 redes de eventos, lo que corresponde a un total de 8 series de tiempo de acuerdo a la ecuación (2.24).

En la figura 4.16 se presenta la distribución del grado de la red de eventos correspondientes para cada distancia para el periodo 2019-12-01. La distribución log-normal que se utiliza para la prueba de bondad de ajuste se construye a partir de la librería *scipy*, en donde se estiman los parámetros de la distribución $\hat{\mu}$ y $\hat{\sigma}$ a partir del conjunto de datos (grado de cada nodo). Si bien los demás meses no se ilustran, este mismo ajuste en los datos se realiza para todas las 8×36 redes de eventos generadas.

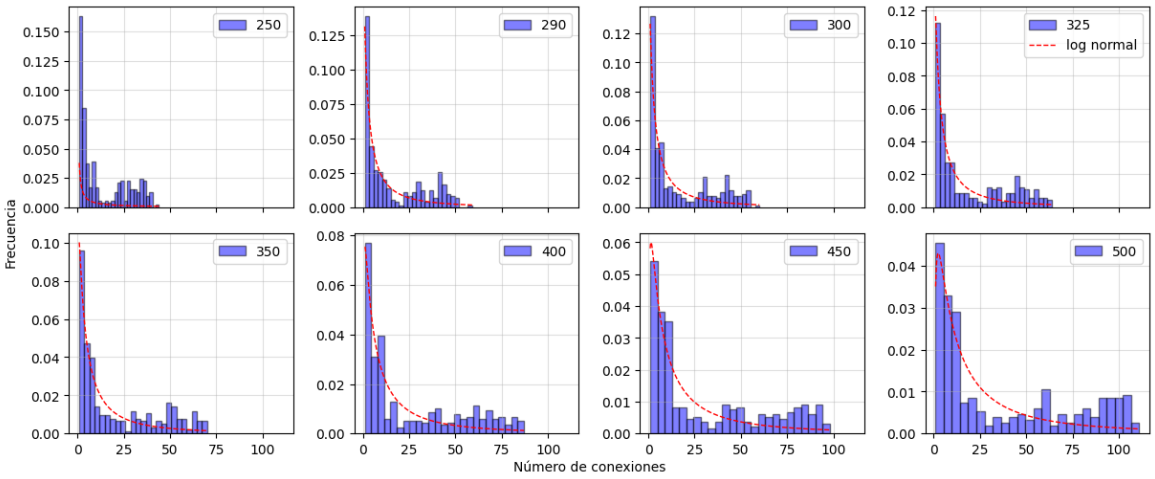


Figura 4.16 Distribución del grado de las redes generadas vs distribución log normal

En el cuadro 4.2 se presentan el estadístico de prueba D^{test} y el p-valor de la prueba de Kolmogorov-Smirnov para determinar si estas distribuciones del grado para cada una de las redes siguen una distribución log-normal. Se utiliza esta prueba perimétrica, ya que se tiene la propiedad de que sí dado un conjunto de datos X, este distribuye $X \sim lognormal(\mu, \sigma^2)$, entonces $\ln(X) \sim N(\mu, \sigma^2)$. La hipótesis nula H_0 corresponde a que los datos sí siguen una distribución log-normal. Los resultados completos se pueden encontrar en el repositorio mencionado al inicio de este capítulo.

Distancia	**Fecha**	**Estadístico**	**p-value**
250	2018-03-01	0.692	0.000
	2020-08-01	0.154	0.002
	⋮	⋮	⋮
	2020-10-01	0.127	0.007
⋮	⋮	⋮	⋮
500	2018-03-01	0.364	0.004
	2018-04-01	0.080	0.158
	⋮	⋮	⋮
	2021-06-01	0.587	0.000

Cuadro 4.2 Valor p para las distintas redes y diferentes distancias

Teniendo en cuenta el número elevado de redes de eventos, y por lo tanto de p-valores, se va a realizar la comparación del p-valor promedio para cada distancia y es sobre este valor que se hará la respectiva prueba. Dichos valores promedios se pueden encontrar en el

cuadro 4.3. Se observa que para las distancias 250, 290, 300, 325, 350 y 450, se rechaza la hipótesis nula de que los datos siguen la distribución log-normal para un nivel de confianza del 0.05. No obstante, para las distancias 400 y 500, no hay evidencia suficiente para rechazar esta hipótesis. Este resultado es muy importante, ya que como presentan Broide y Clauset [29], las redes complejas parecen comportarse de manera similar cuando corresponden a una misma disciplina o área de trabajo, y los resultados obtenidos permiten corroborar lo obtenido por estos autores, en los que las redes sociales se ajustan mejor a una distribución log-normal para la distribución del grado, es decir, los resultados se sustentan en la teoría.

Este resultado contrasta con los resultados obtenidos por Forero, Bohorquez y Renteria [2], ya que ellos utilizan el concepto de *scan statistics* para la definición de esta distancia. Si bien la metodología planteada por los autores tiene un sustento estadístico, el resultado obtenido demuestra que es importante utilizar el contexto del caso de estudio para el entendimiento del mismo como una red compleja. Esto permite además, que metodologías más cualitativas como la Criminología Ambiental sigan teniendo validez, ya que se muestra desde un aspecto cuantitativo como las redes complejas de crímenes mantienen características muy similares independiente de su ubicación geográfica, ya que hacen parte de la familia de redes sociales.

Distancia	**p-valor**	**Distancia**	**p-valor**
250	0.0003	350	0.0221
290	0.0024	400	0.0531
300	0.0018	450	0.0486
325	0.0212	500	0.0660

Cuadro 4.3 p-valor medio para cada una de las distancias

Si bien con los resultados anteriores ya se tiene un sustento suficiente para definir el valor de la distancia, en trabajos futuros podría plantearse realizar pruebas de bondad de ajuste para otros tipos de redes como los dos ejemplos de tipos de redes presentados en la sección 2.3.3, es decir, distribución del grado como una distribución binomial para el caso de las redes de Erdös-Renyí o redes aleatorias, y para una distribución de ley de potencias para el caso de la red libre de escala. No obstante, en este trabajo se realizará una discusión sobre ambos casos.

En el caso de una red con topología de Erdös-Renyí, se podría evaluar el caso límite de la distribución binomial mostrada en la ecuación (2.17), que tiende a una distribución de Poisson. Al evaluar los supuestos de esta distribución, se puede observar lo siguiente:

1. **Independencia:** Dada la naturaleza de los datos, no se puede establecer que los eventos ocurran de manera independiente en un intervalo de tiempo dado. Como se discutió en la sección 2.1.3, los crímenes pueden estar asociados a crímenes pasados, y al mismo contexto social y económico del área geográfica de estudio.

2. **Tasa constante:** La tasa media λ de la distribución de Poisson en que ocurren los eventos, es independiente de cualquier evento. Por lo cual sería necesario estimar dicho parámetro.

3. **Ocurrencias raras:** La distribución Poisson indica que la probabilidad de que ocurra más de un evento en un intervalo de tiempo muy pequeño debe ser cercana a cero. No obstante, la naturaleza de los datos no permite garantizar esto, ya que no existe completa independencia entre los crímenes, además dos crímenes pueden ocurrir en un mismo instante.

Por lo tanto, dada la naturaleza del crímen, no sería adecuado hacer un modelamiento por medio de una red aleatoria, ya que no se pueden garantizar los supuestos de la distribución de Poisson. Se puede descartar que los datos sigan una distribución de este estilo sin realizar una prueba de bondad de ajuste, ya que la naturaleza del problema lo impide.

En el caso de la red libre de escala se podría plantear realizar una prueba de bondad de ajuste para una distribución de ley de potencias, como se muestra en la ecuación (2.21), siguiendo lo establecido por Clauset [36]. Sí bien la metodología se puede explorar más a detalle en un trabajo futuro, vale la pena mencionar que justamente el trabajo anteriormente citado de Broide [29] y el propio autor de la prueba de bondad de ajuste para la distribución de ley de potencias Clauset [29], busca demostrar que las redes libres de escala no son comunes y no se pueden ajustar a cualquier tipo de caso de estudio, y precisamente, las redes sociales son una prueba de esto.

Entonces, para este proyecto y la construcción de la serie de tiempo de redes de eventos, se utilizará una distancia de 500m, ya que es la que tiene el mayor valor p. No se realizaron pruebas con distancias mayores, ya que el área geográfica de Pereira no es muy extensa, y un radio de $500m$ corresponde a un área de aproximadamente $0.7854km^2$ por nodo, por la cual una combinación de muchos nodos, como es el caso de la zona centro, abarcarán un área muy extensa si el radio escogido es mayor, lo cual se puede observar en la figura 4.17, donde se muestra que el radio de 700m ya abarca gran parte de todo el centro de la ciudad.

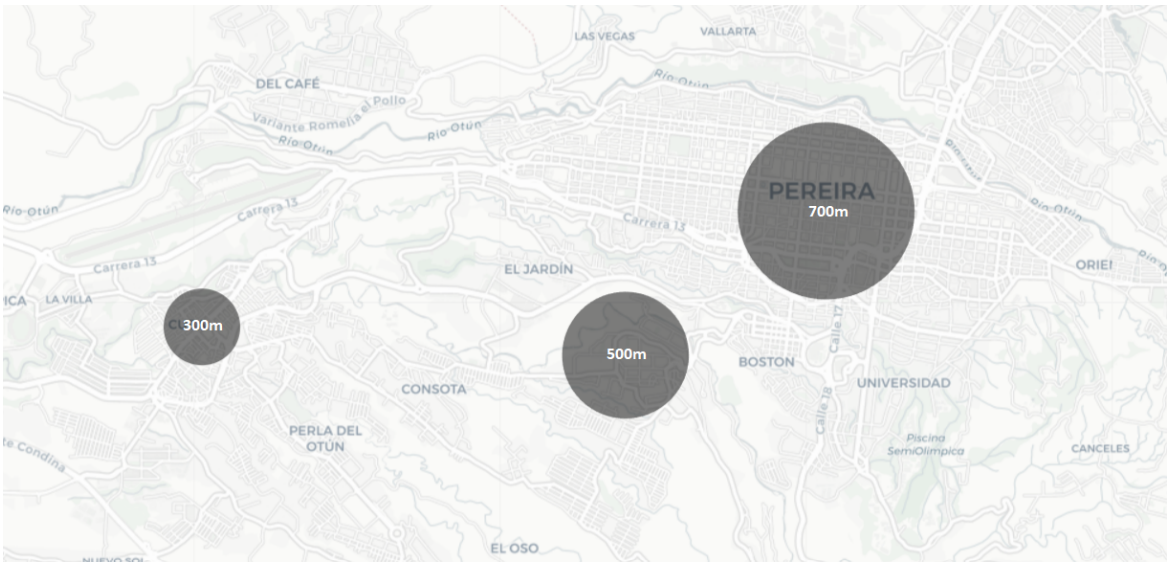
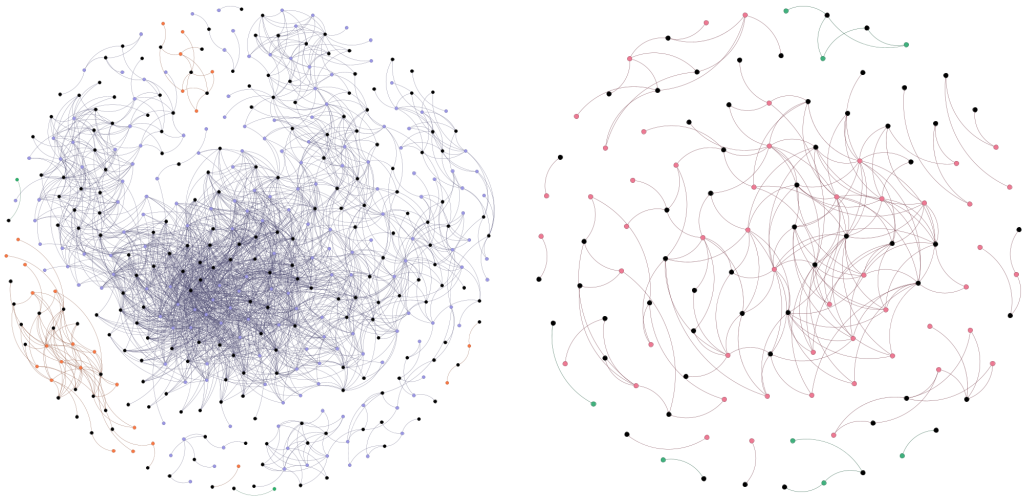


Figura 4.17 Ejemplo de diferentes valores de radio espacial sobre Pereira

4.5.2 Construcción de la red y conteo de subgrafos: Contribución computacional

Una vez seleccionada la distancia, se construyó la red de eventos nuevamente utilizando la metodología descrita en la sección 3.1.4. Como se mencionó anteriormente, se eligió el mes como distancia temporal y 500m como distancia espacial basándose en los resultados previos obtenidos. En la figura **??** se ilustran dos de los grafos obtenidos mediante este método. El color de los vértices indica el cluster al cual pertenecen los datos, en donde se hace evidente que en ambos casos siempre hay un cluster que predomina, y que corresponde a la zona centro de Pereira. Además, se puede observar que para el caso del grafo para la fecha 2019-05-01, se tiene la existencia de una gran componente corresponde a la zona centro de la ciudad. Si bien para el 2020-06-01 se observa un comportamiento distinto, este periodo corresponde a la pandemia por Covid-19, donde se observó una reducción significativa de los reportes de los casos; sin embargo, al visualizar estos datos sobre la ciudad, los nodos con mayor grado se ubican en la zona centro. Esta visualización de los grafos da señales de que puede ser relevante realizar medidas de centralidad sobre las redes de eventos, lo cual se examina al final de este capítulo. Es importante mencionar que si bien no se logra observar, estos grafos son dirigidos.

Una vez construida la red de eventos, se lleva a cabo el conteo de motifs de acuerdo a lo presentado en la sección 3.1.5, es decir, se utilizará la metodología de conteo utilizada por Forero, Bohorquez y Renteria[2], junto con la contribución computacional propuesta en este



(a) (b)

Figura 4.18 (a) Grafo construido para la fecha 2019-05-01. (b) Grafo construido para la fecha 2020-06-01

proyecto, la cual corresponde a la implementación de la Formula de Stirling de acuerdo a lo presentado en la sección 2.5.1. Se hace el conteo por ambas metodologías ya que se realizará una comparación de los resultados para ambos casos, y determinar si la contribución propuesta presenta algún tipo de ventaja con respecto al método propuesto por los autores.

En la figura 4.19 se presenta el conteo de los motifs *out-2-stars, in-2-stars y 2-path* para cada uno de los meses y por ambas metodologías. Se observa un comportamiento similar en todos los conteos de subgrafos, con un pico claro en diciembre de 2019, no obstante el conteo del número de motifs usando la formula de Stirling (figuras b) es mucho mayor. Este resultado se debe a que en primer lugar, la metodología original utilizada por el método ERGEN se basa en una muestra del total de los casos. Sin embargo, el método de Stirling es una estimación del conteo exacto de los subgrafos teniendo en cuenta que el grado del nodo ya es conocido, al igual que la dirección de sus vértices. Como se muestra en la ecuación (2.29), la formula de Stirling es una aproximación para factoriales grandes, lo cual se presenta en este tipo de redes.

Desde el punto de vista computacional, el cálculo mediante la fórmula de Stirling es computacionalmente menos costoso. Esta eficiencia se puede reflejar en el tiempo y consumo de los dos métodos. El método original presentado por Forero, Bohorquez y Renteria [2], tiene una duración aproximada de $8.78 \pm 0.219s$, en cambio la propuesta utilizando la Formula de Stirling tiene una duración de $1.56s \pm 0.0159s$. Adicionalmente, el método por la formula

Mes	*out-2-stars*	*in-2-stars*	*2-path*
2018-04-01	12428	11790	24378
	11787	12182	24071
	⋮	⋮	⋮
	11930	12531	24170
⋮	⋮	⋮	⋮
2021-06-01	13126	13296	27486
	12842	14125	27063
	⋮	⋮	⋮
	13447	13330	27170

Cuadro 4.4 Muestra de los conteos para las distintas permutaciones dentro de cada mes

de Stirling de la manera en que es planteada dentro del repositorio, permite aprovechar el consumo de los recursos del computador. En particular, estos modelos se realizaron en un computador con las siguientes características:

- Memoria RAM: 16gb
- Disco duro: 516 gb SSD
- Procesador: AMD Ryzen 7 4800HS Processor 2.9 GHz (8M Cache, up to 4.2 GHz)
- Núcleos del procesador: 8
- Tarjeta gráfica: GTX 1650 4GB GDDR6

En cuanto a los resultados, sí bien se presentan diferencias en los conteos de los motifs por ambas metodologías, ambos métodos capturan el mismo comportamiento en los datos. Esto es importante, ya que el método ERGEN no solo es un conteo por muestras, sino como ya se mencionó en la sección 2.5, este conteo hace parte de una metodología más robusta para la generalización y entendimiento de los subgrafos y el crímen. Por tal motivo, el hecho de que la formula de Stirling presente un conteo mucho más cercano al real, y además presente una ventaja computacional, hacen que el conteo del método ERGEN por medio de la propuesta dada sea mucho más eficiente para la generalización de los resultados. Por tanto en la siguiente sección se utilizara el método ERGEN en conjunto con el conteo por medio de la Formula de Stirling.

Adicionalmente, al analizar las diferencias en el valor del número de motifs durante el año, estas se deben principalmente al número de nodos y conexiones entre los meses. Como

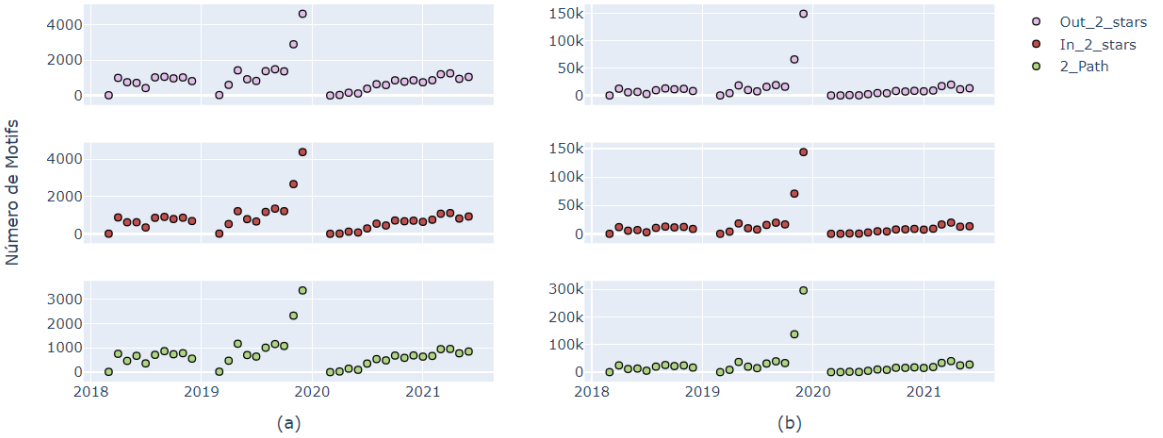


Figura 4.19 (a) Conteo de motifs por medio de la metodología propuesta por Forero A., et al.[2] (b) Conteo de motifs mediante la metodología propuesta.

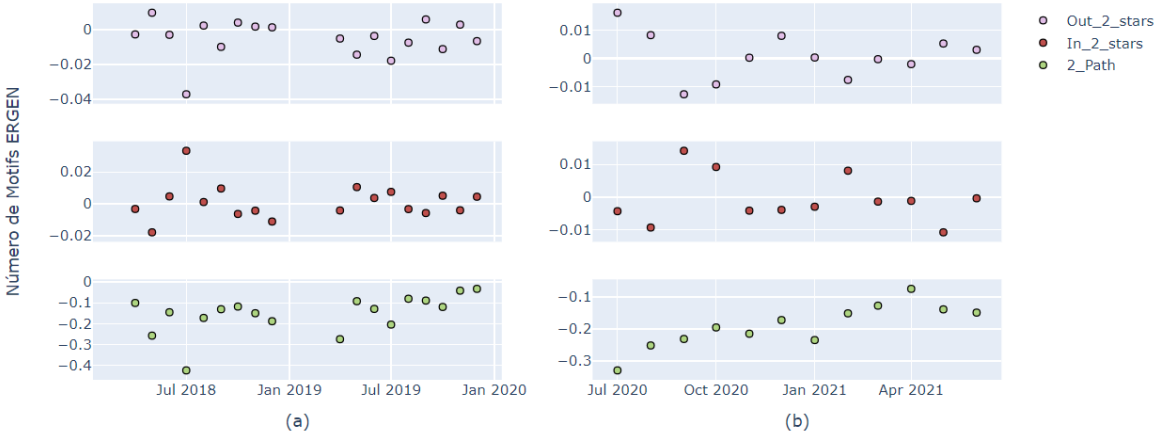


Figura 4.20 (a) Motfis hasta diciembre de 2019.(b) Motifs desde julio de 2020

se mencionó en la sección 4.1, no todos los meses, y en especial los primeros de cada año, presentan los mismos números de denuncias. Es por esto que es difícil realizar una comparación adecuada entre meses, ya que el conteo de subgrafos depende en gran medida del número de nodos de ese mes, es decir, entre más denuncias se hayan recibido, se encontrarán más subgrafos. El método ERGEN logra evitar este enmascaramiento de los datos, y de esta forma, poder comparar los resultados de una manera adecuada, sin depender de que en cada mes él número de denuncias haya sido similar.

4.5.3 Resultados y generalización de conteo de motifs

De acuerdo a lo anterior, se utiliza la metodología ERGEN para generalizar los resultados y reducir el enmascaramiento utilizando la contribución de este proyecto, es decir, mediante la formula de Stirling. Para ello, se realizarán permutaciones de los eventos en cada red de acuerdo a lo descrito en la sección 2.5. En el cuadro 4.4 se presenta una muestra de los conteos realizados para cada una de las permutaciones. Para poder generalizar, se procede a hacer la transformación de los datos mediante la ecuación (2.26) para cada uno de los meses t. Estos resultados se pueden observar en la figura 4.20, que presenta los resultados antes y después del inicio de la pandemia de Covid-19. Cabe mencionar que solo se incluyen los datos desde Julio de 2020 en el caso de los resultados posteriores al covid, ya que los reportes de crímenes fueron muy bajos en meses anteriores y no parece normalizarse hasta mediados de ese mes. Un número muy bajo de denuncias no permite que el método ERGEN logre evitar el enmascaramiento de los datos.

Los resultados de la figura 4.20 corresponden a series de tiempo del conteo de motifs normalizado. Para verificar que estas series de tiempo son estacionarias, se utiliza la prueba de Dickey-Fuller aumentada. Para esto, se define la hipótesis nula H_o : la serie de tiempo es no estacionaria. Los p-valores al aplicar esta prueba se presentan en el cuadro 4.5, en donde se observa que para todos las series menos *out-2-stars* y *in-2-stars* pre-pandemia, no existe evidencia a un nivel de confianza 0.05 para rechazar la hipótesis nula de que las series de tiempo son no estacionarias. No obstante, si se realiza la prueba sobre los datos completos, es decir, sin la separación pre y pos pandemia, se rechaza la hipótesis nula; por lo tanto, las series son estacionarias y es posible calcular una correlación entre ellas. No obstante, este resultado puede carecer de sentido real, ya que la pandemia por Covid-19 ciertamente ha afectado distintos procesos sociales, con lo cual utilizar todos los datos puede estar agregando un sesgo y ruido a las series y la estacionalidad sea una interpretación errónea de la realidad, como ya se ha mencionado anteriormente, y por tanto se aconseja no ignorar la naturaleza de los datos y las particularidades que presenta la base de datos.

Motif	Pre pandemia	Pos pandemia	Datos completos
out-2-stars	0.039	0.959	0.000001
in-2-stars	0.002	0.998	0.000003
2-path	0.998	0.119	0.000002

Cuadro 4.5 Valor-p de la prueba de Dickey-Fuller aumentada

Finalmente, se calcula la correlación entre las series de tiempo estandarizadas, donde los resultados se presentan en el cuadro 4.6. Se puede observar una correlación positiva casi

perfecta entre los subgrafos *out-2-stars* y *in-2-stars*, y en contraste una correlación negativa de estos dos subgrafos con el subgrafo *2-path*. En contraste con el número de subgrafos reales encontrados, y este nivel de correlación, se puede establecer que los motifs *2-path* son mucho más dominantes si se hace el conteo por medio de la formula de Stirling. Además, los subgrafos *out-2-stars* y *in-2-stars* parecen estar altamente relacionados en sus ocurrencias, tal y como se observa por su correlación y el número de casos, siendo siempre similar el conteo.

	out-2-stars	***in-2-stars***	***2-path***
out-2-stars	1		
in-2-stars	0.989	1	
2-path	-0.998	-0.995	1

Cuadro 4.6 Matriz de correlación de las series de tiempo estandarizadas.

De acuerdo a estos resultados, es posible generar propuestas para la toma de decisiones y la mitigación del crimen de la ciudad. Por un lado, los resultados usando un conteo por medio de la formula de Stirling, muestran que los subgrafos *2-path* ocurren con una mayor frecuencia. Por lo tanto, se recomienda que ante la ocurrencia de dos crímenes conectados de la forma en que se presenta este motif, las autoridades inspeccionen 500 metros alrededor de la zona del segundo delito, ya que es bastante probable que el tipo de crimen que se vaya a cometer siga el patrón de un motif *2-path*.

Por otro lado, los resultados una vez aplicado el método ERGEN, sugieren que, si los patrones observados en el crimen corresponden a subgrafos *in-2-stars*, es posible que después se pueda observar un motif *out-2-stars*, y de la misma manera en el caso en que el motif *out-2-stars* ocurra primero. Lo anterior, ya que como se observa en el cuadro 4.6, estos dos tipos de subgrafos están altamente correlacionados. En este sentido, las autoridades pueden responder de manera eficaz cuando observen este tipo de dinámicas, y lograr mitigar el crimen evitando la ocurrencia del siguiente motif.

Para el caso del motif *2-path*, este presenta una correlación negativa con respecto a los otros 2 motifs. Esto sugiere que si los crímenes observados mantienen este patrón, es muy probable que se mantenga y no se evidencien motifs *out-2-stars* o *in-2-stars*. Adicionalmente, dado que el mayor número de motifs corresponde a *2-path*, se puede interpretar que la dinámica del crímen que más se presenta en Pereira no corresponde a dos crímenes que desencadenen un nuevo crímen o un crímen desencadene dos eventos. En cambio, la criminalidad se está moviendo constantemente y parece hacerlo de manera individual. Por

lo tanto, se recomienda a las autoridades estar moviéndose constantemente 500 metros a la redonda de cada crimen una vez han observado la ocurrencia de este subgrafo. Es decir, es importante que el gobierno local le preste atención a bandas criminales que se mueven constantemente por los distintos clusters obtenidos.

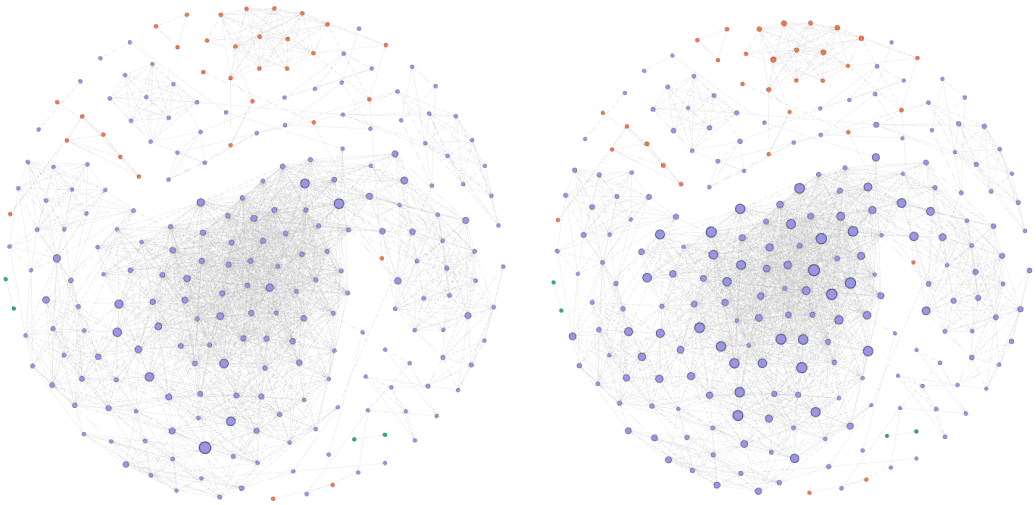
Es importante mencionar que estos resultados parten del supuesto de que tomar la base de datos completa, sin realizar un estudio por separado antes y después de la pandemia de Covid-19, corresponde a la realidad y el comportamiento del crimen antes del Covid sigue manteniéndose. En trabajos futuros puede ser importante realizar una comparación estadística de los crímenes antes y después del covid, y observar si efectivamente se ha recobrado el mismo comportamiento. Para esto, es importante la recolección de más datos, ya que como se observó en la metodología ERGEN, los meses de pandemia en los que hubo un reporte muy bajo de casos, la metodología no funciona correctamente.

Sin embargo, los resultados acá obtenidos pueden ser útiles para las autoridades locales. Ya que estos valores de correlación pueden llegar a observarse sobre el conteo de los motifs en la figura 4.19. Es importante tener en cuenta que las recomendaciones corresponden a una mitigación dentro de cada cluster.

4.5.4 Medidas de centralidad en las redes de eventos

Una vez realizado el análisis de los subgrafos y de acuerdo a lo presentado en la sección 3.1.6, se realiza un análisis de medidas de centralidad y asortatividad sobre los grafos dirigidos obtenidos. Estas medidas corresponden a las medidas presentadas en la sección 2.3.2, junto con medidas de asortatividad que pueden ser importantes dado lo observado en la figura 4.18b.

Una de las primeras medidas a evaluar corresponde a la centralidad de intermediación, que de acuerdo a la ecuación (2.11), determina la frecuencia en que un nodo está entre los caminos más cortos del grafo. Conceptualmente se puede entender como la importancia que tiene el nodo para comunicar otros nodos en el resto de la red. Teniendo en cuenta que esta medida le asigna un valor entre 0 y 1 a cada nodo, los resultados para todos los grafos no se presentan. No obstante, en la figura 4.21a se presenta el caso para la red de eventos del 2019-05-01, en donde el color del nodo corresponde al cluster al cual pertenece, y el tamaño del mismo corresponde al valor de la medida de centralidad. En primer lugar se puede observar que los nodos con mayor grado de centralidad corresponden a nodos en el cluster más grande, que corresponde al centro de la ciudad. Si bien gráficamente este resultado puede dar a entender que las autoridades locales pueden intentar identificar dichos crímenes para



(a) (b)

Figura 4.21 Grafo construido para la fecha 2019-05-01 con atributos para (a) centralidad de intermediación (b) Centralidad de cercanía

desarticular crímenes futuros, la medida de centralidad de intermediación realmente es muy baja para todos los nodos de la red, con un valor máximo de 0.0081 para el caso de estos nodos con mayor tamaño. Este mismo patrón se presenta para el resto de los meses, en donde el valor máximo para un nodo se presenta para el mes 2021-04-01 con un valor de 0.015. Por lo tanto, se puede concluir que no hay nodos específicos que jueguen un rol importante en la red para poder generar una desconexión o desarticulación de los crímenes futuros.

La siguiente medida a evaluar corresponde a la centralidad de cercanía, que de acuerdo a la ecuación (2.10), determina el valor promedio de las distancias geodésicas de un nodo hacia los demás. Sí bien su definición conceptual depende del caso de estudio, en este caso se puede interpretar como el impacto que tiene un primer crímen en la propagación de más crímenes dentro de la red. Esto quiere decir que para valores bajos de la cercanía, los crímenes han tenido un impacto importante dentro de la red de eventos ya que han logrado propagarse.

De la misma manera que para la centralidad de intermediación, los resultados completos no se presentan ya que esta medida le asigna un valor a cada nodo. No obstante, en la figura 4.21b, se presenta nuevamente el caso para la red de eventos del 2019-05-01, donde el tamaño del nodo corresponde al valor de la medida de centralidad. Gráficamente se puede observar que varios nodos pertenecientes al cluster del centro de la ciudad presentan valores más altos de esta medida en comparación al resto de nodos. Para este caso en particular,

el valor máximo de esta medida corresponde a 0.227, que si bien no es muy alto, indica que existen múltiples nodos que logran dicha propagación del crímen. Este mismo patrón se observa para todo el resto de redes de eventos, donde la mayoría de nodos presentan un valores cercanos a 0 para esta medida, y nodos en el centro de la ciudad presentan valores cercanos a 0.25. Contrario a la centralidad de intermediación, este resultado puede llegar a ser más útil para combatir la criminalidad, ya que las autoridades pueden intentar identificar estos nodos que propagan el crímen en la ciudad. Teniendo en cuenta la naturaleza de los datos y la forma en que se construyeron dichas redes, estos nodos corresponden a nodos que ocurren en los primeros días del mes y principalmente en la mañana.

Como una última medida, el cuadro 4.7 presenta el valor de la asortatividad para cada una de las redes de eventos. La asortatividad es una medida que índica cuándo los nodos tienden a unirse a otros que tengan un valor similar a su grado. Se puede observar que los valores obtenidos para todos los meses son similares. En general se tiene que el valor de la asortatividad de los redes está entre 0.15 y 0.35. Sí bien esto significa que las redes en general pueden ser no asortativas, se tiene una tendencia a valores positivos. Este valor de la asortatividad se puede deber a que la mayoría de los crímenes ocurren dentro de la zona centro de la ciudad, en donde los nodos con un grado muy alto se conectarán con otros nodos con grado muy alto, y en los demás clusters de la ciudad se tendrán que nodos con grado bajo se conectan con nodos de grado similar. Desde el punto de vista de las autoridades, esta medida solo refuerza la idea de que es importante centrarse en las zonas con una alta densidad de crímenes.

Fecha	**Asortatividad**
2018-03-01	0.362
2018-04-01	0.337
...	...
2019-09-01	0.291
2019-10-01	0.292
...	...
2020-07-01	0.343
2020-08-01	0.281
...	...
2021-05-01	0.192
2021-06-01	0.241

Cuadro 4.7 Asortatividad de las redes de eventos generadas para una distancia de $D = 500$.

Sí bien estas medidas dan una perspectiva más profunda del comportamiento del crímen en las redes de eventos, como se ha mencionado a lo largo de este proyecto, es importante que constantemente estos resultados se contrasten con el contexto y la naturaleza del caso de estudio, como se mencionó en la sección 2.1.3, ya que esto permite generar conclusiones y estrategias más adecuadas para combatir el hurto en la ciudad de Pereira y Dosquebradas.

Capítulo 5

Conclusiones y trabajo futuro

5.1 Conclusiones

En primer lugar, el análisis exploratorio de los datos del crimen en la ciudad muestra algunos indicadores tempranos e importantes que las autoridades pueden utilizar para mitigar la criminalidad. Se identifica que la zona centro presenta el mayor número de casos, como se observa en las figuras 4.3, 4.4 y 4.5, por lo cual es importante fortalecer la presencial policial en esta zona. Además, el crímen se presenta con mayor frecuencia durante horarios específicos del día como se presenta en las figuras 4.7a y 4.7b, como por ejemplo entre las 9:00 y la 13:00, y entre las 18:00 y las 19:00 para los días lunes, martes, miércoles y jueves. Para el caso de los viernes, sábados y domingos, se presentaría una mayor frecuencia para las noches.

Adicionalmente, al hacer un análisis por modalidad, en la figura 4.9 se observa como el crímen por modalidad tiene tendencias en diferentes horarios; no obstante, la modalidad no presenta alguna relación con la distribución espacial de los eventos. Se observa que el engaño y rompimiento de vidrios tienen un mayor número de casos para las franjas horarias de 9:00 y 13:00 para el primero y 18:00 y 19:00 para el segundo. Las autoridades pueden utilizar estos resultados junto con los resultados para el día de la semana, para establecer controles de vigilancia y control más estrictos en estas franjas horarias, y además fortalecer la seguridad de la zona centro de la ciudad, que está presentando la mayor densidad de hurtos. Adicionalmente, cosquilleo es la modalidad con un mayor número de casos en la ciudad, por lo cual es recomendable que las autoridades se centren especialmente en este tipo de hurto.

Es necesario que estos resultados se complementen con los estudios sociales y económicos del Plan Integral de Seguridad y Convivencia Ciudadana [35] para lograr medidas de miti-

gación más efectivas, de acuerdo al planteamiento de la criminología ambiental presentado en la sección 2.1.3. Esto, teniendo en cuenta que el análisis parece sugerir que la ubicación espacial y el horario del crimen están altamente relacionados con zonas que presentan una alta afluencia de personas por temas económicos. Además, los resultados presentados en la figura 4.6 dejan ver que existe una desigualdad de genero en donde los hombres parecen ser más afectados que las mujeres, lo cual contrasta con la razón de 89,1 hombres por cada 100 mujeres en la población de Pereira y Dosquebradas.

Teniendo en cuenta lo mencionado por Felson y Boba [11] y lo presentado en la sección 2.4, se realiza una clusterización de los datos por medio del método DBSCAN, ya que este presenta algunas ventajas sobre *k-means*. DBSCAN es un método robusto que permite identificar outliers en los datos, y además se corresponde a las dinámicas de la ciudad y la naturaleza de los datos, ya que permite utilizar la distancia de Haversine para el cálculo de la distancia entre eventos, contrario a *k-means* que utiliza la distancia euclidiana y además genera diagramas de Voronoi sobre en los datos. Gracias a los resultados de la figura 4.15 y la metodología de DBSCAN presentada en la sección 2.4.2, se puede concluir que la ciudad presenta dos clusters para los cuales las autoridades locales deben prestar mayor atención. El primero corresponde a la zona centro, la cual en los análisis descriptivos, y como se mencionó anteriormente, se observa que es la zona con mayor densidad y frecuencia de crímenes en cada etapa del año. El segundo corresponde al cluster aledaño a la zona centro en el oeste de la ciudad. Dada la metodología de DBSCAN, este cluster presenta otra zona con un alta concentración de crímenes y que puede ser separado de DBSCAN por una frontera con un menor número de casos.

Adicionalmente, se puede observar que DBSCAN no presenta esta división entre los municipios de Pereira y Dosquebradas, ya que tanto la zona centro como todo el municipio de Dosquebradas forman un único cluster. Como una posible medida de mitigación del crimen en Dosquebradas, las autoridades podrían enfocarse en la frontera con Pereira para disminuir la frecuencia y densidad de crímenes en esta área y observar si esto se refleja positivamente en la reducción de los delitos cometidos en este municipio. Estos resultados pueden estar asociados a las diferentes formas de Yuxtaposición Geográfica mencionados en la sección 2.1.3 propuestos por Cozens et al. [12]. Teniendo en cuenta los resultados de Jimenez [34], es importante que las autoridades de ambos municipios trabajen en conjunto para disminuir tanto el crímen de la zona fronteriza que se presenta como una forma de Proximal GJ, como en el centro de Pereira como una forma de Meso GJ, con el objetivo de mitigar el crimen en todo el municipio de Dosquebradas.

Al definir el radio espacial para la construcción de la red de eventos, se observó que al utilizar distancias de 400m y 500m para la conexión entre nodos, la distribución del grado para las redes de eventos puede ser ajustada a una distribución log-normal con un nivel de confianza del 0.05. Este resultado es importante ya que se sustenta en el trabajo realizado por Broide y Clauset [29], donde los autores concluyen que la distribución del grado de las redes sociales distribuye de esta manera. Esto significa que la mayoría de los eventos tendrán un valor bajo en su grado, es decir, la mayoría de las zonas de la ciudad, a excepción de la zona centro, tienen un número relativamente bajo de delitos. Es necesario mencionar que las redes complejas parecen comportarse de manera similar cuando estas corresponden a una misma área de trabajo, como lo son las redes sociales. Adicionalmente, al hacer un análisis sobre una red aleatoria, se puede observar que los supuestos de la distribución del grado para este tipo de redes no se cumple para la naturaleza de los datos del caso de estudio. Es por este motivo que es importante tener en cuenta la Criminología Ambiental al estudiar el crímen en una ciudad. Este resultado permite que, en caso de replicarse esta metodología para otras ciudades, se pueda modelar la red a través de la escogencia de una distancia que ajuste la distribución del grado a una distribución log-normal.

En el caso de implementarse políticas de seguridad en las zonas con mayor delincuencia, la distribución de los datos criminales debería cambiar de una distribución log-normal a una distribución más uniforme. Esto podría ser un indicador de que la política ha tenido éxito en reducir la delincuencia en todas las áreas, en lugar de concentrarse solo en unas pocas. Por lo tanto, esta distribución puede ser un indicador importante para las autoridades.

En la identificación de subgrafos, se observa que el método ERGEN en conjunto con la contribución de este trabajo que corresponde al conteo mediante la formula de Stirling, presenta una eficiencia computacional en contraste al método utilizado por Forero, Bohorquez y Renteria [2]. Esta eficiencia corresponde a $1.56 \pm 0.0159s$ frente a $8.78s \pm 0.219s$. Si bien estos resultados pueden variar dependiendo de la forma en que se optimice el algoritmo, el método de Stirling es una forma sencilla y efectiva de aproximar los factoriales que aparecen en el conteo de los Motifs *2-path*, *in-2-stars*, *out-2-stars*. En la figura 2.29, se puede observar que el conteo del número de Motifs por medio de la metodología propuesta por los autores, en comparación a la implementación de la formula de Stirling, varia en el número de subgrafos encontrados; no obstante, el método original parece ser solo una escala del método propuesto. Esto se debe a que el método ERGEN usa un conteo por muestras, y la formula de Stirling es una aproximación del conteo exacto de subgrafos. Por lo tanto, ya que la formula de Stirling

presenta la ventaja computacional mencionada y se acerca más al conteo exacto, se utiliza este método dentro del método ERGEN en el análisis de los subgrafos. Es importante resaltar que este método muestra la relevancia de la optimización de algoritmos en redes complejas y en particular para subgrafos, ya que el método propuesto es una alternativa que permite realizar análisis de casos de estudio con un número mucho mayor de datos puesto que se puede realizar una mejor asignación de recursos computacionales.

Con respecto a los resultados observados, es posible generar propuestas para la toma de decisiones que ayuden a mitigar el crímen de la ciudad por parte de las autoridades. Por una parte, dados los tres tipos de subgrafos, es importante que las autoridades se muevan acorde a estos para poder mitigar su ocurrencia. En particular, se observa una mayor frecuencia del subgrafo *2-path*. Por lo tanto, una de las principales recomendaciones consiste en inspeccionar el área a la redonda en un radio de 500 metros alrededor de la zona del segundo delito, de tal forma que este pueda llegar a desencadenar la aparición de este subgrafo. Adicionalmente, la metodología ERGEN nos permite observar que los subgrafos *in-2-stars* y *out-2-stars* están altamente correlacionados. Mitigar la ocurrencia de uno de estos grafos ayudará a que el otro también disminuya. Ante la aparición de un motif con estas características, es importante que las autoridades respondan de manera adecuada en búsqueda de mitigar la aparición de un segundo motif igual, o de su par correlacionado.

En el caso del motif *2-path*, este presenta una correlación negativa con respecto a los otros 2 motifs. Esto sugiere que ante la aparición de un motif *out-2-stars* o *in-2-stars*, no se da la aparición de un motif *2-path*. Por el contrario, dada su frecuencia y este resultado, puede darse la repetición de este mismo motif constantemente. Se recomienda a las autoridades vigilar áreas en un radio de 500 metros alrededor de cada crimen una vez han observado la ocurrencia de este subgrafo. Es decir, es importante que el gobierno local le preste atención a bandas criminales que se mueven constantemente por los distintos clusters obtenidos y no presentan un patrón tipo *in-2-stars* y *out-2-stars*, esto es, se mueven siempre de manera individual. Es importante mencionar que estos resultados de los subgrafos, son un complemento de los resultados de los clusters, el análisis exploratorio y la construcción de la red. Es decir, se llega a este resultado a través de la metodología descrita. Por lo cual, todas las recomendaciones corresponden a mitigar la criminalidad dentro de los clusters observados. Principalmente en la zona centro y la zona oeste de la ciudad de Pereira, teniendo en cuenta las características temporales de los datos mencionadas, y las modalidades de los robos.

Al realizar el cálculo de la centralidad de intermediación de los nodos, se puede concluir que

no existen nodos que jueguen un rol importante en el grafo de tal forma que las autoridades puedan concentrarse en estos casos para generar una desconexión o desarticulación en crímenes futuros. Esto se debe principalmente a que los resultados obtenidos para todos los meses presentan un valor cercano a 0, es decir, los nodos no tienden a comunicar a varios nodos dentro de la red. Por el contrario, en el caso de la centralidad de cercanía, se obtiene que para todos los meses el valor es cercano a 0.25, lo cual es un resultado útil para las autoridades, ya que estas pueden buscar identificar los nodos que propagan el crímen en la ciudad, que pueden corresponder a nodos que ocurren en los primeros días del mes y principalmente en la mañana.

Adicionalmente, la asortatividad obtenida para cada red de eventos varia entre 0.15 y 0.35, de acuerdo a los resultados de el cuadro 4.7. Sí bien no se puede concluir que las redes generadas son asortativas, si existe una tendencia positiva, por lo cual se refuerzan las conclusiones anteriores, en donde se propone que las autoridades se concentren especialmente en las zonas con una alta densidad de crímenes como lo es el centro de la ciudad.

5.2 Trabajo futuro

Este trabajo se puede extender desde los distintos ejes de estudio abarcados. Al tratarse de un caso de estudio con una aplicación en estadística, el alcance del trabajo es limitado; sin embargo, es posible que en futuros proyectos algunos de los métodos utilizados se sustituyan o se reevaluen. En particular, teniendo en cuenta los resultados del análisis descriptivo y exploratorio, es necesario realizar futuros análisis de significancia para determinar si las tendencias observadas en la frecuencia de los crímenes en los distintos meses del año son realmente significativas y deben ser consideradas en la toma de decisiones. Por ejemplo, se debe verificar la tendencia de aumento de crímenes en diciembre en la noche y si esta justifica un ajuste en el control policial, junto con un análisis que permita entender que tipos de modalidades pueden estar relacionadas entre sí y cuáles son independientes por sus características. De esta manera, se podría plantear el análisis de las redes de eventos por separado para cada una de estas modalidades.

Por otro lado, si bien la clusterización de los datos por DBSCAN parece ser una buena metodología de clusterización, existen múltiples metodologías adicionales que quizás tengan un mejor desempeño. DBSCAN tiene la ventaja que se puede utilizar cualquier tipo de métrica para la distancia, contrario a k-means en donde la distancia puede hacer no converger el algoritmo. No obstante, DBSCAN presenta una desventaja en las fronteras de los clusters, ya que los datos que se encuentren en esta zona pueden pertenecer a cualquiera de los clusters

a los que pertenecen. Además, es importante fortalecer los criterios de selección de los parámetros, ya que en este trabajo se han utilizado métodos heurísticos. Examinar métodos más formales ayudaría a tener un mejor sustento de los parámetros utilizados y mejorar los resultados obtenidos.

Adicionalmente, puede ser valioso realizar pruebas de bondad de ajuste en la distribución del grado de las redes de eventos para una distancia determinada, para distribuciones del grado teóricas de otros modelos de redes ya conocidos. De esta manera se pondría nuevamente a prueba los resultados de Broide y Clauset [29].

Para el análisis de los subgrafos en las redes de eventos, es necesario realizar un estudio del comportamiento antes y después del Covid-19. Esto con el objetivo de observar si es posible hablar de una dinámica en el comportamiento del crimen antes del Covid igual a la actualidad. En caso contrario, es importante realizar nuevamente el estudio de los subgrafos teniendo en cuenta esta separación. Es importante que la recolección de los datos por parte de las autoridades locales se pueda realizar correctamente, ya que en este proyecto se evidenció la falta de datos en algunos meses. Adicionalmente, se podrían explorar motifs con topologías diferentes a las utilizadas en este proyecto, siempre y cuando se correspondan a la naturaleza de los datos y se puedan implementar para generar políticas y estrategias enfocadas en la mitigación del crimen de la ciudad.

Adicionalmente se pueden explorar medidas de centralidad o de comunidades adicionales sobre las redes de eventos, ya que esto permite caracterizar y entender aún más el comportamiento del crímen en la ciudad. No obstante, es importante que las medidas escogidas sean acordes al caso de estudio.

Por último, es importante explorar las relaciones que pueden existir entre los clusters obtenidos. Esto es, ver si existe algún efecto de contagio o de difusión de la criminalidad, de forma que sea posible hacer un modelamiento que permita ver cuáles y cómo son los clusters que afectan al resto de la ciudad. Por los resultados observados, es posible que este efecto se esté dando entre la zona centro de la ciudad de Pereira, y el municipio de Dosquebradas, ya que el algoritmo estima que ambos pertenecen al mismo cluster.

Bibliografía

[1] Wikimedia Commons. Modificado por corso de [2] (gnufdl), cc by-sa 4.0 lt;https://creativecommons.org/licenses/by-sa/4.0gt;, via wikimedia commons, 2023.

[2] Alan Miguel Forero Sanabria, Martha Patricia Bohorquez Castañeda, Rafael Ricardo Rentería Ramos, and Jorge Mateu. Identification of patterns for space-time event networks. *Applied Network Science*, 7(1):1–24, 2022.

[3] Toby Davies and Elio Marchione. Event networks and the identification of crime pattern motifs. *PloS one*, 10(11):e0143638, 2015.

[4] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics reports*, 544(1):1–122, 2014.

[5] Laura Lotero, Rafael G. Hurtado, Luis Mario Floría, and Jesús Gómez-Gardeñes. Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes. *Royal Society open science*, 3(10):150654, 2016.

[6] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[7] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[8] Sebastian Wernicke. Efficient detection of network motifs. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(4):347–359, 2006.

[9] Thaddeus Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975.

[10] Rachel Boba. *Crime analysis with crime mapping*. Sage publications, 2016.

[11] Marcus Felson and Rachel Boba. *Crime and everyday life*. Sage, 2010.

[12] Paul Cozens, Terence Love, and Brent Davern. Geographical juxtaposition: A new direction in cpted. *Social Sciences*, 8(9):252, 2019.

[13] Sergey Dorogovtsev, Alexander Goltsev, and José Mendes. Critical phenomena in complex networks. *Reviews of Modern Physics*, 80(4):1275, 2008.

[14] Katarzyna Sznajd-Weron and Jozef Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06):1157–1165, 2000.

[15] Mark Newman. *Networks*. Oxford university press, Ann Arbor, 2018.

[16] Mark Newman, Albert-László Barabási, and Duncan J. Watts. *The Structure and Dynamics of Networks*. Princeton University Press, Princeton, 2006.

[17] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987), 2013.

[18] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, Inc., USA, 2011.

[19] Sergey Dorogovtsev and Jose Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

[20] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex Networks: Principles, Methods and Applications*. Cambridge University Press, USA, 1st edition, 2017.

[21] Linton Freeman. The development of social network analysis. *A Study in the Sociology of Science*, 1, 2004.

[22] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[23] Ernesto Estrada and Philip A Knight. *A first course in network theory*. Oxford University Press, USA, 2015.

[24] Miguel Ángel Quintero. *Ising Model applied to Complex Networks: Opinion formation in dialectic processes*. Tesis de pregrado inédita, Universidad Nacional de Colombia, 2019.

[25] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

[26] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[27] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[28] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[29] Anna Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

[30] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[32] Jacques Dutka. The early history of the factorial function. *Archive for history of exact sciences*, pages 225–249, 1991.

[33] Departamento Administrativo Nacional de Estadística (DANE). Censo nacional de población y vivienda 2018, 2018.

[34] Williams Gilberto Jiménez-García. Hacia una tipología de lugares peligrosos. caso de estudio de la comuna 11 de dosquebradas, colombia. *Revista Criminalidad*, 56(1):133–156, 2014.

[35] Municipio de Pereira. Plan integral de seguridad y convivencia ciudadana, 2020.

[36] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.