INVESTIGACIÓN __X_ PROFUNDIZACIÓN __

ANÁLISIS EXPLORATORIO ESPACIO-TEMPORAL DE LA DELINCUENCIA EN COLOMBIA ASOCIADA AL HURTO DE VEHÍCULOS ENTRE LOS AÑOS 2010-2019 MEDIANTE LA INTEGRACIÓN DE TÉCNICAS DE MINERÍA DE DATOS, GEOESTADÍSTICA Y SIG

Autor

Gabriel Hernán González Buitrago

Ing. Catastral y Geodesta

Tutor

Luis Eduardo Castillo Méndez

Matemático, Magister en Ciencia Estadística

Estudiante del Doctorado en Ciencia Estadística

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Maestría en Ciencias de la Información y las Comunicaciones

Énfasis en Geomática

Bogotá, Colombia

marzo de 2022

Nota de aceptación

_____

_____

_____

_____

_____

presidente del Jurado

_____

Jurado

_____

Jurado

_____

RESUMEN

Esta investigación surge con el objetivo de realizar un análisis exploratorio espacio-temporal de los hechos delictivos ocurridos en Colombia asociados al hurto de vehículos entre los años 2010 al 2019 mediante la integración de técnicas de minería de datos, Geoestadística y Sistemas de información Geográfica con el objetivo de identificar relaciones entre variables asociadas a las denuncias realizadas, caracterizar el delito y determinar las zonas donde se debe priorizar la atención del Estado. La investigación está basada en el método científico y está segmentada en la fase de observación, en donde se realiza un análisis exploratorio de los datos, seguido de la fase de experimentación comprendida por el uso de técnicas de minería de datos, geoestadística y modelamiento estadístico sobre los datos. Herramientas como el algoritmo K -means, árboles de decisión, el cálculo del estadístico I de Moran y un modelo de conteo permiten comprender a nivel municipio el comportamiento del delito

PALABRAS CLAVE

Algoritmo, Colombia, Delincuencia, Estadística, Geografía.

INTRODUCCIÓN

La delincuencia es una de las problemáticas existente en la mayoría de ciudades del mundo, en algunas de éstas la frecuencia de hechos delictivos asociados al hurto de vehículos se presenta en menor medida que en otras, así como la metodología empleada para efectuar el hurto puede variar según la región (Ospina & Correa, 2015). Por otra parte, gracias al crecimiento exponencial de nuevas tecnologías informáticas (nuevos Chipset, mayor capacidad de procesamiento de los equipos y costos menores) surgen conceptos como el Big Data y técnicas de minería de datos, ciencias que permiten transformar datos, aterrizarlos sobre un determinado contexto para convertirlos en información, procesarla, evaluarla y finalmente soportar la toma de decisiones (Katal et al., 2013).

En este documento se emplearán las herramientas de SIG, técnicas de minería de datos y Geoestadística para analizar y determinar relaciones entre múltiples variables asociadas a las denuncias realizadas en todo el territorio colombiano entre los años 2010 al 2019, y consecuentemente, identificar qué medidas o en qué zonas se requiere mayor prioridad por parte de la administración estatal para reducir la tasa de esta tipología de hurtos, con el objetivo de tomar medidas pertinentes para lograr una reducción significativa (Duce & Bandrés, 2001).

1. CONTENIDO

Resumen 3

Palabras Clave 3

Introducción 4

2. Problema de investigación 7

2.1 Planteamiento del problema 8

2.2 Formulación del problema 9

2.3 Sistematización del problema 9

3. Objetivos 10

3.1 Objetivo General 10

3.2 Objetivos Específicos 10

4. Justificación 11

5. Marco de referencia 11

5.1. Marco teórico 11

5.1.1 Método Científico: 11

5.1.2 Geoestadística: 13

5.2 Marco conceptual 15

5.2.1. Base de datos: 15

5.2.2. Análisis Exploratorio de Datos: 15

5.2.3. Algoritmo DBScan: 16

5.2.4. Algoritmo Apriori: 16

5.2.5. Árboles de decisión: 17

5.2.6. Modelo Estadístico: 18

5.2.7. Modelos de regresión con datos de conteo: 18
5.2.8. Modelos lineales generalizados: 18
5.2.9. Modelo de regresión Binomial Negativa: 20
5.3. Marco espacial 21
6. Hipótesis 21
7. Metodología 21
8. Experimentación y Análisis de datos 22
8.1 Identificación de las variables 23
8.2 Pre-Procesamiento de la información 23
8.3 Análisis exploratorio 26
8.3.1 Descripción de variables cualitativas 26
8.3.2 Descripción de variables cuantitativas 38
8.4 Algoritmo K-means 41
8.5. Reglas de asociación (algoritmo apriori) 50
8.6. Árbol de decisión 52
8.7. Geoestadística: 57
8.7.1. Detección de patrones espaciales: 58
8.7.2. Estadístico espacial I de Moran: 60
8.8. Modelamiento estadístico: 67
9. Conclusiones 71
10. Impacto y resultados esperados 73
11. Recomendaciones para trabajos futuros 75
12. Recursos y presupuesto 76
13. Referencias 76

2. PROBLEMA DE INVESTIGACIÓN

El propósito de la investigación consiste en reducir la delincuencia en el país por medio de la generación de resultados concretos que permitan tomar decisiones para reducir hechos delictivos por parte de las autoridades competentes en el país por medio de un análisis exploratorio espacio-temporal de la delincuencia asociada al hurto de vehículos en Colombia en los últimos 10 (diez) años a través de técnicas de minería de datos, Geoestadística y SIG que permita comprender las dinámicas asociadas a la delincuencia en todo el país.

Un artículo publicado en la revista Semana en el mes de marzo de 2018 resalta la oleada de inseguridad que por esos tiempos tenía en pánico a muchos colombianos, y hacían la evaluación de qué tan responsable es la falta de justicia y de autoridad. Según el artículo de la revista Semana llamado ¡Miedo! (2018) en el año 2018 los hechos delictivos, como el asesinato, secuestro, y extorsión se redujeron frente a los años anteriores; el primero ha caído a 24 por cada 100.000 habitantes, uno de los más bajos en 30 años, y el secuestro y la extorsión han bajado cerca del 10 por ciento según datos la Fiscalía y el Ministerio de Defensa.

En el año 2008 se realizó una investigación sobre la delincuencia en Colombia entre los años 1994 y 2007 por funcionarios del Ministerio de Defensa (Quintero & Lahuerta, 2008), en donde se trabajaron tres metodologías para entender la evolución de la criminalidad en el país, teniendo como base la combinación de varios delitos, y no ciñéndose únicamente al homicidio, en dicha investigación se entregó un índice asociado a cada delito en todo el país y concluyen que su investigación tenía limitaciones, pues para hallar un índice integral es necesario hacer un análisis regionalizado y por intervalos de tiempo específicos.

En Estados Unidos, según Bejerano (2018) la Policía de Memphis y Richmond afirmaron que los delitos serios se han reducido en un 30% en ciertas áreas de la ciudad y en un periodo de 12 meses gracias al Big Data. Ambos departamentos policiales utilizan sistemas de IBM para predecir los delitos. La clave está en el algoritmo de PredPol, un startup Californiana que usa Big Data para predecir cómo se comportará la delincuencia.

El cofundador de PredPol, Brantingham afirma “Trabajamos en predecir cuándo y dónde es más probable que ocurran delitos”, y juegan con márgenes de tiempo muy estrechos; las predicciones se hacen para las próximas diez o doces horas, y las áreas marcadas también son muy concretas, alrededor de 150x150 metros, señala Bejerano (2018).

Adicionalmente, aclara nuevamente Brantingham “No decimos que con toda seguridad aquí se producirá un delito. Asignamos una probabilidad de que se produzca un delito en una determinada localización”, Para hacerlo utilizan una materia prima sorprendentemente sencilla. “La única información que usamos son los datos históricos: qué tipo de delito es, dónde ocurrió y cuándo”.

Ramiro Rego, especialista en analítica de IBM, comenta que existe una labor más compleja en la recopilación de datos, y es la necesidad de combinar el contenido de informes policiales, a veces escritos a mano, con información meteorológica (la criminalidad varía si llueve o nieva) y con ciertas coyunturas, como si es día de pagos -en Estados Unidos muchas empresas aún utilizan cheques- (Bejerano, 2018).

Los anteriores artículos tienen algo en común: la Estadística, la medición, el seguimiento y la correlación con variables no relacionadas directamente con los hechos delictivos, permite a las entidades competentes enfocarse en resolver cierto tipo de deficiencias en algunas áreas para reducir la delincuencia en forma paralela.

2.1 Planteamiento del problema

Diagnóstico de la situación actual:

La delincuencia en el país es uno de los mayores problemas que enfrentamos los habitantes y considerando que toda la información de hechos delictivos se encuentra alojada en un único repositorio de información por parte de la Policía Nacional de Colombia, el cual ya ha sido suministrado para desarrollar esta investigación, es posible definir relaciones tanto espaciales como alfanuméricas entre las variables incluidas dentro del conjunto de datos, tales como: Tipo

de delito, Fecha, día, hora, arma empleada, móvil de la víctima, móvil del agresor, departamento, municipio, edad de la víctima, profesión de la víctima, entre otras.

Pronóstico: Teniendo en cuenta lo anterior, la presente investigación tiene como objetivo en primera instancia hacer el análisis exploratorio espacio-temporal entre los años 2010 al 2019 de la delincuencia en el país a nivel municipio usando técnicas de minería de datos, SIG y Geoestadística. Considerando el volumen de datos que serán analizados en esta investigación se prevé el hallazgo de relaciones entre el dataset suministrado por las autoridades que permitirán caracterizar este tipo particular de delincuencia a nivel municipal o regional.

Se presume que en las ciudades más grandes la cantidad de hechos delictivos denunciados sea más alta y que esté fuertemente correlacionado con los índices de pobreza, así como que en las zonas costeras es superior la cantidad de hurtos de motocicletas debido a que es el medio de transporte principal, por encima de los vehículos.

2.2 FORMULACIÓN DEL PROBLEMA

En consonancia con lo anterior, para definir estrategias que ayuden a contrarrestar los diferentes delitos que se presentan en el país, es necesario hacer un análisis exploratorio donde se construyan relaciones entre variables desde tipo alfanumérico hasta de índole espacial, incluyendo una dimensión temporal y generar índices de delincuencia a nivel municipio.

¿Cómo realizar un análisis exploratorio de las denuncias hechas por las personas entre el año 2010 y 2019 integrando herramientas de varios campos de estudio para alcanzar un objetivo específico y eventualmente identificar comportamientos y dependencias entre los datos?

2.3 SISTEMATIZACIÓN DEL PROBLEMA

¿Cómo realizar un análisis exploratorio de las denuncias hechas por las personas entre el año 2010 y 2019 integrando herramientas de varios campos de estudio para alcanzar un objetivo específico?, ¿las variables aportadas suministradas por la Policía Nacional tienen relación entre sí?, ¿la hora y el día (en el mes) genera un cambio en el tipo de arma y el tipo de delito

cometido?, ¿Entre las diferentes categorías de delitos hechas por los ciudadanos existe una relación espacio – temporal?, ¿ante una disminución o aumento abrupto de delitos de cierto tipo es posible relacionarlo con alguna medida tomada por el gobierno Colombiano?, ¿cómo calcular el índice de delincuencia que relacionen la cantidad de delitos reportados a nivel municipio, con variables como: la cantidad de habitantes por municipio, el número de vehículos registrados por municipio y los índices de turismo por temporada de los municipios?.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Proponer, implementar y realizar un análisis exploratorio espacio-temporal de los hechos delictivos ocurridos en Colombia asociados al hurto de vehículo entre los años 2010 al 2019 mediante la integración de Técnicas de minería de datos, Geoestadística y Sistemas de información geográfica.

3.2 OBJETIVOS ESPECÍFICOS

- Implementar un Sistema de Información Geográfica con información disponible en IDECA, el Geoportal del IGAC, y la plataforma de datos abiertos con el fin de generar insumos de tipo espacial y relacionarlos con la información alfanumérica suministrada por la Policía Nacional.
- Estudiar, seleccionar y usar técnicas de minería de datos, para encontrar y definir reglas de asociación entre los datos alfanuméricos asociados a hechos delictivos suministrados por las entidades encargadas de llevar el control en temas de seguridad en la ciudad.
- Buscar relaciones especiales (clustering) entre los datos delictivos georreferenciados suministrados por la Policía Nacional.

4. JUSTIFICACIÓN

Esta investigación tiene como objetivo realizar un análisis de datos exploratorio espacio-temporal de los datos mediante la integración de Técnicas de minería de datos, Geoestadística y Sistemas de información geográfica de los hechos delictivos en Colombia entre los años 2010 al 2019. Esto tiene por utilidad a que las instituciones encargadas del orden púbico enfoquen mejor sus recursos para prevenir que se sigan presentando hechos delictivos que afecten la integridad de las personas. La importancia de esta investigación radica, entonces, en el impacto social que pudiera tener en un futuro ya que la información generada se va a canalizar a las autoridades competentes.

Esta investigación busca aportar al conocimiento existente de técnicas de minería de datos enfocadas en el análisis de información espacial, así como ampliar la perspectiva a otros investigadores acerca de la integración de los sistemas de información geográfica con las Técnicas de minería de datos para el análisis de cualquier tipo de conjunto de datos.

5. MARCO DE REFERENCIA

5.1. MARCO TEÓRICO

5.1.1 Método Científico:

El desarrollo de esta investigación está fundamentado en el método científico, cuya definición indica que es un procedimiento seguido para lograr el conocimiento científico; un aspecto fundamental de este método es el aporta de objetividad a la investigación dado que gracias a la interrelación y evaluación de sus conceptos inherentes se busca la definición de leyes y/o eventualmente composición de teorías, así como la posibilidad de que distintos investigadores pueden averiguar el grado de verdad contenido en las teorías; lo anterior es

congruente con la frase “Una ciencia es una disciplina que utiliza el método científico con la finalidad de hallar estructuras generales” (Bungue, 1919).

El método científico es una metodología para generar nuevo conocimiento, fundamentada en un flujo que inicia con la observación sistemática, reconocimiento del problema, definición de hipótesis, experimentación, resultados y finalmente la comunicación de los resultados. En adición se destaca que las etapas son iterativas (Bungue, 1919).

A continuación, se listan y detallan los elementos que componen el método científico:

5.1.1.1. Observación: Es la adquisición activa de información por medio de la vista.

5.1.1.2. Reconocimiento del problema: Una vez se realiza una observación detallada de los fenómenos presentes en la naturaleza y éstos son llevados a un contexto en específico, el investigador debe centrarse en falencia o elementos susceptibles a mejorar.

5.1.1.3. Medición: Proceso que consiste en la comparación de un objeto o fenómenos basados en magnitudes físicas.

5.1.1.4. Hipótesis: Seguido de las fases anteriores, es posible crear enunciados no verificados, es decir, suposiciones, las cuales una vez refutadas o confirmadas dejarán de ser categorizadas como hipótesis.

5.1.1.5. Experimentación: En esta fase se llevan a cabo diferentes experimentos para comprobar o refutar la tesis.

5.1.1.6. Teorías: Son construcciones racionales que pretenden explicar la realidad basándose en un conjunto de leyes y son generadas a partir de la comprobación de hipótesis.

5.1.1.7. Conclusiones: Consiste en la comunicación de los resultados de una forma estructurada, clara y puntual.

5.1.2 Geoestadística:

En el marco del análisis de la criminalidad en ciudades cada vez ha tomado mayor relevancia la Geoestadística, debido a que décadas atrás era imposible pensar en analizar volúmenes masivos de datos, sin embargo en la actualidad con los avances tecnológicos es viable en términos de tiempo y recursos. En el artículo llamado Space-time Clusters of Crime in Stockholm, Sweden (Uittenbogaard & Vania, 2012) se hace un análisis de la delincuencia en la ciudad de Stockhold en Suiza tomado datos fuentes de la policía de dicha ciudad. Adicionalmente se realiza un análisis multitemporal para determinar si algunos días específicos del mes hacen que cambie la distribución de Hot Spots (valores donde se concentra la delincuencia).

En el artículo llamado Spatial correlations, clustering and percolation-like transitions in homicide crimes se menciona un aplicativo llamado DBScan que permite la identificación de patrones enfocados en la delincuencia que recibe datos como lugar, ocurrencia, armas usadas en el delito. Y de esta manera se genera cartografía con predicciones probabilísticas de los lugares del robo, este software se usa en varios países del mundo en la actualidad (Alves et al., 2015).

No solamente se puede implementar un modelo que permita detectar o estimar la probabilidad de que ocurran crímenes en ciertas zonas de la ciudad, sino que además es posible enfocarse en un fenómeno especifico, como el robo de carros. Para este caso, Ruth, Pierre, Robert P y Vania en su artículo Applying Geostatistical Analysis to Crime Data: Car-Related Thefts in the Baltic States, muestra como a partir de un krigeado, tras la implementación de un semivariograma, ajustar un modelo no lineal, y hacer test de correlación especial local y zonal se detecta que en ciertas zonas es más frecuente la presencia de delitos, y las calcula por medio de una distribución de Poisson (Ruth et al., 2010).

Referente al contexto colombiano, en Zipaquirá municipio de Cundinamarca se han realizado estos tipos de análisis con estadísticas suministradas por la Policía Nacional en formato Kmz (Rodriguez R., 2017) donde al inicio del documento se hace énfasis en toda la estructura del municipio, y después se realiza un análisis enfocado en la correlación

espacial contrastado con prueba de hipótesis al 5% de significancia. Así mismo, Ruth et al. (2010) emplea el test estadístico de Getis –Ord para comparar los resultados un análisis espacial local.

Adicionalmente, en el XXVIII Simposio Internacional de Estadística que tuvo lugar en el municipio de Bucaramanga, Santander se presentó un artículo titulado “Análisis espacio-temporal usando Redes y Datos Funcionales en la dinámica criminal de Medellín para el periodo 2003-2015” realizado por Alan Miguel Forero (2018), en el cual se realizó un análisis similar al que se propone en la presenta investigación enfocado solo en el municipio de Medellín, razón por la cual no fue necesario usar Técnicas de minería de datos; es importante destacar las conclusiones a las que llego el autor tras la conclusión del análisis, como, la conexión de las modalidades de hurtos en Medellín está dada por la ubicación geográfica de los barrios; es decir, en el centro de la ciudad circulan todo tipo de crímenes; Por medio de la estadística se observa que la intervención realizada en 2015 en la ciudad de Medellín tuvo un impacto diferente dependiendo del sexo y la edad de las víctimas: concentró el crimen de mujeres, y de hombres menores a 30 años alrededor del centro, pero no tuvo un efecto sobre los hombres mayores de 30 años. De la conclusión anterior es destacable como se puede relacionar las políticas públicas frente a la delincuencia para evaluar si las medidas tomadas han sido funcionales o no.

Actualmente existen algoritmos como el DBSCAN implementado en la versión 10 Pro de Arcgis que permite realizar análisis Geoestadísticos, donde el usuario orienta el tema a tratar, sin embargo, la plataforma no permite realizar o ingresar parámetros específicos para cada Dataset, por lo que se considera la implementación de todo el trabajo de grado en software como Quantum Gis o R.

5.2 MARCO CONCEPTUAL

5.2.1. Base de datos:

Conjunto de datos almacenados sistemáticamente en un repositorio para su posterior uso. Las bases de datos que además de almacenar información alfanumérica, permiten la manipulación de información geográfica son denominadas bases de datos espaciales, que dentro del ambiente del software desarrollado por ESRI, Arcgis son llamadas Geodatabase (Christian et al., 1998).

La Geodatabase es la estructura de datos nativa para ArcGIS y es el formato de datos principal que se utiliza para la edición y administración de datos. Mientras ArcGIS trabaja con información geográfica en numerosos formatos de archivo del sistema de información geográfica (SIG), está diseñado para trabajar con las capacidades de la Geodatabase y sacarles provecho (Esri, 2018).

5.2.2. Análisis Exploratorio de Datos:

Es un modo de análisis de los datos que utiliza el resumen numérico y visual para explorar datos en busca de patrones no anticipados (Yu, 2010).

En general, el análisis exploratorio de datos se caracteriza por el uso de herramientas o técnicas con mucha carga visual o gráfica, con énfasis en revelar información vital sobre la data examinada. El arsenal correspondiente está compuesto, entre otros, por instrumentos como:

- Diagrama de caja y bigotes (Box-and-whisker): un resumen visual de la distribución (comportamiento) de una variable que provee detalles acerca de si uno o ambos extremos de la distribución contienen valores inusualmente grandes o pequeños (Olivares, 2002).
- Diagrama de tallo y hojas (Stem-and-leaf): es una muestra visual de la distribución de una variable. Se asemeja a un histograma y se usa frecuentemente para mostrar tanto el orden de rangos como la forma de un conjunto de datos en forma simultánea” (Olivares, 2002).

- Diagrama de dispersión (Scatter plot): gráfico que muestra la relación entre variables. Es útil para examinar la dirección, fuerza y forma de la relación (Olivares, 2002).

5.2.3. Algoritmo DBScan:

El algoritmo basado en la densidad para el descubrimiento de clústeres en grandes bases de datos espaciales con ruido (DBSCAN) es un algoritmo de agrupamiento de datos (data clustering) propuesto por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiaowei Xu en 1996 (Schubert et al., 2017).

DBSCAN tiene un enfoque basado en la densidad, modelando los clústeres como cúmulos de alta densidad de puntos. Por lo cual, si un punto pertenece o no a un clúster, debe estar cerca de un montón de otros puntos de dicho clúster.

5.2.4. Algoritmo Apriori:

Algoritmo clasificado como no supervisado y es utilizado para la búsqueda de reglas de asociación, por medio de los siguientes conceptos:

- Soporte: El soporte del ítem o itemset X es el número de transacciones que contienen X dividido entre el total de transacciones.
- Confianza: La confianza de una regla “Si X entonces Y” se define acorde a la ecuación:

$$Confianza\ (X => Y) = \frac{soporte(union(X,Y))}{soporte(X)} \quad (1)$$

donde unión (XY) es el itemset que contienen todos los items de X y de Y. La confianza se interpreta como la probabilidad P(Y|X), es decir, la probabilidad de que una transacción que contiene los items de X, también contenga los items de Y.

La metodología empleada por el algoritmo inicia identificando los items individuales que aparecen en el total de transacciones con una frecuencia por encima de un mínimo establecido por el usuario. A continuación, se sigue una estrategia bottom-up en la que se

extienden los candidatos añadiendo un nuevo ítem y se eliminan aquellos que contienen un subconjunto infrecuente o que no alcanzan el soporte mínimo. Este proceso se repite hasta que el algoritmo no encuentra más ampliaciones exitosas de los itemsets previos o cuando se alcanza un tamaño máximo. Finalmente, se procede a identificar los itemsets frecuentes y, a partir de ellos, crear reglas de asociación (Rodrigo, 2018).

Por otra parte, tenemos estadísticos adicionales como Leverage, el cual puede tomar valores entre (-1,1) y permite identificar si los productos son independientes.

$$Leverage(A => B) = soporte(A => B) - Soporte(A) * Soporte(B) \quad (2)$$

Adicionalmente, se tiene el estadístico de convicción, que ayuda a determinar la dirección de la regla, dado que no siempre es lo mismo A=>B que B=>A, por lo tanto, n valor alto indica que el producto consecuente depende en gran medida del antecedente. Así como si el valor del estadístico es igual a 1, los productos son independientes (Rodrigo, 2018).

$$Convicción(A => B) = \frac{1-Soporte(B)}{1-Confianza(A=>B)} \quad (3)$$

5.2.5. Árboles de decisión:

Los árboles de decisión son uno de los algoritmos más utilizados en Machine Learning porque pueden ser usados para problemas de clasificación y de regresión. En éstos los nodos representan un atributo (Feature), cada línea (Rama) representa una decisión y cada hoja representa la salida de una variable categórica (Ghiasi & Zendehboudi, 2019).
Un árbol puede "aprender" dividiendo el conjunto de origen en subconjuntos según una prueba de valor de atributo. Este proceso se repite en cada subconjunto derivado de una manera recursiva llamada partición recursiva. La recursión se completa cuando el subconjunto en un nodo tiene el mismo valor de la variable de destino, o cuando la división ya no agrega valor a las predicciones. La construcción del clasificador de árbol de decisión no requiere ningún conocimiento de dominio ni configuración de parámetros, y por lo tanto es apropiada para el descubrimiento de conocimiento exploratorio (Ghiasi & Zendehboudi, 2019).
Por otra parte, este tipo de algoritmo se permite realizar tres tipos de aprendizaje:

- Hipótesis de aprendizaje inductivo: Cualquier hipótesis encontrada que clasifique un número grande de datos de entrenamiento clasificará otros elementos no observados.
- Razonamiento deductivo: partiendo de una premisa, se llega necesariamente a una conclusión, sin embargo, no aporta información nueva.
- Razonamiento abductivo: Partiendo del conocimiento de los efectos o síntomas de un fenómeno, se llega a la causa.

5.2.6. Modelo Estadístico:

Son herramientas que permiten realizar predicciones o inferencias mediante fórmulas y ecuaciones matemáticas. Estas ecuaciones están compuestas de variables independientes, por medio de las cuales se espera comprender y de cierta manera cuantificar el comportamiento de la variable dependiente (Kreyszig, 1979).

5.2.7. Modelos de regresión con datos de conteo:

Este tipo de modelos particularmente permiten analizar el comportamiento de variables asociadas a datos de conteo, como el número de accidentes, individuos de una especie, ente otros, frente a los valores del conjunto de variables explicativas (Kreyszig, 1979).

5.2.8. Modelos lineales generalizados:

Esta familia de modelos permite la construcción de modelos con variables de respuestas de tipo numérico y categórico, y considera a diferencia de los modelos lineales otras distribuciones además de la distribución normal. Se destaca que supone la esperanza está relacionada con las variables independientes a través de una función de enlace.

Se considera el supuesto de independencia para las observaciones, sin embargo la distribución puede ser heterocedástica, no se requiere supuesto de homogeneidad de varianzas (Atoche, 2017).

En un MLG, la media μ de la distribución depende de las variables independientes X, a través de la fórmula:

$$E(Y) = \mu = g^{-1}(X^t\beta) \tag{4}$$

Donde E(Y) es el valor esperado de Y, $(X^t\beta)$ es el predictor lineal y g es la función enlace.

La inferencia de este tipo de modelos se realiza mediante el método de Mínimos cuadrados o el método de máxima verosimilitud, el cuál es el más adecuado dado que tiene las propiedades de eficiencia asintótica y consistencia (Atoche, 2017)

El método de máxima verosimilitud consiste en maximizar la verosimilitud para obtener un estimador del vector de parámetros desconocidos $\beta = \beta_0\, \beta_1 \ldots \beta_p$ en el modelo muestral:

$$E(Y|X_i = x_i) = \mu_i = h(x_i'\beta) \quad i = 2, \ldots, n \tag{5}$$

Supuesto que el parámetro de escala θ es conocido y debido a que aparece como factor en la función de verosimilitud, se puede considerar θ=1, sin pérdida de generalidad.

Asumiendo que Y pertenece a la familia exponencial, la función de verosimilitud viene dada por:

$$L(\theta; \mathrm{y}) = f(y; \theta) = \prod_{i=1}^{n} f(y_i; \theta), \textit{ donde } y = (y_1 \ldots y_n)' \tag{6}$$

Por lo tanto la función log-verosimilitud viene dada por:

$$l(\theta, \Phi, \mathrm{y}) = \sum_{i=1}^{n} l_i(\theta, \Phi, \mathrm{y}_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\Phi)} + c(y_i, \Phi) \right\} \tag{7}$$

Se omite la función $c(y_i, \Phi)$ que no depende de θ_i y se inserta la relación $\theta_i = \theta(\mu_i)$ entre el parámetro natural y la esperanza de la i-ésima observación, obteniéndose la función:

$$l(\mathrm{u}_i, \Phi, \mathrm{y}) = \sum_{i=1}^{n} l_i(\mathrm{u}_i, \Phi, \mathrm{y}_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i(\mathrm{u}_i) - b(\theta_i(u_i))}{a(\Phi)} \right\} \tag{8}$$

Bondad del Ajuste:

Las medidas asociadas a la bondad de ajuste permiten medir la discrepancia entre los valores observados y los valores esperados en cualquier modelo. Específicamente para los MLG se puede evaluar la bondad del ajuste por medio del estadístico desviación D:

$$D(y;\hat{\mu}) = 2\Phi\{l(y;y) - l(\hat{\mu};y)\} \tag{9}$$

El estadístico mide la distancia entre el logaritmo de la función de verosimilitud del modelo saturado y el modelo que se está estudiando, de tal manera que un valor pequeño indica que para un número menor de parámetros se obtiene un ajusta tan bueno como cuando se ajusta el modelo saturado (Kreyszig, 1979).

Por otra parte, otro estadístico que permite medir la bondad de ajuste es el coeficiente de determinación R^2, el cual determina la calidad del modelo para replicar los datos y viene dado por la siguiente fórmula:

$$R^2 = 1 - \frac{D(y;\mu)}{D(y;\mu_0)} \tag{10}$$

Donde $D(y;\ \mu)$ y $D(y;\ \mu_0)$ son las funciones de desviación del modelo ajustado y nulo respectivamente.

5.2.9. Modelo de regresión Binomial Negativa:

A pesar de que el modelo de regresión de Poisson es fundamental para modelar datos de conteo y actualmente éste es la base de muchos tipos de modelos de conteo en la actualidad por científicos de datos, en algunas situaciones no ajusta correctamente los datos, específicamente cuando en los datos no coinciden la media y la varianza, por ejemplo, cuando existe sobre dispersión (Atoche, 2017). Para remediar y realizar un mayor ajuste, el modelo paramétrico que se debe usar es el Modelo de la regresión Binomial Negativa.

La distribución binomial negativa estudia la probabilidad de que en una variable aleatoria X se observen *x* fracasos antes del r-ésimo éxito en una serie de experimentos Bernoulli independientes. La variable aleatoria de conteo X sigue una distribución Binomial negativa de la siguiente forma:

$$P(X = x) = \binom{x+r-1}{r-1} p^r (1-p)^x = \frac{\Gamma(x+r)}{x!\Gamma(r)} p^r (1-p)^x \quad con\ x = 0,1,2,3 \ldots y\ r > 0 \tag{11}$$

5.3. MARCO ESPACIAL

El marco de referencia espacial donde se aplicará el desarrollo investigativo corresponde a Colombia, país ubicado en el trópico, en el continente de América que tiene costas sobre los océanos Pacífico y Atlántico. Su territorio es de 1´141.748 kilómetros cuadrados a los cuales se suman la plataforma marina y submarina, para un total de 2.070.408 km^2, tiene una población de 49.7 millones de habitantes a 2018, el idioma oficial es el Castellano y la moneda es el peso colombiano (Roa, 2016).

En Colombia hay 32 departamentos (dirigidos por gobernadores); 1123 municipios (dirigidos por alcaldes); cinco entidades territoriales con una administración especial (Bogotá y las ciudades puerto de Cartagena, Barranquilla, Santa Marta y Buenaventura); las entidades territoriales indígenas; y los territorios colectivos adjudicados a comunidades afrocolombianas en zonas del Pacífico (Sánchez-Torres, 2017).

6. HIPÓTESIS

Por medio de los resultados encontrados tras la implementación del árbol de decisión sobre los datos alfanuméricos asociados a la delincuencia y las diversas pruebas Geoestadísticas sobre información espacial de delincuencia compilada en un sistema de Información Geográfica es posible caracterizar la delincuencia en el país.

7. METODOLOGÍA

La metodología de esta investigación está basada fundamentalmente en el método científico y segmentada en función de los conceptos inherentes a éste. En primera instancia se mencionarán y describirán los conceptos y temáticas que se incluirán en el desarrollo de la investigación, luego en la fase de modelos se realizará el análisis exploratorio

(alfanumérico y espacial) de los datos y posteriormente se aplican técnicas de minería de datos para identificar patrones en las denuncias realizadas, por otro lado, se busca identificar hotspots a nivel municipio usando Geoestadística, específicamente por medio de análisis de datos de área (lattice data).

Finalmente, usando una distribución de conteo se implementará un indicador de nivel de delincuencia que facilite la interpretación del comportamiento de la delincuencia de los últimos años, para eventualmente teniendo en cuenta los resultados obtenidos de las fases anteriores definir hipótesis que serán evaluadas para determinar qué tan congruentes son y de esta manera describir las conclusiones de la presente investigación.

En la siguiente figura se describe de manera general la metodología de la investigación:

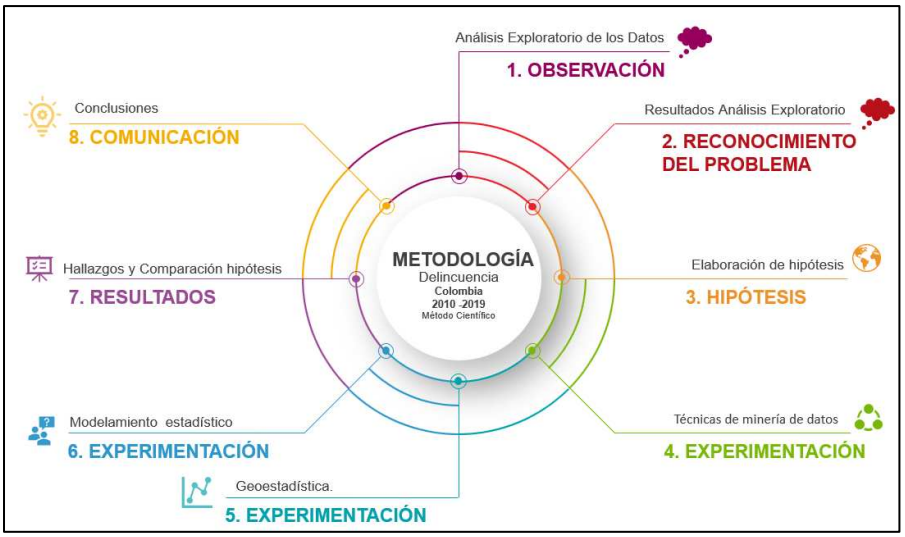


Figura 1. Metodología empleada para el análisis de la delincuencia.

8. EXPERIMENTACIÓN Y ANÁLISIS DE DATOS

El conjunto de datos fue tomado de la página web oficial de la Policía Nacional de Colombia, en el subsitio de Estadística Delictiva y contiene los registros de las denuncias realizadas

por personas nacionales y extranjeras, comprendidas entre los años 2010 y 2019 asociadas a hechos delictivos relacionados con el hurto de vehículos y motocicletas, las cuales tienen asociadas 25 variables tanto de tipo cualitativo como cuantitativo.

8.1 IDENTIFICACIÓN DE LAS VARIABLES

En la tabla 1 se categorizan los 25 tipos de variables existentes en el dataset, se da una corta descripción y un ejemplo particular para cada variable:

Tabla 1. Descripción detallada de variables

TIPO	SUB-TIPO	NOMBRE	DETALLE	EJEMPLO
CUALITATIVA	NOMINAL	Departamento	Nombre del departamento donde ocurrió el hecho delictivo	Cundinamarca
		Municipio	Nombre del municipio donde ocurrió el hecho delictivo	Bogotá D.C.
		Barrio	Nombre del barrio donde ocurrió el hecho delictivo	Alameda
		Zona	Zona donde se presentó el hecho delictivo	Rural
		Clase de sitio	Breve descripción del lugar suministrado poor la victima	Frente al hotel
		Arma empleada	Tipo de arma con la cual se cometió el hecho delictivo	Arma de fuego
		Móvil Agresor	Nombre del vehícuo o la forma en que se movilizaba el agresor	Moto
		Móvil Victima	Nombre del vehícuo o la forma en que se movilizaba la victima al momento del hurto	A pie
		Genero	Genero de la victima	M
		Estado_Civil	Estado civil de la victima	Casado
		Profesion	Profesión de la victima	Ingeniero Industrial
		Clase_Empleado	Breve descripción del tipo de cargo que desempeña la victima	Empleado
		Pais_Nacimiento	País de nacimiento de la victima	Colombia
		COD_DANE	Código según el Departamento Administrativo Nacional de Estadística del Dpto y Municpio	15525
		Clase	Tipo de vehículo que fue hurtado	Camioneta
		Marca	Marca del vehículo hurtado	Renault
		Linea	Línea del vehículo hurtado	Duster
		Color	Color del vehiculo hurtado	Negro
	ORDINAL	Día	Día de la semana en el cual ocurrió el delito	Miércoles
		Escolaridad	Nivel máximo de estudio de la víctima	Pregrado
		Fecha	Fecha en formato DD-MM-YY en la que ocurrió el hecho delictivo	20/01/2018
CUANTITIVA	CONTINUA	Hora	Hora en formato 00:00; 24 horas de cuando se presentó el hecho delictivo	13:24
	DISCRETA	Edad	Edad de la victima	26
		Modelo	Modelo del Vehículo hurtado	2010
		Cantidad	Cantidad de vehiculos interpuestos en la demanda	1

8.2 PRE-PROCESAMIENTO DE LA INFORMACIÓN

El dataset contiene 350.606 registros y 25 variables, para el análisis de la criminalidad en Colombia en los últimos 10 años se tienen 8.765.150 de datos, sin embargo, antes de

realizar el análisis exploratorio de los mismos, se requieren identificar, validar y depurar inconsistencias, omisiones y/o comisiones en la captura de la data.

En la siguiente figura se muestra un resumen de las variables que no tienen información o que las víctimas no reportaron al momento de denunciar, la cual es reemplazada por “Sin Información” en las variables de tipo cualitativas y por “-1000” en las cuantitativas. Del total de datos tan solo el 5,07% no tiene información, equivalente a 444.388 datos.

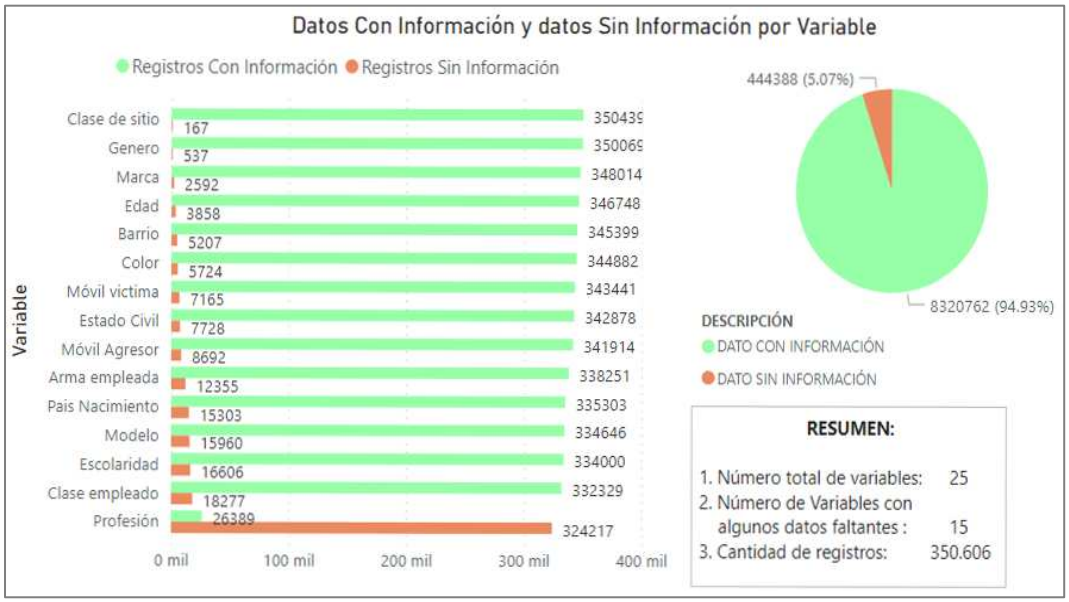


Figura 2. Resumen general del comportamiento de los datos.

En adición, el histograma de la variable Edad refleja que la mayor parte de los datos están concentrados en valores positivos, con una media de 34 años y un coeficiente de variación del 11%. En consecuencia, se plantea completar los valores sin información, en función de la media por municipio por medio de un script en Python (ver figura 3) y de esta manera asegurar que se completará con el dato central de la edad de cada municipio disminuyendo la variabilidad.

```python
#   PRE-PROCESAMIENTO DE LA INFORMACIÓN
import os
import pandas as pd
import matplotlib as plt
import seaborn as sns
from plotly.plotly import iplot
path="C:/Users/Gabriel_Toshiba/Documents/4_semestre_maestria/"

""" Lectura de datos """

data= pd.read_excel(os.path.join(path,"2_Data/","2_Hurto_de_automotores_y_motocicletas_2010_2019_b.xlsx"), sheet_name=0)
preprocesing= pd.read_excel(os.path.join(path,"1_Documento/","Excel_de_data.xlsx"), sheet_name="PRE_dataset")

"""Calculo de edades:"""

data.info()

print(data['Edad'].describe())
plt.figure(figsize=(40, 8))
sns.distplot(data['Edad'], color='y', bins=100, hist_kws={'alpha': 0.4}).set_title("Histograma Variable 'Edad' (Previo al reemplazo de medias)");

#Reemplaznado edadaes por la mediana según el municipio.
categoria = data.groupby(['Departamento'])['Edad'].mean()

tamaño=data.shape
filas=tamaño[0]
columnas=tamaño[1]
for i in range(0,filas):
    if data['Edad'][i]<=0:
        dpto= data['Departamento'][i]
        data_2 = data.replace(data['Edad'][i],categoria['Departamento'==dpto])

sns.distplot(data_2['Edad'], color='g', bins=100, hist_kws={'alpha': 0.4}).set_title("Histograma Variable 'Edad' (Después al reemplazo de medias)");

data_2.to_excel(os.path.join(path,"2_Data/","Hurto_de_automotores_y_motocicletas_2010_2019_corrected.xlsx"), engine='xlsxwriter')
```

Figura 3. Script en Python para completar datos en la variable Edad.

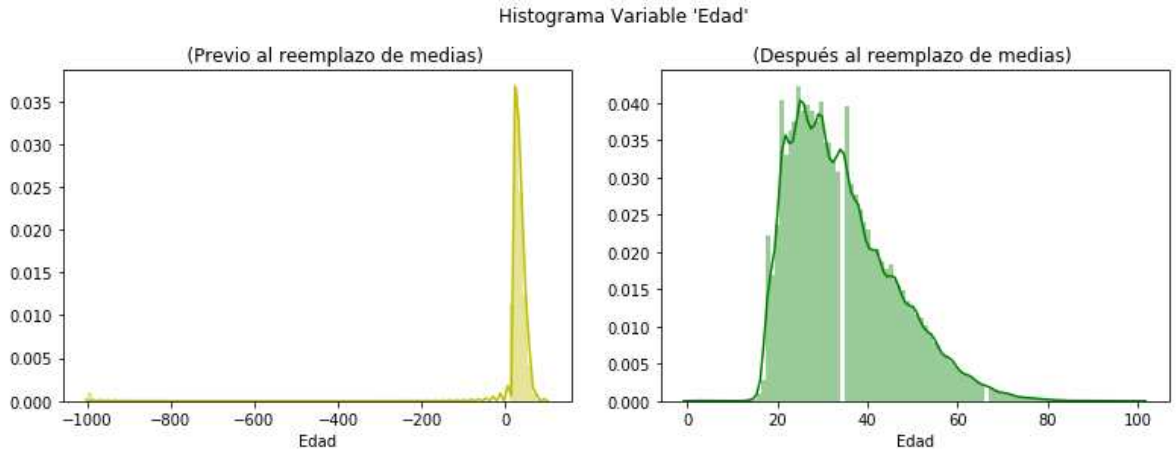


Figura 4. Histograma inicial y final de la variable Edad.

Se destaca el trabajo de la Policía Nacional de Colombia en el cumplimiento a cabalidad de la ley de Transparencia y del Derecho de acceso a la Información Pública Nacional (1712 de 2014), así como la calidad de los datos, por lo que se reduce en tiempos la fase de preprocesamiento de la información.

8.3 ANÁLISIS EXPLORATORIO

8.3.1 Descripción de variables cualitativas

8.3.1.1. Departamento:

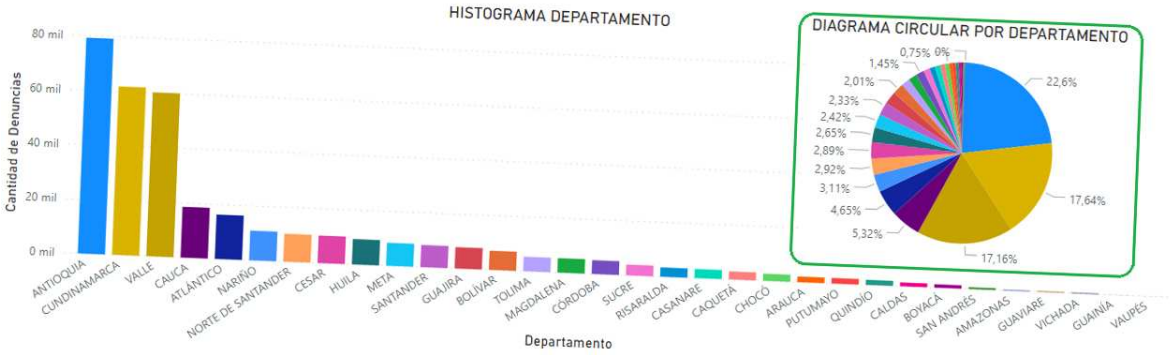


Figura 5. Análisis exploratorio variable Departamento.

En la figura 5 se evidencia que la mayor cantidad de delitos en los últimos 10 años en el país, tuvo lugar en los departamentos de Antioquia con una participación del 22.6% (79.249 hurtos), seguido del departamento de Cundinamarca con un 17.64% (61.841 hurtos) y Valle con un 17.16% (60.166 hurtos). Por otro lado, los departamentos con menor cantidad de delitos asociados al hurto de vehículos en el territorio colombiano son Guainía y Vaupés en donde se reportaron menos de 25 delitos durante los últimos 10 años.

Se destaca que Bogotá D.C como municipio capital está incluido dentro del departamento de Cundinamarca

8.3.1.2. Municipio:

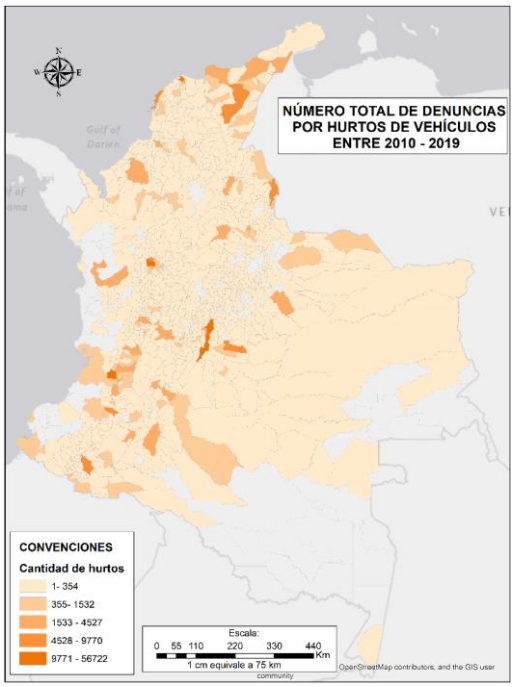


Figura 6. Análisis exploratorio variable municipio

En la figura 6 se detallan las cantidades de hurtos entre los periodos 2010 al 2019 por municipio, en donde se puede visualizar que las zonas del país menos afectadas por este tipo de delitos son los municipios ubicados al sur del país. Por otra parte, se observan concentraciones de municipios con cantidades de hurtos significantes en las zonas norte, sur y centro del país. Los cinco municipios con mayor cantidad de hurtos son Bogotá con 56.722, seguido de Medellín con 56.196, Cali con 44.952, 9.949 en Barranquilla y por último en el municipio de Popayán 8.505 delitos.

8.3.1.3. Día:

Figura 7. Análisis exploratorio variable Día

En los últimos 10 años, los días en que se ha presentado mayor cantidad de delitos asociados al hurto de vehiculos es el sábado con el 14.97%, seguido del viernes con una participación del 14.91% y los días miércoles con un 14.79%.

El día en el que menos se registraron hurtos es el domingo con una participación sobre el total de delitos cercana al 13%.

8.3.1.4. Barrio:

En la siguiente tabla se listan los 20 barrios donde se registró la mayor cantidad de hurtos en los últimos 10 años, y en la figura 8 se relaciona la incidencia porcentual de dichos barrios a nivel de ciudad, en donde destaca Medellín con el 80% de incidencia, lo que a priori permite inferir la existencia de una mayor focalización de criminalidad a comparación de las ciudades de Bogotá y Cali que cuentan con una participación del 10% cada una.

Municipio	Barrio	Cantidad de hurtos
MEDELLÍN (CT)	PRADO C-10	1737
MEDELLÍN (CT)	BOSTON C-10	1452
MEDELLÍN (CT)	ROBLEDO C-7	1378
MEDELLÍN (CT)	BUENOS AIRES C-9	1369
MEDELLÍN (CT)	BELEN C-16	1302
MEDELLÍN (CT)	LA CANDELARIA C-10	1285
MEDELLÍN (CT)	LAURELES C-11	972
MEDELLÍN (CT)	ARANJUEZ C-4	950
MEDELLÍN (CT)	CASTILLA C-5	915
MEDELLÍN (CT)	SIN INFORMACION	894
CALI (CT)	CIUDAD CORDOBA E15	876
BOGOTÁ D.C. (CT)	PATIO BONITO I E-8	810
CALI (CT)	LILI E17	772
MEDELLÍN (CT)	EL CHAGUALO C-10	768
MEDELLÍN (CT)	CAMPO VALDES 1 C-4	762
MEDELLÍN (CT)	CARIBE C-5	756
MEDELLÍN (CT)	GUAYABAL C-15	751
MEDELLÍN (CT)	MANRIQUE CENTRAL No. 1 C-4	749
MEDELLÍN (CT)	VILLA HERMOSA C-8	704
BOGOTÁ D.C. (CT)	CASTILLA LA NUEVA E-8	638

Figura 8. Análisis exploratorio variable barrio.

8.3.1.5. Zona:

En la figura 9 se ve claramente la alta incidencia de hechos delictivos asociados al hurto de vehículos automotores sobre zonas urbanas, con una cantidad de 314.900 delitos en los últimos 10 años, frente a 35.574 delitos efectuados en zonas rurales.

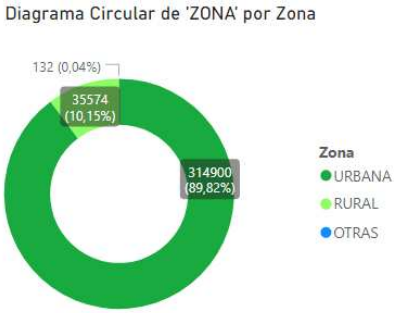


Figura 9. Análisis exploratorio variable zona

8.3.1.6. Clase de Sitio:

Teniendo en cuenta que hay 171 posibles valores para esta variable, por medio del gráfico de Pareto, se determina que el 80% de los sitios donde se cometen los hurtos es en vías públicas y frente a residencias como se detalla en la siguiente figura.

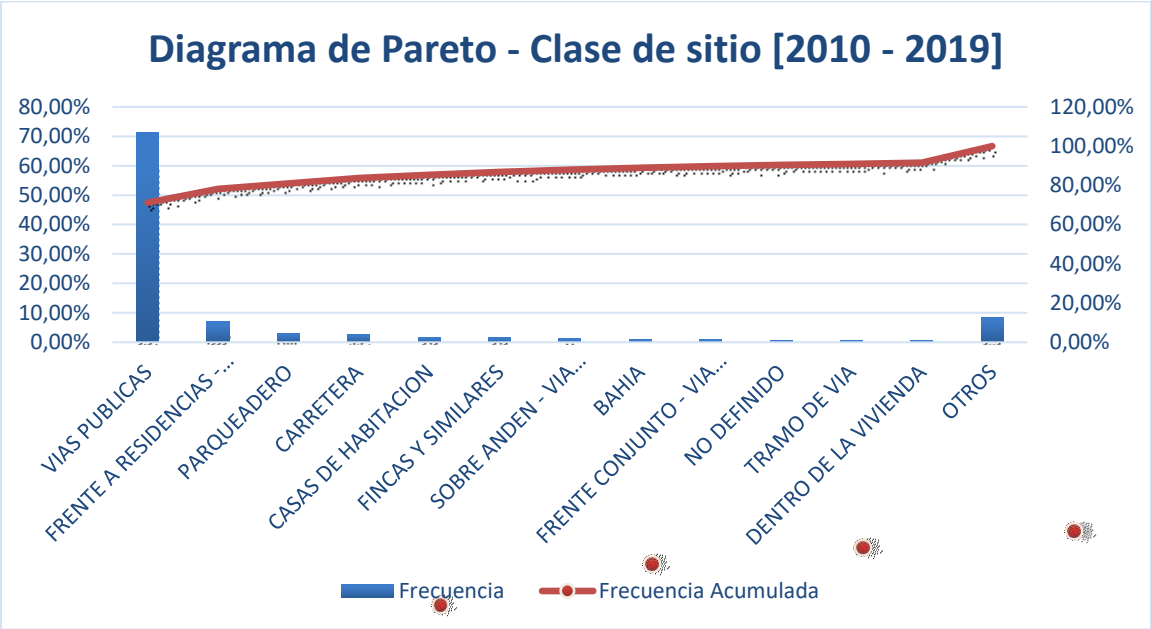


Figura 10. Diagrama de Pareto variable Clase de sitio.

Teniendo en cuenta lo anterior, en el siguiente diagrama de barras apilado se relacionan los departamentos con mayor cantidad de hurtos en las dos clases de sitios más recurrentes, en donde se destacan los departamentos de Antioquia, Atlántico, Boyacá, Cundinamarca y Valle los hurtos sobre vías públicas mientras en el segundo tipo de sitio, correspondiente a “Frente a residencias – Vía pública” en un mayor porcentaje está presente el departamento de Antioquia.

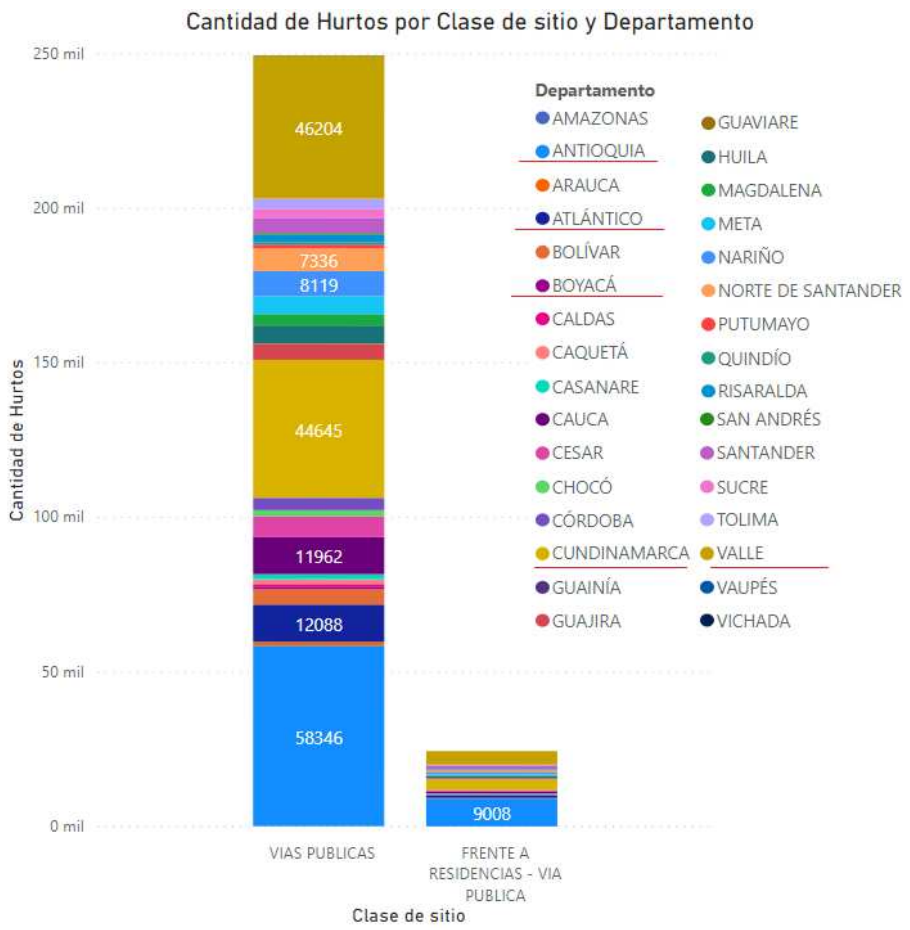


Figura 11. Análisis exploratorio variable clase de sitio y departamento.

8.3.1.7. Arma Empleada:

Por medio de la siguiente figura se relaciona el número total de denuncias realizadas entre el año 2010 – 2019 por hurto de vehículos, aproximadamente el 47% de los hechos fue realizado usando llave maestra y cerca del 31% fue empleando armas de fuego lo que permite inferir que por cada 10 robos, aproximadamente la mitad se ejecutan sin emplear violencia, cifras coincidentes con la afirmación realizada por la Policía Nacional: “el 45 por ciento de los hurtos ocurren cuando se dejan los vehículos parqueados en el espacio público, y el 30 por ciento, mediante la modalidad de atraco” (El Tiempo, 2019).

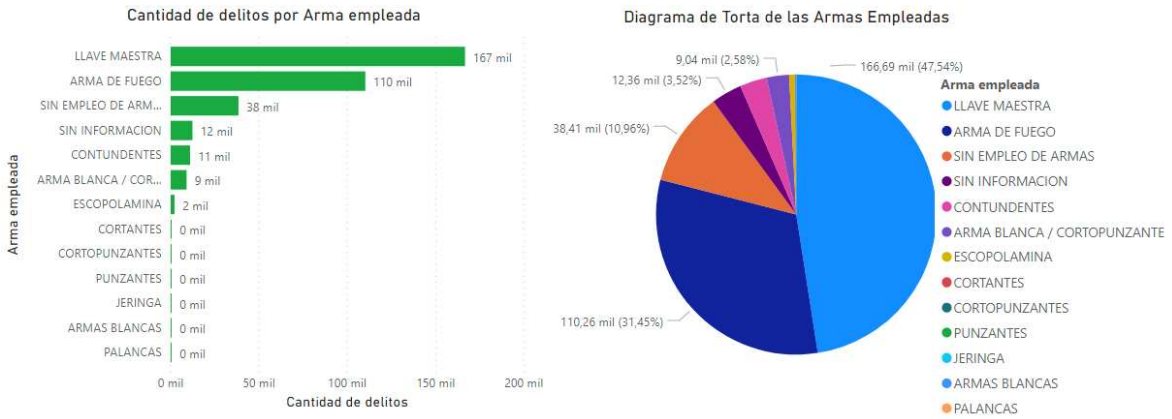


Figura 12. Análisis exploratorio variable arma empleada.

8.3.1.8. Móvil Agresor:

En el diagrama de torta de la siguiente figura 13 se concluye que los delincuentes al efectuar este tipo de hurtos prefieren desplazarse a pie, lo que es congruente con la variable "arma empleada" que relaciona en mayor parte el robo con llaves maestra, es decir, los hechos delictivos con frecuencia se realizan cuando el propietario del vehículo no se encuentra en éste.

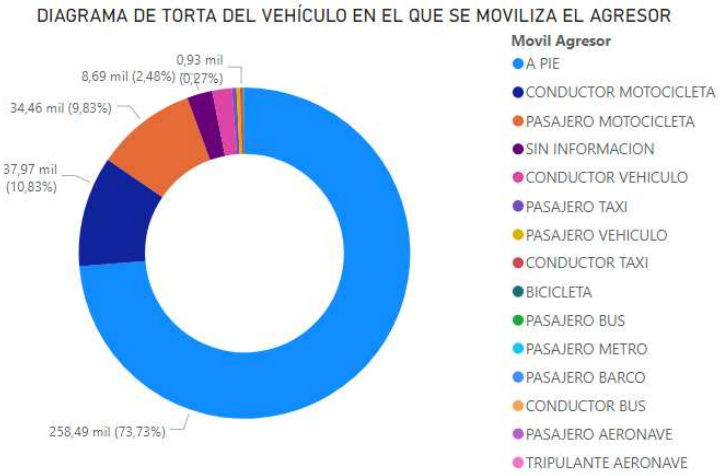


Figura 13. Análisis exploratorio variable móvil agresor.

8.3.1.9. Móvil Víctima:

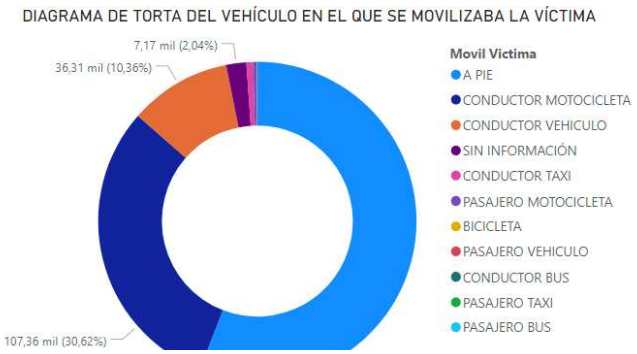


Figura 14. Análisis exploratorio variable móvil victima

En concordancia con la variable anterior los resultados de la figura 14 muestran que cerca del 55% de los casos la víctima no se encontraba al interior del vehículo hurtado, lo cual confirma que un gran porcentaje de todos los hurtos fueron realizados en

la ausencia del propietario. En adición se tiene que alrededor del 30% de los robos fueron realizados por un parrillero.

8.3.1.10. Género:

Significativamente es superior el hurto de vehículos a hombres en el territorio colombiano con una estadística de aproximadamente el 81% de víctimas de género masculino en la última década, sin embargo se debe tener en cuenta que la ocurrencia de este evento es más probable sobre hombres que en mujeres debido a la superioridad de licencias de conducción que han tramitado los hombres (Baraya, 2018).

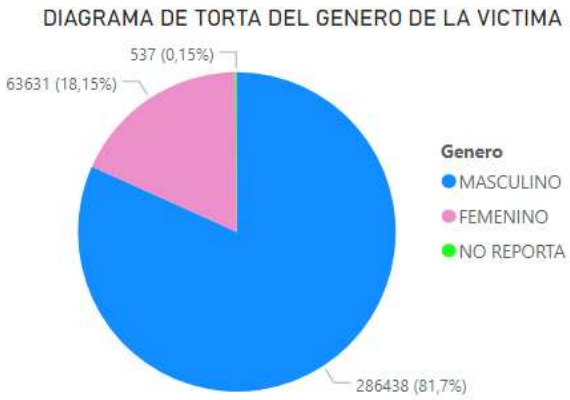


Figura 15. Análisis exploratorio variable género.

8.3.1.11. Estado civil:

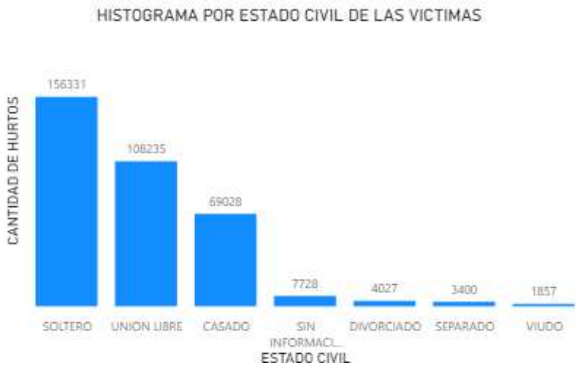


Figura 16. Análisis exploratorio variable estado civil

En el siguiente histograma se muestra que quienes son víctimas con mayor frecuencia son las personas solteras con una participación aproximada del 44%, seguido de las personas en Unión libre con un 30% y en tercer lugar con un 19% las personas casadas.

8.3.1.12. Escolaridad:

Referente al grado de escolaridad que tienen las víctimas, en el siguiente diagrama se observa que alrededor del 60% del total de las víctimas cursaron hasta la secundaria, seguido por cerca de aproximadamente a un 13% de víctimas que cursaron la educación superior.

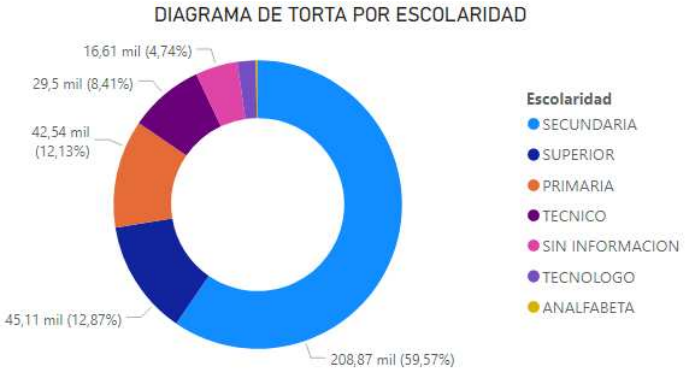


Figura 17. Análisis exploratorio variable escolaridad

8.3.1.13. Profesión:

Figura 18. Análisis exploratorio variable profesión

Durante la última década se registraron hurtos de vehículos y motocicletas a personas con diferentes tipos de profesiones, exactamente a 838 tipos de profesiones u ocupaciones, de las cuales las 3 poblaciones más afectadas fueron las personas con profesiones de Administración de empresas (2.721 víctimas), seguido de Policías (2191 víctimas) y por último profesionales del campo de Derecho (1220 víctimas).

Es importante destacar que esta variable es la que menos información aporta dentro de los datos oficiales, con 324.217 registros “Sin Información”, en otras palabras, se cuenta con apenas el 7.5% de la información de las profesiones del total de las víctimas.

8.3.1.14. Clase empleado:

Por medio del siguiente diagrama de Pareto se identifica que el 80% de las víctimas son empleados particulares con una participación del 51.8% de las denuncias, seguidos de los

independientes con una relación porcentual de denuncias de aproximadamente el 26% y en tercer lugar un 5,2% de las victimas quienes no suministraron información al respecto.

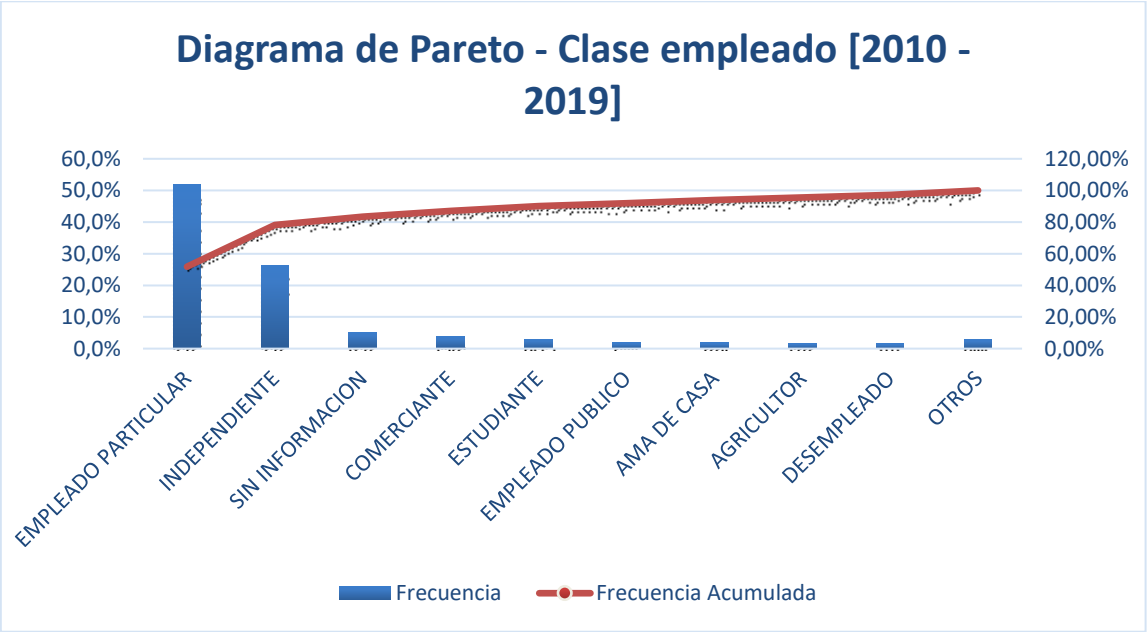


Figura 19. Diagrama *de Pareto variable Clase empleado.*

8.3.1.15. País Nacimiento:

Referente a las víctimas de la delincuencia asociada al hurto de vehículos entre el año 2010 al 2019, se encuentran 50 nacionalidades de víctimas, sobre las cuales pro medios del diagrama de Pareto se identifican que más del 80% de las víctimas son ciudadanos colombianos, exactamente con una participación de alrededor del 95%, seguido de personas que no dieron información con el 4% y en tercera posición víctimas procedentes de Venezuela con 319 denuncias.

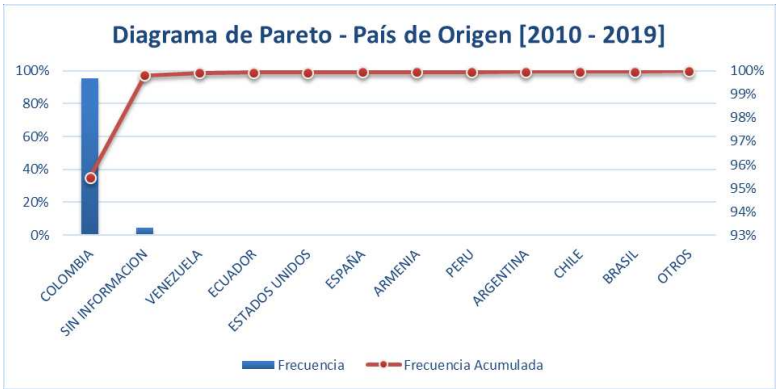


Figura 20. Diagrama de Pareto variable nacionalidad.

8.3.1.16. Clase:

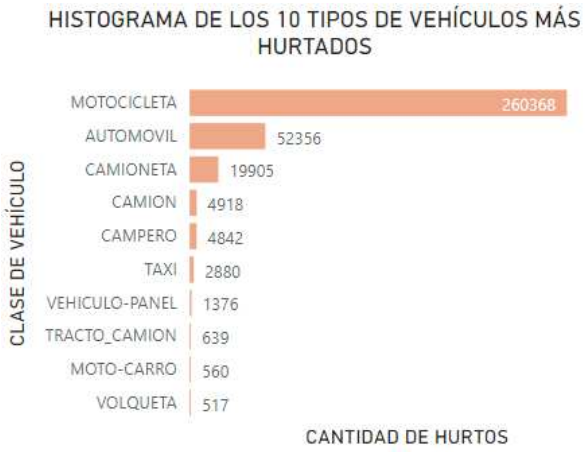


Figura 21. Análisis exploratorio variable Clase

Referente a esta variable en la cual se relaciona el tipo de vehículo hurtado, existen 36 tipologías de vehículos automotores, sin embargo los delitos que se presentaron con mayor frecuencia fueron las motocicletas con un 74% de participación, seguido de los automóviles con un 15% y en tercer lugar las camionetas con aproximadamente un 6% de participación, como se expone en la figura 21.

8.3.1.16. Marca:

Debido a que en la última década han sido hurtados 308 marcas de vehículos, por medio del diagrama de Pareto (*ver figura 22*) se identifican las marcas más hurtadas y se seleccionan las que comprenden hasta el 80% del total, enfocándose el análisis en las 5 primeras marcas en el top: Chevrolet (28.717), Mazda (13.362), Renault (9.611), Hyundai (8.029), Toyota (7.132).

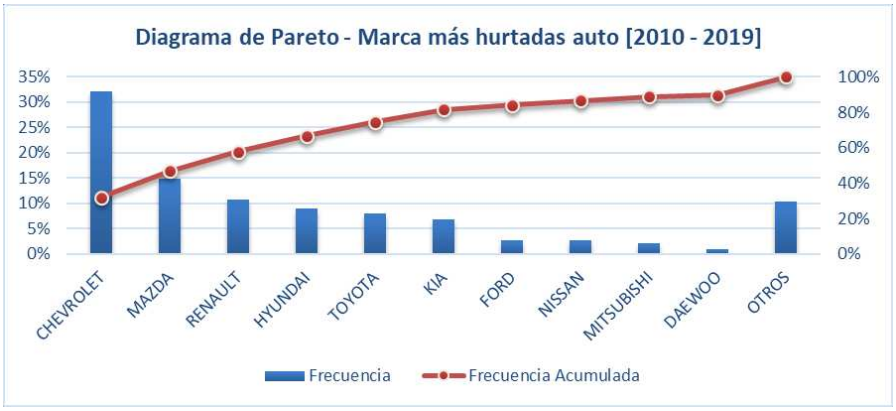


Figura 22. Diagrama de Pareto variable marca de vehículos.

Con respecto a las motocicletas, por medio del siguiente diagrama de Pareto, se evidencia que al 80% de marcas hurtadas en la última década se encuentran 5 marcas: Bajaj (62.888), Yamaha (61.026), Honda (39.449), Suzuki (31.211), AKT (23.804).

Figura 23. Diagrama de Pareto variable marca de motocicletas.

Teniendo en cuenta los diagramas anteriores de Pareto en donde se relaciona un top de 5 marcas de vehículos más hurtadas y 7 para la categoría de motocicletas, en la figura 24 se detallan los histogramas con el top de las 5 marcas más hurtadas para cada tipo de hurto.

Figura 24. Top marcas de vehículos y motocicletas hurtadas

Con respecto al hurto de vehículos la secretaria distrital de Seguridad: “Los carros de gamas baja y media –que representan el 80 por ciento de los robos– son usados para el desguace y la venta de repuestos de segunda, mientras que los modelos de gama alta (20

por ciento restante) son sometidos a un proceso de cambio de documentación para venderlos principalmente en la costa, zonas de frontera y en el eje cafetero." (Baraya, 2019).

En adición, en el marco del hurto de motocicletas las marcas más hurtadas son Bajaj, Yamaha y Honda, igualmente se evidencia que la cantidad de motocicletas alcanza 2 veces la cantidad de hurtos cometidos a vehículos en el mismo periodo de tiempo.

8.3.1.16. Línea:

En consecuencia, con el EDA de la variable anterior, en la figura 25 se exponen los histogramas asociados a los hurtos reportados entre el 2010 al 2019 en todo el territorio nacional diferenciados por motocicleta o vehículo y se destacan que la línea Spark, standard y Sprint son las 3 líneas con mayores estadísticas de hurto de vehículos mientras que en motocicleta destacan las líneas Boxer Ct 100, Standard y sin línea.

Figura 25. Top líneas de vehículos y motocicletas hurtadas

8.3.1.17. Color:

El color con mayores estadísticas de hurto difiere del tipo de vehículo, por un lado, en los autos la mayor cantidad de delitos se han realizado a carrocerías blancas mientras que el robo a motocicletas negras es el color más hurtado.

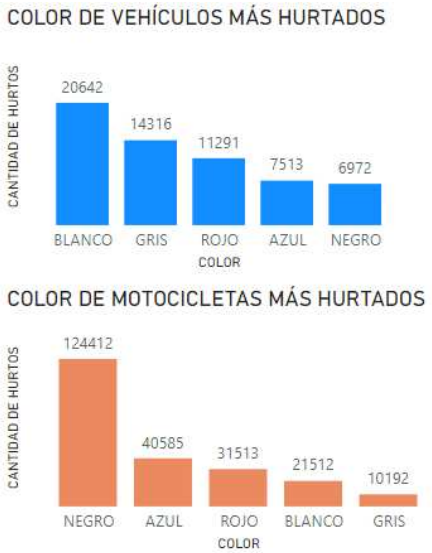


Figura 26. Top colores de vehículos y motocicletas hurtadas

8.3.1.18. Tipo de hurto:

Figura 27. Diagrama de torta por tipo de delito

Por medio de esta variable se evidencia que la mayor cantidad de hurtos a vehículos automotores está focalizada en las motocicletas, con un total de denuncias realizadas entre el año 2010 al 2019 de 260.925, frente a 89.681 denuncias por hurtos de vehículos en el mismo periodo de tiempo.

8.3.2 Descripción de variables cuantitativas

8.3.2.1. Año:

Esta variable representa el año en que se cometió el hecho delictivo. Se evidencia en la figura 28 que en la última década el año donde se ha presentado mayor incremento de denuncias frente al año inmediatamente anterior es el 2011 (43%), seguido del año 2016 (13%). Adicionalmente se destaca la tendencia positiva de aumento de delitos en la última

década, sin embargo es notable que los índices de crecimiento en los últimos 3 años han sido inferiores a los presentados para las vigencias 2015 y 2016.

Figura 28. Diagrama de variaciones y frecuencias relativas de hurtos a vehículos y motocicletas entre 2010 – 2019.

Consecuentemente, en la figura 29 se visualiza el comportamiento de la delincuencia a nivel trimestral durante la última década y se establecen 5 tipos de comportamiento:

SECCIÓN I: Para el año 2010 el trimestre con mayores estadísticas de delincuencias fue el primero, mientras que en los trimestres siguientes disminuyó la cantidad de hurtos y se mantuvo constante.

SECCIÓN II: El comportamiento a nivel trimestral para el periodo de tiempo comprendido entre los años 2011 al 2013 fue similar en las dinámicas de disminución de hurto en el último trimestre de cada año, sin embargo se destaca que la mayor variación (disminución) registrada en la última década se dio en el último trimestre del año 2012.

SECCIÓN III: En contraste con el comportamiento definido en la sección II, para el año 2014 en los dos últimos trimestres de este año los dos últimos trimestres comprenden las más altas tasas de delincuencia a comparación de lo registrado en el primer semestre de este año.

SECCIÓN IV: En esta sección se comprueba la tendencia existente desde 2014 para los años posteriores hasta el 2017, en donde la cantidad de hurtos es mayor en el último semestre del año a comparación de los dos primeros semestres de cada año respectivo.

SECCIÓN V: En el comportamiento de la delincuencia asociada al hurto de vehículos en esta sección, refleja el mismo comportamiento de la sección II, comprendida entre 2011 a 2013 en donde en el último trimestre del año se presenta una disminución importante en la cantidad de hurtos denunciados a comparación de los trimestres anteriores.

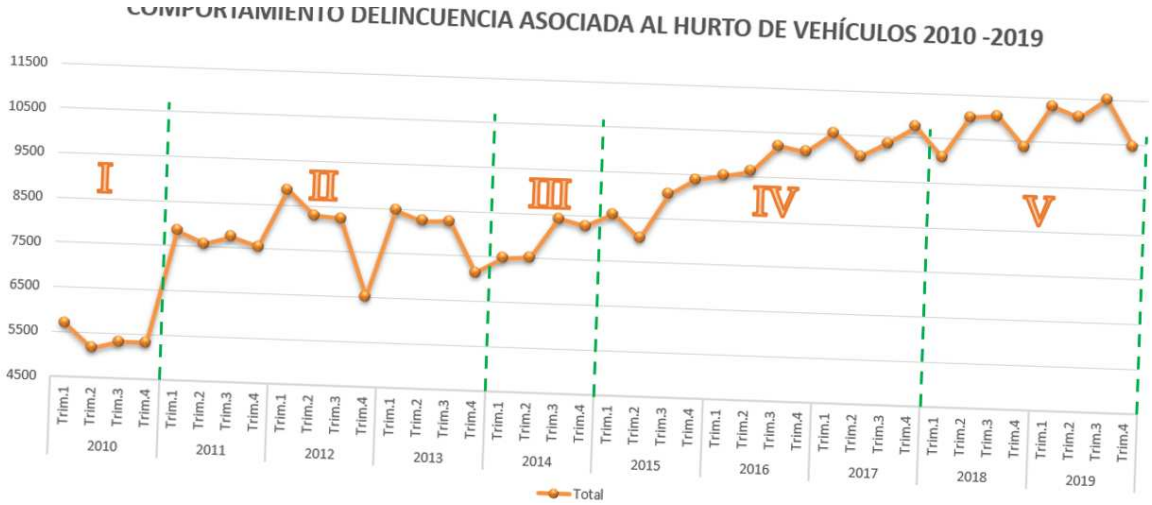


Figura 29. Comportamiento delincuencia asociada al hurto de vehículos entre 2010 – 2019 en Colombia

8.3.2.2. Mes:

El mes donde se presenta la mayor cantidad de hurtos en el país en la última década es el mes de agosto con exactamente 30.605 denuncias, seguido del mes de enero con 30.441 y en tercer lugar el mes de julio con 29.900 denuncias.

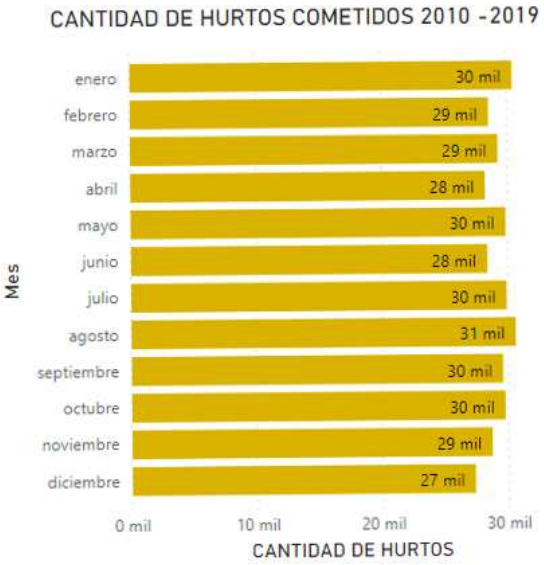


Figura 30. Número de hurtos categorizados por mes de ocurrencia

8.4 ALGORITMO K-MEANS

Debido a que este algoritmo recibe variables cuantitativas, se toman las 343.608 denuncias realizadas entre el 2010 y el 2019 en todo el territorio colombiano y se resumen por municipio, edad media de las víctimas, cantidad de motocicletas y vehículos hurtados, categoría de género con mayor tasa de delincuencia. Lo anterior con el fin de crear un archivo geográfico con extensión SHP, para así espacializar la data de hurtos.

Teniendo en cuenta este algoritmo permite la identificación de clústeres, se añade data adicional por municipio, tales como población e índices de necesidades básicas insatisfechas con fecha del 2018, con el objetivo de encontrar patrones por zonas.

Índice de Necesidades básicas Insatisfechas (NBI - 2018):

La metodología de NBI busca determinar, con ayuda de algunos indicadores simples, si las necesidades básicas de la población se encuentran cubiertas. Los grupos que no alcancen un umbral mínimo fijado, son clasificados como pobres. Los indicadores simples seleccionados, son: Viviendas inadecuadas, Viviendas con hacinamiento crítico, Viviendas con servicios inadecuados, Viviendas con alta dependencia económica, Viviendas con niños en edad escolar que no asisten a la escuela (DANE, 2018).

A continuación, se detallan las definiciones de los indicadores:

Viviendas inadecuadas: Este indicador expresa las características físicas de viviendas consideradas impropias para el alojamiento humano. Se clasifican en esta situación separadamente las viviendas de las cabeceras municipales y las del resto (DANE, 2018) Viviendas con hacinamiento crítico: Con este indicador se busca captar los niveles críticos de ocupación de los recursos de la vivienda por el grupo que la habita. Se consideran en esta situación las viviendas con más de tres personas por cuarto (excluyendo cocina, baño y garaje).

Viviendas con servicios inadecuados: Este indicador expresa en forma más directa el no acceso a condiciones vitales y sanitarias mínimas. Se distingue, igualmente, la condición de las cabeceras y las del resto. En cabeceras, comprende las viviendas sin sanitario o que careciendo de acueducto se provean de agua en río, nacimiento, carrotanque o de la lluvia. En el resto, dadas las condiciones del medio rural, se incluyen las viviendas que carezcan de sanitario y acueducto y que se aprovisionen de agua en río, nacimiento o de la lluvia. (DANE, 2018).

Viviendas con alta dependencia económica: Es un indicador indirecto sobre los niveles de ingreso. Se clasifican aquí, las viviendas en los cuales haya más de tres personas por miembro ocupado y el jefe tenga, como máximo, dos años de educación primaria aprobados. (DANE, 2018).

Viviendas con niños en edad escolar que no asisten a la escuela: Mide la satisfacción de necesidades educativas mínimas para la población infantil. Considera las viviendas con, por lo menos, un niño mayor de 6 años y menor de 12, pariente del jefe y que no asista a un centro de educación formal. (DANE, 2018).

Dado que cada uno de los indicadores se refiere a necesidades básicas de diferente tipo, a partir de ellos se constituye uno compuesto, que clasifica como pobre o con NBI aquellos hogares que estén, al menos, en una de las situaciones de carencia expresada por los indicadores simples y en situación de miseria los hogares que tengan dos o más de los indicadores simples de necesidades básicas insatisfechas. Para estimar la magnitud de la

pobreza en relación con la población, se consideró que las personas que habitaban en viviendas con NBI o en miseria se encontraban en las mismas condiciones de su respectiva vivienda. (DANE, 2018)

Se destaca que para este análisis se extraen los municipios “Área en litigo Cauca – Huila”, “Nariño” y “Albania” dado que no tienen estadísticas del DANE y ninguna denuncia; lo anterior es necesario dado que el algoritmo no funciona con valores nulos, por lo cual el análisis se realizará para 1120 municipios.

Como se mencionó previamente en el marco teórico, el algoritmo k-means es no supervisado y tiene por objetivo encontrar grupos con características similares, dichos grupos se caracterizan a su vez por tener elementos similares entre grupos pero diferentes a los otros grupos; en términos estadísticos se busca maximizar la variación inter-clúser y minimizar la intra-clúster.

Se decide usar este algoritmo debido a su agilidad de implementación y su almacenamiento económico, computacionalmente hablando, dado que solo necesita guardar los k centroides; por otra parte tiene una desventaja y es que se requiere probar para n clústers, sin embargo en el software R se puede modelar para n casos y elegir el número óptimo.

Por medio del software R y las librerías dummies, dplyr y PerformanceAnalytics se carga la información a nivel municipio y las variables que no numéricas se convierten en dicotómicas, como el tipo y genero de arma usada.

Para determinar el número de clúster óptimo, se realizan 10 iteraciones variando el número de clúster y comparando el estadístico de suma de cuadrados intergrupos se define qué el número de clúster optimo es 6, tal como se muestra en la siguiente figura.

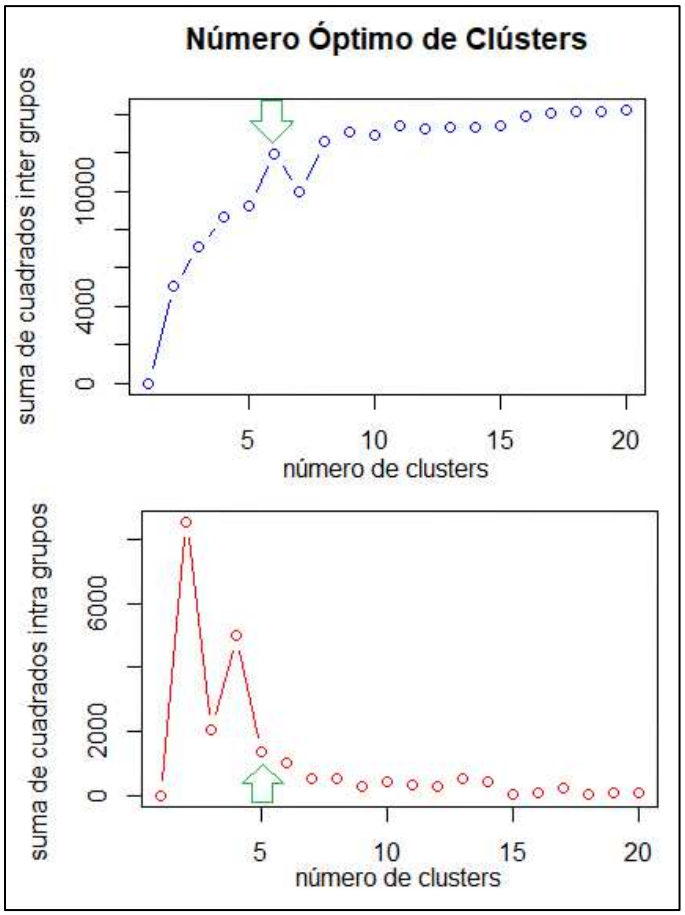


Figura 31. Cálculo de número de clústeres óptimos para el algoritmo K-means.

Buscando el equilibrio entre las iteraciones de la suma de cuadrados Inter grupos e intra grupos se selecciona 6 como el número óptimo de clúster para el conjunto de datos.

Tabla 2. Estadísticas de iteración con 6 clústeres.

N. Clústeres			6
Inercia	Total	totss	16785
	intragrupos	withinss	4848,78
	intergrupos	betweens	11936,21

A partir del resultado anterior, se emplea el algoritmo K-means con el parámetro de clústeres correspondiente a 6. En la figura 32 se muestran espacialmente distribuidos los municipios dentro de los 3 primeros clústeres.

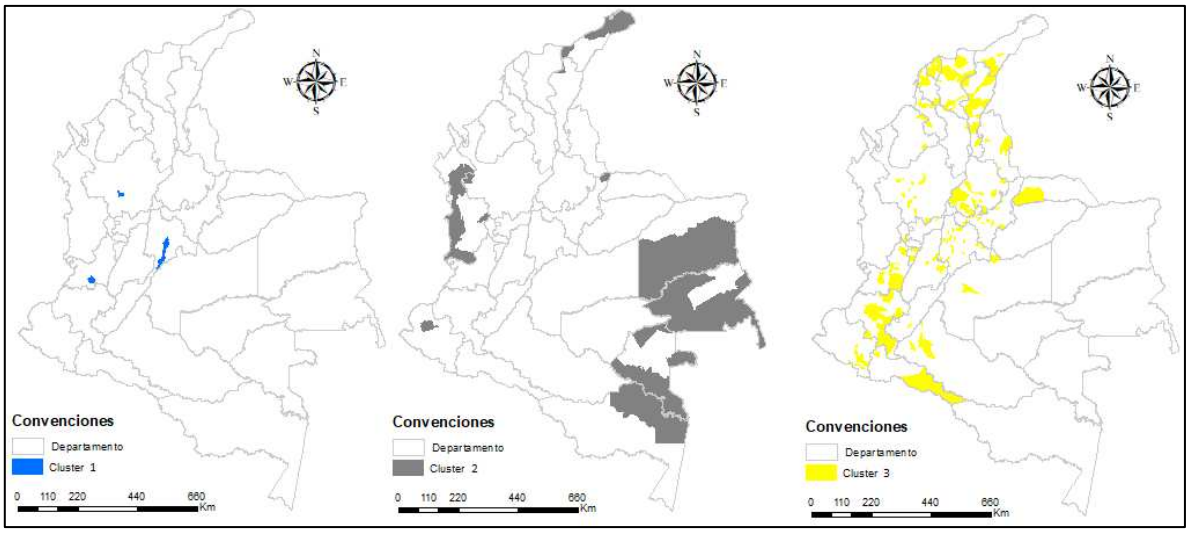


Figura 32. Espacialización de municipios para los clústeres 1, 2 y 3.

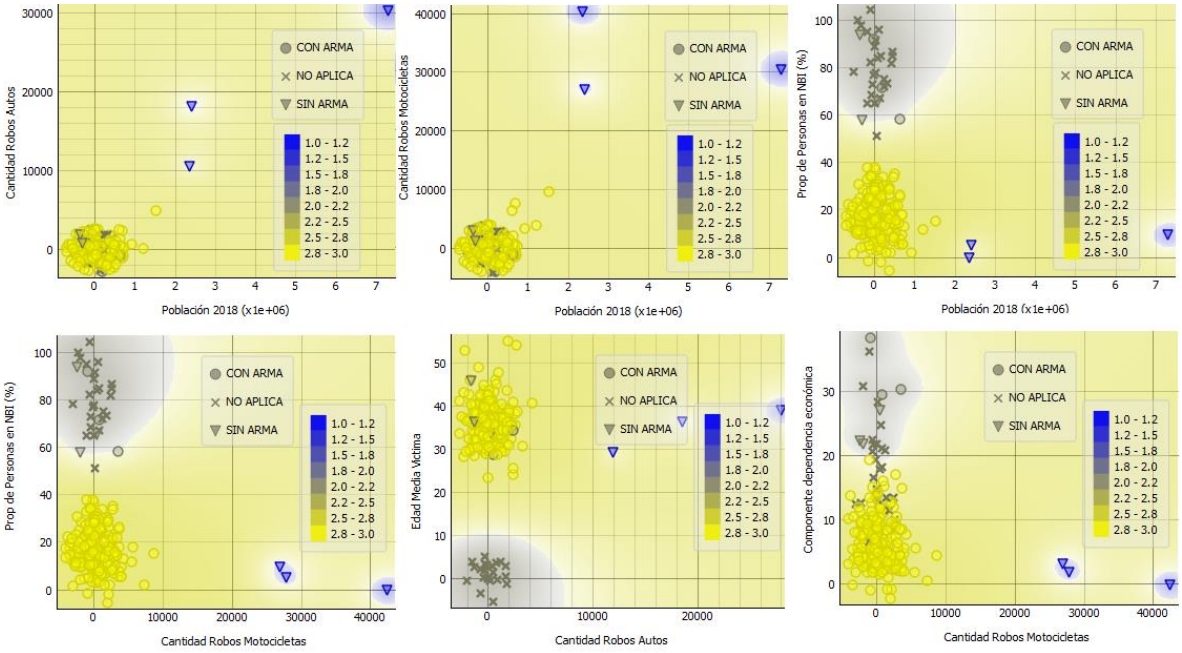


Figura 33. Diagramas de dispersión para clústeres 1, 2 y 3

En el clúster 1 están inscritos los municipios Cali, Medellín y Bogotá, a priori los municipios donde se presentó la mayor cantidad de denuncias en el marco de la delincuencia asociada al hurto de vehículos o motocicletas; en el segundo clúster existen 31 municipios distribuidos espacialmente en los extremos occidente, norte, oriente y sur del país; y en el tercer clúster fueron ajustados 197 municipios, distribuidos espacialmente en la zona centro del país y la costa norte.

En la figura 33 se relacionan 6 diagramas de dispersión en el que se muestran discretizados los 3 primeros clústeres en color azul, gris y amarillo respectivamente, además, en función del tipo de arma más usada cambia la representación del elemento en el diagrama.

En las gráficas superior izquierda y superior central (figura 33), se contrasta la población vs la cantidad de robos y motocicletas respectivamente y se evidencia que las ciudades de Medellín y Cali a pesar de tener una densidad poblacional similar presentan una diferencia porcentual de cerca del 38% de hurtos superior de motocicletas en Medellín, con 43.917 casos reportados frente a los 27.061 reportados en el mismo periodo de tiempo. Para el caso particular de Bogotá, ciudad con la mayor población del país con 7.4 millones de personas aproximadamente para el año 2018, presenta 28.264 denuncias por hurto de motocicletas en la última década, un 35% menos de denuncias asociadas al hurto de motocicletas que Medellín. Por otra parte, el hurto de vehículos tiene un comportamiento inverso entre Medellín y Cali con un total de denuncias de 12.279 y 17.891 respectivamente; para el caso de Bogotá, es la ciudad donde se han presentado la mayor cantidad de hurtos asociadas a vehículos con un total de 28.458 denuncias, cifra cercana al total de denuncias por hurto de motocicletas en la misma ciudad. Se destaca que el delito en la mayoría de las veces para los municipios destacados en el clúster 1 se realizó sin empleo de armas. En adición los municipios clasificados dentro del clúster dos y tres están concentrados en la región que representa poblaciones no superiores a 1 millón de habitantes y cerca de 1.000 denuncias por hurto.

Contrastando en la gráfica superior derecha los municipios clasificados en los clústeres 1-3 para las variables población vs proporción de personas en NBI (%), se observa que los 3 municipios con mayor cantidad de hurtos muestran un porcentaje de Necesidades básicas insatisfechas bajo, inferior al 10%, por otra parte, es notoria la concentración del segundo

clúster, en el que el porcentaje de necesidades insatisfechas es superior al 50%. Los resultados que muestra el diagrama de dispersión son acordes con la realidad colombiana, en donde por diversos temas, ya sea de acceso, existencia de grupos armados, distribución inequitativa de recursos, los municipios ubicados en los extremos del país son las regiones donde sus habitantes tienen las menos favorables estadísticas asociadas a la calidad de vida. En la zona centro del país, referente al clúster 3 se evidencia que el porcentaje de necesidades básica insatisfechas es claramente menor al del clúster 2, sin embargo la cantidad de hurtos es similar, por lo que se concluye que la cantidad de hurtos para los municipios de los clúster 2 y 3 no está asociada con el estadístico de porcentaje de necesidades básicas insatisfechas.

Con respecto a la cantidad de motocicletas hurtadas entre los municipios categorizados en los clústeres 2 y 3 no se observan diferencias altas, el intervalo de número de motocicletas hurtadas está entre 0 a 6.773 en 10 años, dicha relación se muestra en la gráfica inferior izquierda. Se destaca que los municipios ubicados en las zonas extremas del país, correspondientes a los clasificados en el clúster 2 la mayoría de las victimas durante el proceso de la denuncia prefirieron no aportar los datos de su edad y tipo de arma; situación opuesta en los municipios del clúster 3 en el cual la mayoría de los hurtos efectuados por municipio se realizó utilizando armas y el rango de edades de las victimas oscila entre 25 y 50 años aproximadamente. En relación a las ciudades con mayores tasas de delincuencia asociadas al hurto de vehículos, se destaca que las victimas más jóvenes están localizadas en Medellín con una media de 32 años.

En última instancia se observa que el componente de dependencia económica impacta en la misma medida el hurto de motocicletas entre los municipios pertenecientes al clúster 2 y 3, sin embargo se destaca la alta variabilidad de dicho componente en los municipios del clúster 2, en donde por ejemplo el municipio de Uribia en la Guajira refleja un 36% de dependencia económica mientras que Bagadó en Choco muestra un 8%.

De las anteriores gráficas se destaca una alta correlación positiva entre el número de habitantes y el número de hurtos a motocicletas para los municipios de Bello, Cartagena de Indias, Soledad, Cúcuta y Barranquilla por lo que se sugiere enfrentar esta situación aumentando el número de unidades policiales.

De otro modo, a continuación se muestran los resultados asociados a los 3 clústeres restantes

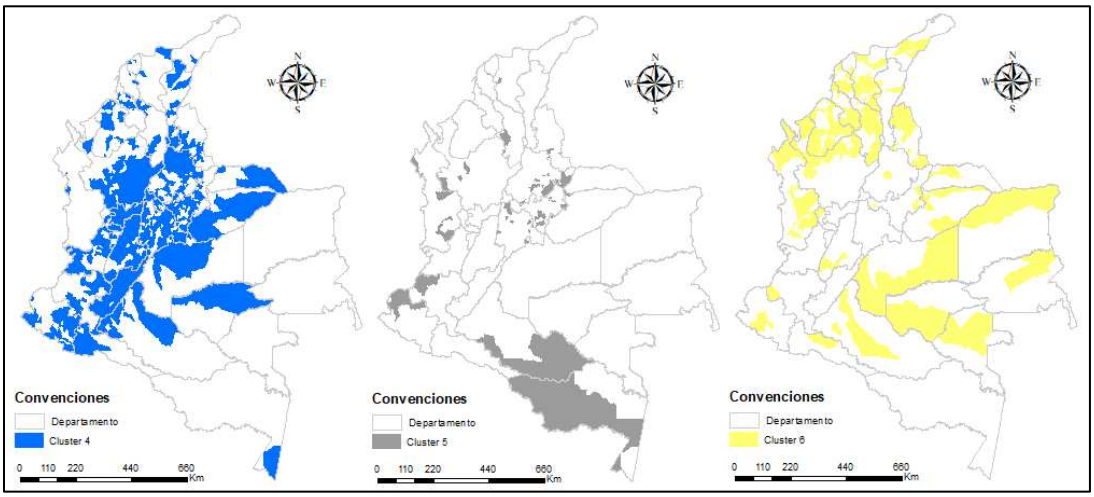


Figura 34. Espacialización de municipios para los clústeres 4, 5 y 6.

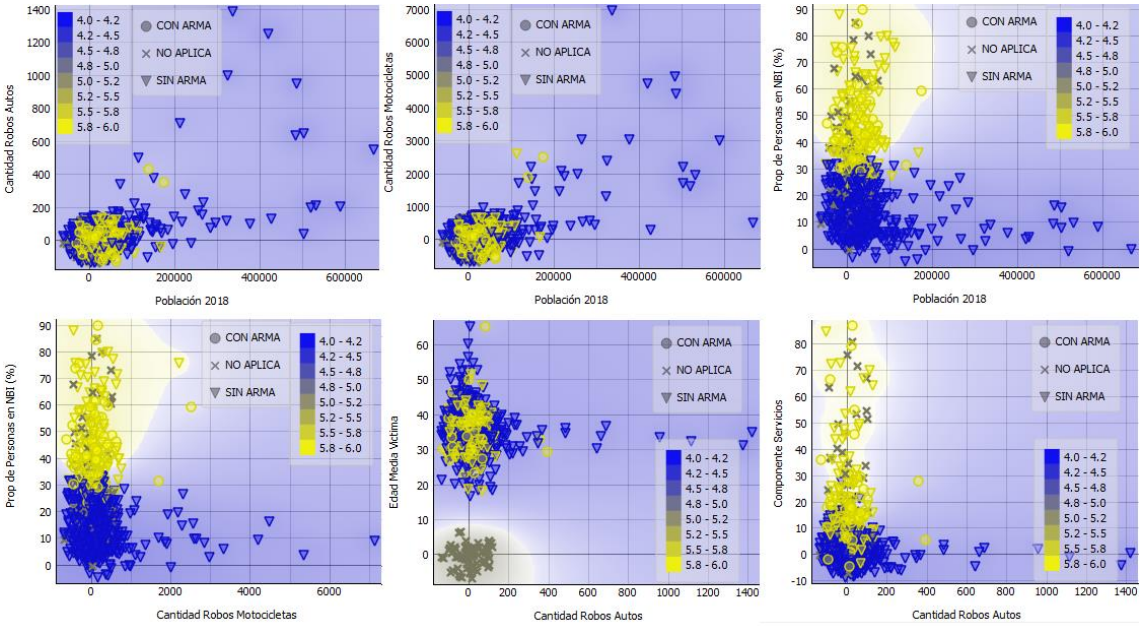


Figura 35. Diagramas de dispersión para clústeres 4, 5 y 6

En consecuencia con el análisis anterior, se describe la clasificación de los municipios en el clúster 4, los cuales en su mayoría están ubicados en la zona centro del país y cerca del

40% tienen poblaciones considerables. En los clústeres 5 y 6 se clasificaron municipios ubicados en zonas cercanas a las costas y en menor medida en zonas centrales al país, cabe destacar que el clúster 6 destaca por tener municipios con bajas tasas de delincuencia y gran parte de denuncias sin la información del tipo de arma con que se cometió el hurto ni la edad de la víctima.

De la gráfica superior izquierda y central (figura 35) se destaca la alta variabilidad de cifras de hurtos de automóviles y motocicletas para los municipios del clúster 4, y se identifica que a mayor población no necesariamente aumenta el número de hurtos, pero si la variabilidad de éstos. En cuanto a los municipios de los clústeres 5 y 6 se encuentran focalizados en la zona que refleja cantidades poblaciones pequeñas y número de hurtos bajos.

Los municipios del clúster 4 se diferencian de los otros dos clústeres por el porcentaje de necesidades básicas insatisfechas, el cual tiene como máximo valor el 35%, mientras que los clústeres restantes se focalizan en porcentajes superiores al 30%, alcanzando un máximo de aproximadamente el 90% en algunos casos, como el municipio de San Jacinto en Bolívar.

De la gráfica inferior izquierda se concluye que el porcentaje de personas con niveles básicos insatisfechos no están relacionados proporcionalmente con los hurtos, sin embargo, se destaca la gran brecha entre los municipios del clúster 4 frente a los dos clústeres restantes.

Como se mencionó anteriormente los municipios del clúster 5 y 6 se diferencian porque los habitantes de este último, no suministraron información acerca de sus edades ni el arma con la cual fueron víctimas del hurto, tal como se puede evidenciar en la gráfica central inferior. Se destaca que en los municipios centrales (clúster 4) los hurtos se cometieron en mayor medida con armas, mientras que en los municipios del clúster 5, es decir los municipios ubicados cerca a los extremos del país y con número de habitantes reducido los delitos se efectuaron empleando algún tipo de arma.

Se destaca el comportamiento atípico de los municipios de Quibdó y Maicao, quienes están clasificados en el clúster 6 y se diferencian de los demás por el hecho de que presentan un porcentaje alto en las necesidades básicas insatisfechas y tienen una relación positiva con la cantidad de hurtos a motocicletas, por lo cual se sugiere solventar dichas problemáticas.

8.5. REGLAS DE ASOCIACIÓN (ALGORITMO APRIORI)

Adicionalmente, por medio de la identificación de asociaciones frecuentes se pretende mostrar la probabilidad de las correlaciones entre las denuncias analizadas, a través de las reglas de asociación establecidas. Éstas están conformadas por un antecedente -si…-, y un consecuente -entonces…-, y a su vez, utilizan los siguientes criterios:

- Soporte: Indicación de la frecuencia con la que los elementos aparecen en las denuncias.
- Confianza: Indica el número de veces que las afirmaciones se consideran verdaderas.

En ese orden de ideas, en esta investigación se emplea el algoritmo Apriori, debido a que no requiere datos etiquetados, su sencilla implementación, y que para el número de denuncias a analizar no tiende a volverse lento o a encontrar muchas asociaciones falsas por falta de data.

La implementación del algoritmo se realiza sobre el software estadístico R, empleando la librería arules. El primer paso es la importación de las denuncias en formato de transacciones, el cual contiene todos los datos de cada denuncia separados por “coma”. Con esta información se procede a generar un diagrama de frecuencias en la figura 36 para identificar gráficamente los 10 eventos, elementos, o datos más recurrentes.

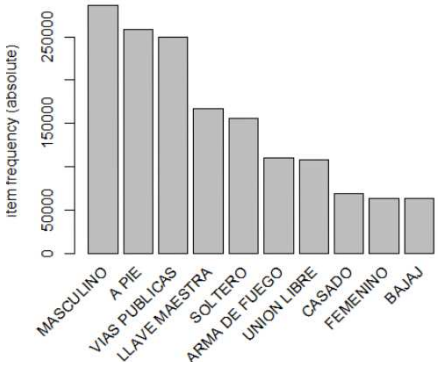


Figura 36. Top 10 datos más recurrentes en transacciones

Seguido de lo anterior, se implementa por medio de la función apriori la generación de las reglas a partir de las transacciones, las cuales posteriormente son depuradas, descartando reglas duplicadas o redundantes y se genera el siguiente grafo con las reglas.

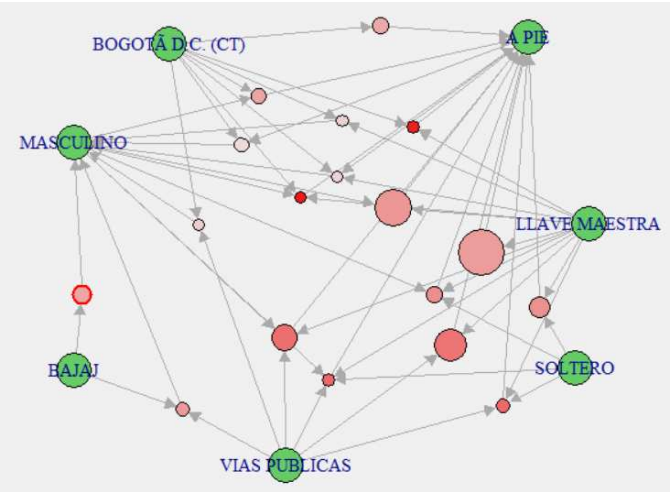


Figura 37. Diagrama de grafos generados a partir de las reglas

Del grafo debemos tener presente que los nodos (círculos verdes) tenemos los datos que forman parte de las reglas, los cuales están conectados por flechas que van conectados del antecedente al consecuente, y en medio de estas fechas se encuentran círculos cuyo tamaño representa el valor del soporte, mientras que el color representa el valor de la confianza. Por otra parte, en la siguiente figura se relacionan las reglas con las cuales se construyó el grafo.

lhs		rhs	support	confidence	coverage	lift	count	leverage	convicton
{BOGOTÁ,LLAVE MAESTRA, MASCULINO}	=>	{A PIE}	0.1030330	0.9634865	0.1069377	1.306842	36124	0.024191772	7.195603
{BOGOTÁ, LLAVE MAESTRA}	=>	{A PIE}	0.1154145	0.9622152	0.1199466	1.305117	40465	0.026982201	6.953506
{LLAVE MAESTRA, MASCULINO, SOLTERO, VIAS PUBLICAS}	=>	{A PIE}	0.1086348	0.9401891	0.1155457	1.275242	38088	0.023447172	4.392785
{LLAVE MAESTRA, SOLTERO, VIAS PUBLICAS}	=>	{A PIE}	0.1342419	0.9400603	0.1428013	1.275067	47066	0.028959656	4.383349
{LLAVE MAESTRA, MASCULINO, VIAS PUBLICAS}	=>	{A PIE}	0.2475343	0.9377715	0.2639601	1.271962	86787	0.052926116	4.222124
{LLAVE MAESTRA, VIAS PUBLICAS}	=>	{A PIE}	0.3013753	0.9363557	0.3218599	1.270042	105664	0.064079808	4.128205
{LLAVE MAESTRA, MASCULINO, SOLTERO}	=>	{A PIE}	0.1547178	0.9285506	0.1666229	1.259456	54245	0.031872828	3.677241
{LLAVE MAESTRA, SOLTERO}	=>	{A PIE}	0.1918592	0.9276799	0.2068162	1.258275	67267	0.039381208	3.632969
{LLAVE MAESTRA, MASCULINO}	=>	{A PIE}	0.3581884	0.9253299	0.3870926	1.255087	125583	0.072799131	3.518632
{BAJAJ, VIAS PUBLICAS}	=>	{MASCULINO}	0.1213413	0.9239239	0.1313326	1.130902	42543	0.014045255	2.405752
{LLAVE MAESTRA}	=>	{A PIE}	0.4388259	0.9229952	0.4754368	1.251920	153855	0.088303717	3.411952
{BAJAJ}	=>	{MASCULINO}	0.1647376	0.9179301	0.1794664	1.123565	57758	0.018117204	2.230054
{BOGOTÁ, MASCULINO}	=>	{A PIE}	0.1316435	0.9176309	0.1434602	1.244644	46155	0.025875545	3.189747
{BOGOTÁ}	=>	{A PIE}	0.1482690	0.9164698	0.1617827	1.243070	51984	0.028992490	3.145408
{BOGOTÁ, VIAS PUBLICAS}	=>	{MASCULINO}	0.1048185	0.8963196	0.1169432	1.097114	36750	0.009278260	1.765235
{A PIE, BOGOTÁ, LLAVE MAESTRA}	=>	{MASCULINO}	0.1030330	0.8927221	0.1154145	1.092710	36124	0.008741763	1.706039

Figura 38. Reglas construidas por medio del algoritmo Apriori

En la figura 38, en la regla #1, se establece que, si los hurtos son cometidos en Bogotá empleando llaves maestras a víctimas de género masculino, entonces el victimario se desplaza a pie; esto con un soporte del 10% del total de denuncias, y una confianza de aproximadamente el 96%. Asimismo, por medio de la regla #6 si en el hurto se emplea llave maestra, en vías públicas, entonces el victimario se desplaza a pie; regla que se soporta con un 30% y una confianza del 93%.

La regla con mayor soporte (43%) identificada es la #11, y con una confianza del 92% indica que, si los hurtos son ejecutados en cualquier parte del país, empleando llaves maestras, el victimario se desplaza a pie. Sin embargo se debe tener presente el contexto de la denuncia, en el cual para aquellos casos en que el denunciante no está presente al momento del hurto, éste indica que el victimario se desplaza a pie.

Se destaca de la regla #12 con una confianza del 91% y un soporte del 16%, que si los hurtos son realizados a motocicletas de marca Bajaj entonces el género de la víctima es masculino. Por otra parte las demás reglas reflejan un comportamiento ya esperado debido a los resultados del análisis exploratorio de la sección anterior.

8.6. ÁRBOL DE DECISIÓN

Otro tipo de algoritmo que se aplica sobre el dataset, son los árboles de decisión, algoritmos que aparte de ser utilizados para la predicción facilitan la interpretación del comportamiento de los datos teniendo en cuenta los resultados del mismo. Cabe destacar que el árbol de decisión solo recibe variables de tipo cualitativas, por lo cual fue necesario categorizar algunas variables numéricas que después de un proceso de cerca de 20 iteraciones añadían información y congruencia al modelo. Las variables insumo para el árbol de decisión son:

Tabla 3. Categorización de variables

NOMBRE VARIABLE	RANGO	VALOR
Hora_categoria	0:00 - 5:59	MADRUGADA
	6:00 - 11:59	MAÑANA
	12:00 - 18:59	TARDE
	19:00 - 11:59	NOCHE

Tipo_hurto	MOTOCICLETA	MOTOCICLETA
	VEHÍCULO	VEHÍCULO
Arma_categoria	ARMA DE FUEGO	CON ARMA
	ARMA BLANCA / CORTOPUNZANTE	
	CONTUNDENTES	
	ESCOPOLAMINA	
	JERINGA	
	CORTOPUNZANTES	
	PUNZANTES	
	CORTANTES	
	ARMAS BLANCAS	
	SIN EMPLEO DE ARMAS	SIN ARMA
	LLAVE MAESTRA	
	SIN INFORMACION	
Movil_agresor_categoria	A PIE	A PIE
	EN VEHÍCULO	EN VEHÍCULO
Clúster	1	1
	2	2
	3	3
	4	4
	5	5
	6	6
Genero	FEMENINO	FEMENINO
	MASCULINO	MASCULINO
	NO REPORTA	NO REPORTA
Edad_categoria	Menor a 30 años	<= 30 A
	De 31 a 50 años	31 -50 A
	Mayor a 51 años	>51 A
Zona	RURAL	RURAL
	URBANA	URBANA
	OTRAS	OTRAS
Día_categoria	LUNES A VIERNES	Entre Semana
	SABADO Y DOMINGO	Fin de semana

Por medio del software Orange Canvas se realizará la clasificación usando un árbol de decisión, el cual tendrá por variable objetivo el “tipo de arma” dado que se busca comprender el comportamiento asociado a la violencia física sobre la que se incurre al momento de efectuar el hurto en los últimos años.

Como primera fase se divide aleatoriamente el archivo con el total de registros en dos datasets, uno con el 70% de los registros y el otro con el porcentaje restante para posteriormente validar el nivel de clasificación del árbol usando validación cruzada.

En la figura 39 se muestran los resultados estadísticos que arroja el árbol de decisión, en donde por ejemplo el AUC (Area Under the curve) con un valor del 0.807, indica un buen nivel de ajuste y la precisión es cercana al 80% para los datos con los que se realizó el entrenamiento.

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.807	0.786	0.776	0.791	0.786

Figura 39. Resultados árbol de decisión (sobre 70% datos)

Posteriormente se realiza la validación cruzada aleatoria con el 30% de los datos que se dividieron al inicio del análisis. Se generan 20 sets de datos aleatorios que comprenden el 66% del total de datos suministrados en el dataset de pruebas y se obtiene un estadístico similar al anterior que refleja el buen nivel de ajuste

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Tree	0.808	0.786	0.777	0.788	0.786

Figura 40. Resultados validación cruzada

Seguido del paso anterior, se construye la matriz de confusión (ver figura 41) en donde se observan el buen nivel de clasificación del árbol obtenido.

		Predicted		
		CON ARMA	SIN ARMA	Σ
Actual	CON ARMA	23123	16849	39972
	SIN ARMA	5676	59533	65209
	Σ	28799	76382	105181

		Predicted		
		CON ARMA	SIN ARMA	Σ
Actual	CON ARMA	80.3 %	22.1 %	39972
	SIN ARMA	19.7 %	77.9 %	65209
	Σ	28799	76382	105181

Figura 41. Matriz de confusión Árbol de decisión

Una vez realizadas las pruebas al árbol de decisión se procede a generar conclusiones a partir de la configuración estructural de éste:

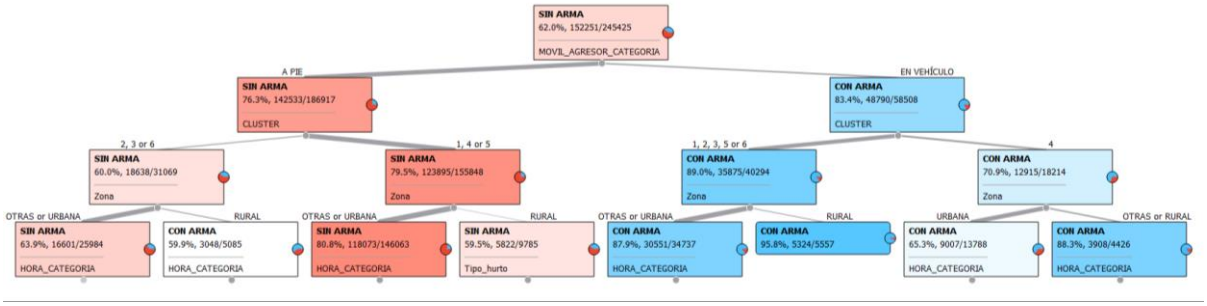


Figura 42. Árbol de decisión generado

Se aclara que las siguientes interpretaciones son basadas en el 70% de los datos aleatorios con los que se generó el árbol de decisión.

El nodo raíz indica que el 62% de los hechos delictivos que fueron realizados entre los años 2010 – 2019 en el territorio colombiano se realizaron sin el empleo de armas. La primera clasificación se hace sobre la manera en que se movilizaba el agresor, e indica que aproximadamente el 83% de las veces en que el victimario se desplaza en vehículos realizó el hurto usando algún tipo de arma, sin embargo cuando el agresor se desplaza a pie, el 76% del total de denuncias reportan que no empleo ningún arma.

Consecuentemente con el análisis previo, nos enfocaremos primero en la ramificación izquierda, la cual refiere a los delitos cometidos donde el victimario se movilizaba a pie: la siguiente clasificación se realiza según el clúster obtenido por medio del algoritmo K-means, e indica que en los municipios ubicados en los clústeres 1 y 4 cerca del 80% de los delitos realizados a pie no se emplearon armas, destacando a su vez que en los sectores rurales desciende la cifra a un 60%, frente a las zonas urbanas en donde el indicador calcula un 80%. En adición, los hurtos cometidos a pie en los municipios ubicados en los clústeres 2, 3 o 6 indican que en la mayoría de casos los delitos cometidos en zonas urbanas fueron realizados sin arma, mientras que en las zonas rurales el comportamiento es inverso y en mayor medida se dio lugar a hurtos con arma en un 60%.

Por otra parte, según la ramificación de la derecha es más representativo el hurto con empleo de algún tipo de arma cuando el victimario se moviliza en un vehículo, lo anterior fundamentado estadísticamente con la participación del 83% del total de denuncias realizadas los últimos 10 años, en las cuales el victimario se desplazaba en un vehículo. Adicional a lo anterior, el 89% de los hurtos realizados usando vehículos como medio de transporte de los asaltantes que se presentaron en los municipios categorizados en los clústeres 1, 2, 3, 5 o 6 fueron realizados con armas de algún tipo, mientras que en los eventos ocurridos en las mismas condiciones en el clúster 4 descienden al 70% el empleo de armas. Existe una situación que llama especialmente la atención debido a la concentrada cantidad de denuncias y es que alrededor del 96% de los hurtos realizados en vehículos, en los clústeres 1, 2, 3, 5 o 6 en zonas rurales se perpetuó usando armas, entretanto en la zona urbana el 88% de los delitos cometidos en vehículos en los mismos municipios también fueron realizados usando armas.

Se destaca en la figura 43 el modelo empleado para la creación del árbol de decisión, así como los test y la respectiva matriz de confusión.

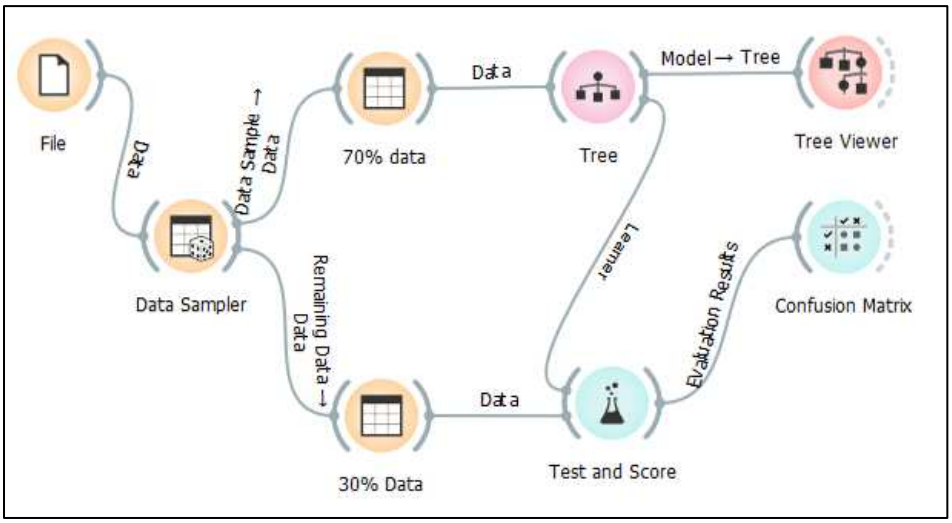


Figura 43.Modelo empleado en Software Orange Canvas para creación del árbol de decisión

8.7. GEOESTADÍSTICA:

A continuación, por medio de herramientas, test y estadísticos se validará la existencia o ausencia de patrones espaciales asociados a los hurtos de vehículos y motocicletas. Sin embargo, dado que hay municipios con mayor cantidad de hurtos debido a su número de habitantes, éstos afectarán en los resultados de los test y mostrarán seguramente que no existen relaciones espaciales.

Para dar solución al posible sesgo que generarán los municipios con mayor número de habitantes, se crearán índices de delincuencia compuestos de la siguiente manera:

$$I_{del_{Vehiculos}} = \frac{Cantidad\ de\ hurtos\ de\ automoviles\ entre\ 2010-2019}{Población\ municipal\ (Año\ 2018)\ Fuente{:}DANE} * 100 \quad (12)$$

$$I_{del_{Motos}} = \frac{Cantidad\ de\ hurtos\ de\ motocicletas\ entre\ 2010-2019}{Población\ municipal\ (Año\ 2018)\ Fuente{:}DANE} * 100 \quad (13)$$

8.7.1. Detección de patrones espaciales:

Una visión de la delincuencia asociada al hurto de vehículos por medio del índice definido anteriormente permite identificar que la región con mayor cantidad y mayor nivel de delincuencia se encuentra al norte de la región Caribe, seguido de la zona Oeste de la región Orinoquía, donde a pesar de que los municipios no presentan los más altos niveles de hurtos, si existe un gran número de municipios afectados (ver figura 44)

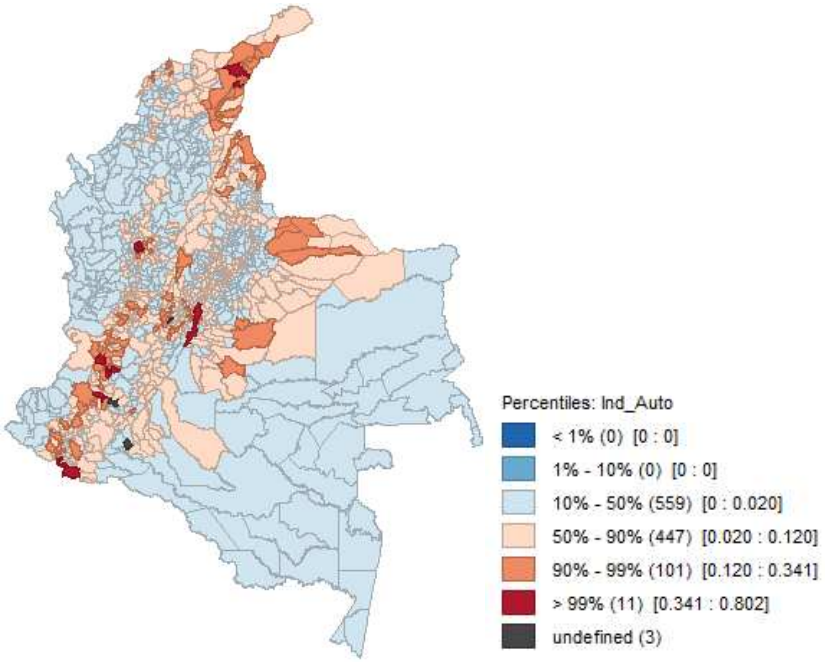


Figura 44. Mapa temático de percentiles asociados a la cantidad de hurtos asociados a vehículos entre 2010 - 2019

En la leyenda del mapa temático de percentiles anterior, entre paréntesis, el número de observaciones de cada categoría, 1121 municipios. Se identifican 11 municipios con índices extremos:

NOM_DEPART	NOM_MUNICI	Ind_Auto <
VALLE DEL CAUCA	CALI	0,802014
LA GUAJIRA	LA JAGUA DEL PILAR	0,714694
ANTIOQUIA	MEDELLÍN	0,505000
CAUCA	POPAYÁN	0,435139
NARIÑO	IPIALES	0,420197
CAUCA	CALOTO	0,404228
CAUCA	CORINTO	0,395476
LA GUAJIRA	SAN JUAN DEL CESAR	0,389751
CUNDINAMARCA	BOGOTÁ, D.C.	0,383592
ANTIOQUIA	ITAGÜÍ	0,366765
ANTIOQUIA	BELLO	0,344270

Figura 45. Municipios con índices extremos de delincuencia.

Otra herramienta para el análisis de la tendencia de una variable geográfica es el diagrama mapa de caja (boxplot), el cual es calculado de los cuartiles y la media de la variable, así como de la obtención de los valores adyacentes superior e inferior, obtenido como el producto de los valores del tercer cuartil por 1,5 veces el rango intercuartílico. Los valores que estén por fuera de estas cotas, son considerados como atípicos.

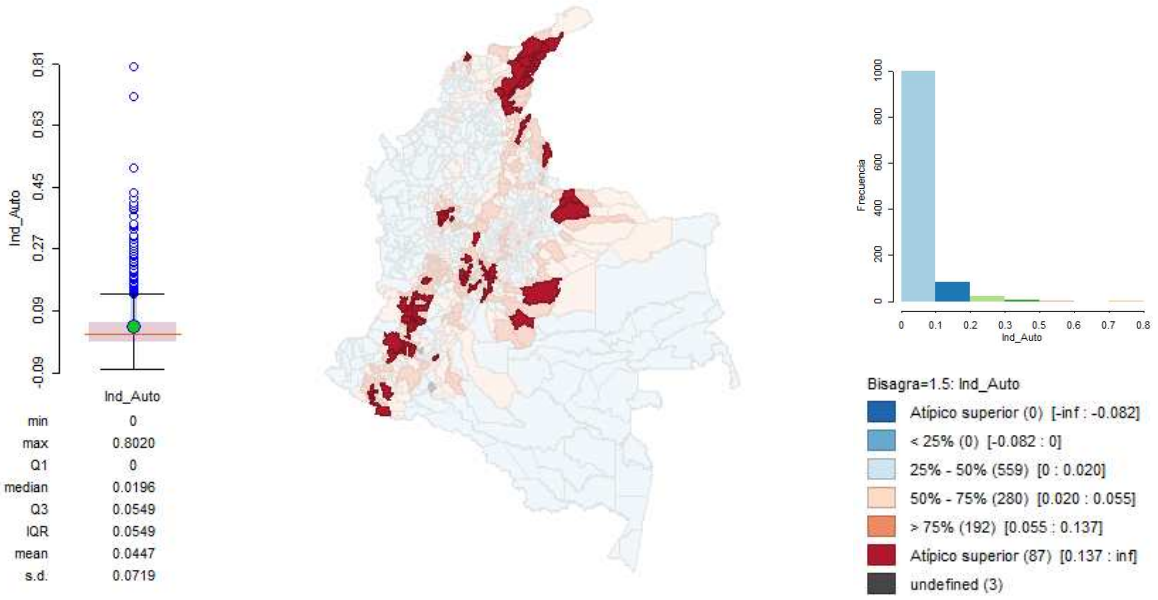


Figura 46. Detección de municipios atípicos

En la figura 46 se evidencia que la media del índice de delincuencia en el país es 0.0447 y la mediana 0.0196 lo que nos lleva a inferir que la distribución de los datos está sesgada a la derecha, conclusión que se comprueba en el histograma, y que a su vez es positiva

dando lugar a la conclusión de que aproximadamente 1.000 municipios tienen una tasa de 0% y reduciendo el número de municipios sobre los cuales las autoridades tendrán que definir medidas para reducir los hurtos. Sin embargo hay presencia de 87 valores atípicos superiores que se salen de la cotas con índices de delincuencia superiores al 0.13, sobre los cuales se deberá prestar puntual atención por parte de las autoridades y las alcaldías.

8.7.2. Estadístico espacial I de Moran:

El fenómeno de asociación espacial entre elementos del espacio puede ser comprendido por medio del diagrama de dispersión de Moran, el cual está subdividido en cuatro cuadrantes, el I y IV indican autocorrelación espacial negativa (valores bajos rodeados por altos y viceversa) y los cuadrantes II y III indican una correlación espacial positiva (bajos rodados de bajos o altos rodeados de altos) (Goovaerts & Jacquez, 2005).

En primera instancia, el test I de moran requiere como insumo una matriz de pesos, que para esta investigación y teniendo en cuenta que el objetivo de este capítulo es identificar las relaciones espaciales, conformadas en conjuntos por la asociación de entidades vecinas se emplean matrices de pesos basadas en la contigüidad de tipo reina de orden 1: Para unidades que comparten un lado en común o un vértice con la región de interés se define $\omega_{ij} = 1$ (Acevedo & Velásquez, 2008).

En la figura 47 se evidencia un ejemplo de los municipios contiguos (blanco) al municipio seleccionado (verde):

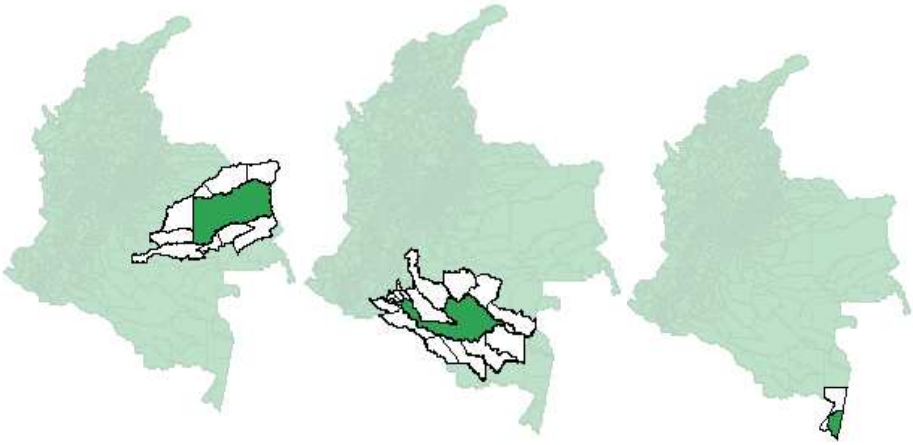
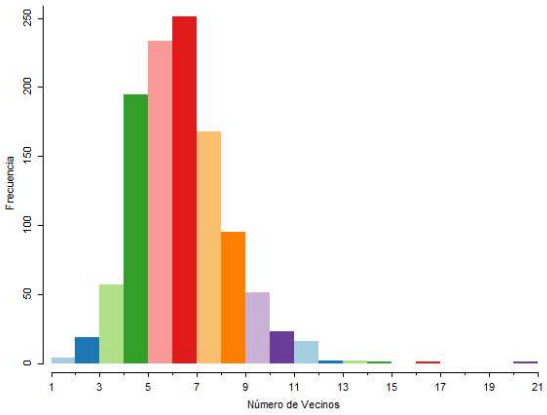


Figura 47. Ejemplo de contigüidad tipo Reina orden 1.

Asociado a la matriz de contiguidades generada se asocian estadísticos como el diagrama de frecuencias que indica que en su mayoria los municipios cuentan con vecindades superiores a 4 entidades , con una media de 5.8 y una mediana muy cercana con un valor de 6 vecinos.



Property	Value
type	queen
symmetry	symmetric
file	Hurtos_Porc_Moto.gal
id variable	ID_ESPACIA
order	1
# observations	1121
min neighbors	1
max neighbors	20
mean neighbors	5,87
median neighbors	6,00

Figura 48. Estadísticas de contigüidad de tipo Reina de orden 1

En la figura 48, el eje x representa la variable “% hurtos” estandarizada y el eje Y representa el retardo espacial de dicha variable estandarizada, es decir, el promedio ponderado de los valores que adopta una variable en el subconjunto de observaciones vecinas a una dada. De la gráfica, es válido interpretar y conocer el grado en que el estadístico I de moran resume la estructura global de asociación lineal existente.

En la siguiente figura se observa que el estadístico I de moran asociado al hurto de vehículos en los últimos 10 años indica que aproximadamente el 50% de los municipios con presencia de hurtos siguen una estructura global de asociación lineal.

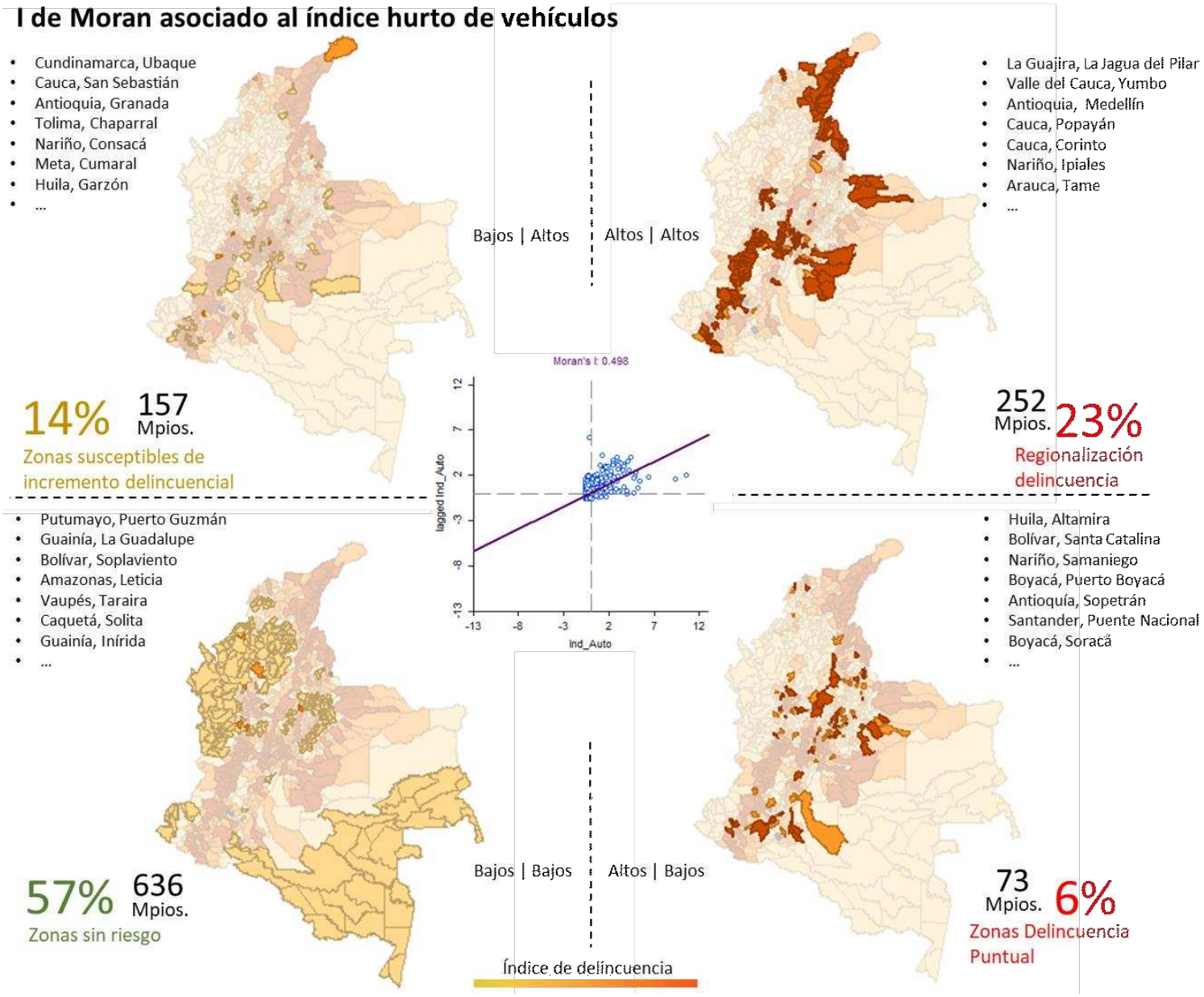


Figura 49. Análisis y estadísticas espaciales hurto vehículos

De la anterior gráfica, se evidencia que por medio del I de Moran se relacionan 4 categorías de municipios en función de las configuraciones espaciales. En el primer cuadrante se encuentran los municipios en donde existen altas tasas delincuencias rodeadas de altas, lo que de otra forma es llamado regionalización de delincuencia (clústeres de tasas de delincuencias alta), correspondientes al 23% de municipios en todo el país distribuidos en las zonas noreste, parte del centro de país y al sur oeste.

Por otra parte, en el cuadrante IV están clasificados todos aquellos municipios que tienen altos índices de delincuencia y los polígonos vecinos presentan un comportamiento opuesto, lo cual representa municipios de delincuencia puntual, en donde se requiere por parte de las fuerzas policiales aumentar el control para evitar que se creen los clústeres de delincuencia descritos en el cuadrante I. Visualmente se aprecia que un alto porcentaje de los municipios en esta categoría colindan o son muy próximos a los clúster de delincuencia [cuadrante I], lo que indica que son municipios que con el pasar de los años se ha venido extendido el radio de efecto de la delincuencia y éstos se encuentran en la periferia de las dichas zonas. El porcentaje de municipios que siguen este comportamiento obedece al 6% del país.

Un comportamiento opuesto al del cuadrante IV es el correspondiente a los 157 municipios del país que se encuentran categorizados en el cuadrante II, los cuales representan el 14% del total del país. Son municipios en donde a pesar de estar rodeados por municipios con altas tasas de delincuencia, el municipio en cuestión no tiene índices de delincuencia tan elevados. Espacialmente se destaca la distribución de estos municipios, los cuales se ubican sobre la zona central del país, cerca de los clústeres regionalizados de delincuencia, lo que indica que son municipios que de no tomarse acciones por parte de las diferencias entidad públicas, en unos años, estás zonas serán absorbidas por los clústeres de delincuencia.

Finalmente, el gráfico generado a partir del I de moran nos indica que el 57% de los municipios presentan un bajo o nulo índice de delincuencia asociado al hurto de vehículos en la última década (cuadrante III). De la misma forma en que se agrupan y crean clústeres de delincuencia, se generan las zonas seguras o clústeres con poca incidencia. En

Colombia en los últimos 10 años hay 636 municipios donde los indicies de delincuencia asociada al hurto de vehículos es muy bajo o nulo.

En la figura 50 se muestra los análisis resultantes del estadístico de Moran asociado al hurto de motocicletas.

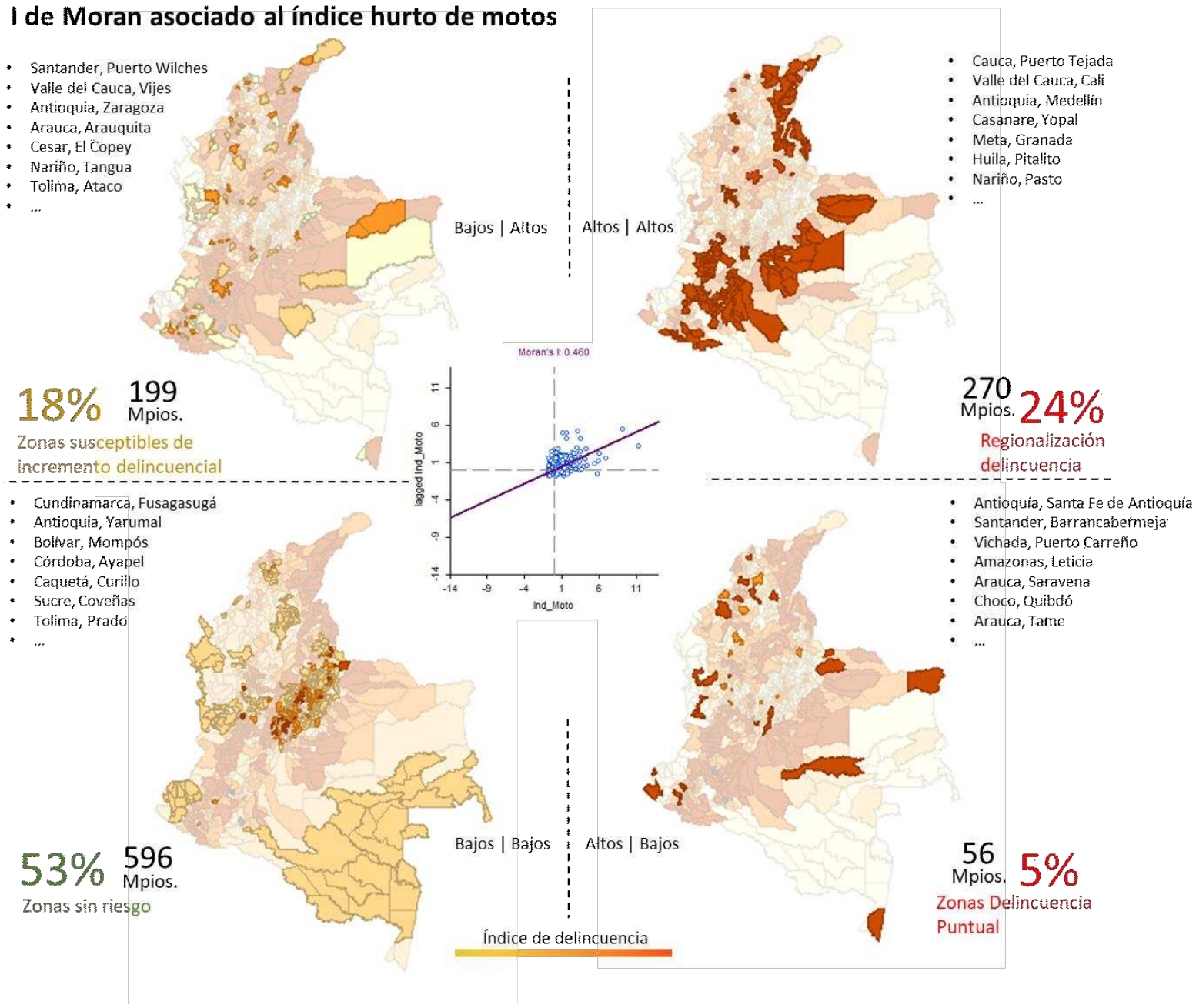


Figura 50. Análisis y estadísticas espaciales hurto de motocicletas

En el cuadrante I se encuentran agrupados los municipios que se encuentran clusterizados, con índices altos rodeados de altos. Corresponden al 24% de los municipios del país, porcentualmente 1% de incremento frente a la agrupación de zonas de delincuencia conglomerada con altos índices de delincuencia asociadas al hurto de vehículos.

Respecto al cuadrante IV, se evidencia geográficamente una distribución de municipios en todas las regiones del país, con mayor presencia en zonas extremas, casos como Leticia, Puerto Carreño, Tumaco, entre otros indican que las autoridades deben implementar estrategias a niveles municipales para reducir el hurto de motocicletas. Tan solo el 5% de municipios en el país presentan este comportamiento, por lo cual es conveniente que se tomen medidas para reducir la tasa de hurtos antes de que en municipios aledaños las cifras comiencen a incrementar.

Las zonas susceptibles de incremento delincuencial asociado al hurto de motocicletas represan un 18%. Esta asociación está conformada por 199 municipios que son ubicaciones que potencialmente se verán afectadas por la delincuencia de motocicletas a menos de que los entes gubernamentales tomen decisiones para solventar esta problemática.

Por último, las zonas sin riesgo corresponden al 53% de los municipios del país, un porcentaje inferior al de los vehículos (57%). Son municipios en donde los índices de delincuencia son bajos, así como el de los polígonos contiguos, en donde se ve que las zonas menos habitadas del país como la zona sureste y sur oeste están categorizadas en este cuadrante.

8.8. MODELAMIENTO ESTADÍSTICO:

Finalmente, por medio de un modelamiento estadístico se encontrará la relación numérica entre las variables analizadas en esta investigación. Teniendo en cuenta la naturaleza de los datos, la variable respuesta será la cantidad de delitos reportados asociadas al hurto de vehículos a lo largo de la última década en el país por municipio, por lo cual se construirá un modelo de regresión de datos de conteo por medio de un modelo lineal generalizado (GLM).

El conjunto de datos se modela en el marco de un modelo de regresión binomial negativo, debido a que la distribución Poisson implica la igualdad de la media y la varianza, así como también no permite una buena estimación si existe sobre dispersión de los datos.

Se descartaron los municipios con datos atípicos, particularmente los 3 que presentan mayor cantidad de denuncias, los cuales fueron clasificados dentro del clúster 1 en la sección 8.4, algoritmo K-means.

La técnica de regresión backward stepwise permitió obtener la mejor combinación de parámetros, seguido de esto las variables no significativas estadísticamente y permanecieron las significativas al 95% de nivel confianza. Los parámetros del modelo fueron estimados usando el método de logaritmo de máxima verosimilitud y los resultados del modelo se encuentran en la figura 48. Se destaca el valor $\theta = 0.7042$, que indica sobre dispersión en los datos evidenciando la justificación del uso de la distribución binomial negativa para el modelamiento.

```
Call:
glm.nb(formula = Y ~ mydata$MOTOS + mydata$POB_2018 + mydata$Prop_de_Personas_en_NBI____ +
    mydata$EDAD_MEDIA + mydata$DUM_GEN_FEMEN, maxit = 1200, init.theta = 0.7041722907,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.6345  -0.9607  -0.5200   0.2752   2.0347

Coefficients:
                                     Estimate    Std. Error z value   Pr(>|z|)
(Intercept)                     -0.9387043404  0.1850123830  -5.074 0.00000039 ***
mydata$MOTOS                     0.0020049170  0.0001167330  17.175    < 2e-16 ***
mydata$POB_2018                  0.0000126197  0.0000008334  15.142    < 2e-16 ***
mydata$Prop_de_Personas_en_NBI____ -0.0082856806  0.0024055923  -3.444   0.000572 ***
mydata$EDAD_MEDIA                0.1252893917  0.0047733497  26.248    < 2e-16 ***
mydata$DUM_GEN_FEMEN            -2.3663009283  0.2811321900  -8.417    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7042) family taken to be 1)

    Null deviance: 3640.5  on 1116  degrees of freedom
Residual deviance: 1224.8  on 1111  degrees of freedom
AIC: 10309

Number of Fisher Scoring iterations: 1


              Theta:  0.7042
          Std. Err.:  0.0278

 2 x log-likelihood:  -10294.9930
```

Figura 51. Parámetros de estimación del modelo

De la figura 51 se evidencia el ajuste de las variables independientes al modelo, con un nivel de confiabilidad estadística al 95%, con lo cual se rechaza la hipótesis nula, es decir los parámetros son significativos estadísticamente.

Para la interpretación del modelo se deben tener en cuenta que la categoría base de la variable cualitativa DUM_GEN_FEMEN es Masculino. Asimismo, en función de los parámetros estimados por el modelo se tiene que un aumento de la variable Número de Motos hurtadas, población y edad media producen un incremento en la cantidad de total de hurtos en los municipios. Por otra parte, la variable género Femenino disminuye la cantidad total de hurtos. Se observa que un aumento en la variable proporción de personas en NBI produce una disminución en el número de delitos asociados a hurtos de vehículos.

Se verifica que las variables del modelo no tienen problemas de multicolinealidad, tal como se muestra en la siguiente figura.

```
> vif(negbin)
                    mydata$MOTOS   3.545197
                mydata$EDAD MEDIA   1.025405
            mydata$DUM_GEN_FEMEN   1.006416
                 mydata$POB_2018   3.546435
mydata$Prop_de_Personas_en_NBI.   1.036424
```

Figura 52. Test de factor de inflación de varianza.

En la figura 52 se evidencia que el valor init.theta es igual a 0.7041. Por lo tanto el valor estimado de θ es la inversa, 1.42. Con lo cual se concluye que existe sobre dispersión en los datos y efectivamente es conveniente usar la distribución Binomial negativa en vez de la distribución Poisson para la construcción del modelo de conteo.

Por otra parte, el estadístico criterio de información Akaike (AIC) muestra un valor significativamente alto, sin embargo dado que el rango de éste está inscrito en el intervalo (-∞, ∞) no es posible afirmar el grado de ajuste del modelo. Se destaca que el valor de 2 veces el logaritmo de máxima verosimilitud es negativo, lo cual es un indicio positivo de que el modelo ajusta los datos.

En consecuencia, se calculan estadísticos adicionales (figura 53) para evaluar el ajuste del modelo que permiten concluir que éste tiene un grado aceptable de ajuste dado su estadístico Pseudo R2 equivalente a 0.72, con lo cual la variabilidad de la variable Total Hurtos es explicada en aproximadamente un 72% por las 5 variables independientes (Cantidad de motos hurtadas, población del municipio [vig. 2018], Proporción de personas en NBI, Edad Media de la víctima en el municipio y género más afectado en el municipio (dummy femenino) respectivamente). Adicionalmente, las variaciones entre las desviaciones nos indican un buen ajuste de los datos.

```
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML        r2CU
-5147.4962807 -5869.7454952  1444.4984289     0.1230461     0.7256072   0.7256270
```

Figura 53. Estadísticos para evaluar ajuste del modelo

Por otra parte, en la figura 54 se relacionan las gráficas asociadas a los residuales del modelo. Residuals vs Fitted: En esta gráfica se identifica que no existe un patrón lineal sin embargo se destaca que los residuales están visualmente distribuidos en la misma cantidad por encima y por debajo del ajuste del modelo; Normal Q-Q: Se evidencia que los datos no siguen del todo una distribución normal; Scale-Location: Están visualmente distribuidos en la misma cantidad por encima y por debajo del ajuste del modelo; Residuals vs Leverage: Se evidencia presencia de Outliers en la observación 1008 y 1009.

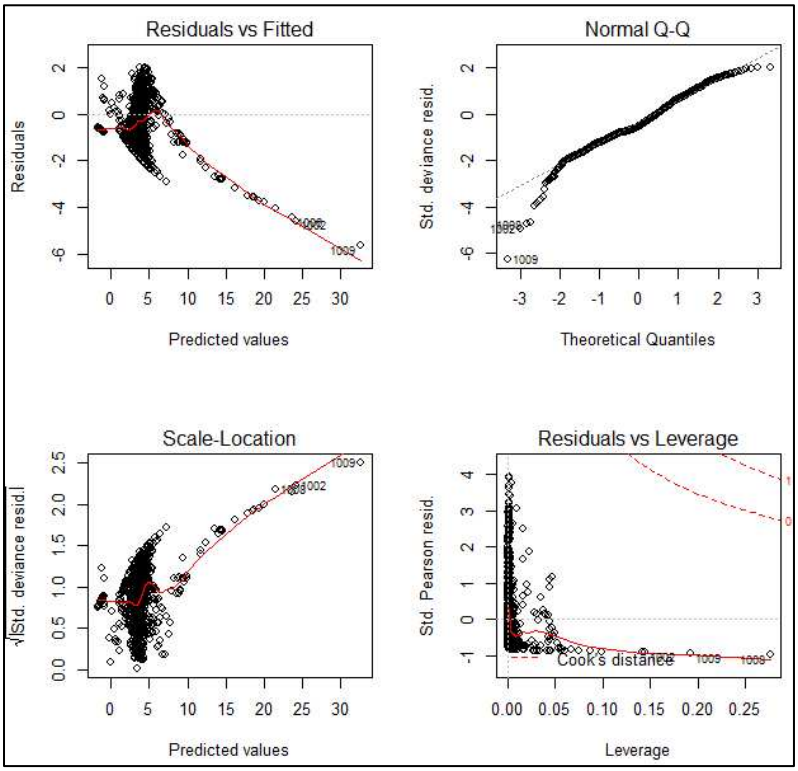


Figura 54. Residuales del modelo

9. CONCLUSIONES

De acuerdo a los objetivos planteados, se construye satisfactoriamente un sistema de información geográfica compuesto por varios niveles de información, entre los cuales se destacan los datos alfanuméricos de las denuncias, información del número de habitantes y estadísticos de las necesidades básicas insatisfechas, por medio del cual se centraliza de manera ordenada y estructurada la información, para posteriormente ser procesada y analizada por medio de las diversas ciencias que se integran en esta investigación.

Como consecuencia de lo expuesto en este documento, se comprueba que la hipótesis planteada es verdadera, toda vez que por medio de la integración de varias disciplinas se logra caracterizar, comprender y analizar el comportamiento de la delincuencia asociada al hurto de vehículos durante una década en el país. De esta manera, se comprueban por medio de los resultados obtenidos de los algoritmos las siguientes afirmaciones:

> Se evidencia que las variables “zona” y tipo de arma empleada están correlacionadas, dado que según el árbol de decisión y el análisis exploratorio de los datos se interpreta que existen porcentajes de participación entre el 88% y 95% de hurtos empleando armas en zonas rurales para ciertos municipios, mientras que en zonas urbanas la tendencia baja del 65% al 87%.
>
> En virtud de lo estudiado, es posible sostener que gracias a las reglas de asociación creadas con el algoritmo Apriori se evidencia con una confianza superior al 90% en varias reglas, que cuando se realizan hurtos empleando llaves maestras, el victimario de desplaza a pie, sin embargo se debe tener presente que en aquellos hurtos donde el denunciante no está presente, al momento de denunciar se relaciona en el campo “móvil del agresor“ que éste se desplaza a pie. Por otra parte, particularmente para motocicletas de marca Bajaj se evidencia con un soporte del 13% y una confianza del 91% este hurto afecta a la población masculina, y a su vez se resalta que del análisis exploratorio se concluye que Bajaj es la línea de motocicletas más hurtadas en la última década. Sin embargo, la evidencia presentada nos lleva a concluir que la generación de las reglas generadas por el algoritmo Apriori no provee resultados diferentes o

destacables frente a los que se obtienen a través de los demás algoritmos, por lo que se recomienda en próximas investigaciones añadir más variables al conjunto de datos.

Asimismo, se comprueba que la hipótesis: “las NBI están relacionadas en forma proporcional positivamente con el hurto de vehículos” es falsa, dado que según lo encontrado tras realizar la segmentación de los datos con el algoritmo K-Means y analizar los municipios agrupados (clustering) con menores NBI se evidencian de medias a altas estadísticas de hurto, mientras que en municipios con mayores NBI se encontró sorpresivamente que el hurto de vehículos es menor. Esto posiblemente se debe a que en poblaciones con mayores NBI, la adquisición de vehículos es mucho menor que en las grandes ciudades o centros poblados principales, los cuales tienden a tener NBI de bajas a medias.

Por otra parte, el algoritmo K-means permite afirmar que la violencia con que se comete el hecho delictivo está condicionada por el número de habitantes del municipio, debido a que en los municipios con un número medio y alto de habitantes la tendencia es efectuar el robo sin emplear armas. Se resalta la especial atención que deben prestar las autoridades a los municipios agrupados en el clúster 3 debido a que son regiones con poblaciones principalmente pequeñas en cantidad de habitantes y donde en la mayoría de hechos fueron empleadas armas, con lo cual la probabilidad de atentar contra la vida de las victimas es superior.

Adicionalmente, se concluye por medio del árbol de decisión que en el área rural los robos tienden a ser mayormente empleando armas de fuego, a diferencia de las zonas urbanas donde destaca el uso de llaves maestras. A su vez, el móvil en que se moviliza el agresor está asociado con el modo de efectuar el hurto, dado a que si se moviliza en algún vehículo es más probable el empleo de armas que si se moviliza a pie, por lo cual los entes competentes deberían incrementar las requisas en zonas rurales. Con respecto a la configuración de vecindad entre la delincuencia de los municipios, se recomienda a la administración municipal y a las autoridades proyectar en los planes de desarrollo dentro de sus pilares o ejes transversales la disposición de recursos enfocados en la reducción de la delincuencia y para las

zonas donde se ha regionalizado la delincuencia plantear estrategias conjuntas, así como sumar esfuerzos en aquellas zonas susceptibles del incremento delincuencial.

10. IMPACTO Y RESULTADOS ESPERADOS

En la última década el hurto asociado a vehículos en el territorio colombiano ha incrementado, un dato importante es que de cada 10 vehículos hurtados 7 son motocicletas; asimismo, el top 10 de barrios donde se reportan más denuncias a lo largo de toda la década analizada están ubicados en Medellín, con lo cual por parte de la administración estatal deben iniciarse procesos de vigilancia y control especial sobre estas zonas.

Se destaca que para combatir la delincuencia no solo basta de la fuerza policial, sino también es responsabilidad de la administración estatal, toda vez que, según los parámetros del modelo de conteo, se evidencia que a medida que la población aumenta, la cantidad de hurtos incrementará, por lo tanto este fenómeno se puede tornar imposible de manejar para la autoridad a menos de que se creen planes de gobiernos que fortalezcan aspectos como la economía, educación, salud, y tejido social de los Colombianos para desde allí reducir la propensión a cometer estos hechos delictivos.

Por otra parte, desde una perspectiva más teórica se encuentra que los resultados son congruentes en su mayoría con lo esperado desde que se da inicio en la fase de exploración de datos. Por una parte, el empleo del mecanismo de iteración para determinar el número óptimo de clústeres a generar por parte del algoritmo K-means, permitió generar un número justo de subconjuntos para eventualmente analizar el comportamiento de las variables de una manera precisa, así como también partiendo de la naturaleza propia del algoritmo, el cual es no supervisado y tiene por objeto la agrupación de elementos con características similares, se destaca favorablemente su aplicación en esta investigación.

Por otro lado, el algoritmo Apriori implementado a pesar de sus bondades en rendimiento para el conjunto de datos estudiado y generación de reglas de asociación con confianzas superiores al 89%, no aportó reglas que permitieran comprender con mayor profundidad

relaciones entre los hechos delictivos, más allá de las obtenidas en principio gracias al Análisis exploratorio y el análisis realizado a los clústeres generados por K-means, por lo tanto, este algoritmo se toma en la investigación como un mecanismo más confirmatorio que de descubrimiento.

Por contrario, el árbol de decisión favoreció al entendimiento de una manera visual del comportamiento del delito en cuanto al empleo de armas por parte del victimario, con lo cual se válida éste como una herramienta muy poderosa por cuanto permite al interesado de una manera dinámica y soportada cuantitativamente comprender las dinámicas entre los datos suministrados por los denunciantes. Asimismo, se resalta que se logra construir un árbol de decisión con un estadístico AUC muy bueno, correspondiente a 0.807, el cual fue validado posteriormente por medio de validación cruzada con 20 datasets aleatorios, y se obtiene un estadístico similar al anterior que refleja el buen nivel de ajuste. En adición a lo anterior, como limitante del árbol de decisión se tiene que solo recibe información de tipo cualitativa, sin embargo, este hecho no represento un gran problema debido a que la mayoría de las variables eran cualitativas, y las pocas cuantitativas fueron dicotomizadas empleando librerías del software R.

Igualmente, el desarrollo realizado empleando Geoestadística fue altamente favorable debido a que gracias a la implementación de un indicador para limitar el sesgo que se genera en algunas ciudades por el número de habitantes, se logra detectar las correlaciones espaciales existentes a nivel municipal a lo largo del país y clasificar los municipios de manera puntual según el tipo de influencia que presentan frente a sus vecinos más cercanos.

11. RECOMENDACIONES PARA TRABAJOS FUTUROS

Partiendo de las principales limitaciones de la investigación se destaca que no se tuvieron en cuenta los datos asociados al número de vehículos recuperados en cada municipio por las autoridades debido a la dificultad para obtener los mismos, los cuales podrían generar afirmaciones adicionales en cuanto a la caracterización del delito, tales como la tasa de vehículos recuperados. Además, en el contexto del estudio Geoestadístico desarrollado se hubiese podido realizar un análisis más detallado si la Policía Nacional suministrará la coordenada de la ubicación puntual donde se comete el hurto, sin embargo, las políticas actuales por seguridad del denunciante limitan compartir este nivel de detalle con la ciudadanía.

Con respecto a los algoritmos empleados, para trabajos futuros se recomienda emplear algoritmos de varios tipos para comparar los resultados, así como el rendimiento de éstos, por ejemplo, en cuanto a la creación de las reglas de asociación, además de usar el algoritmo Apriori, se sugiere emplear algoritmos como el Eclat, AIS y SETM, que presentan metodologías diferentes y mejoras en rendimiento frente al empleado en esta investigación.

Finalmente, se debe tener en cuenta que debido a la emergencia sanitaria que se presentó en 2020 por consecuencia del COVID-19, para próximos trabajos se debe tener en cuenta que las cantidades globales posiblemente presenten variaciones, dado que potencialmente en aquellos municipios donde se limitó la movilidad de las personas, este tipo particular de hurtos decrecieron ampliamente, mientras que una vez estas medidas desaparecían los hurtos probablemente aumentaban progresivamente. Asimismo, sería relevante en próximas investigaciones comprobar si a partir del periodo post-COVID los hurtos aumentaron o disminuyeron y comparar estos niveles con estadísticos sociales como las necesidades básicas insatisfechas por municipio.

12. RECURSOS Y PRESUPUESTO

Esta investigación se desarrolló en 1 año, con tiempos de trabajos medios de 6 horas semanales de dedicación y sesiones para consultas y revisiones periódicas mensuales con el tutor.

Debido a que la información será obtenida de fuentes públicas, no se incurrirán en costos para la adquisición de los datos. En adición, el recurso físico tecnológico empleado ha sido un computador con una memoria RAM de 8gb, un procesador Intel Core i7 de 8va generación y unidad de disco solido SSD de 500gb para poder procesar los datos rápidamente y hacer múltiples pruebas de validación cruzada.

13. REFERENCIAS

Acevedo, I., & Velásquez Ceballos, E. (2008). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. *Ecos de Economía No. 27. Medellín*, 9–34.

Alves, L. G. A., Lenzi, E. K., Mendes, R. S., & Ribeiro, H. V. (2015). Spatial correlations, clustering and percolation-like transitions in homicide crimes. *EPL*, *111*(1). https://doi.org/10.1209/0295-5075/111/18002

Atoche Calzada, P. (2017). *Modelos de regresión con datos de conteo: aplicación a competiciones deportivas*.

Baraya, A. (2018). Así están las mujeres al volante en Colombia. *Revista Motor*. https://www.motor.com.co/actualidad/industria/mujeres-volante-colombia-demografia/30837

Baraya, A. (2019). Se disparó el hurto de vehículos. *El Tiempo*.

https://www.motor.com.co/actualidad/industria/disparo-hurto-vehiculos/31853

Bejerano, P. (2018). La herramienta que permite saber cuándo y dónde se va a producir un delito | Tecnología | EL PAÍS. *Diario El País*. https://elpais.com/tecnologia/2017/03/09/actualidad/1489078250_691655.html

Bungue, M. (1919). La ciencia: su método y su filosofía (Vol 1). *Laetoli*.

Christian, S., Curtis, E., Michael, B., & Shashi, K. (1998). Jensen, Dyreson, hlen, James Gadia, . . . Yannis Mitsopoulos. *23 de 05 de The Consensus Glossary of Temporal Database ConceptsFebruary Version*, *1 SRC-B*, 367–405.

DANE. (2018). Necesidades Básicas insatisfechas (NBI). Metodología NMBI. *Colombia*.

Duce, D. T. M., & Bandrés, E. (2001). Delincuencia y acción policial: Un enfoque económico. *Revista de Economía Aplicada*, *9*(27), 5–34.

El Tiempo. (2019). Hurto de vehículos, un delito en ascenso desde el 2016. *El Tiempo*.

Esri. (2018). *Arcgis Help*. https://desktop.arcgis.com/es/arcmap/10.3/manage-data/geodatabases/what-is-a-geodatabase.htm

Forero, M. A. (2018). Análisis espacio-temporal usando redes y datos funcionales en la dinámica criminal de Medellín para el periodo 2003 - 2015. *Memorias Simposio Internacional de Estadística. Universidad Nacional de Colombia*.

Ghiasi, M. M., & Zendehboudi, S. (2019, November 17). *Decision tree-based methodology to select a proper approach for wart treatment*. Computers in Biology and Medicine. https://www.sciencedirect.com/science/article/pii/S001048251930109X

Goovaerts, P., & Jacquez, G. M. (2005). Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models. *J Geograph Syst*, *7*, 137–159. https://doi.org/10.1007/s10109-005-0154-7

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: Issues, challenges, tools and Good

practices. *2013 6th International Conference on Contemporary Computing, IC3 2013*, 404–409. https://doi.org/10.1109/IC3.2013.6612229

Kreyszig, E. (1979). Introducción a la estadística matemática: principios y métodos. *Editorial LIMUSA 1979*, 505.

Olivares, J. P. (2002). ANÁLISIS EXPLORATORIO Y ANÁLISIS CONFIRMATORIO DE DATOS. *Espacio Abierto*, *11*. https://www.redalyc.org/pdf/122/12211106.pdf

Ospina, J., & Correa, M. (2015). El hurto de automotores bajo el enfoque prospectivo: una caracterización actual y futura de la actividad criminal en Colombia. *Revista Criminalidad*, *57*(3), 9–25.

Quintero, C., & Lahuerta, P. (2008). D., & Y. (Junio de . *Un Índice de Criminalidad Para Colombia*, *50*(1 SRC-BaiduScholar FG-0), 37–58.

Roa, C. F. R. (2016). *Colombia: Energy Policy (Electricity)*. Dundee: CEPMLP–University of Dundee.

Rodrigo, J. (2018, June). *Reglas de asociación y algoritmo Apriori con R*. https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion

Rodriguez R., L. A. (2017). *Generación del mapa de crimen del municipio de Zipaquira mediante herramientas geoestadisticas y SIG*. https://1library.co/document/q5w6omrq-generacion-mapa-crimen-municipio-zipaquira-mediante-herramientas-geoestadisticas.html

Ruth, K., Pierre, G., Robert, P., & Vania, C. (2010). H., & Applying Geostatistical Analysis to Crime Data: Car-Related Thefts in the Baltic States. *Geographical Analysis*, *42*(1 SRC-BaiduScholar FG-0), 53–57.

Sánchez-Torres, R. M. (2017). Desigualdad del ingreso en Colombia: un estudio por departamentos. *Cuadernos de Economía*, *36*(72), 139–178.

Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., Xu, X., & Kriegel, H.-P. (2017). DBSCAN

Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst*, *42*(3). https://doi.org/10.1145/3068335

Semana. (2018). ¡Miedo! *Revista Semana*. https://www.semana.com/nacion/articulo/miedo-inseguridad-en-colombia-delincuencia/555836/

Uittenbogaard, A., & Vania, C. (2012). *Space-time Clusters of Crime in Stockholm,Sweden. RES . 4*(10.5539/res.v4n5p148 DOI-10.5539/res.v4n5p148 SRC-BaiduScholar FG-0).

Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*, *3*(1), 9–22. https://doi.org/10.21500/20112084.819