# ANÁLISIS ESPACIOTEMPORAL DE LA RELACIÓN ENTRE LAS INFRACCIONES Y LOS ACCIDENTES DE TRÁNSITO EN LA CIUDAD DE CALI, COLOMBIA 2021-2022

*Christian Fernando Grisales Cárdenas – Cod: 8992772*
*Fabián Andrés Castro Salazar – Cod: 8937795*
*Gustavo Andrés Moreno Collazos – Cod: 8937795*

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

Director:
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 19 DE 2025

# FICHA RESUMEN

**TÍTULO DEL PROYECTO:** Análisis espaciotemporal de la relación entre las infracciones y los accidentes de tránsito en la ciudad de Cali, Colombia 2024.

1. **ÁREA DE TRABAJO:** Sector Gubernamental

2. **TIPO DE PROYECTO**: Aplicado

3. **ESTUDIANTES**: Christian Fernando Grisales Cárdenas, Fabián Andrés Castro Salazar y Gustavo Andrés Moreno Collazos.

4. **CORREO ELECTRÓNICO**: chrisfgc@javerianacali.edu.co, cfabian@javerianacali.edu.co y gustavomoreno1015@javerianacali.edu.co

5. **DIRECCIÓN Y TELEFONO:** Calle 18 # 118 – 250, Universidad Javeriana – Cali, Colombia (3147647789), (3227778327), (3168079612)

6. **DIRECTOR:** David Arango Londoño

7. **VINCULACIÓN DEL DIRECTOR:** Profesor Planta, Departamento de Ciencias Naturales y Matemáticas, Pontificia Universidad Javeriana Cali.

8. **CORREO ELECTRÓNICO DEL DIRECTOR**: david.arango@javerianacali.edu.co

9. **OTROS GRUPOS O EMPRESAS:** Grupo de investigación Estadística y Matemática Aplicada (EMAP)

10. **PALABRAS CLAVE** (al menos 5): Análisis espacial, infracción de tránsito, siniestro vial, dato georreferenciado, función de intensidad, mapa de calor, proceso puntual espacial, modelo predictivo.

11. **FECHA DE INICIO:** 23 de Julio del 2024

12. **DURACIÓN ESTIMADA:** 10 meses

**13. RESUMEN:**

El presente documento es resultado del proyecto aplicado, requisito esencial para optar por el título de Maestría en Ciencia de Datos de la Pontificia Universidad Javeriana Cali. En él se propone la aplicación de modelos estadísticos para procesos puntuales espaciales con el fin de analizar la distribución geográfica de los siniestros viales con fatalidad, en relación con diversas covariables asociadas a infracciones de tránsito, en el área urbana de Santiago de Cali. La información utilizada

fue proporcionada por la Secretaría de Movilidad e incluye datos georreferenciados de siniestros y reportes de infracciones correspondientes a los años 2021 y 2022. Con un enfoque teórico-práctico, guiado por la metodología CRISP-DM, se integró el conocimiento académico en la solución de una problemática social crítica como lo es la mortalidad por accidentes de tránsito. Inicialmente, se establecieron los fundamentos conceptuales de los procesos puntuales espaciales. Luego, se llevó a cabo una exploración de los datos espaciales y se desarrollaron modelos de intensidad de puntos, utilizando tanto enfoques estadísticos clásicos como algoritmos de aprendizaje automático. Entre estos, el modelo de *bosques aleatorios* presentó el mejor desempeño según las métricas MAE, RMSE y $R^2$. Los resultados evidencian que los siniestros mortales están significativamente asociados con infracciones como *conducir bajo los efectos del alcohol o sustancias psicoactivas*, así *como ignorar señales de pare o semáforos en cruces viales*. Además, se identificaron zonas de alto y bajo riesgo de fatalidad en la ciudad, lo cual permitió generar recomendaciones de intervención en infraestructura y programas de educación vial.

# TABLA DE CONTENIDO

## LISTA DE FIGURAS

**Pág.**

## LISTA DE TABLAS

**Pág.**

## LISTA DE ECUACIONES

**Pág.**

# INTRODUCCIÓN

Según la Organización Mundial de la Salud, cada año, mueren producto de accidentes de tránsito 1.9 millones de personas en el mundo, es decir, una persona cada dos minutos. Si bien es cierto que en Colombia se registró una disminución del 12% entre los años 2022 y 2023, esta cifra es muy alta al registrar 8.405 personas fallecidas indicando que diariamente en Colombia mueren 23 personas por accidentes de tránsito. Sin embargo, es aún más alarmante el hecho de saber que diferentes estudios han demostrado que en un alto porcentaje (más del 90%) los accidentes se pueden evitar, entendiendo que las principales causas están asociadas a conducir bajo los efectos del alcohol o sustancias psicoactivas y el exceso de velocidad [1].

Las estadísticas también muestran que Cali tuvo una disminución del 4% en las fatalidades, al comparar los años 2022 y 2023 que, de acuerdo con los reportes de la Agencia Nacional de Seguridad Vial, finalizó el año con 322 muertes [2]. Lo anterior, significa que en la ciudad fallece una persona todos los días por accidentes de tránsito y donde el actor vial con mayor incidencia es el motociclista, seguido de los peatones. Respecto a las lesiones, Cali mostró un comportamiento diferente al presentar un aumento, pasando de 1885 a 1997 accidentes convirtiéndose la siniestralidad en uno de los principales problemas de salud pública de la ciudad.

De acuerdo con lo expuesto anteriormente, y considerando la alta incidencia de los accidentes de tránsito en la ciudad de Cali, los cuales, en su mayoría, son evitables al estar asociados a factores comportamentales y decisiones humanas de último momento, se planteó como objetivo el desarrollo de un modelo que permita analizar la relación entre las infracciones viales y los accidentes de tránsito con resultado fatal. A través de este enfoque, se buscó identificar patrones que permitan inferir la posible influencia de la reincidencia en determinados tipos de infracciones sobre la ocurrencia y severidad de los siniestros viales.

Finalmente, la posibilidad de georreferenciar los puntos con mayor concentración de infracciones y de accidentes de tránsito ofrece un valor estratégico significativo para la gestión urbana. Esta información espacial puede ser utilizada por las instituciones municipales como insumo clave para la toma de decisiones en materia de inversión en infraestructura vial, intervenciones en puntos críticos o fortalecimiento del control por parte de las autoridades. En este marco, se desarrolló un modelo que permitió construir mapas de intensidad espacial, los cuales visualizan la relación entre la reincidencia en infracciones y la ocurrencia de accidentes fatales.

# 1. CONTEXTUALIZACIÓN DEL PROYECTO

## 1.1 Definición del problema

### 1.1.1 Planteamiento del problema

De acuerdo con la agencia de noticias de las Naciones Unidas [1], los accidentes de tránsito dejan un saldo de 1,3 millones de personas muertas y más de 50 millones de personas heridas de gravedad cada año. Esta cifra llevó a que las Naciones Unidas definieran el decenio 2011 – 2020 para ejecutar un plan que llevara a una disminución progresiva de los accidentes y los costos asociados que esto representa. Después del cierre del periodo, el año 2020 cerró con 100,000 casos más que el 2011, lo que indica que las acciones que se han tomado a nivel global no han sido suficientes. Para Colombia, aunque en el último reporte se demuestra que, en especial en 2023, ha disminuido un 12% frente a 2022 [3], esta cifra representa un total de 8,405 personas fallecidas, lo que indica que diariamente mueren 23 personas, en su mayoría entre 25 y 34 años. Estas muertes representan solo una parte de la eventualidad, ya que en muchos accidentes las personas quedan con lesiones graves que no les permiten volver a realizar sus tareas de rutina. De igual forma, existen choques sin lesiones que generan impactos en la economía por la pérdida de actividad productiva y los gastos cuantiosos en salud pública.

De acuerdo con estadísticas publicadas por la Agencia Nacional de Seguridad Vial [2], la ciudad de Cali, con corte a abril del 2024, presenta un aumento del 8,81% de siniestros viales respecto al mismo periodo del 2023. En el año 2023 se presentaron 101 fallecidos y lo que va corrido del 2024 se han presentado 102 fallecidos. Por su parte, se presentaron 421 casos con lesionados en el 2023 y para el 2024 van 466 personas lesionadas. Sin embargo, estas estadísticas no incluyen los choques simples, ya que no se pueden cuantificar fácilmente debido a un cambio en la regulación. Según esta nueva normativa, en casos sin lesiones o fatalidades, los terceros deben tomar fotografías para hacer las reclamaciones directamente a las aseguradoras, evitando así bloqueos en la ciudad y la necesidad de la presencia de un agente de tránsito, lo que dificulta la cuantificación de los costos por pérdida de productividad al no contar con datos oficiales. Por lo tanto, el análisis se centró en los accidentes con fatalidad.

Con la información disponible sobre comparendos por infracciones de tránsito de la Secretaría de Movilidad de la ciudad de Cali, se busca predecir principalmente la ocurrencia de un siniestro. La evidencia ha demostrado que, a mayor velocidad, hay una mayor probabilidad de que las personas involucradas sufran lesiones graves o fallezcan [4]. Basándose en este antecedente, se desarrolló un modelo que relaciona las infracciones de tránsito con la probabilidad de accidentes fatalidades. Para el desarrollo del proyecto, se utilizaron técnicas de análisis de datos geoespaciales y estadística espacial en las zonas de ocurrencia. Además, se identificó la relación con la imposición de comparendos para determinar los puntos más críticos y proponer soluciones que reduzcan la probabilidad de siniestros. Esto incluye la geolocalización de cada punto para identificar las mejores oportunidades de intervención en infraestructura, pedagogía, o acciones de seguimiento y control por parte de las autoridades.

### 1.1.2 Formulación del problema

El problema principal que se abordó en este proyecto estuvo enfocado en determinar cómo las infracciones de tránsito impuestas por las autoridades competentes guardaron relación con los accidentes de tránsito con personas fallecidas en la ciudad de Cali para los años 2021 y 2022.

### 1.1.3 Sistematización

- ¿De qué manera se puede predecir la accidentalidad a partir de los comparendos impuestos por infracciones de tránsito en la ciudad?
- ¿Qué tipo de relación existe entre la dirección de la imposición del comparendo y la dirección de la ocurrencia del accidente?
- ¿Cómo se pueden identificar patrones en la fecha, hora y lugar de los accidentes de tránsito en la ciudad?
- ¿Qué información se puede obtener sobre la cantidad de comparendos, su tipología y su ubicación?
- ¿Cómo se puede construir un modelo de mapas de intensidad que relacione las infracciones con los accidentes de tránsito utilizando los insumos proporcionados por la maestría?
- ¿De qué manera se pueden visualizar y entregar resultados pertinentes desde la academia y las instituciones públicas de tránsito de Cali?

## 1.2 Objetivos

### 1.2.1 Objetivo General

Desarrollar un modelo predictivo de la accidentalidad vial a partir del análisis espacio-temporal de los comparendos, con el propósito de identificar patrones de riesgo y reducir la ocurrencia de fatalidades en los siniestros de tránsito.

### 1.2.2 Objetivos Específicos

1. Preparar la base de datos con la información de las direcciones de la ocurrencia de las infracciones proporcionada por la Secretaría de Movilidad de Cali.
2. Realizar una exploración espacial y temporal de las infracciones y de los accidentes de tránsito en la ciudad de Cali.
3. Construir un modelo de mapas de intensidad con la relación entre infracciones y los accidentes de tránsito.
4. Visualizar y desplegar los resultados obtenidos después de la ejecución del modelo construido.
5. Evaluar el desempeño del modelo desarrollado.

## 1.3 Marco de Referencia

Para dar un entendimiento más profundo sobre el estado actual del problema de investigación, a continuación, se presenta el marco de referencia, que está desarrollado por los componentes del marco teórico y antecedentes. El marco teórico reúne los conceptos fundamentales y las principales técnicas utilizadas para comprender la problemática de la accidentalidad de tránsito en la ciudad de Cali. Asimismo, proporciona el sustento conceptual necesario para la investigación, enfocándose en las soluciones basadas en modelos predictivos que se construyen a partir del análisis de las infracciones de tránsito registradas. Por su parte, los antecedentes recopilan estudios previos y trabajos relevantes que abordan problemas o técnicas similares.

### 1.3.1 Marco Teórico

De acuerdo con las cifras mencionadas en la descripción del problema y estadísticas actualizadas de la Organización Mundial para la Salud, anualmente en el mundo mueren aproximadamente 1,19 millones de personas, y entre 20 y 50 millones sufren traumatismos no mortales pero que provocan una posible discapacidad [5]. Este marco teórico busca mostrar los estudios que se han realizado en el mundo tratando de encontrar una correlación entre algunas variables. Dependiendo de los años en los que se inició la investigación se puede ir variando entre análisis basados más en la infraestructura y en el diseño de las vías hasta llegar a estudios más recientes que han encontrado en la conducta de las personas un factor determinante en la ocurrencia de los accidentes de tránsito. Es así como describimos los siguientes estudios que lo demuestran:

1. Maycock y Hall propusieron en su momento un modelo de predicción de accidentes de tránsito mediante técnicas de regresión lineal en donde los parámetros de entrada fueron el tráfico (coches, peatones y ciclistas), el ancho de las vías y el diseño de los accesos dando como resultado el total de accidentes particularizados por tipo de accidentes [6].
2. Para 1995, el trabajo realizado por Arndt y Troubeck propone un modelo basado en técnicas de regresión lineal múltiple con variables independientes que tienen que ver con el comportamiento de los conductores en lugar de las características geométricas de la vía; la respuesta de los modelos varía para cada tipo de accidente al incluir variables como la velocidad y la consistencia [7].
3. En 1997, Dia y Rose hicieron un estudio con el desarrollo y evaluación de una red neuronal para la detección de incidentes en las autopistas usando datos de campo en la que encontraron como conclusión que existe una gran viabilidad al utilizar datos reales para la detección de incidentes a través de modelos. El estudio evidenció que la detección puede ser rápida, confiable y que además disminuyó la tasa de falsos positivos al aplicar umbrales de decisión más altos durante los picos del día [8].
4. El modelo en Estados Unidos fue desarrollado en el documento NCHRP de 2007 en el que se incluyeron variables independientes como la intensidad del tráfico, el número de accesos a la vía y el número de carriles, se intentó establecer un modelo basado en la velocidad, pero la regresión resultó inadecuada [9].
5. Naurois y Bourdin realizaron en 2019 un estudio donde se buscó predecir la somnolencia de un conductor a través de redes neuronales debido a las implicaciones que esto representa en la ocurrencia de los accidentes de tránsito. Se encontró información valiosa con la instalación

de cámaras y sensores que medían el cierre de los párpados, la mirada y los movimientos de la cabeza, así como el tipo de conducción. Los resultados arrojaron datos con una precisión de hasta 5 minutos, sin embargo, se requiere utilizar un conjunto más amplio de datos que pueda tomar en cuenta mediciones fisiológicas con muchas más personas, rangos de edad y diferentes horas del día lo que implica tomar en cuenta muchas más variables [10].

Los estudios mencionados anteriormente representan solo una parte del trabajo investigativo realizado en torno a la accidentalidad vial. Estos evidencian una evolución en el enfoque analítico: desde una perspectiva centrada en la infraestructura, hacia un abordaje más integral que incluye factores comportamentales, como la velocidad y fisiológicos, como la somnolencia [11]. Ninguno de estos enfoques debe descartarse; por el contrario, todos los hallazgos deben considerarse para enriquecer el análisis. Esta integración, sin embargo, conlleva la necesidad de incorporar una gran cantidad de variables, lo que daría lugar a modelos complejos, pero potencialmente más robustos y explicativos en relación con los datos modelados.

A continuación, se definen algunos conceptos y técnicas relevantes para el desarrollo del proyecto, abordando y contextualizando diferentes puntos de análisis.

#### 1.3.1.1 Accidentes de tránsito.

De acuerdo con el Código Nacional de Tránsito, un accidente de tránsito es un evento involuntario generado por al menos un vehículo en movimiento, que causa daños a personas y bienes involucrados en el e igualmente afecta la normal circulación de los vehículos que se movilizan por la vía o vías comprendidas en el lugar. De acuerdo con esta definición la mayoría de los accidentes de tránsito son causados por una mala conducción en la que se identifican varios factores de riesgo en los que se encuentran:

- Exceso de velocidad.
- Distracción.
- Fatiga.
- Consumo de alcohol y/o sustancias psicoactivas.
- Infraestructura y/o ambiente.
- Problemas asociados a vehículos.

#### 1.3.1.2 Tipos de accidentes de tránsito.

Dentro de los tipos de accidentes de tránsito existe una clasificación de colisiones que no se basa en la fuerza del impacto, sino en el tipo de objeto o material impactado, los tipos de choques se pueden clasificar en:

- Duros: Como un poste de concreto reforzado.
- Blandos: Como el cuerpo humano.

Además, según la dirección del impacto, la colisión se clasifica en:

- Frontal
- Lateral
- Trasera

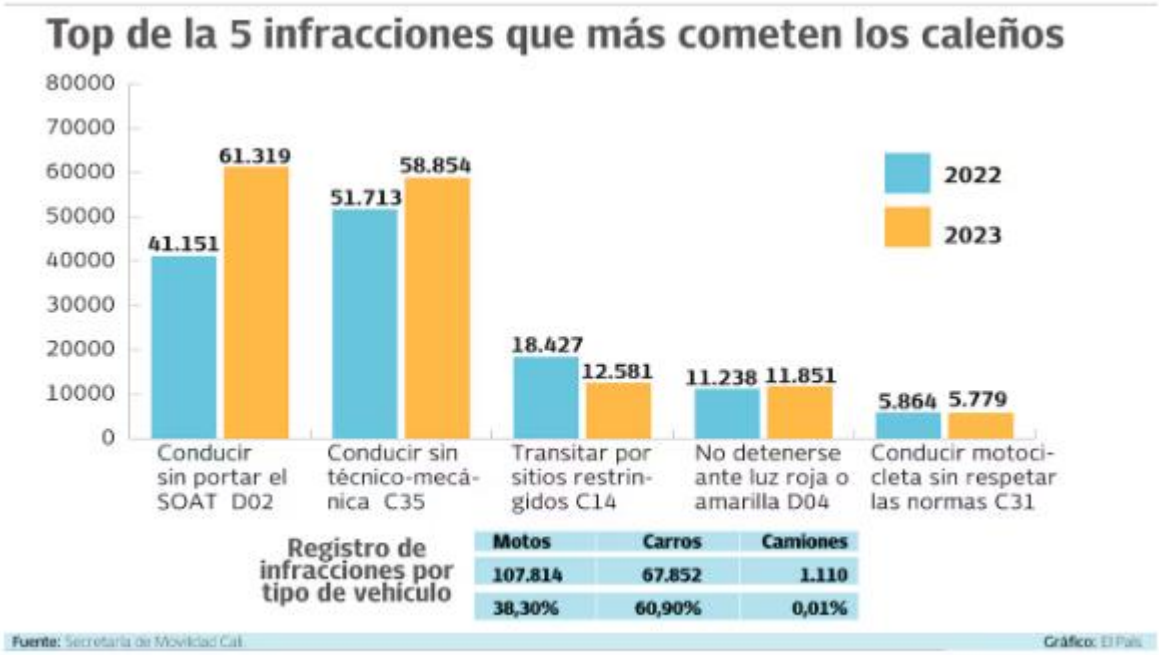
El impacto que puede generar mayor afectación en un accidente de tránsito es el que se genera de manera frontal y contra un objeto duro como un poste de concreto, los choques, respecto al choque con material blando como el cuerpo humano es de resaltar que es el actor vial más vulnerable, seguidos por los ciclistas y motociclistas (especialmente si no llevan equipo de protección).

#### 1.3.1.3 Infracciones de tránsito.

Las infracciones son una pequeña muestra de las conductas que se generan en la vía por cada uno de los actores viales. En este caso, las infracciones las podemos dividir en leves y graves dependiendo de las consecuencias que la omisión o falta de alguna pudiera llegar a provocar. Podemos tomar como ejemplo no respetar la señal de PARE. En este caso, cometer esta infracción genera una sanción económica, sin embargo, si omitir esta señal genera un accidente de tránsito esta es una causal para dictaminar posibles responsabilidades penales, lo que convertiría esta infracción en una omisión grave a las normas de tránsito.

En Cali, las infracciones más recurrentes en este momento se discriminan de la siguiente manera:

***Ilustración 1.*** *Top de infracciones que más cometen los caleños.*




*Top de las 5 infracciones que más cometen los caleños. | Foto: El País*

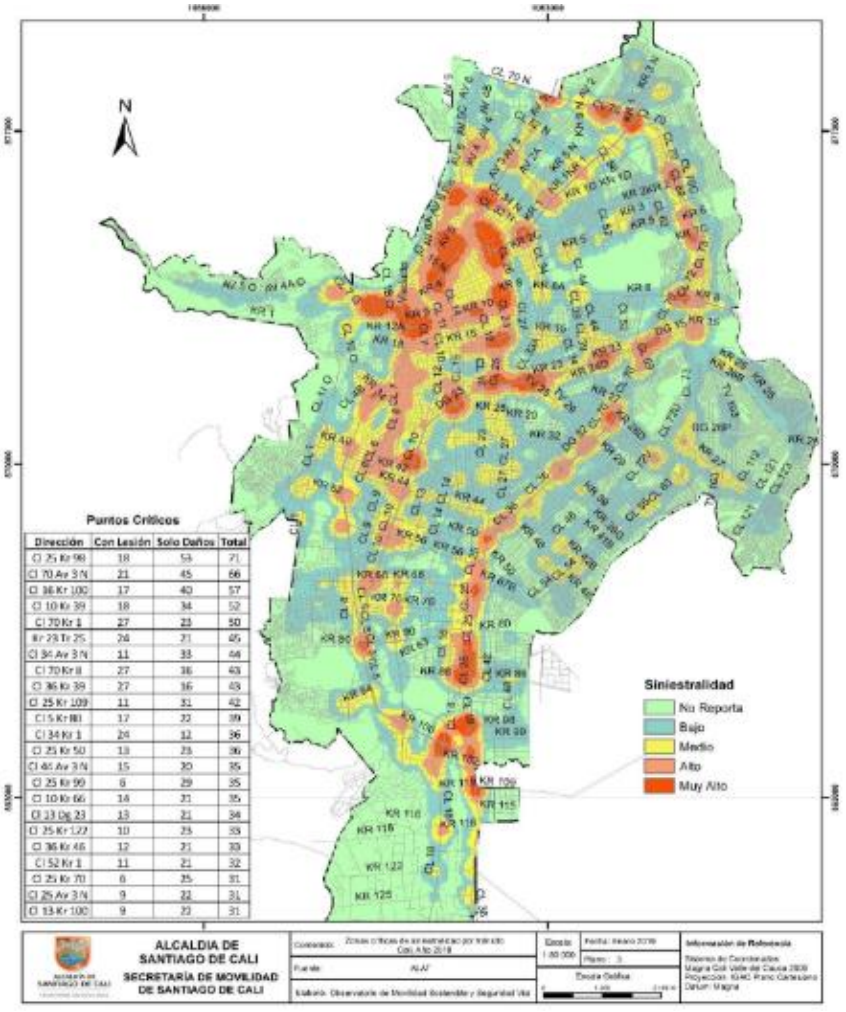
**Fuente:** Secretaría de Movilidad Cali

El anterior gráfico muestra que hay un componente importante en las conductas que generan infracciones, así como una posible relación con los siniestros. De igual forma, se evidencia un aumento en temas de responsabilidad al adquirir un vehículo automotor como tener un seguro obligatorio y realizar la revisión técnico-mecánica.

### 1.3.2 Análisis georreferenciado, geolocalizado o espaciotemporal

La geolocalización de los accidentes de tránsito es un insumo importante a la hora de tomar decisiones que puedan impactar en la reducción de los siniestros en los que puedan resultar personas fallecidas. Es por esto que, Cali cuenta con un observatorio de movilidad y seguridad vial que a su última actualización muestra el mapa de la ciudad y los *hotspot* en donde se deben desarrollar las acciones específicas. A continuación, se presenta el mapa correspondiente, con datos actualizados hasta el año 2018.

***Ilustración 2.*** *Análisis georreferenciado de siniestralidad en Cali.*




**Fuente:** Secretaría de movilidad de Santiago de Cali.

#### 1.3.2.1 Estadística espacial

Los fenómenos de la naturaleza, en su gran mayoría, contienen variables de interés que son observadas a través del espacio, el tiempo o a su vez espaciotemporalmente. Sin embargo, los supuestos de la estadística clásica no son suficientes para el estudio de este tipo de eventos. El análisis y modelado de datos espaciales debe involucrar una estructura de correlación, lo cual da origen a la estadística espaciotemporal, que permite encontrar predicciones de las variables en lugares donde no existen observaciones, determinar cómo se relacionan las variables de interés en determinadas ubicaciones, o extender las teorías de modelos de regresión y hallar patrones de ocurrencia de eventos [12].

Los métodos estadísticos de análisis de datos espaciales pueden variar de acuerdo con la naturaleza del fenómeno, del dominio espacial o del conjunto índice [12]. Surgiendo así los tres (3) enfoques de la estadística espacial conocidos como geoestadística, datos de área y patrones espaciales de puntos. Para el presente trabajo, se abordó este último enfoque, donde se propuso el modelamiento de procesos puntuales espaciales con el fin de identificar la distribución de la mortalidad en relación con las infracciones de tránsito en la ciudad de Cali, Colombia para los años 2021 y 2022.

#### 1.3.2.2 Patrones de Puntos Espaciales o *Spatial Points Pattern (SPP)*

El análisis de patrones puntuales espaciales ("*SPP*" por sus siglas en inglés) es una de las ramas de estudio de la estadística espacial, que, junto a la geoestadística y el estudio de procesos espaciales discretos o análisis de área, se encarga de identificar y modelar la distribución espacial de un conjunto de eventos en un espacio geográfico o dominio [13].

Un proceso puntual aleatorio es entendido como un proceso estocástico cuyas realizaciones consisten en un conjunto numerable de puntos (eventos) distribuidos sobre una región del espacio continuo [14]. Las coordenadas espaciales (longitud y latitud), son denominadas eventos [15] y el principal objetivo de análisis es comprobar si los puntos exhiben algún tipo de patrón:

a. *Patrón Aleatorio*: puntos distribuidos aleatoriamente en el espacio.
b. *Patrón Regular*: la distancia media entre los puntos suele ser constante.
c. *Patrón Agregado:* existen aglomeraciones de los puntos en el espacio.

El análisis estadístico de los SPP, en su primera etapa, buscan determinar si la intensidad de ocurrencia de los eventos es constante en el área de observación, mostrando algún esquema de agregación o inhibición [12]. Se parte del supuesto de que el conjunto de puntos está aleatoriamente ubicado en el espacio, y para su comprobación se realizan dos tipos de medidas:

- Medidas de primer orden que cuantifican la medida del proceso $\lambda(S)$, es decir el número de evento por unidad del área y,
- Estadísticas de segundo orden que describen la correlación espacial de los eventos $\lambda(S_i, S_j)$.

#### 1.3.2.3 Aleatoriedad Espacial Completa (AEC) o *Complete Spatial Randomness (CSR)*

El supuesto de aleatoriedad espacial completa plantea que los eventos se distribuyen de manera uniforme dentro del área analizada y que cada uno ocurre de forma independiente respecto a los demás. Esto significa que la densidad esperada de eventos por unidad de superficie es constante en toda la región, sin que los puntos muestren preferencia por ubicarse en zonas específicas [12]. Este concepto permite evaluar dos condiciones clave: la homogeneidad, todas las subregiones tienen la misma probabilidad de que ocurra un evento y la ausencia de interacción, la aparición de un evento no influye en la probabilidad de ocurrencia de otros en el espacio.

El método de conteo por cuadrantes para Aleatoriedad Espacial Completa se utiliza para identificar si los puntos son completamente aleatorios en el área de interés. Consiste en dividir la ventana de observación en subregiones llamadas cuadrantes de igual área. Con ello, se realiza un conteo de los puntos que caen en cada subregión. Los conteos deberían ser igual o cercanos en cada subregión (homogénea), de lo contrario grandes cambios sugieren tendencias espaciales [13]. Una de las principales desventajas de esta técnica está en que el tamaño de la partición de los cuadrantes puede afectar los resultados.

Para contrastar la hipótesis de homogeneidad espacial, se recomienda emplear la prueba de Chi-Cuadrado ($X^2$). Esta se basa en el total de eventos observados, el área completa de la región de análisis, la intensidad estimada del patrón y el número esperado de eventos por subdivisión espacial. Un valor de *p* significativamente bajo lleva al rechazo de la hipótesis nula, lo que sugiere que la intensidad no es uniforme a lo largo del espacio y que podría haber interacción entre los eventos. En tal caso, se aconseja estimar la intensidad mediante métodos no paramétricos.

Cuando se rechaza la hipótesis de AEC, las alternativas son: *patrón regular* o *patrón agregado*. Para ello, existen pruebas de aleatoriedad gráficas basadas en la distancia del evento más cercano (Función F), distancia al vecino más cercano (Función G), distancia entre eventos (Función H) y la distancia media entre eventos (Función K), la estimación de la densidad del *Kernel* y las estadísticas de escaneo [13].

#### 1.3.2.4 Estimación de la Intensidad

La intensidad es la densidad promedio de puntos en un área, mide la abundancia o frecuencia de los eventos registrados por los puntos. La intensidad puede ser constante (uniforme u homogénea) o puede variar de un lugar a otro (no uniforme o inhomogénea) [16]. Dicho de otra forma, la intensidad está definida como el valor esperado del número de puntos por unidad en un área.

Existen diversas estrategias para estimar la función de intensidad en procesos puntuales. Los métodos no paramétricos, en particular, se fundamentan únicamente en la localización de los eventos en el espacio, sin suponer un modelo subyacente. Entre los enfoques más utilizados se encuentran los conteos por cuadrantes, que dividen el área de estudio en subregiones y calculan la densidad de puntos por unidad de superficie, y la estimación por Kernel, que aplica una función de suavizamiento sobre los puntos para obtener una representación continua de la intensidad espacial.

**a. Estimación de intensidad por modelado *Kernel***

La estimación de la densidad por modelado Kernel permite la representación espacial de los puntos que, a nivel general, se verá reflejada en mapas de calor. Este término hace referencia al área geográfica donde existe alta concentración de los hechos, eventos o fenómenos.

Los estimadores de intensidad de *Kernel* constituyen una alternativa metodológica para la estimación de densidades espaciales, fundamentada en técnicas de interpolación que permiten generar una representación continua del fenómeno, mediante funciones suavizadas aplicadas sobre una malla ráster, en la cual cada celda adquiere un valor específico asociado a la densidad estimada en esa localización. Este procedimiento se inicia a partir de una distribución espacial de puntos con atributos asociados, ya sea medidos directamente o inferidos a partir de promedios agregados en torno al centroide de un área. Posteriormente, cada punto es conceptualizado como el centro de un volumen tridimensional cuya extensión está determinada por un alcance espacial definido. A partir de esta estructura, se calcula la densidad estimada en cada ubicación del espacio, lo que permite su representación en una cartografía continua que refleja gradientes de concentración del fenómeno estudiado [16].

**b. Función K (distancia entre puntos)**

La función K de Ripley, propuesta en 1976, permite clasificar un patrón de puntos como aleatorio, agrupado o regular, permite evaluar la dependencia entre ubicaciones a diferentes distancias. La función es cero, si el patrón es puntos es completamente aleatorio [17].

Para patrones de puntos espaciales agrupados, es probable que cada evento este rodeado de otros eventos. Por tanto, los valores pequeños de la distancia serán relativamente grandes. Para patrones de puntos regulares, es probable que cada evento este rodeado de espacio vacío. Para determinar si los valores de una función K son relativamente grandes o pequeños, podemos comparar la función K del patrón de puntos espaciales observado con la función K de un proceso de Poisson homogéneo (CSR) [17].

**1.3.2.5 Modelado estadístico de datos espaciales**

El aprendizaje estadístico se centra en el uso de modelos estadísticos y computacionales para identificar patrones en los datos y, a partir de ellos realizar predicciones [18]. El modelado estadístico considera no solamente la descripción de los datos, si no que realiza una evaluación inferencial sobre la población muestreada a fin de cuantificar las relaciones entre variables, estimación de parámetros poblacionales y predicción de resultados de observaciones [19].

Los modelos estadísticos y el muestreo se basan en el concepto de probabilidad. Al enfrentarse a datos que provienen de un muestreo aleatorio (espacial) con interés en estimar medias o totales, un enfoque basado en el diseño que asume aleatoriedad en las ubicaciones de la muestra es el enfoque de análisis más directo. Si las observaciones no se muestrearon aleatoriamente, o si es de interés predecir valores en ubicaciones específicas (mapeo), se requiere un enfoque basado en modelos [19]. El rendimiento predictivo de los modelos se evaluará mediante validación cruzada

espacial (CV), que tiene en cuenta el hecho de que los datos geográficos son espaciales [20].

#### 1.3.2.5.1 Modelo de regresión lineal

La regresión lineal fue propuesta por primera vez por Sir Francis Galton (1822–1911). Este término se utilizó en su momento para describir una observación, y era que padres muy altos tenían hijos más bajos y la mayoría de los padres muy bajos tenían hijos más altos que ellos. La tendencia en la estatura era hacia la estatura promedio; esto en su momento se denominó regresión a la media.

La regresión lineal, en su proceso básico, busca modelar la relación entre dos variables ajustando una ecuación lineal a los datos observados, donde una se considera la variable dependiente y una o más variables independientes. La regresión lineal es una de las metodologías de ciencia de datos más antiguas, pero también el método más fácil para crear una función que explique y prediga el valor de la variable objetivo [21]. La regresión lineal es ampliamente utilizada en el aprendizaje supervisado para predecir una respuesta cuantitativa. Los mismos son de gran interés debido a su simplicidad, su posibilidad de implementación y la facilidad de la interpretación.

#### 1.3.2.5.2 Modelo de regresión de Poisson

El modelo de regresión de Poisson fue introducido a partir de la distribución de Poisson planteada por Siméon Denis Poisson en el siglo XIX. La regresión de Poisson es el modelo de regresión de conteo más simple, debido a que las distribuciones de conteo usualmente tienen una distribución de Poisson y esta tiende a ajustar mejor los datos que la regresión lineal. Como resultado, las relaciones predictivas con una variable dependiente pueden examinarse como en la regresión lineal, pero sin los problemas de tener distribuciones normales [22].

Para el análisis del proyecto, el modelo de regresión resulta una opción viable teniendo en cuenta que el número de accidentes de tránsito es una variable de conteo, solo puede tomar valores enteros no negativos, el número de eventos que ocurren en un intervalo de tiempo se ajusta bien a una distribución de Poisson y, por último, permite una interpretación intuitiva de los efectos de las infracciones sobre los accidentes.

#### 1.3.2.5.3 Modelo de regresión Gaussiano

El modelo de regresión gaussiano utiliza una distribución normal para modelar el parámetro de despeje. Implica el uso del kernel, como el kernel exponencial al cuadrado (*squared exponential kernel*), junto con parámetros que definen el modelo. Algunas de las razones por las que este modelo puede utilizarse en contextos como el desarrollado en este proyecto tienen que ver con su simplicidad, familiaridad, capacidad de manejo en muestras pequeñas y supuestos que, si se cumplen, podrían generar un mejor ajuste. Aunque presenta limitaciones, como la carga computacional, en un contexto de análisis exploratorio inicial puede facilitar la comunicación clara de los resultados [23], [24].

#### 1.3.2.5.4 Modelo Random Forest

El modelo Random Forest es un clasificador de Machine Learning supervisado que comprende la estructura de un árbol. Está basada en los árboles de decisión que son una estructura simple que buscan dividir los datos en subconjuntos de datos con unas características de entrada para predecir un valor objetivo [25]. Esta idea viene también de otras que se intentaron en los años 90 como el Bagging pero no fue sino hasta el año 2001 que fue formalizado y popularizado.

Dentro de las fortalezas que puede tener Random Forest para el proyecto está su alta capacidad de precisión y rendimiento predictivo, al poder combinar múltiples árboles de decisión. Otra de sus mayores ventajas es evitar el sobreajuste gracias al uso del bagging y la selección aleatoria de características. Finalmente, destaca su capacidad para identificar las variables más influyentes en el análisis, lo que resulta útil para la interpretación de modelos complejos en contextos como el análisis de accidentes de tránsito.

#### 1.3.2.5.5 Modelo Gradient Boosting

El modelo Gradient Boosting es un algoritmo de aprendizaje automático supervisado que se puede utilizar tanto para problemas de clasificación como de regresión. Su principal fortaleza es generar un modelo final a partir de muchos modelos individuales. En Gradient Boosting, tras construir los aprendices débiles, se comparan las predicciones con los valores reales, y la diferencia representa la tasa de error del modelo [26].

Dado que es un modelo muy potente que construye modelos de forma secuencial corrigiendo los errores del anterior, es una buena opción para encontrar relaciones complejas y mejorar el rendimiento predictivo. En este caso, la variable dependiente es el número de accidentes de tránsito, y las variables independientes corresponden a infracciones de tránsito y sus características que las hacen relevantes [27].

#### 1.3.2.5.6 Modelo SVM

Una máquina de vectores de soporte (SVM) es un método de aprendizaje automático que organiza individuos en un hiperplano para calcular una función que separa mejor los datos según los niveles de una variable categórica. Son algoritmos utilizados para clasificación, regresión y detección de valores atípicos. En su esencia, buscan encontrar el hiperplano óptimo que mejor separe o se ajuste a los puntos de datos en un espacio multidimensional [28].

Utilizar SVM puede representar desafíos importantes a la hora de modelar el número de accidentes de tránsito, dado que está diseñada principalmente para variables continuas y no discretas. Sin embargo, podría ser muy útil si se reformula el problema como uno de clasificación, por ejemplo, clasificando los accidentes según su gravedad.

#### 1.3.2.5.7 Red neuronal artificial

Las redes neuronales artificiales son un modelo computacional inspirado en la estructura y función del cerebro humano. Las neuronas están organizadas en capas y cada neurona tiene un peso asociado. Son un método eficaz para resolver problemas con relaciones no lineales sin necesidad de conocer la forma exacta de la función [29].

Las redes neuronales constituyen uno de los pilares del aprendizaje automático y el aprendizaje profundo. El proceso de entrenamiento suele seguir la siguiente secuencia:

- Propagación hacia adelante: los datos de entrada alimentan la capa de entrada, se calculan los pesos y los sesgos, y se produce una salida.
- Cálculo de la función de pérdida, comparando la predicción con el valor real.
- Propagación hacia atrás (backpropagation) para ajustar pesos y sesgos.
- Actualización de los pesos y repetición del ciclo hasta que la pérdida se estabiliza o el modelo converge [30].

### 1.3.2.6 Mapeo con modelos de regresión no espacial y aprendizaje automático

Los modelos de regresión u otros modelos de aprendizaje automático (*machine learning*) se pueden aplicar a datos espaciales y espaciotemporales de la misma manera que se aplican para predecir nuevas observaciones en problemas no espaciales [19].

El análisis de datos espaciales, mediante técnicas de aprendizaje automático, requiere enfoques que incorporen las dependencias inherentes al espacio geográfico, como la autocorrelación espacial, en la que las observaciones próximas tienden a presentar similitudes. Si estas relaciones espaciales no se tienen en cuenta durante el entrenamiento y validación de los modelos, se corre el riesgo de incurrir en fuga de información. Un ejemplo común es la partición aleatoria de los datos espaciales en conjuntos de entrenamiento y prueba, lo cual puede derivar en que los datos de prueba se encuentren espacialmente próximos a los de entrenamiento. Esta proximidad compromete el supuesto de independencia entre ambos conjuntos, lo que puede conducir a una sobrestimación del rendimiento del modelo y una capacidad limitada de generalización. En respuesta a estos desafíos, se han desarrollado diversos métodos orientados a integrar explícitamente las estructuras y relaciones espaciales en la construcción y validación de modelos predictivos [18].

### 1.3.2.7 Modelo predictivo

Es un modelo matemático o estadístico que busca predecir el valor de una variable de interés basándose en datos históricos y patrones observados mediante el uso de Inteligencia Artificial (Machine Learning, Ciencia de Datos, etc.). Normalmente, el evento a predecir es futuro, sin embargo, el modelado predictivo se puede aplicar a eventos desconocidos sin importar su temporalidad [31].

#### 1.3.2.8 Validación Cruzada Espacial

La validación cruzada espacial (*CV espacial*) es una técnica que considera la naturaleza geográfica y espacialmente auto correlacionada de los datos. A diferencia de la validación cruzada tradicional, que divide aleatoriamente los datos en subconjuntos de entrenamiento y prueba, la CV espacial reconoce que las observaciones espaciales cercanas pueden no ser independientes, lo que constituye una violación crítica de los supuestos de independencia estadística. En contextos espaciales, una división aleatoria puede generar conjuntos donde puntos de entrenamiento están contiguos a los puntos de prueba, afectando negativamente la capacidad del procedimiento para detectar sobreajuste, ya que un modelo podría estar ajustándose al ruido local y no generalizando adecuadamente. La validación cruzada espacial mitiga este sesgo, ya que permite estimar de manera más realista el rendimiento del modelo frente a nuevos datos distribuidos espacialmente, y mejora la capacidad para evaluar su generalización, robustez y estabilidad en entornos con autocorrelación espacial [32].

#### 1.3.2.9 Métricas de evaluación de los modelos

$\boldsymbol{RMSE}$ como el error cuadrático promedio de la predicción. El error cuadrático medio (RMSE) es la raíz cuadrada de la media del cuadrado de todo el error. Su uso es muy común y se considera una excelente métrica de error de propósito general para predicciones numéricas [33]. Su formulación es la siguiente:

***Ecuación 1.*** *Raíz del Error Cuadrático Medio*

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Donde yi son las observaciones, ŷi los valores predichos de una variable y n el número de observaciones disponibles para el análisis. El RMSE es una buena medida de precisión, pero solo para comparar errores de pronóstico de diferentes modelos o configuraciones de modelos para una variable en particular, y no entre variables, ya que depende de la escala. [34]

$\boldsymbol{R^2}$ como el porcentaje que explica el modelo de la validación. El R2 es una medida estadística que indica la proximidad de los datos a la línea de regresión ajustada. También se conoce como coeficiente de determinación o coeficiente de determinación múltiple para la regresión múltiple. La definición de R2 es bastante sencilla: es el porcentaje de la variación de la variable de respuesta que explica un modelo lineal [35].

***Ecuación 2.*** *Coeficiente de Determinación – $R^2$ (R cuadrado)*

$$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

El $R^2$ siempre está entre 0 y 100%. Un 0 % indica que el modelo no explica ninguna de las variaciones de los datos de respuesta en torno a su media, por su parte un 100 % indica que el modelo explica toda la variabilidad de los datos de respuesta en torno a su media. En general, cuanto mayor sea el R-cuadrado, mejor se ajusta el modelo a los datos. [35]

$\boldsymbol{MAE}$ (Mean Absolute Error) como el error absoluto medio. El error absoluto medio es una métrica popular porque, al igual que con el error cuadrático medio (RMSE), las unidades del valor de error coinciden con las unidades del valor objetivo previsto. A diferencia del RMSE, los cambios en el MAE son lineales y, por lo tanto, intuitivos. El MSE y el RMSE penalizan más los errores mayores, inflando o aumentando el valor del error medio debido al cuadrado del valor del error. En el MAE, los diferentes errores no se ponderan más o menos, sino que las puntuaciones aumentan linealmente con el aumento de errores. La puntuación del MAE se mide como el promedio de los valores de error absolutos. El absoluto es una función matemática que hace que un número sea positivo. Por lo tanto, la diferencia entre un valor esperado y un valor previsto puede ser positiva o negativa y necesariamente será positiva al calcular el MAE [34]. El valor del MAE se puede calcular de la siguiente manera:

***Ecuación 3.*** *Error Absoluto Medio*

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left|Y_i - \widehat{Y_i}\right|$$

### 1.3.3 Antecedentes

A nivel global hay una variedad de estudios que buscan identificar patrones en accidentes de tráfico que permitan mitigar su ocurrencia, el trabajo realizado por los investigadores Ben Laoula, Elfhaim, El Midaoui, Youssfi y Bouattane realizado en 2023 buscaba identificar áreas de riesgo e infracciones comunes en la ciudad de Maryland con datos del año 2022 a través del test de Kolmogorov-Smirnov en las curvas de distribución de las infracciones de tránsito [36], se diferencia del presente trabajo en que no identifica relaciones directas o indirectas entre las infracciones y los accidentes, el estudio aporta una aproximación metodológica a cómo abordar problemas con variables espaciales.

En el estudio realizado por Ze-Hao Jiang, Xiao-Guang Yang, Tuo Sun, Tao Wang y Zheng Yang en el 2021 buscaba identificar patrones entre infracciones de tránsito y accidentes de tránsito en intersecciones señalizadas en dos ciudades chinas utilizando modelos de regresión lineal múltiple [37]. El estudio es cercano a lo que el proyecto pretende, pero difiere en el alcance, ya que solo estudia intersecciones con adecuadas señales viales en la ubicación geográfica, además teniendo en cuenta las diferencias socioculturales entre China y Colombia y las herramientas usadas, ya que

el estudio fue mayormente realizado usando STATA 12.

Algunas investigaciones se han realizado en la región con temáticas similares, el trabajo realizado por González y Prada en 2016 buscaba determinar la relación entre la implementación de cámaras de fotodetección y los accidentes de tránsito [38], el estudio se diferencia del presente en que sólo se centra en los impactos de las cámaras de fotodetección y únicamente en los puntos donde se encuentran ubicadas dichas cámaras a través de una metodología de análisis bayesiano empírico, pero brinda un marco de referencia contextual de los accidentes de tránsito en la ciudad.

El trabajo de Gómez Angulo, Osorio Henao y Plazas Albornoz desarrollo un ejercicio estadístico de identificación de causas de accidentes de tránsito en la doble calzada Buga-Tuluá entre 2017 y 2019 [39], se diferencia del presente trabajo en que el ejercicio que se realiza es exploratorio usando estadística descriptiva y realizado en una zona no urbana, y el aporte que brinda es la caracterización de accidentes en una vía "rápida" relativamente cercana a la zona de estudio.

Cardona J.C [40] desarrolló en el 2023 un modelo predictivo de zonas de riesgo espacio temporal de accidentes en la ciudad de Manizales utilizando técnicas de aprendizaje automático. Las técnicas como Facebook Prophet, Light Gradient Boosting y redes neuronales recurrentes (Long Short-Term Memory, LSTM), se utilizaron con base en su capacidad demostrada en el estado del arte en esta problemática y su capacidad de procesar datos y patrones a fin de conocer el comportamiento de los accidentes de tránsito en esa ciudad. De igual forma, se utilizaron métricas como MAPE y RMSE, para evaluar el rendimiento y la precisión de las predicciones para realizar una comparación entre las técnicas y modelos.

Por su parte, Bao, J., Liu, P., y Ukkusuri, S. V en 2019 desarrollaron una arquitectura espacio temporal de aprendizaje profundo para la predicción de riesgo de accidentes a corto plazo en la ciudad de Nueva York [41]. Como resultado se propuso una red de memoria a largo plazo convolucional espaciotemporal (STCL-Net) que incluía modelos de análisis diario y semanal dando solución a problemas de análisis, que generalmente se realizaban en periodos anuales. Para este análisis se utilizaron diferentes fuentes de datos como: datos de accidentes, datos de viajes en taxi, atributos de la red de carreteras, características de uso del suelo, datos de población y datos meteorológicos. La arquitectura de aprendizaje profundo propuesta está fusionada por múltiples capas CNN, capas LSTM y capas ConvLSTM, y podría integrar variables explicativas recopiladas a partir de datos de diversas fuentes.

Dávila García realizó en el 2020 un diagnóstico de los puntos críticos de siniestralidad vial en la ciudad de Santiago de Cali para los periodos comprendidos entre el 2016 y 2018 [42]. Utilizó como fuente, datos otorgados por la Secretaría de Movilidad del municipio y creó un geo codificador con capas espaciales que ubicaban cada uno de los eventos de tránsito. Ese análisis espacial se realizó mediante la herramienta "Densidad de *Kernel*", que otorga los puntos calientes (*hotspot*) los cuales emitieron patrones espaciales para cada año y la respetiva ubicación de las localizaciones más críticas.

# 2. PREPARACIÓN DE LA BASE DE DATOS

## 2.1 Comprensión del Negocio

La fase de comprensión del negocio en la metodología CRISP-DM tiene como objetivo fundamental establecer una comprensión clara del problema que se busca resolver desde una perspectiva operativa y contextual. En este proyecto, centrado en la identificación y análisis de la accidentalidad de tránsito en la ciudad de Cali, dicha comprensión parte del reconocimiento del impacto social y humano que representan los siniestros viales, así como de la necesidad institucional de diseñar estrategias preventivas más efectivas a partir de datos.

A nivel local, la ciudad de Cali ha registrado una alta incidencia de accidentes de tránsito, lo que ha motivado la búsqueda de soluciones basadas en evidencia. El problema no se limita únicamente a las cifras de siniestralidad, sino que involucra una compleja interacción de factores asociados al comportamiento de los actores viales, las condiciones de la infraestructura y el cumplimiento de las normas de tránsito. En este contexto, las infracciones representan un insumo crítico, ya que reflejan patrones de conducta que pueden estar directamente asociados a la ocurrencia de accidentes.

El objetivo del proyecto es construir modelos predictivos que permitan identificar zonas de mayor riesgo de siniestralidad a partir del análisis espacial y temporal de infracciones y accidentes, aportando herramientas útiles para la toma de decisiones por parte de las autoridades de movilidad. Este enfoque implica no solo una revisión técnica y estadística, sino también una comprensión sistémica del problema: cómo interactúan las variables espaciales, temporales y comportamentales, y qué acciones pueden derivarse de los hallazgos para mejorar la seguridad vial.

Además, se reconoce que la calidad de los datos, la estructura institucional y la capacidad tecnológica del municipio son elementos clave para el éxito del proyecto. Así, en esta fase se identificaron restricciones operativas, se definieron los actores involucrados y delimitaron los objetivos del modelo dentro de los márgenes reales de acción, enfatizando en la utilidad práctica del análisis para la gestión de la movilidad urbana.

## 2.2 Comprensión de los datos

La fase de comprensión de los datos en CRISP-DM tiene como objetivo explorar, familiarizarse y evaluar la calidad de los datos disponibles para determinar su utilidad en el desarrollo del modelo. En este proyecto, para el desarrollo de la investigación, se recibieron dos bases de datos provenientes de la Secretaría de Movilidad de la ciudad de Cali, una primera base de datos en Excel que contiene los datos de infracciones y comparendos entre los años 2021 y 2022, y otra con los accidentes de tránsito registrados por la entidad entre los años 2018 y 2022.

La primera base de datos, “Siniestros 2018 al 2022”, cuenta con un total de 55.092 registros y 58 variables

categóricas. Dentro de sus variables, se observaron características de la accidentalidad en la zona metropolitana de la ciudad de Cali, la dirección donde hubo el accidente, su ubicación geográfica en longitud y latitud, la hora, la fecha, la edad, comuna, tipo de lesión, entre otros.

Por otra parte, se tuvo una segunda base de datos con la información relacionada con las infracciones de tránsito interpuesta por los agentes de tránsito y cámaras de fotomultas registradas entre el 2021 y 2022. Se contó con un total de 510.000 registros y 26 variables categóricas. En esta misma base de datos, existía información de la ubicación geoespacial de las 38 fotomultas que funcionaban en los periodos 2021 y 2022 en la ciudad de Santiago de Cali.

El lenguaje de programación usado es *R* en el ambiente *RStudio.* Sin embargo, con el apoyo de algunas herramientas complementarias del lenguaje de programación Python, se logró realizar el análisis de datos correspondiente. Para el manejo básico de la información en R se usaron las librerías *dplyr* (sintaxis y eficiencia de código) y r*eadxl* (lectura de archivos de Excel), mientras que para la manipulación de datos espaciales se usaron las librerías *sp* (SpatialPolygons) y *sf* (SimpleFeatures).

### 2.3 Preparación de los datos

La fase de preparación de los datos en CRISP-DM implica transformar, limpiar y estructurar los datos disponibles para que estén listos para el análisis y modelado. En este proyecto, se llevó a cabo un proceso minucioso de depuración y estandarización de dos bases de datos clave: una sobre infracciones de tránsito y otra sobre accidentes de tránsito, ambas suministradas por la Secretaría de Movilidad de Cali.

Para hacer los registros de las dos bases de datos comparables, lo primero que se decidió fue reducir el registro temporal de la base de datos de accidentes de tránsito al mismo registro temporal de la base de datos de infracciones, es decir, las dos bases deben quedar entre el primer día del año 2021 y el último día del año 2022. Adicionalmente, tras definir claramente qué variables podrían ser útiles para el estudio, se eliminaron aquellas que no resultaban útiles para el propósito quedando depurada cada una de las bases y con las siguientes variables para el desarrollo del proyecto:

***Tabla 1.*** *Variables inicialmente seleccionadas*

| Variables de la BD de accidentes | | | Variables de la BD de infracciones | | |
|---|---|---|---|---|---|
| **Temporales** | **Espaciales** | **Categóricas** | **Temporales** | **Espaciales** | **Categóricas** |
| FECHA_EVEN | COOR_X | TIPO_VEHIC | FECHACOMP | lat | CODIGOINFR |
| AÑO | COOR_Y | TIPO_EVENT | MES | long | DESCRINFRACCION |
| MES_EVENTO | DIRECCION_ | | ANO | DIRECCION...21 | CLASE |
| DIA_EVENTO | COD_COMUNA | | DIASEM | COMUNA | SERVICIO |
| | NOMBRE | | HORACOMP | | |
| | | | HORA | | |
| **Total Variables: 11** | | | **Total Variables: 14** | | |

**Fuente:** Elaboración propia

Tras fijar el rango temporal de las observaciones y seleccionar las variables del estudio, las bases de datos resultantes quedaron con las siguientes dimensiones, para la base de datos de infracciones con 14 variables

y 510.694 registros, y la base de datos de accidentes con 11 variables y 20.050 registros. Se procedió a revisar los datos faltantes en las variables de coordenadas (COOR_X y COOR_Y para la base de datos de accidentes; y lat y long para la base de datos de infracciones), encontrando que solo tres registros de la base de datos de infracciones no los tenían. Dado lo crítico de esta información, fue necesario remover esos registros. Luego, se verificó los registros faltantes en las otras variables y se obtuvo la siguiente información:

***Tabla 2.*** *Datos faltantes en las bases de datos.*

| Variables de la BD de accidentes | | Variables de la BD de infracciones | |
|---|---|---|---|
| **Variable** | **Datos Faltantes** | **Variable** | **Datos Faltantes** |
| **Todas las variables** | 0 | COD_COMUNA | 0 |
| | | NOMBRE | 9474 |
| | | TIPO_VEHIC | 430 |
| | | Todas las demás variables | 0 |

**Fuente:** Elaboración propia

Estos datos sirvieron de insumo para evaluar cómo avanzar en la preparación de los datos tras la exploración inicial, ya que antes de tomar una decisión de cómo tratar los datos faltantes se requirió conocer el contexto de dichas variables. Se procedió entonces a evaluar la consistencia de las variables categóricas.

Uno de los principales retos en esta etapa fue abordar las inconsistencias semánticas y estructurales en las variables categóricas. Por ejemplo, la variable “CLASE” en la base de infracciones presentaba 41 categorías distintas, muchas de ellas redundantes o poco representativas. Para facilitar el análisis, se agruparon los vehículos en nueve grandes categorías, tales como automóviles, motocicletas, vehículos de carga, transporte masivo, entre otros, conservando la granularidad necesaria para un análisis posterior más detallado.

En cuanto a la variable “SERVICIO”, se identificó una categoría nula que fue imputada utilizando la moda, debido a la alta prevalencia del servicio “particular”. Asimismo, se realizó una depuración importante sobre las variables “CODIGOINFR” y “DESCRINFRACCION”, excluyendo aquellas infracciones de carácter documental —como la ausencia del SOAT o la revisión técnico-mecánica— para concentrarse en las conductas de riesgo que tienen una relación directa con los accidentes.

La georreferenciación también fue un aspecto central en esta etapa. Se abordaron inconsistencias en la variable “COMUNA”, la cual contenía valores inválidos o nulos. Se realizó una corrección utilizando archivos shapefile oficiales de las divisiones territoriales de Cali, lo que permitió reasignar las comunas a partir de las coordenadas geográficas de cada registro. Los registros ubicados fuera del perímetro urbano fueron eliminados por no aportar valor al análisis.

En la base de datos de accidentes se aplicaron correcciones similares, aunque esta mostró una mayor consistencia general. Variables temporales como “MES_EVENTO” y “DIA_EVENTO” fueron corregidas a partir de la información contenida en la fecha del evento, y también se depuraron errores en las asignaciones geográficas. Se mantuvo sin modificación la variable “TIPO_VEHIC”, pese a algunos datos

faltantes, por no afectar significativamente los análisis posteriores.

Finalmente, ambas bases de datos fueron transformadas al formato SpatialPolygons, permitiendo su integración con herramientas de análisis espacial y visualización cartográfica. El resultado de este proceso fue una base de infracciones con 139.417 registros y 17 variables, y una base de accidentes con 19.756 registros y 12 variables, listas para ser utilizadas en las siguientes fases del proyecto, como el análisis exploratorio, la modelación y la validación.

A continuación, se describen los hallazgos más relevantes.

### 2.3.1 Consistencia de la base de datos de infracciones.

### 2.3.2 Variable "CLASE"

La variable "CLASE" categoriza las infracciones por el tipo de vehículo que cometió la infracción. El principal reto es que existen demasiadas categorías, la revisión de la variable arrojo 41 categorías distintas. Se consideraron diversas opciones y se decidió que, si bien puede mantenerse para cuando se requieran descripciones detalladas, por propósitos prácticos se creó una variable adicional para tener categorías con un agrupadas para análisis más sencillos. Al considerar las categorías de la variable "CLASE", se decidió agrupar los vehículos en 9 categorías, que se describen a continuación:

***Tabla 3.*** *Categoría de vehículos.*

| Categorías de Vehículos | | | |
|---|---|---|---|
| **Categoría "Tipo de Vehículo"** | **Categoría "Clase"** | **Categoría "Tipo de Vehículo"** | **Categoría "Clase"** |
| **Automóvil** | AUTOMOVIL, AUTOMOVIL TURISMO, CAMIONETA, CAMPERO, MINI VAN, MINI MPV | Bicicletas | BICICLETA, CICLOMOTOR |
| **Buses y busetas** | BUS, BUSETA, MICROBUS, MINIBUS | Maquinaria | ABONADORA, APILADOR, APLANADORA, COMPACTADORA, CORTADORA, GRUA, MAQUINARIA AGRICOLA, MAQUINARIA INDUSTRIAL, MINI EXCAVADORA, MONTACARGAS, RETROEXCAVADORA, VOLQUETA |
| **Motocicletas** | MOTOCICLETA | Otros | AMBULANCIA, DESCONOCIDA, MOTOCARRO, TRACCION ANIMAL |
| **Transporte Masivo** | BUS ALIMENTADOR, BUS ARTICULADO, BUS PADRON | Vehículo de carga | CAMION, REMOLQUE, SEMIREMOLQUE, TRACTO/CAMION |
| **Vehículo ligero** | CUADRICICLO, CUATRIMOTO, MOTOCICLO, MOTOTRICICLO, TRICIMOTO | | |

**Fuente:** Elaboración propia

### 2.3.3 Variable “SERVICIO”

La variable servicio categoriza quienes son usuarios del vehículo; si el transporte es público, privado, o de alguna categoría especial. Al revisar la variable servicio se encontraron 5 categorías, que se presentan a continuación:

***Tabla 4.*** *Tipo de servicio.*

| Tipo de Servicio | | | | |
|---|---|---|---|---|
| **Null (inconsistencia)** | **OFICIAL** | **OTRO** | **PARTICULAR** | **PÚBLICO** |
| **1757** | 851 | 27 | 483.451 | 24.608 |

**Fuente:** Elaboración propia

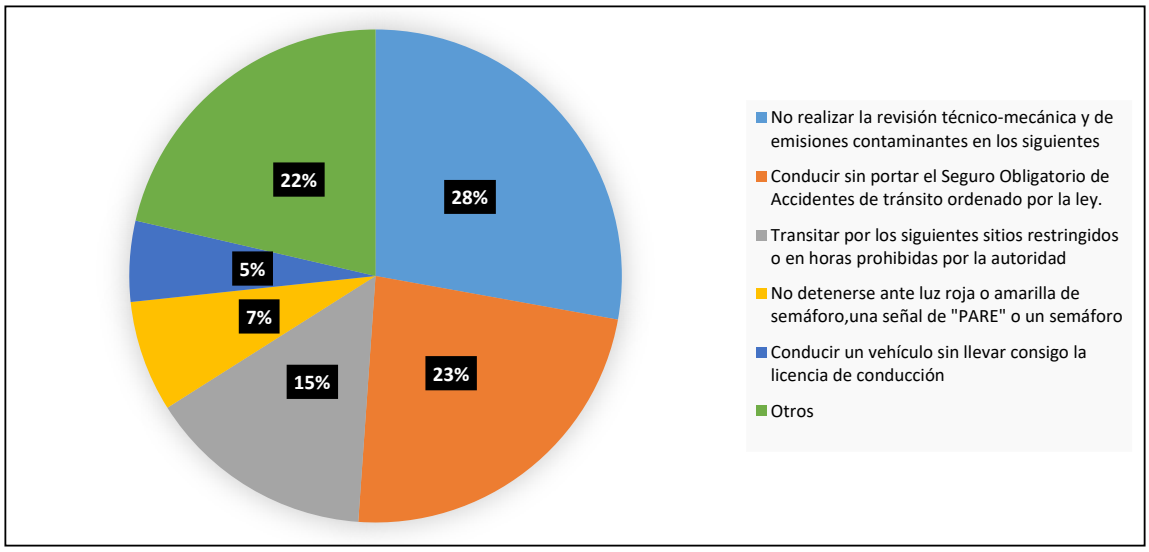
La categoría “null” no debería existir, así que, evaluando la naturaleza de la variable, se decide imputar por la moda, dada la predominancia de la categoría “PARTICULAR”.

### 2.3.4 Variables "CODIGOINFR" y "DESCRINFRACCION"

Las dos variables están asociadas a la conducta que generó la infracción, siendo "codigoinfr” el código tipificado en el Código Nacional de Tránsito y el segundo, la descripción de la conducta; hay 96 categorías diferentes. A diferencia de lo realizado con la clase de vehículos, se optó por mantener las categorías, ya que hay algo más importante que detallar en esta categoría y son las conductas que pueden caracterizarse como “conductas riesgosas”, las cuales pueden derivar en un accidente de tránsito, en tanto que las infracciones por “otras infracciones” como documentación en regla, infracciones ambientales, etc., no son el propósito de este estudio.

De acuerdo con el análisis exploratorio de los datos, la mayoría de las infracciones impuestas por las autoridades están relacionadas con aspectos documentales requeridos para la circulación en la ciudad, como la falta de licencia de conducción o de SOAT. No obstante, el presente proyecto se enfoca en analizar las infracciones derivadas de conductas de los conductores y su posible relación con los accidentes de tránsito. Por esta razón, las infracciones de tipo documental han sido excluidas del análisis inicial. A continuación, se presenta una gráfica que resume los resultados obtenidos.

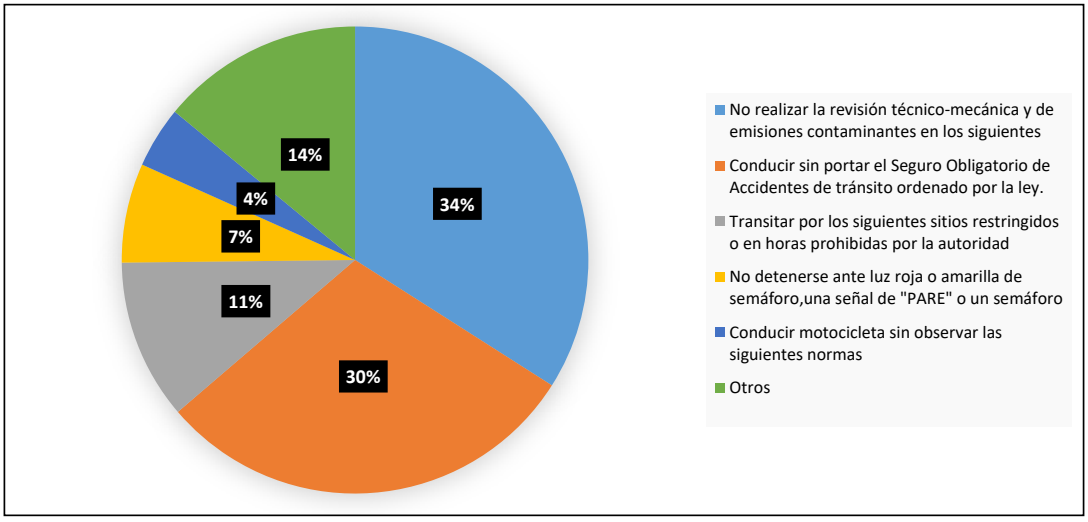
***Ilustración 3.*** *Tipo de infracciones en 2021*




**Fuente:** Elaboración propia

Para el año 2021 un poco más del 50% estuvieron asociadas a temas documentales como no tener vigente la revisión técnico-mecánica, ni el SOAT, para el año 2022 la proporción de estas infracciones llegó a casi un 60% del total como se muestra a continuación.

***Ilustración 4.*** *Tipo de infracciones en 2022*

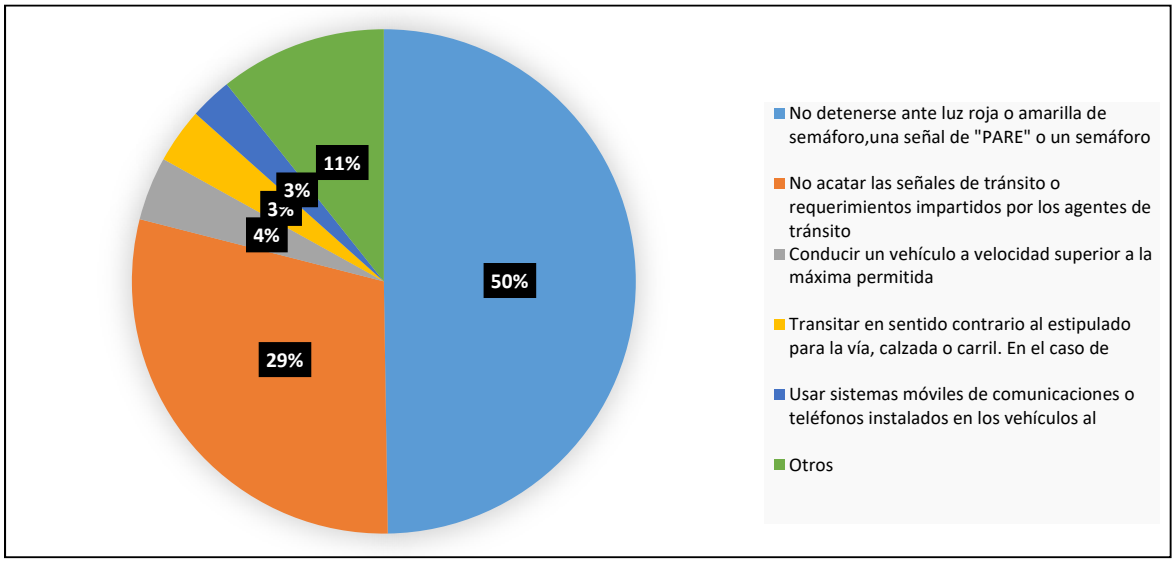



**Fuente:** Elaboración propia

Tras inspeccionar las distintas categorías se decidió eliminar los registros que correspondían a 61 categorías que no encajan en la tipificación de “conducta riesgosa”, permaneciendo así 35 categorías diferentes.

Al final las infracciones más críticas son las que se muestran a continuación, para el año 2021, tres infracciones son los 83% del total asociadas a no detenerse ante la luz roja, no acatar las señas de tránsito y conducir un vehículo con exceso de velocidad.

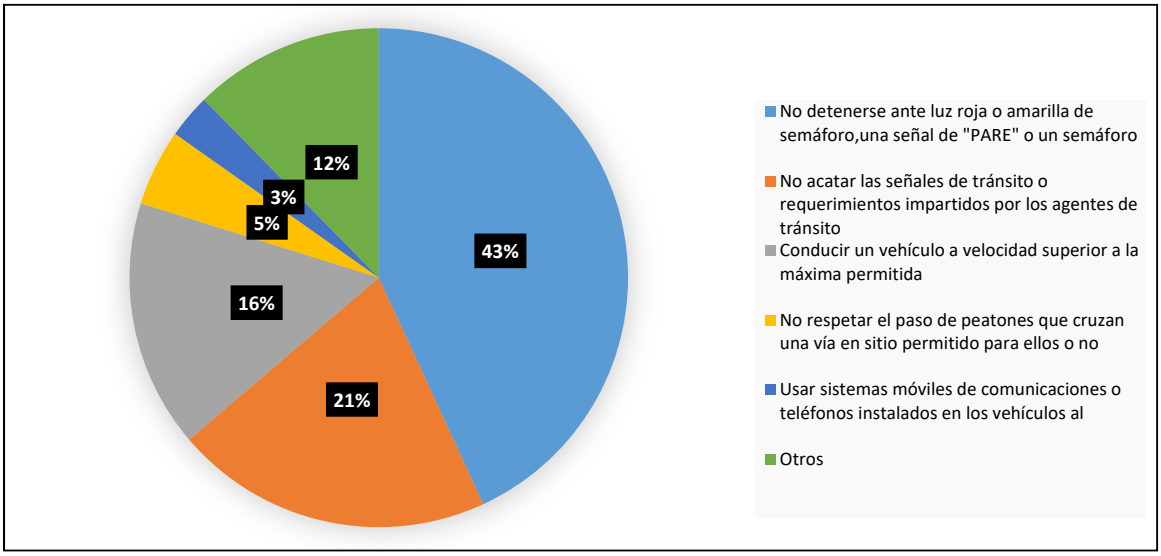
***Ilustración 5.*** *Tipo de infracciones en 2021 después de depuración.*




**Fuente:** Elaboración propia

Para el año 2022 la descripción de las infracciones es la misma, sin embargo, la proporción cambia significativamente mostrando un aumento en la omisión de la señal de pare y no acatar las señales de tránsito, sin embargo, se muestra una disminución importante en la infracción por excesos de velocidad.

***Ilustración 6.*** *Tipo de infracciones en 2022 después de depuración.*




**Fuente:** Elaboración propia

### 2.3.5 Variable " COMUNA"

La variable “COMUNA” corresponde a una subdivisión administrativa y geográfica de la ciudad de Cali que existió para la época en que los registros fueron realizados, corresponde a la zona urbana de la ciudad, y es complementaria a los corregimientos, que fue la subdivisión administrativa y geográfica de la zona rural, Cali contaba con 22 comunas y 15 corregimientos, los registros de la base de datos presentaban las siguientes inconsistencias:

***Tabla 5.*** *Categorías con inconsistencias en la variable “COMUNA”.*

| Categorías con inconsistencias en la variable “COMUNA” | | |
|---|---|---|
| **CALI** | **Null (inconsistencia)** | **ZONA DESCONOCIDA** |
| **10.253** | 113.242 | 10.446 |
| **Registros correctos: 376.753** | | |

**Fuente:** Elaboración propia

Se encontró una gran cantidad de registros con problemas. Las categorías identificadas en la *tabla 5* no deberían existir, así que se procedió, partiendo de las variables de coordenadas, reasignar las divisiones administrativas, teniendo en cuenta la zona rural usando shapefiles obtenibles desde la página web de datos abiertos del municipio de Cali. Con esto, se procedió a unir los shapefiles de las comunas (zona urbana) y corregimientos (zona rural) a través del software R, y se realizó la reasignación de las divisiones administrativas encontrando que solo 1582 datos permanecieron sin asignación. Al revisarlos se encuentra que son datos de zonas límites con los municipios aledaños, especialmente Yumbo y Palmira, por practicidad se decidió eliminar estos registros.

Finalmente, al revisar los datos se observa que la zona rural exceptuando Pance (855 registros) tienen muy pocos registros, por lo que se deciden eliminarlos. Pance se mantiene ya que las vías Panamericana y Cañasgordas tienen un importante número de registros útiles para el estudio.

### 2.3.6 Validación final de la base de datos de Infracciones

Tras finalizar la revisión se encuentra que ya no hay datos faltantes, y se cuenta con una base de 139.417 registros y 17 variables, el aumento de variables se debe a la creación de la variable tipo vehículo, y a la conversión de la base de datos al formato de SpatialPolygons, lo que permitirá graficarlo con otras herramientas en análisis posteriores.

### 2.3.7 Consistencia de la base de datos de accidentes.

Esta base de datos resultó mucho más consistente que la base de datos de infracciones, encontrándose pocos errores, resultando así en correcciones puntuales de los casos hallados.

### 2.3.8 Variable " MES_EVENTO"

La variable “MES_EVENTO” registra el mes en que ocurrió el accidente. Al revisar la consistencia entre las variables temporales de la base de datos de accidentes, se identificó una única inconsistencia: 9.474 registros presentaban el valor “0” como mes, lo cual no es válido. Para corregir este error, se reasignó el

mes correspondiente a dichos registros utilizando la información contenida en la variable “FECHA_EVEN”.

### 2.3.9 Variable " DIA_EVENTO"

La variable “DIA_EVENTO” registra el día de la semana en el que ocurrió el accidente. De este se halló un único registro errado y al igual que en anterior caso, se usó la variable “FECHA_EVEN” para reasignar el día correspondiente.

### 2.3.10 Variable " COD_COMUNA"

Al igual que en la base de datos de infracciones, se presentaron registros inconsistentes con las divisiones administrativas. Se encontraron 9475 registros cuya comuna era “0”, que no existió nunca en la realidad, por lo que, se procedió a reasignar las comunas y corregimientos con el shapefile creado en el ejercicio realizado con la base de datos de infracciones.

Tras la reasignación del mes a partir de la variable “FECHA_EVEN”, se identificó que 41 registros permanecían sin asignación válida. Estos casos correspondían a puntos ubicados en los límites del municipio, por lo que fueron eliminados de la base de datos. Adicionalmente, al igual que en el tratamiento de la base de datos de infracciones, se excluyeron los registros correspondientes a zonas rurales, con excepción del corregimiento de Pance, debido a su mayor representatividad dentro del conjunto de datos.

### 2.3.11 Validación final de la base de datos de accidentes

Finalizada la revisión, se identificó que únicamente la variable “TIPO_VEHIC” presenta 422 datos faltantes. No obstante, se decidió no intervenir estos valores, ya que no afectan la medición ni el análisis de las demás variables. La base de datos resultante cuenta con 19.756 registros y 12 variables. El aumento en el número de variables se debe a la conversión de la base al formato SpatialPolygons, lo cual permitirá su representación gráfica y el uso de herramientas geoespaciales en análisis posteriores.

# 3. EXPLORACIÓN ESPACIOTEMPORAL DE INFRACCIONES Y ACCIDENTES

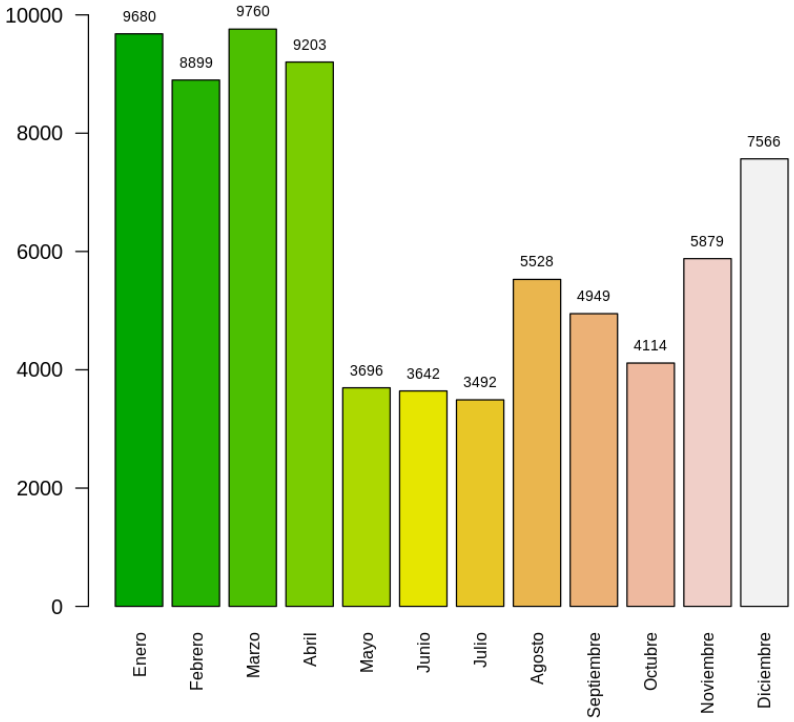
Con la base de datos depurada y lista para el análisis, se dio inicio al análisis exploratorio, enfocado en identificar relaciones causales potenciales a partir de combinaciones entre variables. A continuación, se presentan los hallazgos más relevantes obtenidos en esta etapa inicial.

## 3.1 Exploración de la base de datos de infracciones.

### 3.1.1 Exploración Temporal

Se revisaron los datos hallados para las variables temporales a fin de ver su comportamiento. Al explorar los registros por mes se encontró la siguiente distribución.

***Ilustración 7.*** *Frecuencia de infracciones por mes.*

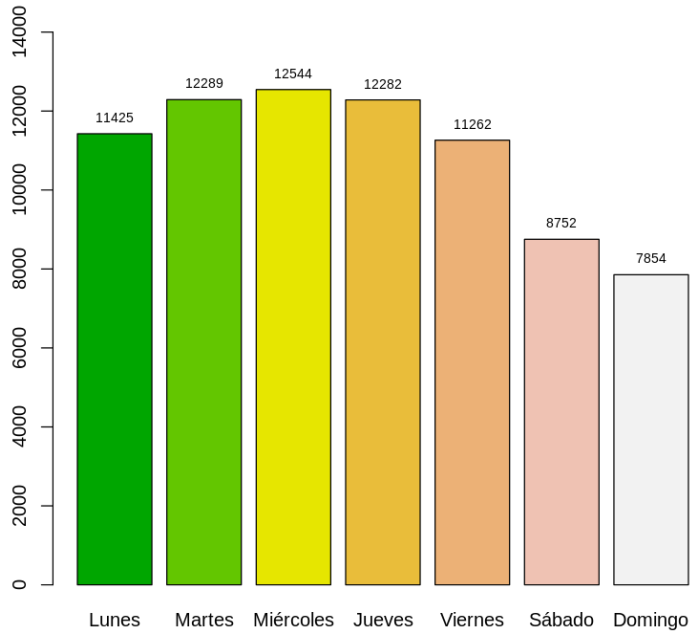



**Fuente:** Elaboración propia

Los datos presentan un comportamiento que debe ser contextualizado en el progresivo desescalamiento de las medidas tomadas por la pandemia de COVID durante 2021 y 2022 lo que hace difícil sugerir que el patrón presentado sea un comportamiento estacional.

Adicionalmente se revisó el comportamiento de las infracciones a lo largo de los días de la semana, en la que se obtuvo la siguiente distribución:

***Ilustración 8.*** *Frecuencia de infracciones por día de la semana.*

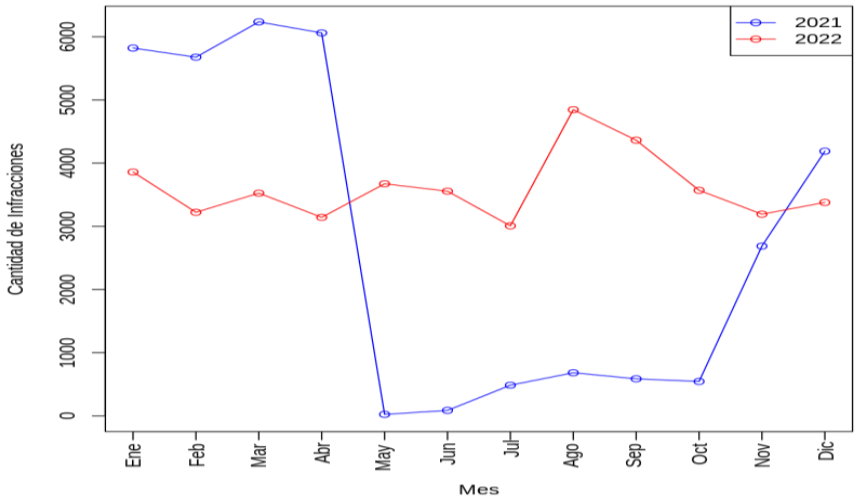



**Fuente:** Elaboración propia

En este caso los datos si presentan un patrón de comportamiento explicable naturalmente, los fines de semana se presentan menos infracciones que en los días laborables, probablemente producto de un menor tráfico esos días.

A continuación, se presenta el comportamiento de las infracciones a lo largo de los años 2021 y 2022.

***Ilustración 9.*** *Línea de tiempo por año de las infracciones.*



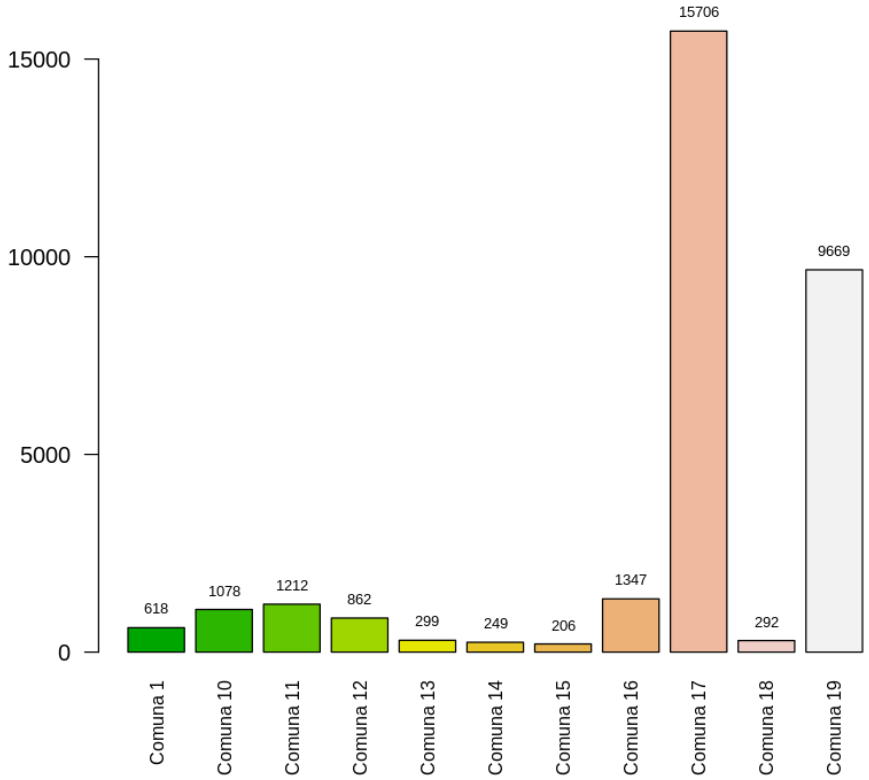

**Fuente:** Elaboración propia

El año 2022 presentó un comportamiento que podría calificarse como estable. Sin embargo, el año 2021 presenta una caída importante, llegando prácticamente a 0 en mayo, y recién vuelve a ubicarse en niveles normales a partir de noviembre de ese año, se resalta que en 2021 todavía estaban vigentes varias de las restricciones de la pandemia de Covid-19, y empezaba a registrarse una reactivación progresiva de las actividades civiles y económicas.

### 3.1.2 Exploración Espacial

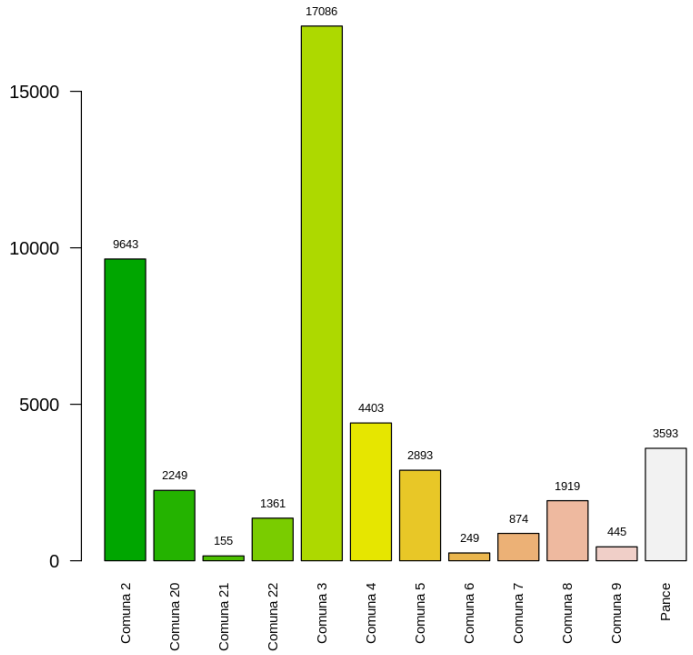
Se realizó un ejercicio similar respecto de la distribución espacial de las infracciones, los datos registrados presentan la siguiente distribución:

***Ilustración 10.*** *Frecuencia de infracciones por división administrativa.*



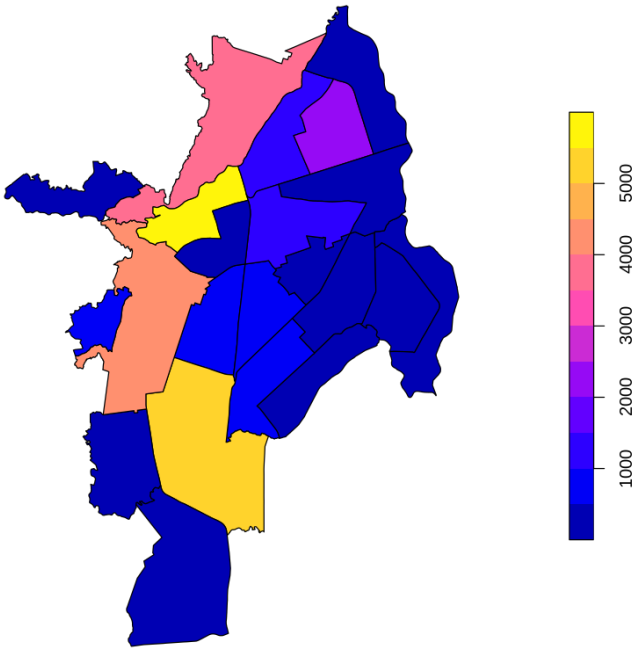

**Fuente:** Elaboración propia

***Ilustración 11.*** *Frecuencia de infracciones por división administrativa (incluyendo Pance)*



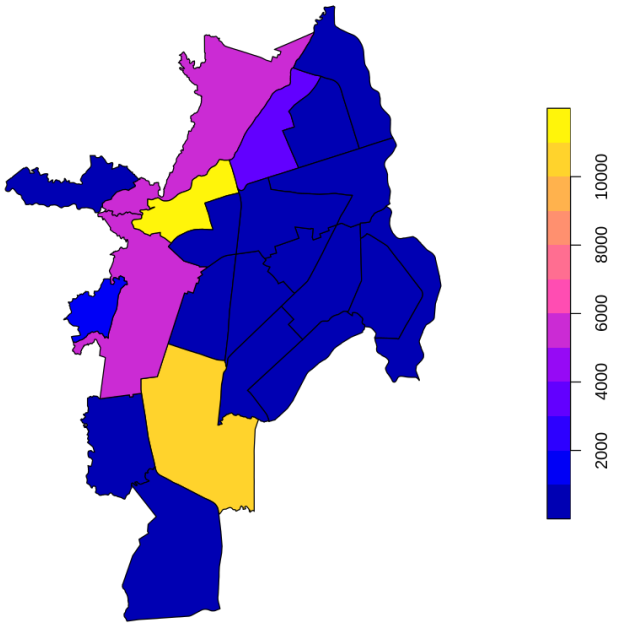

**Fuente:** Elaboración propia

***Ilustración 12.*** *Mapa de infracciones año 2021.*




**Fuente:** Elaboración propia

***Ilustración 13.*** *Mapa de infracciones año 2022.*




**Fuente:** Elaboración propia

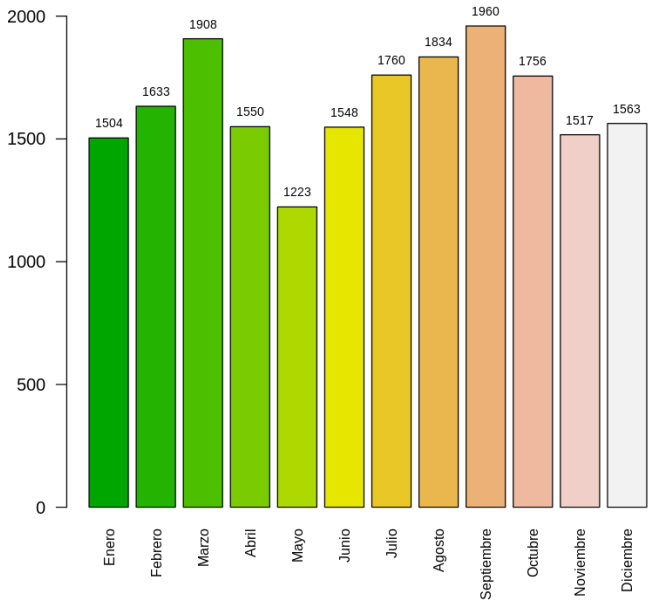
Los datos muestran una marcada concentración de las infracciones sobre las comunas 3 y 17, la comuna 3 está sobre lo que se conoce como el centro histórico de la ciudad y concentra una buena parte de la actividad comercial de la ciudad, con un tráfico pesado durante la mayor parte del día, la comuna 17 está al sur donde se ubican varios centros comerciales, varias urbanizaciones y la Universidad del Valle, además de ser paso obligado de quienes viajan hacia el sur de país por vía terrestre, y si bien su tráfico no es tan lento y pesado como en la comuna 3, circulan una buena cantidad de vehículos y su tráfico puede tornarse muy lento en horas pico. Las comunas 2 (que limita con la zona industrial de Yumbo y la comuna 3) y la comuna 19 (que limita con la comuna 3 y tiene varios puntos de interés como el estadio Pascual Guerrero y varias de las más importantes clínicas y centros médicos de la ciudad, además de la sede alterna de la Universidad del Valle) también presentaron un alto índice de infracciones.

## 3.2 Exploración de la base de datos de accidentes.

### 3.2.1 Exploración Temporal

Se revisaron los datos hallados para las variables temporales a fin de ver su comportamiento. Al explorar los registros por mes se encontró la siguiente distribución:

***Ilustración 14.*** *Frecuencia de accidentes por mes.*

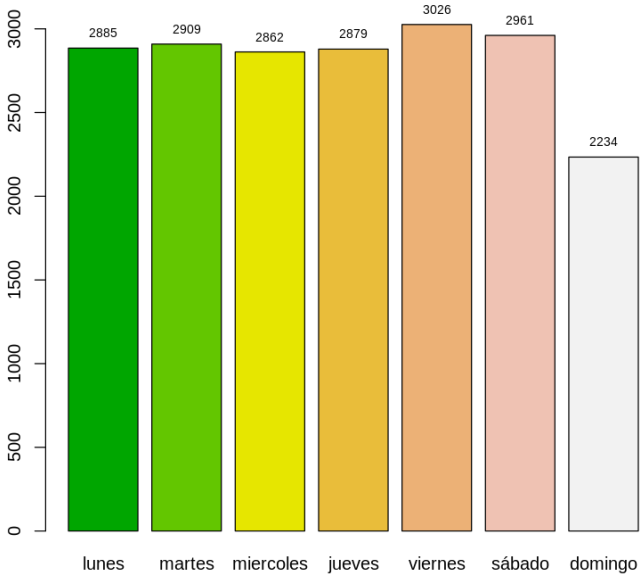



**Fuente:** Elaboración propia

A diferencia de lo ocurrido con lo ocurrido con las infracciones, los accidentes se distribuyeron de forma algo más uniforme, con picos en los meses de marzo y septiembre, y una caída notoria en el mes de mayo.

Con respecto a los accidentes de tránsito respecto de los días de la semana en la ciudad de Cali, se registró el siguiente comportamiento:

***Ilustración 15.*** *Frecuencia de accidentes por día de la semana.*

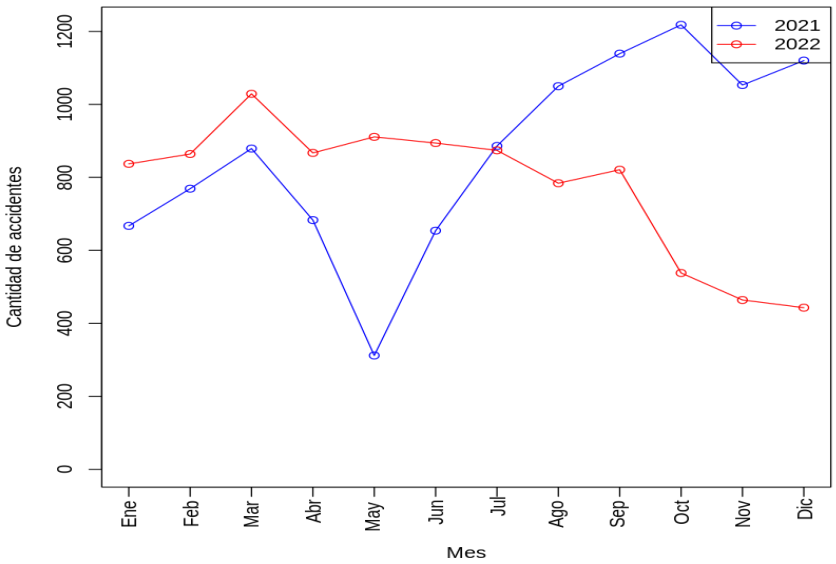



**Fuente:** Elaboración propia

El domingo, al igual con lo que ocurría con la base de datos de infracciones, tiene un comportamiento menor que el resto de la semana, no ocurre lo mismo el sábado, que parece de hecho crecer muy ligeramente con respecto a los días de semana.

A continuación, se presenta el comportamiento de los accidentes a lo largo de los años 2021 y 2022.

***Ilustración 16.*** *Línea de tiempo por año de los accidentes.*



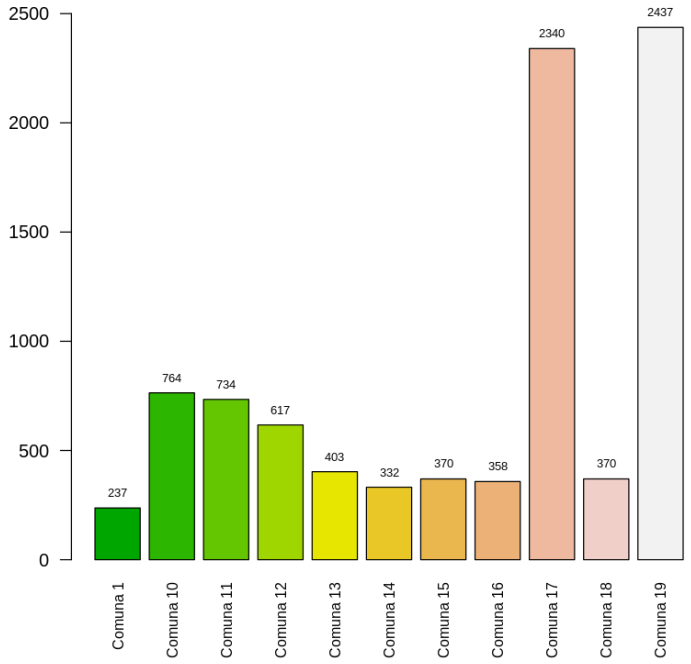

**Fuente:** Elaboración propia

Para el año 2022 el comportamiento de los accidentes es estable, e incluso puede notarse un decaimiento a medida que se acerca al final del año, mientras que el para el año 2021, se observa una reducción importante en el mes de mayo, esto ocurrido en el marco de las medidas tomadas durante la pandemia de Covid-19, pero una vez que se alcanzó ese mínimo los accidentes crecieron fuertemente los meses posteriores a medida que se realizaba la reactivación de las actividades civiles y económicas.

### 3.2.2 Exploración Espacial

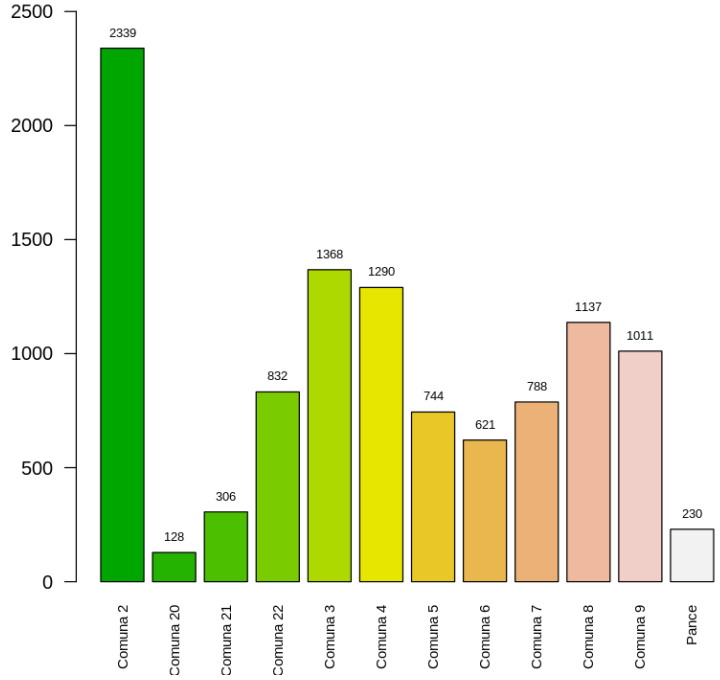
Al realizar el mismo ejercicio, pero respecto de la distribución espacial de los accidentes, los datos registrados presentan la siguiente distribución:

***Ilustración 17.*** *Frecuencia de accidentes por división administrativa.*



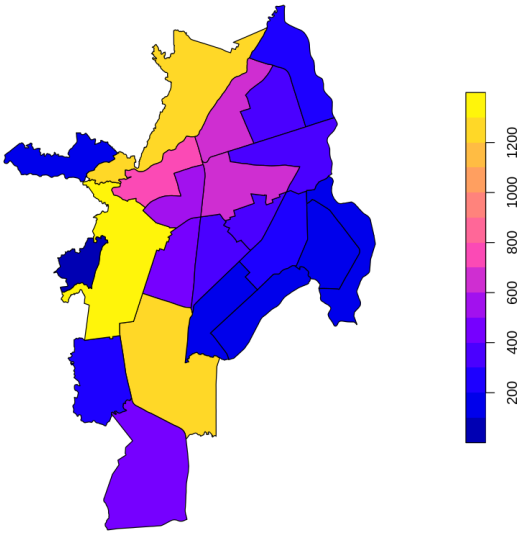

**Fuente:** Elaboración propia

***Ilustración 18.*** *Frecuencia de accidentes por división administrativa (incluido Pance)*




**Fuente:** Elaboración propia

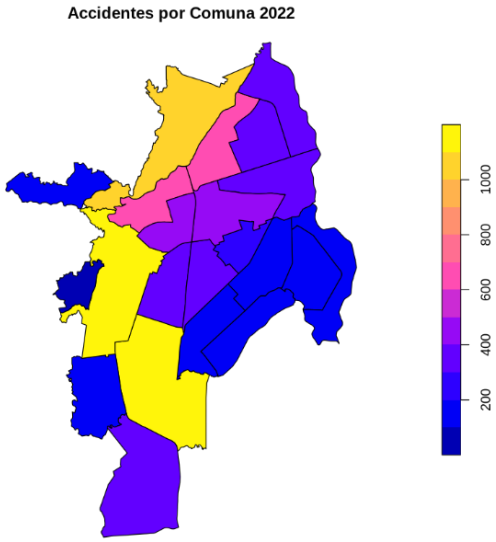
***Ilustración 19.*** *Mapa de accidentes año 2021.*




**Fuente:** Elaboración propia

La comuna 17 que había tenido un alto índice de infracciones también tiene uno de los más altos índices de accidentalidad, en tanto que la participación de las comunas 2 y 19 sube notoriamente en el índice de accidentalidad comparado con el gráfico de infracciones, la comuna 3 tiene un curioso comportamiento de alto índice de infracciones, pero accidentalidad cercana a la media.

***Ilustración 20.*** *Mapa de accidentes año 2022.*




**Fuente:** Elaboración propia

# 4. MAPAS DE INTENSIDAD DE INFRACCIONES VS ACCIDENTES

## 4.1 Descripción de las ubicaciones de las cámaras de fotodetección.

Los registros de la secretaría de movilidad de la ciudad de Cali indican que en la ciudad existían 38 cámaras de fotodetección antes del Paro Nacional en el 2021, las cuales en su mayoría fueron derribas por los manifestantes que se movilizaron para esas fechas. Al mes de noviembre del 2023, se habla de que se habían logrado instalar nuevamente cerca del 50% y a la fecha de realización del presente documento, la Agencia Nacional de Seguridad Vial de Colombia estipula la autorización de 42 cámaras de fotodetección en la capital del Valle del Cauca.

Estos dispositivos tienen como función seguir el monitoreo electrónico en cuanto al exceso de velocidad, cruzar el semáforo en rojo o amarillo, salir cuando el vehículo tiene pico y placa y no tener al día la revisión técnico-mecánica o SOAT al día. A continuación, la tabla 6 presenta la lista de principales infracciones detectadas por cámaras de foto multa.

***Tabla 6.*** *Lista de infracciones que detectan las cámaras de foto detección.*

| Código de la Infracción | Descripción de la infracción |
|---|---|
| **C03** | Bloquear una calzada o intersección con un vehículo, salvo cuando el bloqueo obedezca a la ocurrencia de un accidente de tránsito. |
| **C14** | Transitar por sitios restringidos o en horas prohibidas por la autoridad competente |
| **C35** | No realizar la revisión técnico-mecánica y de emisiones contaminantes en el plazo legal establecido o cuando el vehículo no se encuentra en condiciones técnico-mecánicas o de emisiones, aun portando los certificados correspondientes. |
| **D02** | Conducir un vehículo sin portar el Seguro Obligatorio de Accidentes de Tránsito (SOAT) que es un requisito legal para circular. |
| **C24** | Se aplica a conductores de motocicletas que no cumplen con las normas de tránsito establecidas en el Código Nacional de Tránsito. |
| **C29** | Conducir un vehículo a velocidad superior a la máxima permitida. |
| **D04** | No detenerse ante una luz roja o amarilla de semáforo, una señal de "PARE" o un semáforo intermitente en rojo. |
| **D05** | Conducir un vehículo en zonas prohibidas como aceras, plazas, vías peatonales, separadores, bermas, zonas verdes o vías especiales para vehículos no motorizados. |
| **D07** | Conducir realizando maniobras altamente peligrosas, siempre y cuando ponga en peligro a las personas o cosas y que constituya conductas dolosas o altamente imprudentes. |

**Fuente:** Elaboración propia con base en el Programa de Servicios de Tránsito, Alcaldía de Cali. 2024

Analizar la relación entre las infracciones de tránsito y la ocurrencia de accidentes es fundamental para mejorar la seguridad vial. Factores humanos, así como las condiciones del vehículo y la vía, son las principales causas de los siniestros automovilísticos. Para el desarrollo del proyecto, hemos estipulado la clasificación de cuatro bloques fundamentales divididos en Infracciones que afectan la seguridad del conductor y otros usuarios de la vía, Infracciones relacionadas con el estado del vehículo, Infracciones relacionadas con el comportamiento del conductor y otras infracciones con impacto indirecto, y en relación con la base de datos obtenida, se han detectado las siguientes infracciones:

- Infracciones relacionadas con la seguridad del conductor y otros vehículos
- Infracciones relacionadas con el estado del vehículo
- Infracciones relacionadas con el comportamiento del conductor

De acuerdo con información consolidada de la Secretaría de Movilidad de Santiago de Cali y la Agencia Nacional de Seguridad Vial [43], se puede deducir que para la ciudad de Cali existen un total de 38 cámaras de fotomultas autorizadas y que, para la fecha del presente trabajo, se encuentran funcionando.

Sus ubicaciones se describen en la siguiente tabla:

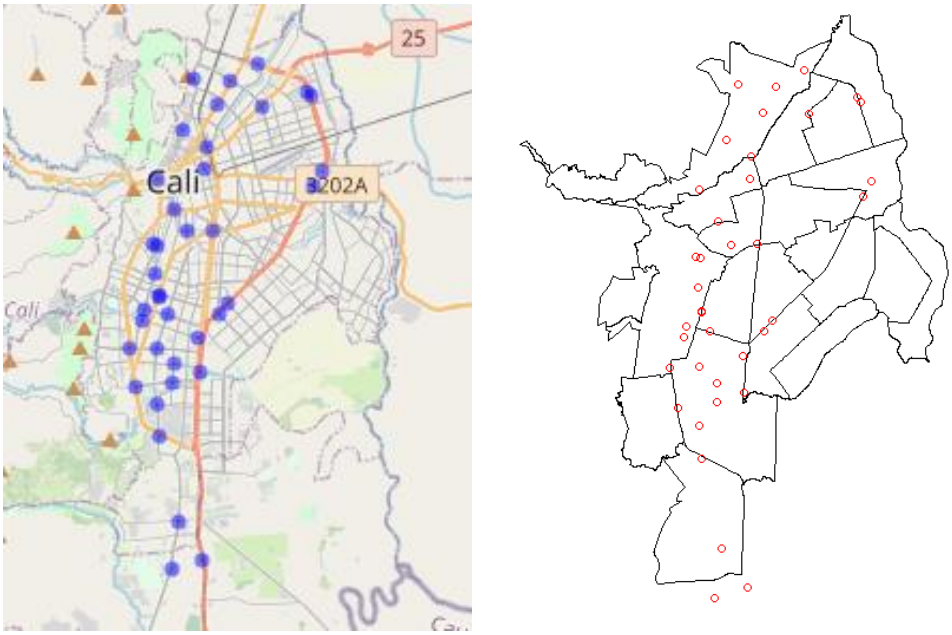
***Tabla 7.*** *Ubicaciones de las cámaras de foto detección.*

| Dirección | Latitud | Longitud |
|---|---|---|
| Avenida 2 Con Calle 6N | 3,452564 | -76,537934 |
| Avenida 3N Con Calle 40 | 3,476221 | -76,5182469 |
| Avenida 6N Con Calle 26 | 3,4678576 | -76,5293658 |
| Avenida 6N Con Calle 47 | 3,4846512 | -76,5260089 |
| Calle 10 Con Carrera 15 | 3,4427762 | -76,531974 |
| Calle 10 Con Carrera 44A | 3,4152051 | -76,537151 |
| Calle 10 Con Carrera 45 | 3,4149691 | -76,5371256 |
| Calle 13 Con Carrera 100 | 3,3703348 | -76,5371272 |
| Calle 13 Con Carrera 23 | 3,4356575 | -76,5281992 |
| Calle 13 Con Carrera 50 | 3,4093168 | -76,534435 |
| Calle 13 Con Carrera 66 | 3,39840597 | -76,5378339 |
| Calle 13 Con Carrera86 | 3,3802098 | -76,5377124 |
| Calle 14 Con Carrera 70 | 3,3933259 | -76,532425 |
| Calle 14 Con Carrera 80 | 3,3873908 | -76,532488 |
| Calle 18 Con Carrera 122 | 3,3428313 | -76,5309674 |
| Calle 18 Con Carrera 130 | 3,3274078 | -76,53311 |
| Calle 23 Con Carrera 23 | 3,4359883 | -76,5200453 |
| Calle 25 Con Avenida 2N | 3,4628735 | -76,521979 |
| Calle 25 Con Carrera 73 | 3,3905044 | -76,5240686 |
| Calle 36 Con Carrera 128 | 3,3307345 | -76,5229133 |
| Calle 36 Con Carrera 42B | 3,4124376 | -76,5152203 |
| Calle 5 Con Carrera 80 | 3,3855963 | -76,5444215 |
| Calle 52 Avenida 3N | 3,4839147 | -76,5141141 |
| Calle 52 Con Carrera 1 | 3,4756242 | -76,5041172 |
| Calle 6 Con Carrera 29 | 3,4320396 | -76,53883 |
| Calle 7 Con Carrera 29 | 3,4314865 | -76,5376072 |
| Calle 70 Avenida 2An | 3,4891759 | -76,5055641 |

| | | |
|---|---|---|
| Calle 70 Con Carrera 1A-12 | 3,4806964 | -76,4894095 |
| Calle 70 Con Carrera 1B | 3,4792551 | -76,488393 |
| Calle 70 Con Carrera 7Mbis | 3,4552027 | -76,484845 |
| Calle 70 Con Carrera 9 | 3,4502579 | -76,4875828 |
| Calle 9 Con Carrera 38 | 3,4224253 | -76,538248 |
| Calle 9 Con Carrera 50 | 3,4108353 | -76,5417392 |
| Carrera 46 Con Calle 36 | 3,4093953 | -76,5178873 |
| Carrera 5 Con Calle 23 | 3,4560002 | -76,5223654 |
| Carrera 56 Con Calle 18A | 3,4018078 | -76,524469 |
| Carrera 56 Con Calle 9 | 3,407457 | -76,5424192 |
| Carrera 66 Con Calle 5 | 3,3981553 | -76,5468447 |

**Fuente:** Elaboración propia.

***Ilustración 21.*** *Ubicación geoespacial de las cámaras de foto multas en Cali, Colombia.*




**Fuente:** Elaboración propia

### 4.2 Análisis de Patrones Puntuales Espaciales

El análisis de patrones espaciales se define como el proceso de examinar representaciones abstractas de patrones espaciales del mundo real para obtener información sobre los procesos subyacentes. Cada punto representado corresponde a la ocurrencia de un evento en un lugar específico. [44]

Teniendo en cuenta esta definición y entendiendo que los datos que componen este análisis tanto en los accidentes como en las infracciones tienen las características de la información para ser analizadas y es

que tanto las infracciones como los accidentes se registran en coordenadas espaciales (x, y), se puede utilizar el método de patrones puntuales.

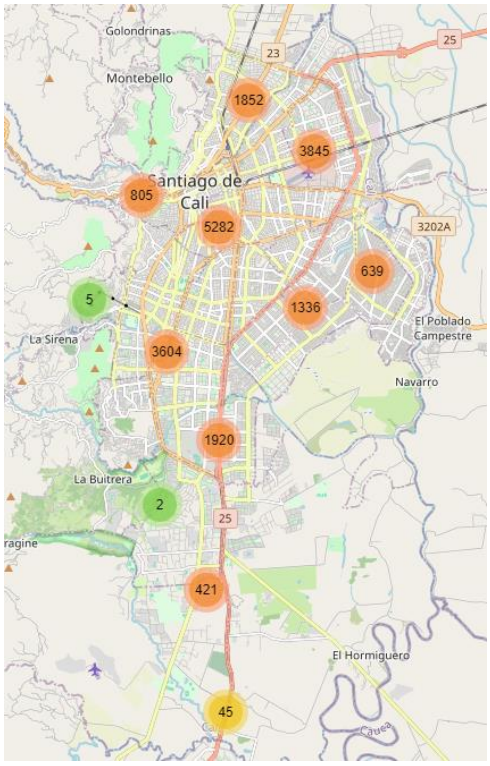
Para realizar este análisis, se verificaron los siguientes puntos.

1. Validar que todos los eventos tuvieran coordenadas espaciales (x, y).
2. Representar los eventos como un patrón de puntos espaciales.
3. Evaluar la aleatoriedad espacial a través de la prueba chi-cuadrado ($X^2$) para verificar si los patrones se distribuyen aleatoriamente.
4. Evaluar la aleatoriedad a través del gráfico de la función K de Ripley.
6. Generar los mapas de intensidad para visualizar la concentración espacial de los eventos.

### 4.3 Análisis de intensidad de los siniestros

Tomando como referencia lo mencionado en el punto anterior, se da por ejecutado el punto 1 ya que las bases de datos en la exploración inicial mostraron que cuentan con la información y las coordenadas respectivas para realizar el análisis. Posterior a esto, se realizó un filtro para los años 2021 (10,430) y 2022 (9,326), con un total de registros de: 19,756 y se procedió a graficar la frecuencia de los accidentes en cada zona de la ciudad de Cali como se muestra en la siguiente gráfica.

***Ilustración 22.*** *Frecuencia de siniestros agrupados en la división administrativa de Cali 2021 – 2022*




**Fuente:** Elaboración propia

De manera preliminar se puede visualizar que la mayor concentración en los accidentes está en las zonas centro y noroccidental de la ciudad.

Siguiendo cada uno de los puntos mencionados en el punto anterior continuamos con la aplicación del test chi-cuadrado. La prueba de chi-cuadrado se aplica para evaluar si la intensidad de los puntos tiene un comportamiento o no homogéneo, bajo el supuesto nulo de aleatoriedad (*Complete Spatial Randomness* - CSR) para patrones espaciales.

La prueba es calculada bajo el siguiente estadístico:

**Ecuación 4**. *Estadístico Chi-cuadrado de Pearson*

$$\chi^2 = \sum \frac{(observado - esperado)^2}{esperado}$$

Con el software *R y RStudio*, se realizó la prueba de chi-cuadrado con la función *quadrat.test*, la cual arrojó el siguiente resultado:

```
Chi-squared test of CSR using quadrat counts
data:  patron_siniestros
X2 = 52048, df = 24, p-value < 2.2e-16
alternative hypothesis: two.sided
Quadrats: 5 by 5 grid of tiles
```

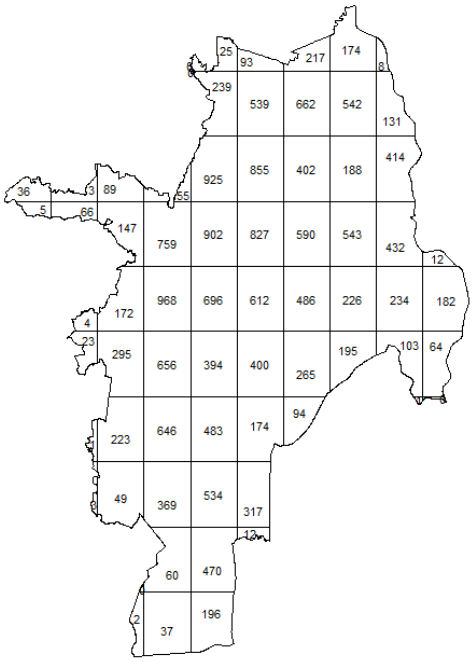
El valor del estadístico *p-valor < 2.2e-16* indica que hay una diferencia altamente significativa respecto a la hipótesis de aleatoriedad completa, lo que nos indica que la probabilidad de que esta distribución haya ocurrido de manera aleatoria es baja. Esto quiere decir que, se descarta la hipótesis nula de que los eventos de accidentalidad con personas fallecidas sean distribuidos de manera aleatoria.

### 4.3.1 Conteo de siniestros por cuadrantes

El conteo de siniestros por cuadrantes consiste en dividir la región de estudio en rectángulos de igual tamaño (cuadrantes) y contabilizar el número de puntos dentro de cada uno. La Ilustración 23 muestra el resultado de este análisis aplicado a los accidentes de tránsito.

Se observa una concentración significativa en el centro de la ciudad, donde algunos cuadrantes registran entre 800 y 900 eventos, en contraste con otras zonas que presentan menos de 50 registros. Esta visualización forma parte del análisis chi-cuadrado y confirma que la distribución espacial de los accidentes no es aleatoria, lo cual se ve reflejado en el bajo valor del *p-value* obtenido.

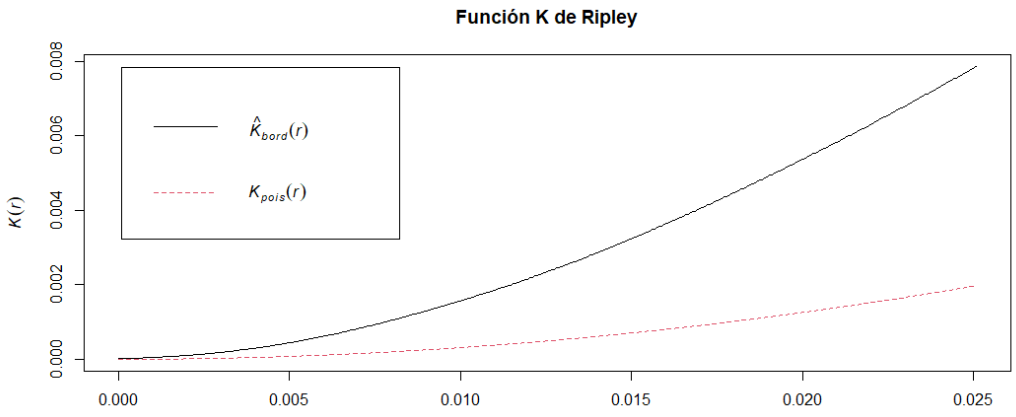
***Ilustración 23.*** *Conteo de accidentes fatales por cuadrante*




**Fuente:** Elaboración propia

Lo anterior, sugiere que la distribución de los siniestros con personas fallecidas no ocurre de manera aleatoria. Sin embargo, el test de cuadrantes no permite identificar la naturaleza exacta del patrón, es decir, si se trata de agrupamiento o regularidad. Por ello, se procedió a completar este análisis con herramientas como la Función K de Ripley, que permite explorar la interacción espacial entre eventos a diferentes escalas y caracterizar el patrón como agregado o regular. En R, se encuentra como la función *Kest.*

***Ilustración 24.*** *Función K de Ripley de accidentes*



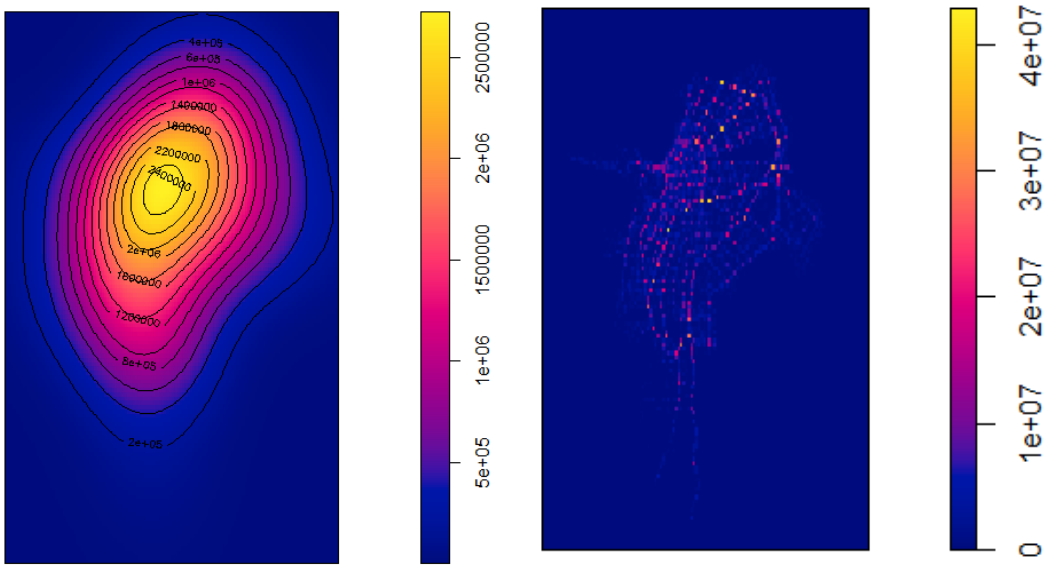

**Fuente:** Elaboración propia

El gráfico de *K de Ripley* es una herramienta utilizada comúnmente para análisis de patrones puntuales espaciales. En donde, dependiendo de cómo se comporta la función K empírica frente a la esperada bajo la CSR, podemos clasificarla como aleatoria, agregada o regular.
De la gráfica 17, se pudo concluir que:

- Se confirmaron los resultados de la prueba de chi-cuadrado.
- Se evidencia que la curva empírica $\widehat{K}(r)$ (líneas solidas) se desvían del valor teórico esperado $K_{Poisson}(r)$ bajo el supuesto de que los puntos están completamente aleatorios (líneas continuas).
- Lo anterior confirma que, los eventos se encuentran agrupados.
- Existen zonas específicas con mayor concentración de casos.

Los siguientes gráficos son mapas de intensidad realizados en *R* con la función *density.ppp*. El software permite ajustar la función sigma, para obtener un resultado visual deseado.

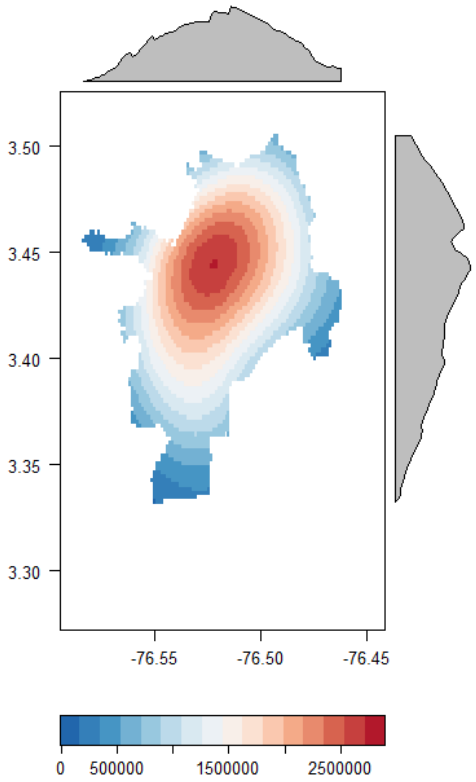
***Ilustración 25.*** *Mapas de intensidad de Kernel Isodensidad y Gaussiano σ=2e-05*




**Fuente:** Elaboración propia

Los mapas de intensidad de Kernel permiten visualizar como los accidentes se concentra en la parte del centro histórico y el oeste de la ciudad, así como una visualización rápida de los puntos críticos de las vías de la ciudad.

***Ilustración 26.*** *Mapa de densidad de Kernel gradiente*




**Fuente:** Elaboración propia

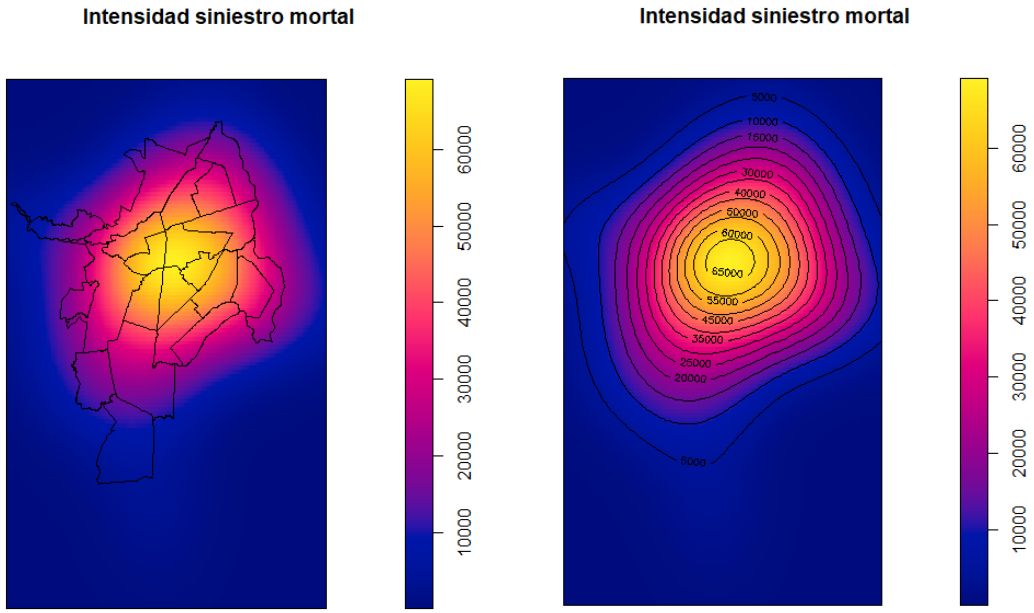
De acuerdo con los gráficos de intensidad mostrados en las figuras anteriores se puede ver un registro importante de los eventos concentrados principalmente en la zona centro de la ciudad que corresponde principalmente a las comunas 2, 3, 4, 9 y 19.

Los tres gráficos anteriores evidencian distintos enfoques de visualización mediante mapas KDE (Kernel Density Estimation), cada uno con un propósito analítico específico. El mapa KDE con contornos de isodensidad, permite una interpretación desde una perspectiva más matemática y cuantitativa. El mapa KDE con escala de color continua, en el cual se incorporan coordenadas geográficas para facilitar la ubicación espacial precisa de los eventos. El mapa KDE espacial suavizado, delimitado por comunas y barrios, que ofrece una lectura territorial más contextualizada.

### 4.3.2 Densidad de los siniestros mortales

Los siguientes gráficos muestran la intensidad espacial de siniestros viales mortales para la ciudad de Cali. Los mapas de calor y líneas de contorno permiten identificar visualmente estas áreas críticas, donde se presentan con mayor frecuencia los eventos fatales. Comunas 8, 9, 10 y 11 como zonas de mayor intensidad y zonas centrales, y específicamente en avenidas principales como la autopista sur oriental en los cruces con la calle 25 y transversal 25.

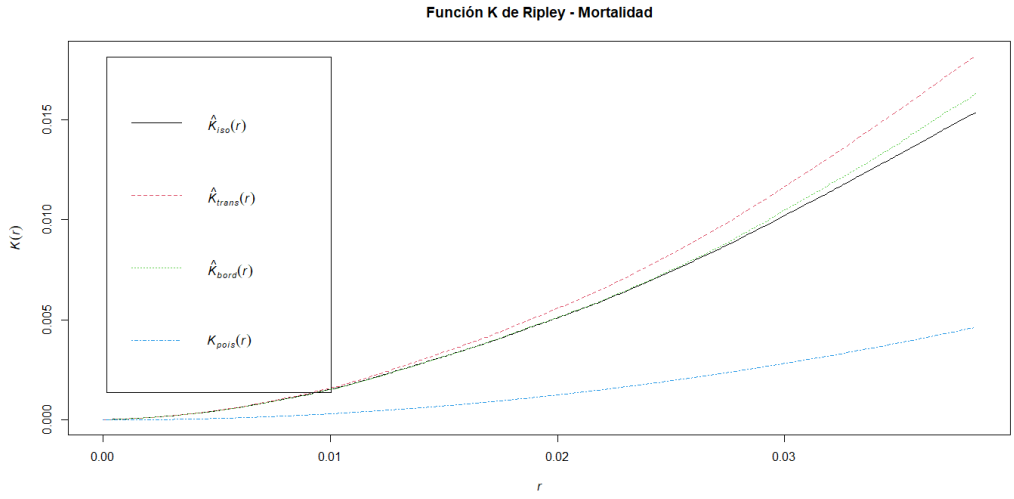
***Ilustración 27.*** *Mapa de densidad de siniestralidad mortal*




**Fuente:** Elaboración propia

La siguiente gráfica corresponde a la función *K de Ripley* donde la curva observada se encuentra por encima del intervalo de simulaciones aleatorias (líneas punteadas), lo que indica que los siniestros mortales presentan un patrón de agrupamiento espacial significativo. Es decir, las muertes no ocurren de manera aleatoria en el espacio urbano, sino que tienden a concentrarse en determinadas zonas.

***Ilustración 28.*** *Función K de Ripley de siniestralidad mortal*

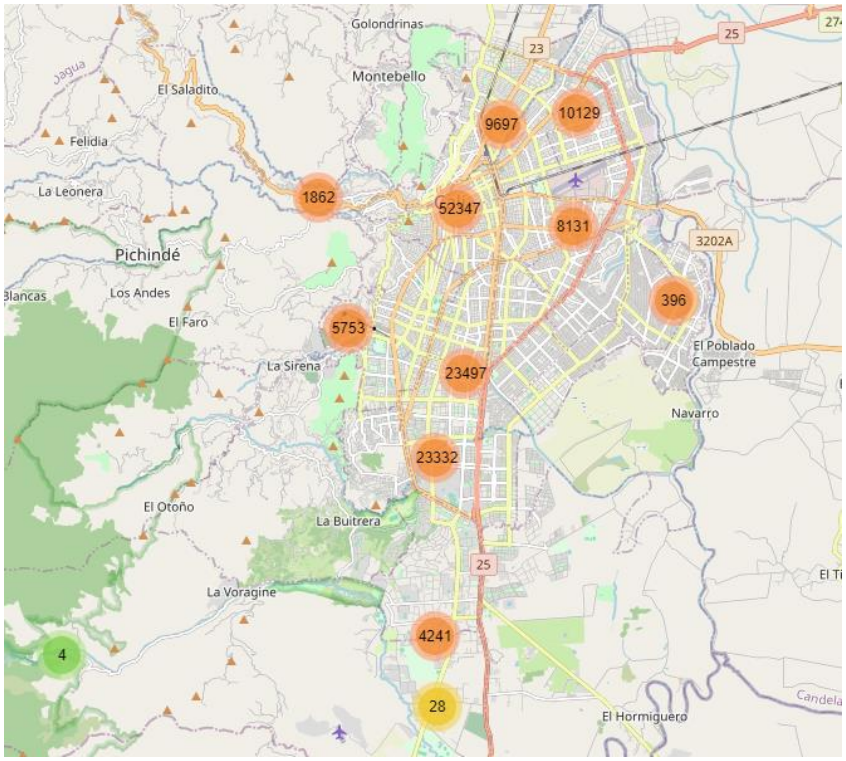



**Fuente:** Elaboración propia

### 4.4 Análisis de intensidad de las infracciones

Para el análisis de intensidad en las infracciones, se continuó con el filtro de la base de datos para los años 2021 (61,576) y 2022 (77,842), con un total de registros de: 139,417. A continuación, se procedió a graficar la frecuencia de las infracciones en cada zona de la ciudad de Cali como se muestra en las siguientes gráficas.

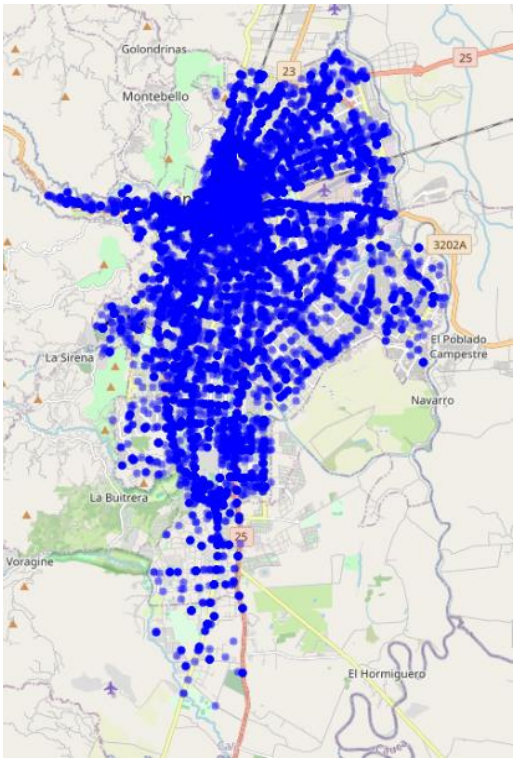
***Ilustración 29.*** *Frecuencia de infracciones agrupadas en la división administrativa de Cali 2021 - 2022*




**Fuente:** Elaboración propia

El mapa de infracciones agrupadas muestra similitud con la Figura 22, donde la mayor cantidad de infracciones se ven en el centro y sur de la ciudad. El mapa de distribución de infracciones valida lo observado en el mapa de infracciones agrupadas y permite ver espacialmente todas las zonas de la ciudad donde se han impuesto comparendos de tránsito.

***Ilustración 30.*** *Mapa de distribución de infracciones*




**Fuente:** Elaboración propia

Al realizar la aplicación de la prueba de chi-cuadrado en la información espacial de las infracciones, se encontró el siguiente resultado:

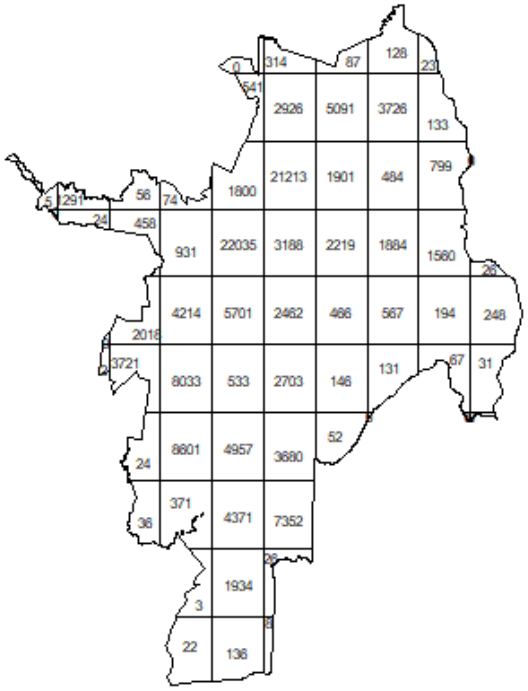
```
Chi-squared test of CSR using quadrat counts
data:  patron_multas
X2 = 538330, df = 24, p-value < 2.2e-16
alternative hypothesis: two.sided
Quadrats: 5 by 5 grid of tiles
```

El valor *p-value* es muy bajo (< 2.2e-16), lo que indica que tanto los accidentes como con las infracciones de tránsito, no se distribuyen aleatoriamente en el espacio. Se evidencia un patrón espacial que indica una probabilidad de agrupamiento.

### 4.4.1 Conteo de infracciones por cuadrantes

Nuevamente, se divide la región de estudio en cuadrantes de igual tamaño, y se realiza el conteo de puntos en cada rectángulo como se observa en la siguiente ilustración. La mayor cantidad de infracciones, de acuerdo con el conteo por cuadrantes, indica un elevado número hacía las zonas centro y occidente de la ciudad con valores superiores a los 20 mil registros.

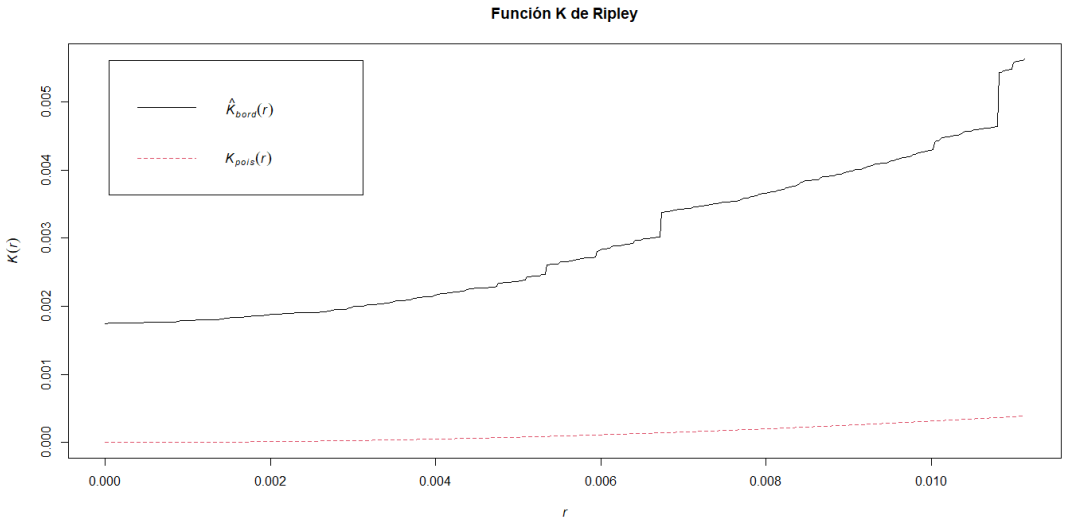
***Ilustración 31.*** *Conteo de infracciones por cuadrante*




**Fuente:** Elaboración propia

Realizando el conteo por cuadrantes para las infracciones, se ve empíricamente una concentración de estas en las zonas cercanas al centro de la ciudad.

***Ilustración 32.*** *Gráfico función K de Ripley de infracciones*



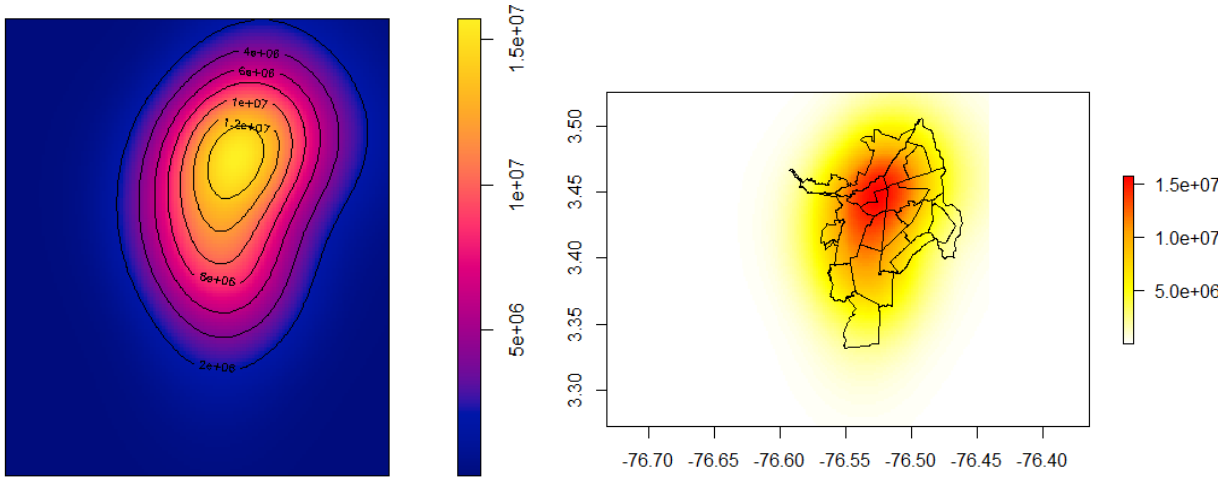

**Fuente:** Elaboración propia

Con el anterior gráfico de la función de K de Ripley, se concluyó que:

- Se confirmaron los resultados de la prueba de chi-cuadrado.
- Se evidencia que la curva empírica $\widehat{K}(r)$ (líneas solidas) se desvían del valor teórico esperado $K_{Poisson}(r)$ bajo el supuesto de que los puntos están completamente aleatorios (líneas continuas). Lo anterior confirma que, los eventos se encuentran agrupados.
- Existen zonas específicas con mayor concentración de casos.

Los siguientes gráficos muestran una estimación de la intensidad inicial, es decir, las zonas de mayor agregación a nivel espacial de las infracciones. Esto con el fin de entender el comportamiento del patrón a nivel espacial.

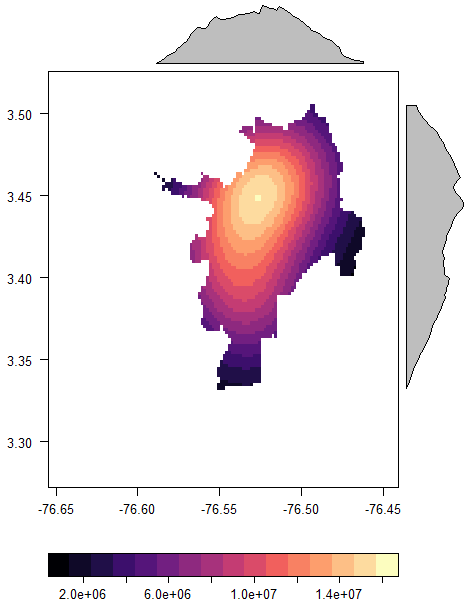
***Ilustración 33.*** *Mapas de intensidad de Kernel Isodensidad de infracciones en el área administrativa*




**Fuente:** Elaboración propia

El mapa de intensidad de Kernel de las infracciones, corrobora las observaciones anteriores al apuntar a una gran concentración de eventos de infracciones en las zonas cercanas al centro histórico de la ciudad.

***Ilustración 34.*** *Mapa de densidad de Kernel gradiente*




**Fuente:** Elaboración propia

Al igual que la intensidad observada en los análisis de accidentes, las infracciones también presentan una concentración predominante en la zona centro de la ciudad. Para la identificación espacial, se presentan igualmente los tres tipos de visualizaciones basadas en mapas KDE (Kernel Density Estimation):

- Mapa KDE con contornos de isodensidad,
- Mapa KDE con escala de color continua y,
- Mapa KDE suavizado con delimitación por comunas

Estas representaciones permiten identificar zonas de mayor densidad de infracciones y comparar su comportamiento espacial con el de los accidentes de tránsito.

# 5. MODELADO ESPACIAL Y VISUALIZACIÓN DE RESULTADOS

## 5.1 Modelado estadístico de datos espaciales

El modelamiento estadístico de procesos puntuales se realizó a través de la intensidad de puntos usando modelos lineales y no lineales para representar la relación con un conjunto de covariables espaciales, determinando que el número de hechos de mortalidad por unidad de área (intensidad) guarda relación con las infracciones. Es decir que, el modelamiento del proyecto estuvo orientado en conocer cómo la variable infracciones ($X_n$) afecta la mortalidad ($Y$ ). Con ello, se pudo plantear un modelamiento clásico llevado a un campo espacial por medio de la creación de imágenes que contienen información del evento en cada uno de sus pixeles, expresado por la siguiente ecuación:

**Ecuación 5**. *Modelo clásico y fundamental en estadística*

$$Y = X_1 + X_2 + X_3 + \cdots + X_n$$

O, de una forma más compacta:

$$Y = \sum_{i=1}^{n} X_i$$

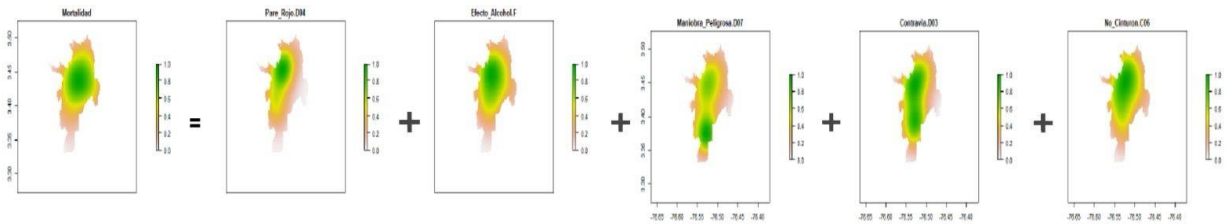
Donde,

$Y$ como variable dependiente (*mortalidad en siniestros viales*) y,

$X_n$ como las variables independientes o predictoras (*infracciones de tránsito*)

El modelo desarrollado en este proyecto consistió en inferir la mortalidad de siniestros viales con base en las variables predictoras elegidas, lo que quiere decir que la operación de unificación de las bases de datos se realizó a partir de las imágenes para determinar el modelamiento predictivo. El método consistió en construir la imagen de la densidad de la variable mortalidad en siniestros viales, junto con las imágenes de las variables predictoras más influyentes en relación con las infracciones de tránsito, tal se observa en la figura 35. Posterior a ello, se realizó la transformación a imágenes tipo ráster, su estandarización y escalamiento de pixeles para luego, aplicar los modelos estadísticos con sus respectivas métricas, para finalmente realizar predicciones sobre el que tuviese un mejor rendimiento.

***Ilustración 35.*** *Modelamiento de la intensidad por imágenes ráster*



### 5.2 Elección de las variables para el modelo

Para la elección de las variables, se tuvo en cuenta las infracciones más representativas a nivel vial. En un principio, la estipulación de estas variables no mostró alta importancia en el aporte del modelo o existía una alta colinealidad con otras variables, por lo que fue necesario establecer un criterio que estuviese relacionado con la mortalidad.

En anteriores gráficos, se presentaron las infracciones que más se cometen en la ciudad. Sin embargo, se encuentran algunas que no están directamente relacionadas con el objeto de estudio como lo fueron el vencimiento del seguro obligatorio SOAT, las luces y componentes eléctricos, entre otros.

Dado lo anterior y mediante el uso de la función *vif* (Variance Inflation Factor o Factor de Inflación de la Varianza) en el software *R*, se pudo interpretar estos resultados con mayor facilidad en la elección de las variables. Esta función sirvió para detectar la colinealidad de las variables independientes (predictoras) en el modelo de regresión; mide cuanto se incrementa la varianza del coeficiente estimado de una variable debido a la colinealidad con las otras variables predictoras.

Si una variable tiene un VIF alto, significa que está altamente correlacionada con otras variables independientes, lo que puede volver inestables las estimaciones del modelo.

***Tabla 8.*** *Interpretación de la función VIF*

| VIF | Interpretación |
|---|---|
| **1** | No hay colinealidad |
| **1 – 5** | Colinealidad moderada (normalmente aceptable) |
| **> 5** | Alta colinealidad (precaución) |
| **> 10** | Colinealidad severa (problema grave) |

Los primeros resultados indicaban una colinealidad severa en variables elegidas como no detenerse ante una luz roja o amarilla (semáforo) o una señal de “pare” y no respetar el cruce peatonal (12.51), además no aportaban significativamente en el desarrollo de los modelos. Con ello, fue necesario realizar un juego de variables que permitiera agrupar la mayor información posible sin perder calidad y sin recurrir a métodos como el análisis de componentes principales (PCA) que haría crear nuevas variables perdiendo la interpretación en la originalidad de las variables.

Finalmente, procedió con la elección de algunas de las variables que, aunque estuviesen altamente correlacionadas, fueron teóricamente las más interesantes y directamente relacionadas con el objeto estudio. Con ello, se decidió por el uso las siguientes variables:

1. C06 - No utilizar el cinturón de seguridad
2. D03 - Transitar en sentido contrario al estipulado para la vía, calzada o carril.
3. D04 - No detenerse ante luz roja o amarilla de semáforo o una señal de "PARE".
4. D07 – Conducir realizando maniobras altamente peligrosas e irresponsables que pongan en peligro a las personas o las cosas

5. F - Conducir bajo los efectos del alcohol o bajo los efectos de sustancias psicoactivas.

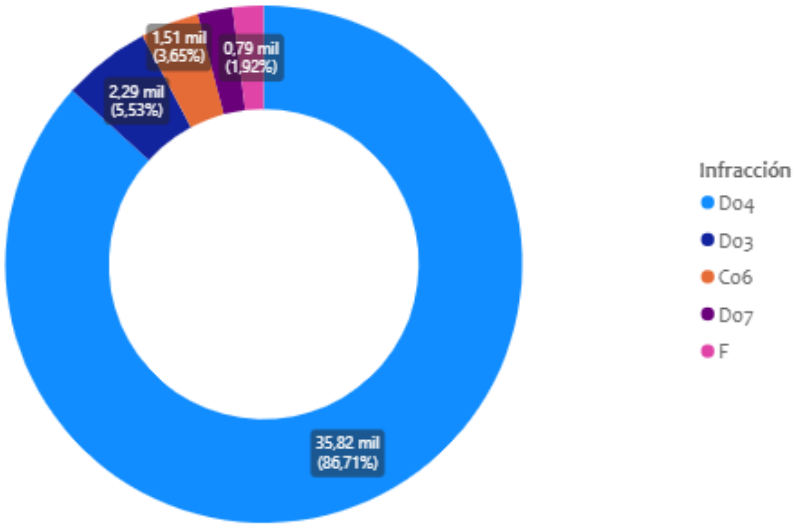
***Tabla 9.*** *Participación de las infracciones (variables) elegidas.*

| Código de la Infracción | Descripción Infracción | Participación No. | Participación % |
|---|---|---|---|
| D04 | No detenerse ante luz roja o amarilla de semáforo o una señal de "PARE" | 35819 | 86,71% |
| D03 | Transitar en sentido contrario al estipulado para la vía, calzada o carril | 2286 | 5,53% |
| C06 | No utilizar el cinturón de seguridad por parte de los ocupantes del vehículo | 1506 | 3,65% |
| D07 | Conducir realizando maniobras altamente peligrosas | 906 | 2,19% |
| F | Conducir en estado de embriaguez | 794 | 1,92% |
| **Total** | | **41311** | **100,00%** |

**Fuente:** Elaboración propia

El siguiente gráfico muestra la incidencia de cada una de las infracciones en la participación del número de infracciones totales.

***Ilustración 36.*** *Participación de las Infracciones (variables) elegidas.*




**Fuente:** Elaboración propia

Estos resultados de colinealidad se dan puesto que los eventos predictores, tienden a ubicarse espacialmente en los mismos puntos de la mortalidad, lo cual podría ser indicador de un buen ajuste para el modelo predictivo.

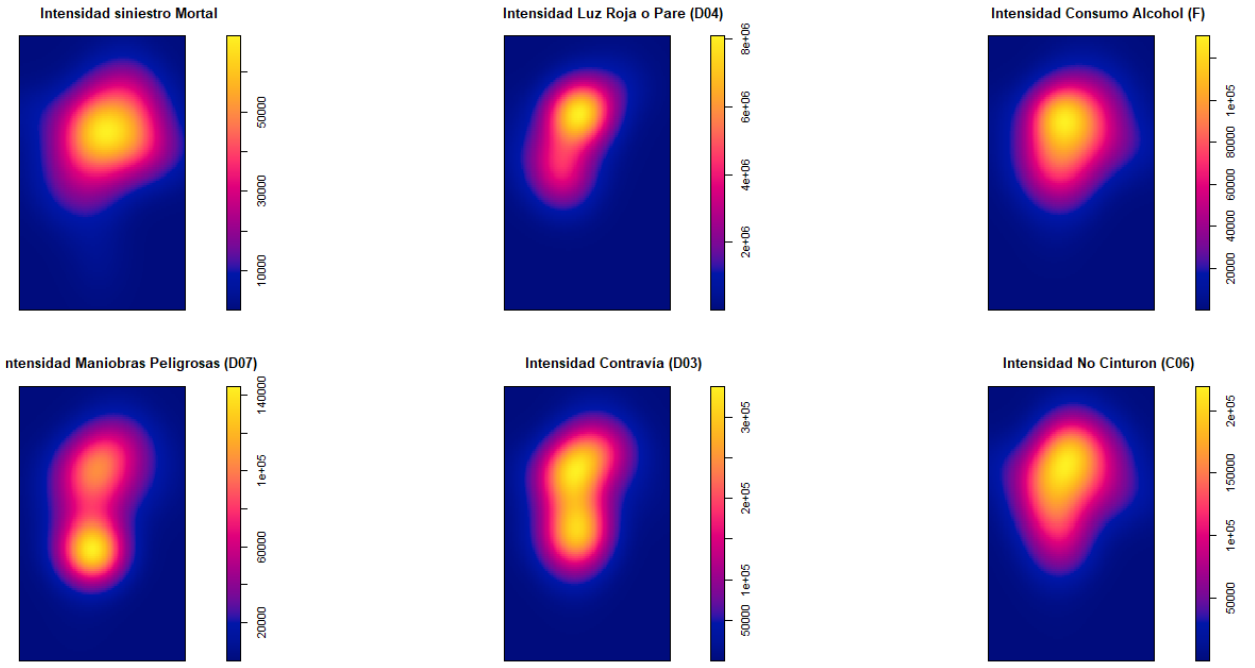
*Tabla 10. Resultados de la función VIF e interpretación.*

| Variable | VIF | Interpretación |
|---|---|---|
| **Luz Roja o Pare [D04]** | 9.32 | Alta colinealidad |
| **Efecto Alcohol [F]** | 14.73 | Muy alta colinealidad |
| **Maniobra Peligrosa [D07]** | 4.92 | Moderada colinealidad |
| **Contravía [D03]** | 12.71 | Muy alta colinealidad |
| **No Cinturón [C06]** | 27.82 | Extremadamente alta colinealidad |

**Fuente:** Elaboración propia

Antes de exportar las imágenes en formato ráster, se realizó una visualización previa general de los resultados de las intensidades, donde se logró observar la distribución espacial de información por medio de mapas de calor. Eso sirvió para confirmar los supuestos de agrupación establecidos para el patrón.

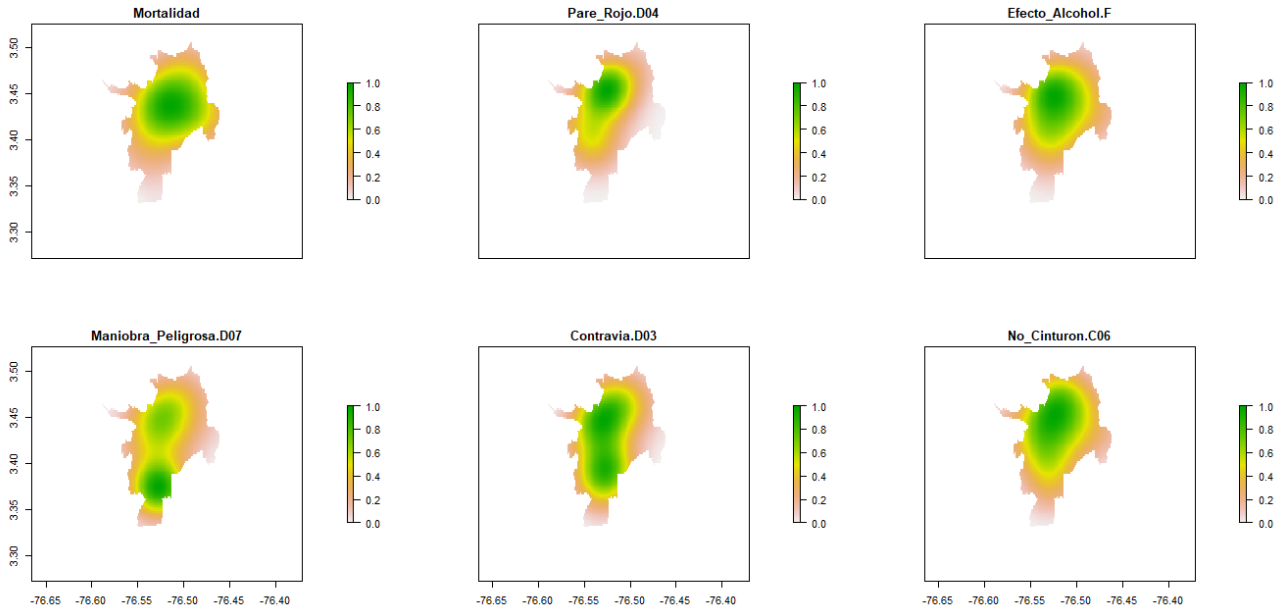
*Ilustración 37. Estimación de la intensidad para la mortalidad e infracciones viales*




**Fuente:** Elaboración propia

Luego, se realizó una visualización con escala estandarizada y delimitada con corte para la ciudad de Cali. Estas imágenes son las que posteriormente se utilizaron para la realización del modelamiento estadístico para patrones puntuales.

***Ilustración 38.*** *Intensidad para la mortalidad e infracciones viales (ráster)*




**Fuente:** Elaboración propia

### 5.3 Estimación de modelos de regresión no espacial y aprendizaje automático

Las técnicas de aprendizaje automático, especialmente de aprendizaje supervisado, buscan predecir el valor de un atributo de salida a partir de un conjunto de atributos de entrada previamente establecidos; lo que resulta muy útil para desarrollar tareas de clasificación o regresión en aras de una predicción, en el caso del proyecto, espacial.

Dada la existencia de agrupaciones en el patrón espacial tanto de la mortalidad como de los predictores, se ajustaron algunos modelos para procesos puntuales que permitieron comparar las estimaciones realizadas. Estos permitieron tener un acercamiento para entender el comportamiento que tienen los modelos de la estadística clásica y los modelos predictivos avanzados de aprendizaje automático. En ellos, se aplicó el modelamiento espacial con las imágenes ráster escaladas y estandarizadas, y un método de validación cruzada como control en su entrenamiento.

Para adaptar los datos espacio-temporales a los modelos de aprendizaje, se implementó un proceso de conversión que permitió estructurar una matriz de características multivariadas a nivel de cuadrante geográfico (espacial) y por año (temporal). Cada observación representa un cuadrante por año, y se le asociaron variables independientes como la densidad de población, velocidad promedio, uso del suelo, presencia de infraestructura vial, entre otras. La variable dependiente corresponde al número de muertes por accidentes de tránsito registrado en dicho cuadrante durante el año correspondiente. Esta transformación permitió aplicar modelos estadísticos y de machine learning que requieren una estructura de datos tabular y sin correlación directa espacial o temporal no modelada.

La elección de los modelos permitió comparar distintos enfoques predictivos, desde los más clásicos, como

la regresión lineal, hasta los más avanzados basados en técnicas de aprendizaje automático. Esto permitió observar cómo varía el desempeño de los modelos al incorporar mayor complejidad y capturar la no linealidad presente en ciertos patrones espaciales específicos.

Se utilizaron tres (3) criterios principales para la selección de los modelos de análisis, los cuáles se detallan a continuación:

- Capacidad para manejar tanto variables continuas como datos de conteo: Se utilizaron modelos que pudieran procesar adecuadamente distintos tipos de variables, incluyendo cantidades absolutas (como el número de accidentes o infracciones) y datos continuos. Modelos como regresión de Poisson son apropiados para conteo, mientras que Random Forest o regresión lineal manejan variables continuas con eficacia.
- Nivel de interpretabilidad y aplicabilidad en contextos institucionales: Dado que los resultados pueden ser usados por entidades públicas para diseñar políticas de movilidad, se priorizó la inclusión de modelos que permitieran interpretar y comunicar claramente las relaciones entre variables, como la regresión lineal o Random Forest, este último por su capacidad de mostrar la importancia de las variables predictoras.
- Habilidad para capturar relaciones lineales y no lineales: Se incluyeron modelos que pudieran representar relaciones simples y complejas, ya que el comportamiento de los accidentes no siempre sigue patrones lineales. Por ello se combinaron modelos clásicos (como la regresión) con modelos más sofisticados (como redes neuronales o Gradient Boosting), capaces de descubrir relaciones ocultas en los datos.

La siguiente tabla muestra comparativamente los modelos de análisis utilizados para el desarrollo del proyecto, desde estadística clásica hasta machine learning.

***Tabla 11.*** *Comparativa de modelos de predicción espacial.*

| Modelo | Tipo de Modelo | Descripción | Aplicaciones Comunes |
|---|---|---|---|
| **Regresión Lineal (LM)** | Estadístico clásico | Modelo estadístico que estima la relación lineal entre una variable dependiente continua y una o más variables independientes mediante mínimos cuadrados. | Relaciones lineales, predicción económica |
| **Modelo Lineal Generalizado - Poisson** | Estadístico | Extensión de la regresión lineal para datos de conteo, donde la variable dependiente sigue una distribución de Poisson y se modela la tasa de ocurrencia de eventos. | Accidentes, número de eventos, epidemiología |
| **Bosques Aleatorios** Random Forest -RF | Ensamble / Árboles | Algoritmo de aprendizaje automático basado en árboles de decisión que combina múltiples árboles entrenados sobre subconjuntos aleatorios para mejorar la precisión y reducir el sobreajuste. | Clasificación, regresión, selección de variables |

| | | | |
|---|---|---|---|
| **Máquina de Soporte Vectorial** Support Vector Machine (SVM) | Aprendizaje supervisado | Modelo supervisado que encuentra el hiperplano óptimo que separa clases (o ajusta una función en regresión) maximizando el margen entre los datos más cercanos. | Clasificación, predicción con bordes no lineales |
| **Gradient Boosting (GBM)** | Ensamble / Árboles | Técnica de ensamble que construye modelos secuenciales, donde cada nuevo modelo corrige los errores del anterior, combinando sus predicciones para mejorar el rendimiento general. | Regresión compleja, competencias de ML |
| **Red Neuronal Simple (NNET)** | Aprendizaje profundo | Modelo inspirado en el cerebro humano, compuesto por capas de nodos conectados que transforman la entrada mediante funciones de activación no lineales, útil para capturar patrones complejos. | Series de tiempo, patrones complejos |

**Fuente:** Elaboración propia

### 5.4 Elección de parámetros e hiperparámetros

En los modelos de aprendizaje supervisado, los parámetros son aquellos valores internos que el algoritmo aprende directamente del proceso de entrenamiento, como los coeficientes en una regresión lineal. En cambio, los hiperparámetros son configuraciones externas al modelo que deben definirse antes del entrenamiento y que afectan directamente su desempeño, como la profundidad de los árboles, el número de iteraciones o el tipo de kernel.

Para este proyecto, el ajuste de hiperparámetros se realizó utilizando la función train() del paquete caret en R, la cual por defecto emplea una búsqueda basada en grilla (grid search) junto con validación cruzada de 5 particiones (k-fold CV con k=5), definida mediante trainControl(method = "cv", number = 5).

Además, para los modelos más sensibles al escalado como máquina de soporte vectorial y redes neuronales, se incorporaron técnicas de preprocesamiento de centrado y escalado mediante el argumento preProcess = c("center", "scale").

A continuación, se detallan los parámetros e hiperparámetros utilizados para los modelos:

#### 5.4.1. Regresión Lineal Clásica (lm)

- Parámetros aprendidos: Coeficientes de regresión.
- Hiperparámetros: Este modelo no tiene hiperparámetros ajustables en caret más allá del preprocesamiento y control de validación.

**5.4.2. Modelo Lineal Generalizado - Poisson (glm, family = "poisson")**

- Parámetros: Coeficientes del modelo.
- Hiperparámetros:
  - Familia de distribución: poisson.
  - No posee hiperparámetros adicionales ajustables vía caret.

**5.4.3. Modelo Lineal Generalizado - Gaussiano (glm, family = "gaussian")**

- Hiperparámetro: family = "gaussian".

**5.4.4 Random Forest (rf)**

- Hiperparámetros principales ajustados automáticamente por caret:
  - *mtry:* número de predictores considerados en cada división.
  - *ntree:* número de árboles (por defecto en randomForest es 500).
  - *nodesize:* tamaño mínimo de nodos terminales.

**5.4.5. Máquinas de Vectores de Soporte (SVM Radial) (svmRadial)**

- Preprocesamiento obligatorio: centrado y escalado.
- Hiperparámetros ajustados:
  - C: parámetro de penalización (cost).
  - sigma: ancho del kernel radial (RBF).

**5.4.6. Gradient Boosting Machines (gbm)**

- Hiperparámetros ajustados automáticamente:
  - n.trees: número de árboles.
  - interaction.depth: profundidad de los árboles.
  - shrinkage: tasa de aprendizaje.
  - n.minobsinnode: número mínimo de observaciones en nodos terminales.
  - verbose = FALSE se incluyó para controlar la salida en consola, no afecta el modelo.

**5.4.7. Red Neuronal Simple (nnet)**

- Preprocesamiento obligatorio: centrado y escalado.
- Hiperparámetros ajustados:
  - size: número de neuronas en la capa oculta.
  - decay: valor de regularización (peso del término de penalización).
  - linout = TRUE: modelo para regresión.
  - trace = FALSE: suprime la salida durante el entrenamiento.

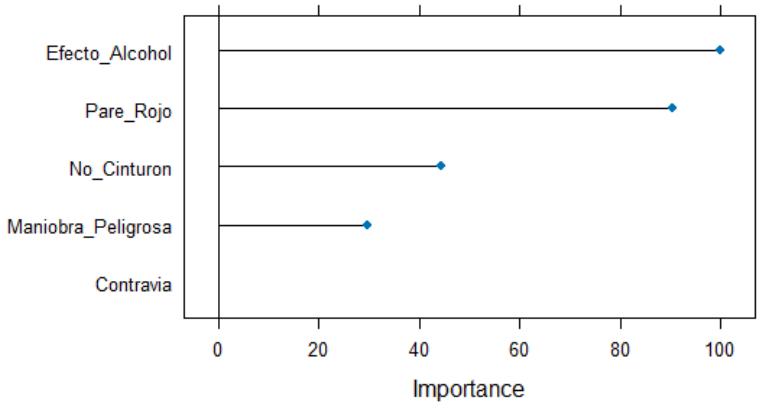
### 5.5 Resultados preliminares

Los resultados obtenidos para cada uno de los modelos mostraron un comportamiento similar, con buenos niveles de ajuste y desempeño satisfactorio en las métricas de evaluación aplicadas. No obstante, con fines de análisis preliminar e interpretación de resultados, se seleccionaron dos modelos representativos para examinar con mayor detalle la importancia relativa de las variables predictoras. A continuación, se presentan los hallazgos correspondientes:

#### 5.5.1 Modelo Lineal Generalizado de la Familia Poisson

El siguiente gráfico muestra qué tan influyente fue cada predictor en el modelo lineal generalizado de la familia Poisson. La métrica utilizada es la importancia relativa, basada en el valor absoluto de los coeficientes estandarizados. Para este modelo, las variables *conducir bajo los efectos del alcohol* y el *no obedecer una señal de pare o semáforo en rojo*, son los factores más peligrosos asociados a siniestros. El no uso del cinturón o la realización de maniobras peligrosas, también influyen en el resultado dado su impacto al modelo. Por su parte, la variable contravía no tiene un efecto significativo en la predicción de muertes en este contexto específico.

***Ilustración 39.*** *Importancia de las variables en modelo Poisson*




**Fuente:** Elaboración propia

A continuación, se puede observar el resumen de los resultados del coeficiente, observamos la influencia que tienen ciertas infracciones sobre la probabilidad de ocurrencia de un siniestro con desenlace fatal.
Algo importante hay que destacar es que un coeficiente positivo aumenta la tasa esperada del evento (accidente), y uno negativo la "disminuye" o infiere negativamente en el establecimiento de la relación lineal, es decir que actúan de acuerdo con su contexto.

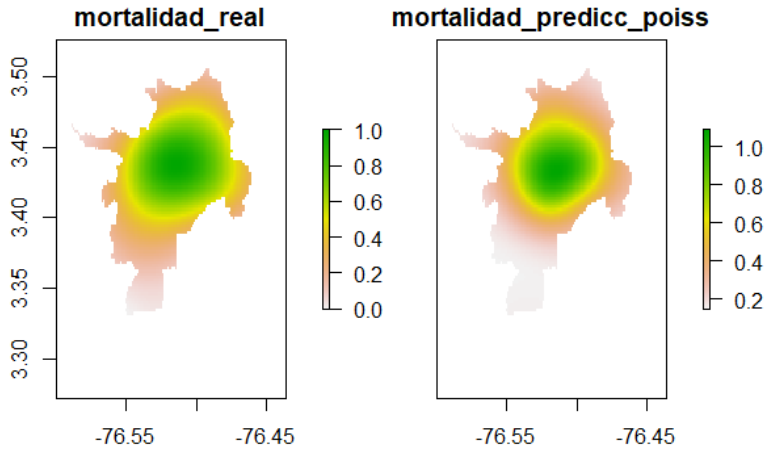
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.81644    0.08746 -20.770  < 2e-16 ***
Pare_Rojo         -1.75749    0.27995  -6.278 3.43e-10 ***
Efecto_Alcohol     2.28469    0.33154   6.891 5.53e-12 ***
Maniobra_Peligrosa -0.63251    0.27103  -2.334  0.01961 *
```

```
Contravia           -0.11576    0.27902  -0.415  0.67824
No_Cinturon          1.69853    0.51719   3.284  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 640.520  on 4153  degrees of freedom
Residual deviance:  61.996  on 4148  degrees of freedom
```

El modelo de Poisson estimó la incidencia de muertes en función de las variables seleccionadas. Se observó que tanto el intercepto como algunas variables, como la presencia de señal de pare o luz roja, las maniobras peligrosas y el conducir en contravía, presentaron coeficientes negativos y estadísticamente significativos, lo que sugiere que, en los cuadrantes con mayor intensidad de estos eventos, la mortalidad tiende a disminuir. En principio, este resultado podría parecer contraintuitivo y, por ello, debe interpretarse con cautela, considerando la necesidad de validaciones adicionales y el uso de datos más detallados y complementarios. No obstante, en el contexto del análisis espacial desarrollado en este proyecto, este comportamiento podría explicarse por la alta correlación que algunas variables mantienen con la mortalidad, así como por la limitada capacidad de los modelos lineales para capturar relaciones complejas en este tipo de datos.

Un acercamiento a un ejemplo visual, expresado en imágenes ráster, de la comparativa entre la mortalidad real y la predicción de acuerdo con este modelo, sería el de la siguiente ilustración. Aquí se observa como a nivel de intensidad espacial hay mayor distorsión de la imagen predicha con la real.

***Ilustración 40.*** *Predicción modelo Poisson*




**Fuente:** Elaboración propia

### 5.5.2 Modelo Bosques Aleatorios o Random Forest

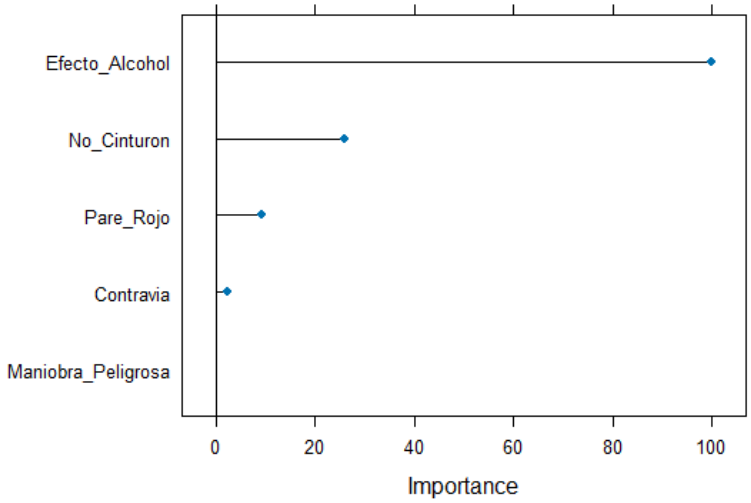
En el caso del modelo de Bosques Aleatorios, se identificó nuevamente la variable *conducir bajo los efectos*

*del alcohol* como el factor con mayor peso en la predicción de fatalidades viales, en concordancia con lo observado en el modelo de Poisson. Otras variables, como el *no uso del cinturón de seguridad*, *el incumplimiento de señales de pare o semáforo en rojo* y la *circulación vial en sentido contrario*, también contribuyen al modelo, aunque en menor magnitud. En contraste, la *realización de maniobras peligrosas* no presenta un efecto significativo en la predicción de muertes en este contexto específico.

Los resultados detallados del modelo se presentan a continuación en forma de cuadro y visualización gráfica.

| | Overall |
|---|---|
| Efecto_Alcohol | 100.000 |
| No_Cinturon | 25.923 |
| Pare_Rojo | 9.383 |
| Contravia | 2.196 |
| Maniobra_Peligrosa | 0.000 |

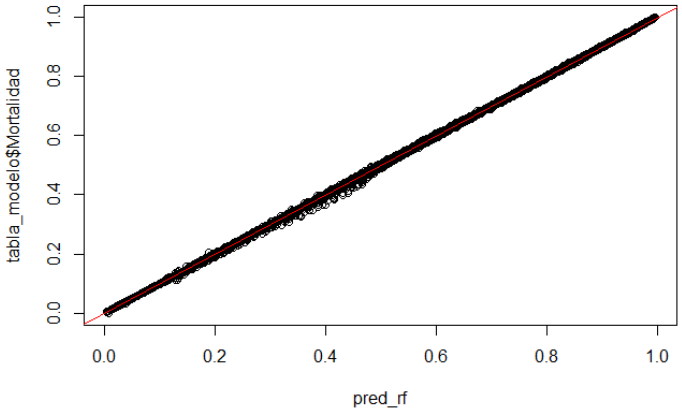
***Ilustración 41.*** *Importancia de las variables en modelo de Bosques Aleatorios*




**Fuente:** Elaboración propia

Adicionalmente, el siguiente gráfico ilustra el nivel de ajuste del modelo respecto a las predicciones realizadas. En la mayoría de los modelos de aprendizaje automático evaluados, se observa una correspondencia consistente con los valores reales de mortalidad, lo cual indica un bajo error de predicción y una ausencia significativa de sesgo sistemático.

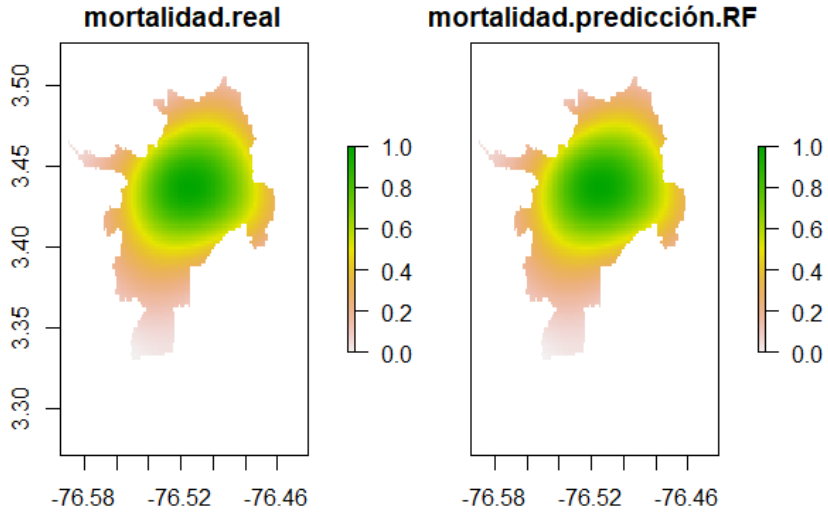
***Ilustración 42.*** *Ajuste del modelo de Bosques Aleatorios*




**Fuente:** Elaboración propia

Continuando con el acercamiento visual de la comparativa entre la mortalidad real y la predicción del modelo *random forest*, se puede observar en la siguiente imagen.

***Ilustración 43.*** *Predicción modelo Arboles Aleatorios RF*

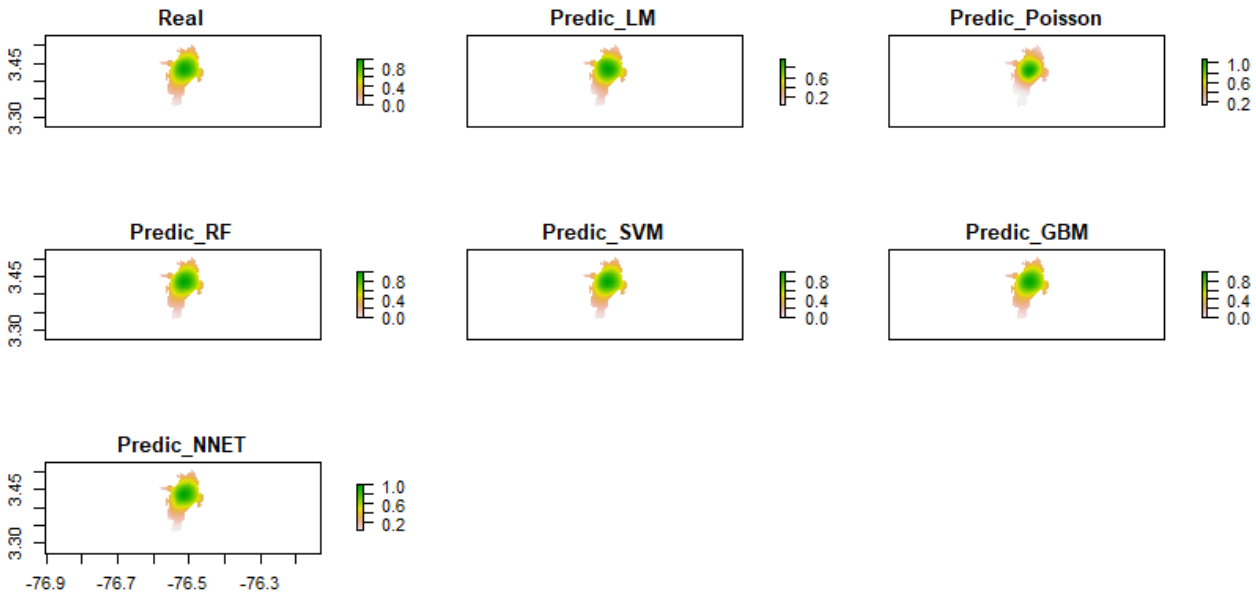



**Fuente:** Elaboración propia

En general, los resultados obtenidos muestran una alta consistencia entre la mayoría de los modelos evaluados. La infracción *conducir bajo los efectos del alcohol* se destaca como la más influyente, superando ampliamente al resto en términos de importancia predictiva. Otras variables, como *no respetar la señal de pare en rojo*, *no utilizar el cinturón de seguridad* y *circular en contravía*, también presentan una contribución significativa dentro del comportamiento estimado por los modelos.

Asimismo, se llevó a cabo una comparación visual preliminar de las imágenes ráster generadas por cada modelo, con el objetivo de evaluar el grado de aproximación de las predicciones respecto a los patrones espaciales de la mortalidad real observada en los siniestros viales. A continuación, se presentan los resultados obtenidos:

***Ilustración 44.*** *Comparación ráster de las predicciones de mortalidad por modelo*




**Fuente:** Elaboración propia

La visualización muestra que las predicciones de los modelos parecen ajustarse adecuadamente a la descripción del problema, faltando establecer con indicadores la certeza de estas predicciones.

### 5.6 Visualización de las predicciones

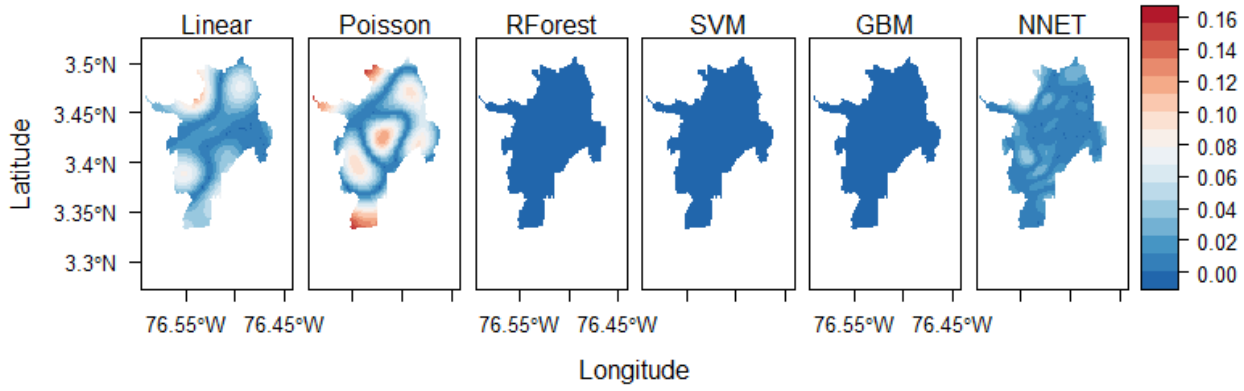
A continuación, se propuso una evaluación espacial del rendimiento de los modelos predictivos a partir de los archivos tipo ráster. La estimación del error espacial asociado a cada modelo se logra mediante el cálculo de la diferencia absoluta entre el mapa real y cada uno de los mapas generados por los modelos.

La siguiente imagen muestra los mapas de error absoluto entre las predicciones de los modelos evaluados (Lineal, Poisson, Bosques Aleatorios, Maquina de Soporte Vectorial y Gradient Boosting Machine, y Redes Neuronales) y un mapa real de referencia. El gradiente de colores indica la magnitud del error: tonalidades rojas representan errores más altos, mientras que los azules oscuros indican errores bajos o cercanos a cero.

Visualmente, los modelos que presentan un mejor desempeño son Random Forest, Suport Vector Maquine y Gradient Boosting Machine, ya que muestran superficies casi completamente azules, lo que sugiere una

alta precisión espacial y errores muy reducidos en todas las áreas geográficas. Por el contrario, el modelo que exhibe el peor desempeño es Poisson.

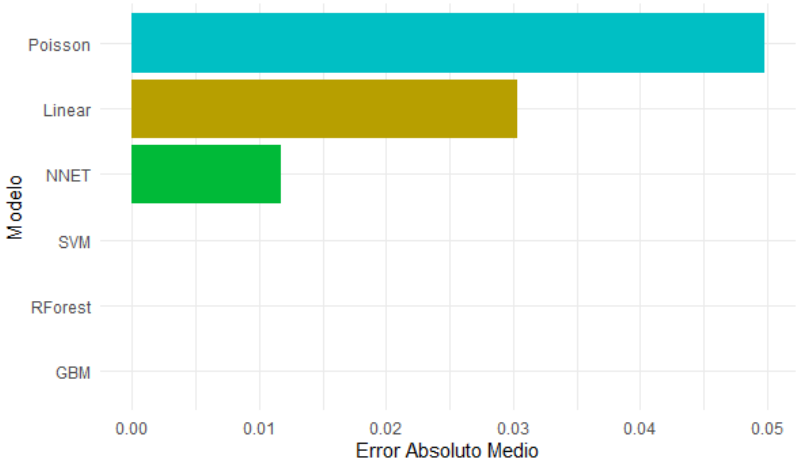
***Ilustración 45.*** *Visual del error absoluto por modelo*




**Fuente:** Elaboración propia

El gráfico presenta una comparación visual del Error Absoluto Medio (MAE) de los modelos evaluados, lo cual permite identificar de forma rápida cuáles ofrecen mejores aproximaciones a la mortalidad observada. Se observa que el modelo de regresión de Poisson es el que presenta el mayor error promedio, lo que indica menor precisión en sus predicciones. Le sigue el modelo lineal clásico, que también muestra un error relativamente alto. Por otro lado, modelos como la red neuronal (NNET), SVM, Random Forest y Gradient Boosting (GBM) exhiben errores absolutos considerablemente más bajos. Esto sugiere que estos modelos capturan mejor la complejidad de los datos y producen predicciones más cercanas a los valores reales de mortalidad.

***Ilustración 46.*** *Cálculo del error absoluto por modelo*




**Fuente:** Elaboración propia

# 6. EVALUACIÓN DEL MODELO

## 6.1 Métricas de evaluación

Para el desarrollo del presente proyecto aplicado, se tuvo en cuenta la ejecución de diferentes modelos estadísticos y de aprendizaje automático en un contexto espacial con el fin de predecir la mortalidad. Las métricas de evaluación fueron el *Error Cuadrático Medio* (RMSE), el *Error Absoluto Medio* (MAE) y el Coeficiente de determinación ($R^2$).

***Tabla 12.*** *Comparación de los modelos en cada métrica de evaluación.*

Evaluación de modelos

| Modelo | RMSE | R2 | MAE |
|---|---|---|---|
| Lineal | 0.038038383 | 0.9783687 | 0.030296050 |
| Poisson | 0.060547503 | 0.9463339 | 0.049773096 |
| Gaussian | 0.038038383 | 0.9783687 | 0.030296050 |
| Random Forest | 0.004218049 | 0.9997460 | 0.002711351 |
| Gradient Boosting | 0.026758268 | 0.9894383 | 0.019371777 |
| SVM | 0.016859803 | 0.9960506 | 0.014624100 |
| Red Neuronal | 0.015214299 | 0.9965887 | 0.012110845 |

**Fuente:** Elaboración propia

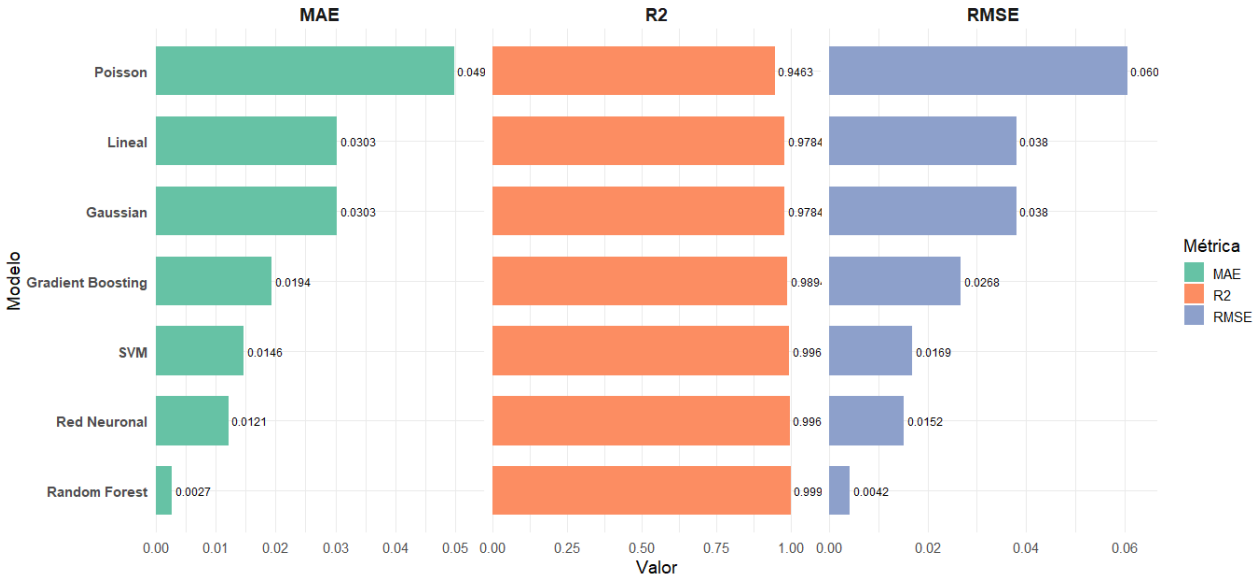
El Error Cuadrático Medio (RMSE) representa la magnitud promedio de la desviación entre los valores predichos por el modelo y los valores observados; por tanto, valores más bajos indican un mejor desempeño predictivo. En este sentido, el modelo *Random Forest* se destaca como el más preciso entre los evaluados, con un RMSE de 0.0042, lo que significa que, en promedio, sus predicciones se desvían solo 0.0042 unidades respecto a la mortalidad real. En contraste, el modelo *Poisson* presentó el RMSE más elevado, reflejando un menor ajuste.

Por otro lado, el coeficiente de determinación ($R^2$), que expresa la proporción de la variabilidad explicada por el modelo, permite evaluar su capacidad explicativa. Un valor cercano a 1 indica un alto poder predictivo, mientras que un valor cercano a 0 sugiere escasa utilidad. En los modelos analizados, nuevamente *Random Forest* obtuvo el mejor resultado, explicando el 99.67% de la variación en la mortalidad, mientras que el modelo *Poisson* mostró el menor ajuste, resultado esperable dada su estructura y la naturaleza estandarizada de los datos.

Finalmente, el Error Absoluto Medio (MAE) que refleja el promedio de los errores absolutos entre predicción y realidad, también señala al modelo *Random Forest* como el de mejor desempeño, con un MAE de 0.002711, lo cual refuerza su superioridad en términos de precisión y estabilidad predictiva.

A continuación, se presenta una visualización comparativa del desempeño de los modelos evaluados, ordenados del menor al mejor rendimiento y desempeño, según los valores obtenidos en las métricas de evaluación: RMSE, $R^2$ y MAE. Esta representación gráfica permite observar de manera integrada la efectividad relativa de cada modelo, facilitando la identificación de aquellos con mayor capacidad predictiva y un menor error.

***Ilustración 47.*** *Comparación visual de los modelos en las métricas de evaluación*




**Fuente:** Elaboración propia

La evaluación extendida de modelos presentada en la siguiente tabla incluye, además de las métricas tradicionales como el RMSE, el MAE y el $R^2$, otras métricas que permiten una comprensión más detallada del comportamiento de los errores. Entre estas se encuentran:

- El **MedAE (Error Absoluto Mediano)** que mide el valor central de los errores absolutos, menos sensible a valores atípicos que el MAE,
- El **MaxAE (Error Absoluto Máximo):** que representa el mayor error cometido por el modelo, útil para evaluar su comportamiento en casos extremos,
- El **MAPE (Error Porcentual Absoluto Medio)** que expresa el error relativo en porcentaje, facilitando la interpretación comparativa entre escalas y,
- El **Residual_SD (Desviación estándar de los residuales)** que cuantifica la dispersión de los errores en torno a la media, indicando la estabilidad del modelo.

En comparación con la evaluación de la visualización previa, los resultados se mantienen consistentes, confirmando que el modelo *Random Forest* presenta el mejor desempeño general con el menor RMSE,

MAE y desviación residual, así como el mayor coeficiente $R^2$. Sin embargo, la evaluación extendida permite observar que modelos como Red Neuronal (NNET) y SVM también ofrecen un rendimiento notable, con errores bajos y estabilidad aceptable, lo cual puede ser relevante en contextos donde se priorice la reducción de errores extremos o la consistencia en las predicciones.

***Tabla 13.*** *Evaluación extendida de los modelos.*

| Modelo | RMSE | MAE | MedAE | MaxAE | MAPE | R2 | Residual_SD |
|---|---|---|---|---|---|---|---|
| Lineal | 0.0380 | 0.0303 | 0.0255 | 0.1137 | | 0.9784 | 0.0380 |
| Poisson | 0.0605 | 0.0498 | 0.0461 | 0.1557 | | 0.9452 | 0.0606 |
| Gaussian | 0.0380 | 0.0303 | 0.0255 | 0.1137 | | 0.9784 | 0.0380 |
| RF | 0.0042 | 0.0027 | 0.0018 | 0.0357 | | 0.9997 | 0.0042 |
| GBM | 0.0268 | 0.0194 | 0.0134 | 0.1080 | | 0.9893 | 0.0268 |
| SVM | 0.0169 | 0.0146 | 0.0147 | 0.0639 | | 0.9958 | 0.0168 |
| NNet | 0.0152 | 0.0121 | 0.0109 | 0.0767 | | 0.9965 | 0.0151 |

**Fuente:** Elaboración propia

A partir de los resultados obtenidos en las distintas métricas de evaluación, se puede establecer que el modelo *Random Forest* demuestra un desempeño notablemente superior frente al resto de los modelos considerados. Este algoritmo exhibe los valores más bajos de error (RMSE, MAE, MedAE y MaxAE) y el coeficiente de determinación ($R^2$) más alto, lo que indica tanto una alta precisión en sus predicciones como una gran capacidad explicativa respecto a la variación de la mortalidad observada.

Este rendimiento puede atribuirse a las ventajas estructurales del modelo *Random Forest*, el cual no parte de supuestos de linealidad, permitiendo capturar relaciones complejas y no evidentes entre las variables. Además, es un modelo robusto frente a la colinealidad entre predictores y no requiere un preprocesamiento riguroso de los datos, como el escalado o la transformación de variables. Su capacidad para aprender interacciones de manera automática entre múltiples variables predictoras, sin necesidad de que sean especificadas de forma explícita, lo convierte en una herramienta poderosa para el modelado de fenómenos multivariados en contextos como la predicción espacial de la mortalidad.

***Tabla 14.*** *Observaciones finales de los modelos en relación a las métricas de evaluación.*

| Modelo | Observaciones finales. |
|---|---|
| Random Forest (RF) | Difícil de interpretar, pero excelente para predicción. |
| Red Neuronal (NNet) | Gran precisión, errores pequeños. Un poco más disperso en MaxAE |
| SVM | Buen equilibrio en todas las métricas, levemente menor que NNet |
| GBM | Modelo competitivo, simple y robusto. Algo menos preciso en MaxAE |

| | |
|---|---|
| Lineal | No capta relaciones complejas |
| Poisson | Inadecuado para variable continúa estandarizada. Peor rendimiento global. |

**Fuente:** Elaboración propia

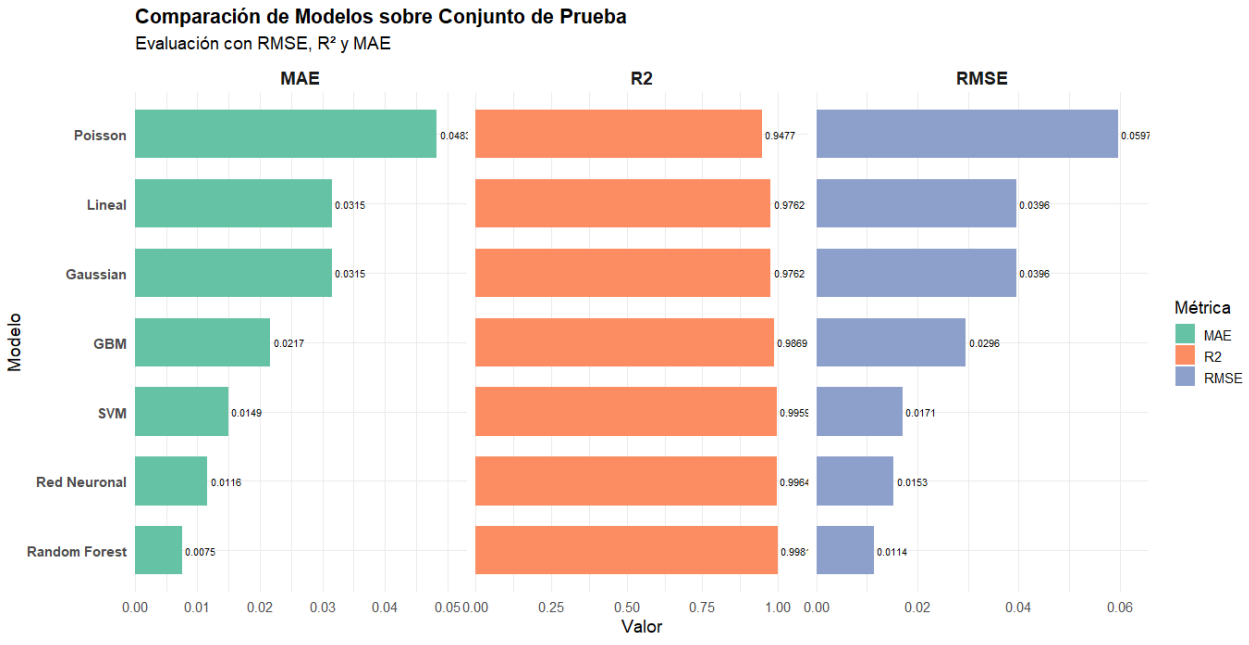
### 6.2 División de la base de datos en conjuntos de entrenamiento y prueba

A continuación, se realizó una partición del conjunto de datos en una proporción 80/20, asignando el 80% de los registros al conjunto de entrenamiento y el 20% restante al conjunto de prueba. Esta división permitió ajustar los modelos con una muestra amplia y evaluar su desempeño sobre datos no utilizados en el proceso de entrenamiento inicial.

Además, se implementó un esquema de validación cruzada de 5 pliegues (5-fold cross-validation) durante el entrenamiento, con el fin de reducir el riesgo de sobreajuste y obtener una estimación más robusta y generalizable del rendimiento de cada modelo. Este procedimiento consistió en dividir el conjunto de entrenamiento en cinco subconjuntos, entrenar el modelo en cuatro de ellos y validar en el quinto, repitiendo el proceso de forma rotativa.

Los resultados obtenidos mediante este enfoque fueron evaluados con las mismas métricas de desempeño previamente utilizadas (RMSE, MAE, $R^2$, entre otras).

***Ilustración 48.*** *Comparación de los modelos en cada métrica de evaluación (conjunto de prueba)*




**Fuente:** Elaboración propia

Los resultados obtenidos tras aplicar la evaluación extendida a los modelos sobre el conjunto de prueba permiten confirmar las tendencias observadas durante la etapa de entrenamiento. En términos generales,

los modelos basados en técnicas de aprendizaje automático continúan mostrando un desempeño superior frente a los modelos estadísticos tradicionales.

El modelo de Red Neuronal (NNET) se posiciona como el más preciso, al presentar el menor RMSE (0.0153), MAE (0.0116) y desviación estándar de residuales (0.0153), además de un alto coeficiente de determinación ($R^2 = 0.9964$), lo que indica un excelente ajuste general y una baja variabilidad en los errores. Le sigue de cerca el modelo SVM, con métricas también favorables y un $R^2$ de 0.9955.

El modelo de Random Forest (RF), que había sido destacado en la fase de entrenamiento, mantiene un buen desempeño, aunque se ve superado en esta etapa por NNET y SVM. No obstante, sigue destacando por su robustez, especialmente al considerar su bajo MAPE (0.0219) y buen control del error máximo absoluto.

Por el contrario, los modelos Poisson y lineales muestran un desempeño significativamente inferior. En particular, el modelo Poisson presenta los mayores errores absolutos y el valor más bajo de $R^2$ (0.9455), lo cual confirma sus limitaciones para capturar la complejidad del fenómeno modelado, posiblemente debido a la estructura de conteo y a la estandarización de los datos.

Evaluación extendida de modelos sobre conjunto de prueba

| Modelo | RMSE | MAE | MedAE | MaxAE | MAPE | R2 | Residual_SD |
|---|---|---|---|---|---|---|---|
| Lineal | 0.0396 | 0.0315 | 0.0270 | 0.1145 | 0.1660 | 0.9760 | 0.0396 |
| Poisson | 0.0597 | 0.0483 | 0.0435 | 0.1497 | 0.3365 | 0.9455 | 0.0597 |
| Gaussian | 0.0396 | 0.0315 | 0.0270 | 0.1145 | 0.1660 | 0.9760 | 0.0396 |
| RF | 0.0114 | 0.0075 | 0.0051 | 0.0723 | 0.0219 | 0.9980 | 0.0113 |
| GBM | 0.0296 | 0.0217 | 0.0157 | 0.1081 | 0.0813 | 0.9866 | 0.0295 |
| SVM | 0.0171 | 0.0149 | 0.0148 | 0.0570 | 0.0679 | 0.9955 | 0.0171 |
| NNet | 0.0153 | 0.0116 | 0.0091 | 0.0624 | 0.0626 | 0.9964 | 0.0153 |

**Fuente:** Elaboración propia

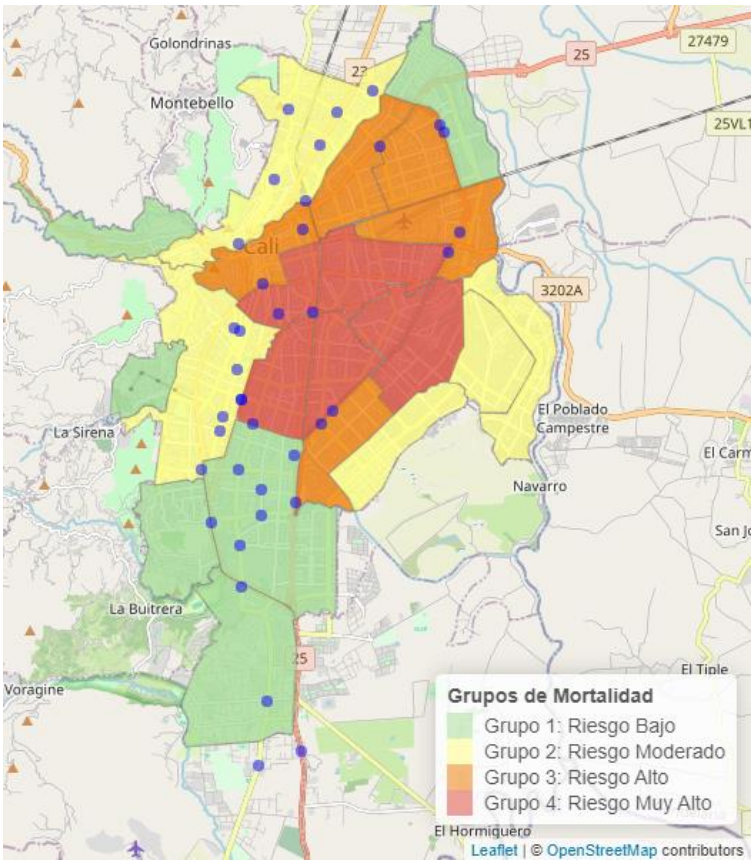
Un alto desempeño en la mayoría de los modelos, incluso con niveles de explicabilidad superiores al 90% ($R^2$), puede resultar sorprendente e incluso generar sospechas. Sin embargo, es importante considerar la naturaleza de la variable dependiente, la mortalidad, la cual no presenta gran variabilidad ni dispersión. Su intensidad se calcula mediante un conteo espacialmente agregado, lo que implica una alta frecuencia de eventos por cuadrante. Esto permite suponer que dicho tipo de variable tiende a exhibir patrones definidos y una alta correlación con sus variables predictoras, lo cual facilita que los modelos capturen adecuadamente la variabilidad. Además, esta alta capacidad de ajuste se ve favorecida por el uso de modelos no lineales o no paramétricos, capaces de representar relaciones complejas e interacciones implícitas, especialmente cuando se acompaña de un adecuado escalamiento y preprocesamiento espacial.

También podría pensarse que los modelos presentan sobreajuste. Sin embargo, la división del conjunto de datos en entrenamiento (80%) y prueba (20%), la aplicación de validación cruzada espacial, la evaluación del desempeño en los datos de prueba, así como el uso de múltiples métricas de evaluación, permiten confiar en la robustez de los resultados obtenidos. La adecuada selección de variables contribuye a una predicción más precisa. Esto no implica que el modelo sea infalible o mágico, sino que el fenómeno en estudio presenta una estructura subyacente suficientemente determinable.

### 6.3 Predicción espacial de las zonas de riesgo mortal

La representación cartográfica obtenida mediante el modelo predictivo seleccionado (*Random Forest*) permitió identificar espacialmente los niveles esperados de mortalidad en las distintas comunas de la ciudad de Cali, en superposición con los puntos azules que indican las cámaras de foto detección. A través de la clasificación en cuatro grupos de riesgo (bajo, moderado, alto y muy alto), el mapa evidenció un patrón geográfico diferenciado en la distribución de la mortalidad predicha. Las comunas clasificadas en los grupos de mayor riesgo se localizan principalmente en el área central y suroriental de la ciudad, mientras que las zonas periféricas tienden a concentrarse en los grupos de menor riesgo.

***Ilustración 49.*** *Categorización del riesgo de mortalidad*




**Fuente:** Elaboración propia

# 7. CONCLUSIONES Y TRABAJOS FUTUROS

## 7.1 Conclusiones

En los procesos puntuales espaciales, el hecho de predecir donde va a ocurrir el próximo accidente mortal, a veces es poco intuitivo y muy complejo de desarrollar. Se pueden estimar métricas y realizar validaciones cruzadas de los datos, sin embargo, dar un número exacto resulta ser complicado. Un primer acercamiento a esta problemática se puede realizar con la predicción de aquellas zonas de riesgo de que ocurra un evento, categorizándolas con indicadores de mayor o menor probabilidad, pues los accidentes, por naturaleza de orden vial, siempre tienden a estar aglomerados. Por ello, es importante poder identificar aquellas áreas más propensas a realizar una intervención de tránsito y movilidad.

Los hallazgos del documento refuerzan la ventaja de los modelos no paramétricos y basados en aprendizaje automático para tareas de predicción como la estimación espacial de la mortalidad, donde es probable que existan relaciones no lineales e interacciones no explícitas entre las variables predictoras. El análisis de patrones puntuales espaciales puede establecerse como una metodología importante en la gestión y evaluación del riesgo en zonas de alta intensidad de siniestros, ya que además de identificar los lugares de ocurrencia también se logra comparar con covariables espaciales relacionadas a este problema.

El proyecto permitió descubrir que, para el área espacial y los años de estudio, la variable *conducir bajo los efectos del alcohol y no respetar la luz roja en un semáforo o una señal de pare, son* las variables que mayor significancia tienen en los modelos analizados, especialmente con el modelo de *Bosques Aleatorios*. Además, los resultados mostraron que las comunas 8, 9, 10, 11, 12 y 13 tienen un mayor riesgo de siniestros con desenlace de víctimas fatales. No obstante, el accidente mortal muchas veces no es inmediato, puede existir algún tiempo de hospitalización o rezago, que indique un reporte de fatalidad mucho tiempo después del evento. Por ello, antes que vigilar la mortalidad, es necesario priorizar las zonas de mayor consumo de licor y riesgo de manejar en esos estados. Algo llamado como un “comportamiento espacial para zonas potencialmente peligrosas en siniestros viales”, donde se logren orientar decisiones en salud pública, priorizando las intervenciones focalizadas y diseñando estrategias preventivas que respondan a las condiciones específicas de cada sector urbano.

La idea es monitorear para entender cuáles son los detonantes más inmediatos, antes de que ocurra el accidente mortal. Tomar las infracciones que, puedan brindar la inmediatez para hacer seguimiento e intervención al evento de manera efectiva, conociendo que después de ese tipo de infracciones puede concluir en un desenlace fatal. Lo anterior, podría estar configurado con la integración de las cámaras de foto detección actual, cuyo monitoreo en tiempo real a infracciones como exceso de velocidad o pare en rojo, brinden a las fuerzas de orden público una herramienta para la optimización de recursos en seguridad vial, que reduzca la incidencia y gravedad de los accidentes, salvando vidas y reduciendo costos económicos. Sin embargo, los resultados evidenciaron que estas cámaras de foto detección no están actualmente relacionadas con aquellas zonas de alto y muy alto riesgo, como lo son las zonas del centro oriente de la ciudad.

## 7.2 **Trabajos futuros**

El proyecto desarrollado queda abierto para diferentes trabajos futuros de alto impacto tanto en lo social, económico y el impacto en la siniestralidad vial que al final se traduce en salvar vidas. Los próximos trabajos podrían tener como principio los siguientes puntos:

1. En seguridad vial se habla constantemente de los riesgos viales y la búsqueda permanente y dinámica de la identificación de los riesgos. Al ser este un factor dinámico se puede construir un modelo espacio temporal que se actualice en tiempo real con la información recolectada de las infracciones y que permita predecir los accidentes y las acciones preventivas que se pueden implementar.

2. Integración con la información que ya se obtiene en tecnologías como las mismas cámaras de fotomultas u otras tecnologías que permiten identificar infracciones como omitir una señal de semáforo o manejo peligroso, al hacer esta integración se podría tener información automatizada y relacionada con el punto anterior al permitir actualización dinámica de los factores de riesgo y mitigar la ocurrencia de accidentes.

3. Intervenciones en zonas críticas al tener información en tiempo real, permite ejecutar actividades de prevención y control por parte de las autoridades competentes haciendo una previa clasificación de las infracciones que muestran una mayor relación con las fatalidades.

4. Generación de mapas de calor interactivos y que puedan ser socializados con todos los usuarios, campañas de comunicación que genere conciencia tanto en el punto como en el tipo de infracción buscando evitar que existan accidentes mortales.

5. Al implementar estos modelos de predicción, los mismos con una metodología de registro de infracciones y la respectiva localización puede ser fácilmente implementado en cualquier ciudad y/o municipio de Colombia y aportar en los objetivos de la ONU en cuando a la disminución de muertes producto de accidentes de tránsito fatales.

6. Finalmente, se propone a las entidades públicas encargadas de la movilidad y el tránsito adoptar un enfoque basado en la gobernanza de datos para el análisis de información. Este modelo permitiría definir desde el inicio los objetivos de la recolección de datos, facilitando así su aprovechamiento estratégico y maximizando su utilidad en la formulación de políticas y toma de decisiones.

# 8. REFERENCIAS BIBLIOGRÁFICAS


[1] A. F. Burlacu, «Las muertes por siniestros de tránsito en el mundo siguen siendo inaceptablemente altas, y repensar la movilidad es la única solución,» Banco Mundial Blogs, 16 05 2023. [En línea]. Available: https://blogs.worldbank.org/es/voices/las-muertes-por-accidentes-de-transito-en-el-mundo-siguen-siendo-inaceptablemente-altas.

[2] Medicina Legal. Gobierno de Colombia, «Lesiones Accidentales,» Centro Referencia Nacional Sobre Violencia, [En línea]. Available: https://www.medicinalegal.gov.co/documents/20143/49222/Muertes+Transito.pdf.

[3] Gobierno de Colombia, «Observatorio Estadísticas. Cifras año en curso,» Agencia Nacional de Seguridad Vial, [En línea]. Available: https://www.ansv.gov.co/es/observatorio/estad%C3%ADsticas/cifras-ano-en-curso.

[4] Organización Panamericana de la Salud, «La velocidad y los siniestros viales,» 08 05 2017. [En línea]. Available: https://www.paho.org/es/documentos/hoja-informativa-velocidad-siniestros-viales.

[5] Organización Mundial de la salud, «Traumatismos causados por el tránsito,» Organización Mundial de la salud, 13 12 2023. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries.

[6] G. Maycock y R. D. Hall, «Accidents at 4-ARM Roundabouts,» The National Academies of Sciences, Engineering, and Medicine, US, 30 09 1985. [En línea]. Available: https://trid.trb.org/View/214685.

[7] O. K. Arndt y R. J. Troutbeck, «Relationship between roundabout geometry and accidents rates,» The National Academies of Sciences, Engineering, and Medicine, US, 12 06 2000. [En línea]. Available: https://onlinepubs.trb.org/onlinepubs/circulars/ec003/ch28.pdf.

[8] D. Hussein y G. Rose, «Development and evaluation of neural network freeway incident detection models using field data,» Transportation Research Part C: Emerging Technologies, Volume 5, Issue 5, Pages 313-331, 31 07 1997. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X97000168.

[9] National Academies of Sciences, Engineering, and Medicine, US, «Roundabouts in the United States,» The National Academies Press, 2007. [En línea]. Available: https://doi.org/10.17226/23216.

[10] C. J. de Naurois, C. Bourdin , A. Stratulat , E. Diaz y J.-L. Vercher , «Detection and prediction of driver drowsiness using artificial neural network models,» Accident Analysis & Prevention, Volume 126, Pages 95-104, 06 12 2017. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0001457517304347.

[11] J. L. Rubio Martín, «Modelos de predicción de accidentes,» Glorietas.com, [En línea]. Available: https://glorietas.com/glorietas/seguridad/modelos-de-prediccion-de-accidentes/.

[12] M. Bohorquez, «Estadística espacial y espacio-temporal para campos aleatorios escalares y funcionales. [Notas de clase.],» 04 11 2024. [En línea]. Available: https://drive.google.com/file/d/1dg09eClOJeCmqMvwDG9fylfuUTU_gXfX/view. [Último acceso: 2025].

[13] Y. B. Caballero Pérez, «Test de aleatoriedad para procesos puntuales espaciales basado en el cálculo de la dimensión fractal,» Repositorio Universidad Nacional, 05 2017. [En línea]. Available: https://repositorio.unal.edu.co/bitstream/handle/unal/62197/52901124_2017.pdf;jsessionid=5B9A742E456C3CBE8288D8B14BABF3F0?sequence=1 . [Último acceso: 2025].

[14] J. Valero Zorraquino, «Una breve introducción a la simulación de patrones puntuales espaciales,» Facultad de Ciencias, Universidad de Granada, 09 2021. [En línea]. Available: https://masteres.ugr.es/estadistica-aplicada/sites/master/moea/public/inline-files/TFM_ValeroZorraquino.pdf. [Último acceso: 2025].

[15] L. D. Ramírez Sánchez y W. Pineda Rios, «Modelos de procesos puntuales para la identificación espacial de los sismos ocurridos en Colombia,» 2019. [En línea]. Available: https://repository.usta.edu.co/server/api/core/bitstreams/8854874a-c117-4dc5-954f-f60c769c0862/content. [Último acceso: 12 05 2025].

[16] A. Baddeley, «Analysing spatial point patterns in R [Version 4.1],» CSIRO and University of Western Australia, 12 2010. [En línea]. Available: https://research.csiro.au/software/wp-content/uploads/sites/6/2015/02/Rspatialcourse_CMIS_PDF-Standard.pdf. [Último acceso: 2025].

[17] G. D. Buzai y E. Montes Galbán, «Estadística espacial:Fundamentos y aplicación con sistemas de información geográfica,» Universidad Nacional de Luján, 2021. [En línea]. Available: https://geofabio.com/wp-content/uploads/2022/03/2021_buzai_montes-galban_espacialidades_9_v1_1.pdf. [Último acceso: 2025].

[18] P. Moraga, «Spatial Statistics for Data Science: Theory and Practice with R,» Chapman & Hall/CRC Data Science Series, 2023. [En línea]. Available: https://www.paulamoraga.com/book-spatial/index.html. [Último acceso: 2025].

[19] R. Bivand y J. Nowosad, «CRAN Task View: Analysis of Spatial Data. Version 2025-05-03,» 03 05 2025. [En línea]. Available: https://cran.r-project.org/web/views/Spatial.html. [Último acceso: 2025].

[20] E. Pebesma y R. Bivand, «Spatial Data Science: With Applications in R (1st ed.),» Chapman and Hall/CRC, 2023. [En línea]. Available: https://r-spatial.org/book/. [Último acceso: 2025].

[21] A. Shobha y S. Rangaswamy, «Machine Learning,» de *Handbook of Statistics*, Elsevier, 2018, pp. 197-228.

[22] J. Elhai y P. Calhoun, Statistical procedures for analyzing mental health services data, Psychiatry Research, 2008.

[23] C. E. Rasmussen y C. Williams, Gaussian Processes for Machine Learning, Cambridge, Massachusetts: MIT Press, 2006.

[24] K. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, Massachusetts: MIT Press, 2012.

[25] P. Dutta, S. Paul y A. Kumar, «Chapter 25 - Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19,» de *Electronic Devices, Circuits, and Systems for Biomedical Applications*, Academic Press, 2021.

[26] J. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Stanford,

CA: Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[27] T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York, NY: Springer, 2009 (2ª edición).

[28] N. Cristianini y J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambrigde: Cambridge University Press, 200.

[29] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Upper Saddle River, NJ: Pearson Education, 2009.

[30] I. Goodfellow, Y. Bengio y C. Aaron, Deep Learning, Cambridge, MA: MIT Press, 2016.

[31] Generación Automática de Modelos de Conocimiento S.L. (GAMCO), «Modelo predictivo. Concepto y definición,» Generación Automática de Modelos de Conocimiento S.L., 2021. [En línea]. Available: https://gamco.es/glosario/modelo-predictivo/. [Último acceso: 2025].

[32] R. Lovelace, J. Nowasad y J. Muenchow, «Geocomputation with R,» 28 04 2025. [En línea]. Available: https://r.geocompx.org/. [Último acceso: 2025].

[33] P. Schneider y F. Xhafa, «Anomaly Detection and Complex Event Processing over IoT Data Streams,» de *Chapter 3 - Anomaly detection: Concepts and methods*, Academic Press, 2022, pp. 49-66.

[34] D. Christie y S. P. Neill, «8.09 - Measuring and Observing the Ocean Renewable Energy Resource,» de *Comprehensive Renewable Energy (Second Edition)*, Elsevier, 2022, pp. 149-175.

[35] Minitab, LLC, «Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?,» Minitab, LLC, 2024. [En línea]. Available: https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit. [Último acceso: 11 05 2025].

[36] E. M. Ben Laoula , O. Elfahim, M. El Midaoui , M. Youssfi y O. Bouattane, «Traffic violations analysis: Identifying risky areas and common violations,» Heliyon, Volumen 9, Edición 9,, 2023. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S2405844023062667.

[37] Z.-H. Jiang, X.-G. Yang, T. Sun, T. Wang y Z. Yang, «Investigating the relationship between traffic violations and crashes at signalized intersections: An empirical study in China,» Journal of Advanced Transportation, 04 2021. [En línea]. Available: https://www.researchgate.net/publication/350929897_Investigating_the_Relationship_between_Traffic_Violations_and_Crashes_at_Signalized_Intersections_An_Empirical_Study_in_China.

[38] J. F. González y S. I. Prada, «Cámaras de fotodetección y accidentalidad vial. Evidencia para la ciudad de Cali,» Revista Desarrollo Y Sociedad, 1(77), 131-182, 01 07 2016. [En línea]. Available: https://revistas.uniandes.edu.co/index.php/dys/article/view/6689.

[39] H. E. Gómez Angulo, M. F. Osorio Henao y P. Plazas Albornoz, «Evaluación de causas probables de accidentes de tránsito emitida por el Organismo de Tránsito que opera en la doble calzada Buga- Tuluá entre los años 2017 y 2019,» Repositorio Institucional Unilibre, 2021. [En línea]. Available: https://repository.unilibre.edu.co/handle/10901/23491.

[40] J. C. Cardona Álvarez, «Modelo predictivo de zonas de riesgo espacio temporal de accidentes de tráfico en la ciudad de Manizales,» Repositorio Digital Universidad de Caldas, 13 07 2023. [En línea]. Available: https://repositorio.ucaldas.edu.co/handle/ucaldas/19537?show=full.

[41] J. Bao, P. Liu y S. V. Ukkusuri , «A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data,» Accident Analysis & Prevention, Volume 122, Pages 239-254, 01 2019. [En línea]. Available: https://www.sciencedirect.com/science/article/pii/S0001457518303877?via%3Dihub.

[42] D. A. Dávila García, «Diagnóstico de puntos críticos de siniestralidad vial en Santiago de Cali durante el periodo de 2016 al 2018 mediante herramientas SIG,» Repositório Institucional, Universidad de Manizales, 2022. [En línea]. Available: https://ridum.umanizales.edu.co/server/api/core/bitstreams/985adbf4-dcc6-4b24-a7f5-2c44c4a33668/content.

[43] Agencia Nacional de Seguridad Vial de Colombia (ANSV), «Cámaras de fotodetección autorizadas,» Agencia Nacional de Seguridad Vial de Colombia (ANSV), 2023. [En línea]. Available: https://fotodeteccion.ansv.gov.co/ubicaciones-aprobadas.html. [Último acceso: 01 03 2025].

[44] A. Getis, «Encyclopedia of Social Measurement, Páginas 627-632,» Kimberly Kempf-Leonard, 2005. [En línea]. Available: https://www.sciencedirect.com/topics/computer-science/spatial-pattern-analysis. [Último acceso: 20 04 2025].

[45] F. M. Giorgi, C. Carmine y D. Mercatelli, «The R language: An engine for bioinformatics and Data Science,» Life (Basel), 27 04 2022. [En línea]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9148156/. [Último acceso: 2025].

[46] R. Bivand y J. Nowosad, «spatstat: Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests,» 2025. [En línea]. Available: https://cran.r-project.org/web/packages/spatstat/index.html. [Último acceso: 2025].

[47] C. A. Engel, «Introduction to R,» 16 10 2023. [En línea]. Available: https://cengel.github.io/R-intro/. [Último acceso: 2025].

[48] R. J. Hijmans, «Spatial Data Science with R and “terra”,» 30 11 2023. [En línea]. Available: https://rspatial.org/index.html. [Último acceso: 2025].

[49] Python Software Foundation, «The Python Wiki - Spanish Latin,» 11 02 2021. [En línea]. Available: https://wiki.python.org/moin/SpanishLanguage. [Último acceso: 10 05 2025].

[50] K. J. Millman y M. Aivazis, «Python for Scientists and Engineers,» Computing in Science & Engineering, 03 2011. [En línea]. Available: https://www.computer.org/csdl/magazine/cs/2011/02/mcs2011020009/13rRUx0xPMx. [Último acceso: 10 05 2025].