PRIMERA ENTREGA SECCIONES 2, 3 Y 4 - SEGURIDAD VIAL EN BOGOTÁ

CAMILO ENRIQUE PEÑUELA ESPINOSA

PONTIFICIA UNIVERSIDAD JAVERIANA

ANALÍTICA DE DATOS

FERNEY MALDONADO LOPEZ

BOGOTÁ D.C. 2025

Tabla de contenido

SECCIÓN 2 - INTRODUCCIÓN

2.1. PANORAMA GLOBAL DE LA SINIESTRALIDAD VIAL

La seguridad vial constituye un reto global y local de alta relevancia. La Organización Mundial de la Salud (OMS) reporta que cada año mueren más de 1,35 millones de personas en siniestros viales, lo que equivale a una muerte cada 25 segundos y convierte a los accidentes de tránsito en la principal causa de muerte entre los 15 y 29 años (OMS, 2018). En Colombia, y en especial en Bogotá, la situación es crítica: en 2022 se registraron 536 muertes, un aumento del 16,5% frente al año anterior, con un accidente grave cada 41 minutos (Infobae, 2023). Estos datos evidencian que la problemática persiste a pesar de iniciativas internacionales como Visión Cero y los Decenios de Acción de la ONU.

Tradicionalmente, la gestión de seguridad vial ha sido reactiva, basada en reportes históricos para actuar después del accidente. Sin embargo, los avances en Big Data, inteligencia artificial y analítica de datos permiten un cambio de paradigma hacia la prevención, anticipando escenarios de riesgo y orientando políticas públicas más efectivas (Butt & Shafique, 2025). Los datos históricos de accidentalidad constituyen la base para identificar patrones, aunque su análisis aislado resulta insuficiente para prever nuevos eventos. Por ello, organismos internacionales promueven el aprovechamiento de fuentes múltiples como tránsito, clima, condiciones viales y comportamiento de conductores para alimentar modelos que alerten sobre riesgos inminentes.

Este proyecto busca aplicar técnicas descriptivas y predictivas a los siniestros viales en Bogotá, identificando factores de riesgo y zonas críticas. El objetivo es generar recomendaciones basadas en evidencia que reduzcan víctimas fatales y lesionados, combinando el rigor académico con la pertinencia social, y aportando insumos útiles para la toma de decisiones en la ciudad.

SECCIÓN 3 - TRABAJOS RELACIONADOS Y ESTUDIOS SIMILARES REALIZADOS

3.1. ENFOQUES TRADICIONALES VS. MODERNOS EN LA PREDICCIÓN DE ACCIDENTES

Las metodologías para modelar y predecir accidentes de tráfico pueden dividirse, grosso modo, en dos enfoques: los modelos estadísticos tradicionales y las técnicas modernas basadas en aprendizaje automático o inteligencia artificial. A continuación, se describen ambos, destacando sus características y diferencias.

3.2. MODELOS ESTADÍSTICOS CONVENCIONALES

En el campo de la ingeniería de transporte y la seguridad vial, desde hace décadas se emplean modelos estadísticos para estimar la frecuencia de siniestros en función de diversas variables. Estos modelos incluyen típicamente regresiones de tipo Poisson o binomial negativa para modelar la ocurrencia de choques en segmentos viales o intersecciones, así como modelos logísticos u Ordered Probit/Logit para analizar la gravedad de los accidentes (por ejemplo, probabilidad de que un accidente sea fatal o cause lesiones severas). Un ejemplo representativo es el conjunto de herramientas provisto por el Highway Safety Manual (HSM) de AASHTO, ampliamente utilizado en EE. UU., que ofrece funciones predictivas basadas en datos históricos para distintos tipos de vías y condiciones.

Estos enfoques tradicionales han sido valiosos para entender relaciones generales: por ejemplo, cómo el volumen de tráfico, el diseño geométrico de la vía o la presencia de ciertos elementos de seguridad correlacionan con la tasa de accidentes. En Chile, un estudio técnico resaltó la necesidad de contar con herramientas predictivas de accidentes para evaluar proyectos viales, proponiendo modelos basados en el método Bayesiano Empírico (Empirical Bayes). Dicho estudio combinó cinco años de datos de siniestros en 5 ciudades chilenas con información de las características físicas y operativas de vías (tramos e intersecciones) para desarrollar modelos que explicaran la ocurrencia de accidentes según diversos atributos. Tras probar numerosas formulaciones, lograron calibrar modelos separados para tramos e intersecciones urbanas, con ajustes estadísticos comparables a los reportados internacionalmente y sensibles a variables relevantes (p. ej., volumen de tráfico, tipo de control en intersección, etc.). Este enfoque clásico,

apoyado en fundamentos estadísticos, ha sido la base de muchas políticas de seguridad vial, permitiendo identificar puntos críticos (zonas con alta siniestralidad) y estimar los beneficios en reducción de accidentes al implementar ciertas mejoras.

No obstante, los modelos estadísticos tradicionales tienen limitaciones importantes. Suelen asumir relaciones lineales o predefinidas entre variables y resultados, y pueden pasar por alto la naturaleza compleja y no lineal de los factores que contribuyen a los accidentes. Por ejemplo, un modelo logístico puede identificar si la presencia de lluvia incrementa la probabilidad de accidentes graves, pero quizás no capture interacciones sutiles entre hora del día, tipo de vía y comportamiento del conductor simultáneamente. Asimismo, dependen de la calidad y exhaustividad de los datos oficiales: registros policiales, sanitarios o de seguros que en muchos países presentan subregistro o información limitada. Estos límites de representatividad y calidad de las bases de datos históricas deben ser tomados en cuenta al plantear análisis predictivos tradicionales, especialmente en regiones donde la recolección de datos de siniestros no es homogénea ni 100% confiable. En resumen, si bien los métodos convencionales aportan insights valiosos, a menudo carecen de capacidades predictivas robustas ante relaciones no lineales o big data, y su precisión puede verse comprometida por datos incompletos.

3.3. TÉCNICAS DE APRENDIZAJE AUTOMÁTICO E INTELIGENCIA ARTIFICIAL

El aprendizaje automático (machine learning, ML) y la inteligencia artificial (IA) han irrumpido con fuerza en el campo de la seguridad vial en años recientes, ofreciendo herramientas más flexibles y poderosas para la predicción de accidentes. A diferencia de los modelos paramétricos tradicionales, las técnicas de ML pueden descubrir patrones complejos en los datos sin asumir a priori una forma funcional específica. Esto resulta especialmente útil para capturar las relaciones no lineales y de alta dimensionalidad típicas de los factores de riesgo viales (interacciones entre condición climática, densidad de tráfico, tipo de vehículo, comportamiento humano, etc.).

Diversos algoritmos de ML se han aplicado a problemas de seguridad vial, desde métodos de clasificación supervisada (p. ej., árboles de decisión, bosques aleatorios, máquinas de vector soporte, redes neuronales artificiales, gradient boosting) hasta

enfoques no supervisados o de deep learning. En general, la literatura reporta que los enfoques basados en árboles de decisión y sus ensambles suelen brindar muy buen desempeño predictivo para datos de accidentalidad. Por ejemplo, estudios en distintos países han encontrado que modelos de Random Forest alcanzan precisiones del 73–75% al clasificar la severidad de accidentes (por ejemplo, distinguir accidentes fatales de no fatales). En un caso, al aplicar cuatro algoritmos de ML a datos de choques de motocicleta en Ghana, el modelo Random Forest obtuvo la mejor exactitud (≈73.9% de acierto). De manera similar, en Dubai se logró alrededor de 75% de precisión usando múltiples técnicas, destacando la importancia de factores como el exceso de velocidad y la imprudencia en la severidad de los accidentes. Algoritmos más sofisticados como XGBoost (eXtreme Gradient Boosting) también han mostrado resultados prometedores; un estudio comparativo en Riyadh (Arabia Saudita) reportó que XGBoost alcanzó un 95% de exactitud en la clasificación de severidad de accidentes en un dataset local de dos años. Aunque esta cifra tan alta puede deberse a las características específicas del conjunto de datos, demuestra el potencial de las técnicas de ensamble moderno. Igualmente, se han explorado redes neuronales profundas y modelos híbridos. Por ejemplo, investigadores en Reino Unido propusieron modelos de Deep Forest (un enfoque que combina bosques aleatorios en múltiples capas) para predecir la gravedad de accidentes usando datos de tráfico de Leeds, obteniendo mejoras frente a modelos tradicionales.

En cuanto a la predicción de la frecuencia o probabilidad de accidentes en cierto lugar y periodo, las técnicas de ML también se emplean para superar las limitaciones de los modelos estadísticos. Por ejemplo, algoritmos de clasificación pueden identificar si un segmento específico es de alto riesgo (accidentado) o bajo riesgo, o si en la siguiente hora/día habrá un accidente en cierta zona dada la información disponible. También se han utilizado modelos de series de tiempo y redes neuronales recurrentes para pronosticar el conteo de accidentes a futuro en una región, incorporando patrones estacionales o tendencias.

Una ventaja clave de la IA/ML es su capacidad para integrar gran variedad de fuentes de datos heterogéneas. Mientras que un modelo tradicional podría requerir un conjunto de

variables limitadas y limpias, un sistema de ML podría procesar datos masivos de muy distinta índole (historias de accidentes, datos meteorológicos, flujos vehiculares en tiempo real, incluso datos geoespaciales o de smartphones) para extraer señales de riesgo. Estas capacidades permiten desarrollar sistemas de alertas tempranas y de análisis predictivo en tiempo real, que serían impensables bajo el paradigma analítico clásico. En síntesis, los enfoques modernos basados en aprendizaje automático representan un salto cualitativo en la predicción de siniestros viales, al ofrecer mayor precisión predictiva y adaptabilidad. No obstante, también presentan desafíos, como la necesidad de grandes cantidades de datos de calidad, la interpretabilidad de los modelos y su integración en la toma de decisiones públicas.

3.4. VARIABLES CLAVE Y FUENTES DE DATOS PARA MODELOS PREDICTIVOS

Tanto en modelos tradicionales como en los de ML, la selección de variables y datos de entrada es crítica para un buen desempeño predictivo. Numerosos estudios han investigado qué factores son más determinantes en la ocurrencia y severidad de los accidentes. A continuación, se resumen algunas categorías de variables comúnmente empleadas y su relevancia:

- **Características de la vía e infraestructura:** Incluyen el tipo de vía (autopista, vía urbana, zona escolar, etc.), la geometría (curvas, pendientes), la presencia de intersecciones o rotondas, el estado de la vía, iluminación, señalización y elementos de seguridad (barreras, demarcación). Estas variables suelen ser fuertes predictores de la frecuencia de accidentes. Por ejemplo, un estudio encontró que la categoría de carretera (especialmente si es autopista) y la ausencia de iluminación nocturna fueron factores clave asociados a mayor severidad de accidentes. Asimismo, en modelos predictivos chilenos basados en el método Bayesiano Empírico, se evidenció sensibilidad a diversos atributos de infraestructura al explicar la ocurrencia de siniestros en intersecciones semaforizadas vs. no semaforizadas.
- **Volumen y condiciones de tráfico:** La intensidad de circulación (número de vehículos que pasan por una vía) es directamente proporcional a la exposición al

riesgo; por tanto, se integra en casi todos los modelos de predicción de accidentes. También se consideran la composición del tráfico (porcentaje de vehículos pesados, motocicletas, etc.) y factores como la congestión o velocidad promedio. Por ejemplo, un alto flujo vehicular mezclado (camiones, transporte público, autos, motos) suele correlacionarse con puntos de alta accidentalidad, denominados "puntos rojos" por autoridades de tránsito. El análisis de patrones horarios de accidentes en Bogotá mostró claramente que los siniestros siguen las horas pico de tráfico: con tres picos marcados en la mañana temprano, mediodía y la tarde-noche, coincidiendo con los horarios de mayor congestión.

- **Condiciones climáticas y ambientales:** Factores como la lluvia, temperatura, niebla, hielo o iluminación natural influyen notablemente en el riesgo. Varios modelos incorporan datos meteorológicos para mejorar sus predicciones. Por ejemplo, un modelo predictivo desarrollado para la ciudad de Bogotá incluyó variables climatológicas (temperatura y precipitación) para evaluar cómo el clima afecta la frecuencia de siniestros. En ese estudio, se encontró que las condiciones meteorológicas adversas podían aumentar la probabilidad de siniestros en ciertos lugares y momentos. Del mismo modo, en Buenos Aires se analizó la influencia de factores estacionales y climáticos (junto a factores demográficos) para explicar variaciones en la cantidad de accidentes diarios.
- **Factores humanos y del vehículo:** El comportamiento del conductor es quizá el factor más complejo de modelar. Variables indicativas de riesgo humano incluyen exceder la velocidad permitida, conducir bajo efectos del alcohol u otras sustancias, el uso del cinturón de seguridad, distracciones (ej. uso del celular) o historial de infracciones previas. Estudios han intentado cuantificar cómo la ocurrencia de ciertas infracciones se correlaciona con mayor probabilidad de accidente. Por ejemplo, investigaciones en España plantearon que las distracciones y las infracciones en la conducción son áreas de riesgo conductual que, combinadas con otros factores, pueden servir para predecir la accidentalidad y perfilar conductores de alto riesgo. Igualmente, variables como la edad y género del conductor, tipo de vehículo (automóvil, motocicleta, transporte de carga) o antigüedad del vehículo pueden ser relevantes. Un estudio con datos masivos de

Tailandia destacó que entre los predictores clave de accidentes graves figuraban el exceso de velocidad, el hecho de ser conductor hombre, y si el accidente ocurría de día o de noche con iluminación insuficiente.

- **Datos geoespaciales y del entorno:** La localización exacta de los siniestros permite incorporar variables espaciales: densidad de población en la zona, proximidad a centros comerciales o escuelas, características del entorno vial (por ejemplo, presencia de cruces peatonales, semáforos, bares en la zona -indicador indirecto de posible conducción bajo efectos del alcohol-, etc.). Con el auge de sistemas de información geográfica, es posible cruzar accidentes con mapas de infraestructura y actividades humanas para encontrar "hotspots" y factores contextuales de riesgo.
- **Fuentes de datos no tradicionales:** En años recientes, han cobrado importancia datos alternativos como los provenientes de redes sociales y aplicaciones móviles. Estos pueden servir tanto para alimentar modelos predictivos en tiempo casi real, como para mejorar la detección de incidentes. Un ejemplo notable es el uso de datos de la aplicación Waze o de la red social Twitter (ahora "X") para anticipar accidentes. Investigadores en Bogotá desarrollaron un modelo de IA que combinó reportes ciudadanos de Waze/Twitter con datos históricos de accidentalidad y clima para predecir accidentes viales en ubicaciones específicas de la ciudad. Estas plataformas proveen información muy rápida, reportada por testigos en tiempo real, lo cual complementa a los registros oficiales que suelen tener retrasos. La integración de estos datos permitió identificar lugares con alta ocurrencia de incidentes (p. ej. alrededores de ciertas estaciones de transporte) y demostrar el potencial de nuevas fuentes para mejorar la rapidez y precisión de la predicción.

Esto ha motivado colaboraciones entre instituciones (p. ej., observatorios viales, universidades, empresas tecnológicas) para recopilar y compartir datos con el fin de mejorar los modelos.

3.5. ESTUDIOS Y APLICACIONES EN DISTINTOS CONTEXTOS

A lo largo de la última década se ha acumulado un número creciente de estudios de caso, proyectos piloto y aplicaciones prácticas de modelos predictivos de accidentes de tránsito, tanto en América Latina como en otras regiones del mundo. A continuación, se presentan ejemplos representativos que ilustran cómo se han implementado y evaluado estos modelos en diversos contextos.

EXPERIENCIAS EN AMÉRICA LATINA (INCLUYENDO COLOMBIA)

En Colombia y otros países latinoamericanos, la aplicación de modelos predictivos de siniestralidad vial ha ido ganando terreno gradualmente, aunque con desafíos en cuanto a disponibilidad de datos. Un caso reciente en Bogotá (Colombia) fue liderado por investigadores de la Universidad Nacional de Colombia, quienes desarrollaron un modelo de predicción de accidentes apoyado en técnicas de aprendizaje profundo. Este modelo integró datos históricos de accidentalidad en la ciudad con información climatológica (temperatura, lluvia) y datos obtenidos de redes sociales de tráfico (Waze y Twitter) para predecir la probabilidad de ocurrencia de accidentes en ubicaciones y franjas horarias específicas. Los resultados de este trabajo identificaron varios "puntos rojos" en Bogotá – intersecciones con alta circulación de vehículos mixtos (particulares, transporte público, motos, carga) y también elevado flujo peatonal – donde la probabilidad de accidentes graves es significativamente alta. Ejemplos de estos puntos críticos incluyeron intersecciones muy transitadas como la Avenida Carrera 72 con Calle 6 (zona de estación de TransMilenio Marsella), la Autopista Sur con Calle 68 Sur, entre otras. Además, se corroboró que la distribución temporal de los accidentes en la ciudad refleja las horas pico de movilidad mencionadas (tres picos diarios). Este proyecto demostró el potencial práctico de los modelos predictivos: sus creadores señalan que la herramienta podría ayudar a las autoridades locales a mejorar la seguridad vial en Bogotá y optimizar la asignación de recursos de emergencia, concentrándose en las zonas y momentos de mayor riesgo. Cabe resaltar que esta iniciativa aprovechó fuentes de datos innovadoras (redes sociales) que complementan los datos oficiales, mostrando una vía viable para sortear la escasez de datos tradicionales.

Otro esfuerzo en Colombia es el desarrollo de modelos de accidentalidad para vías interurbanas. Por ejemplo, en la concesión Devinorte (carretera al norte de Bogotá), se ha explorado un modelo predictivo para víctimas no fatales, con el objetivo de caracterizar tramos peligrosos y proponer soluciones de ingeniería. También universidades colombianas han realizado análisis de minería de datos sobre accidentes en Bogotá y Cundinamarca, identificando variables con mayor incidencia en la gravedad de los accidentes (como el tipo de vehículo, el factor humano o la condición de la vía).

En Ecuador, se llevó a cabo un proyecto destacado a través de la comunidad AI Saturdays, enfocado en la ciudad de Guayaquil. Allí se buscó responder a la pregunta: "¿Se puede crear un sistema que permita prevenir accidentes graves y/o fatales en la ciudad?". Como objetivo general, el equipo planteó implementar un modelo de Machine Learning para la estimación temprana de accidentes de tránsito graves o fatales mediante el análisis de datos históricos de siniestros en Guayaquil. Se partió de un dataset proporcionado por la Agencia de Tránsito y Movilidad (ATM) local, con información detallada de accidentes entre 2018 y 2021. Tras limpiar y seleccionar variables relevantes (basándose en literatura científica sobre factores asociados a accidentes), se optó por desarrollar y comparar varios algoritmos: una máquina de soporte vectorial (SVM), un Random Forest y un modelo de Gradient Boosting. Estos modelos se entrenaron para clasificar la severidad del accidente (por ejemplo, sin lesionados vs. lesionado grave vs. fatal). Según reportan los autores, los algoritmos elegidos se encuentran entre los más adecuados para problemas de clasificación multiclase de este tipo. Los resultados mostraron que es factible predecir con cierta anticipación dónde es más probable que ocurra un accidente grave, visualizando estos pronósticos en un mapa interactivo web. Este proyecto, aunque de carácter experimental, sienta bases para que agencias de tránsito en Ecuador (ATM Guayaquil, Comisión de Tránsito del Ecuador, etc.) evalúen la incorporación de estas herramientas de IA en sus operaciones futuras.

En Argentina, una tesis de maestría (Universidad Torcuato Di Tella, 2022) abordó la predicción de siniestros viales en la Ciudad Autónoma de Buenos Aires (CABA). El objetivo fue explotar masivamente la información de accidentes ocurridos entre 2015 y 2018 en la ciudad, usando técnicas analíticas para extraer conocimiento y proponer

acciones preventivas. Como primer paso, se analizó cuáles factores climáticos, demográficos o estacionales podían incidir en la ocurrencia de accidentes. Luego, en la fase de modelado, se implementó un modelo de Random Forest para predecir la cantidad de siniestros por unidad de tiempo (día, hora, mes). Los resultados obtenidos permitieron identificar patrones útiles para las autoridades en la formulación de políticas de prevención. Este estudio resaltó la utilidad de los métodos de ML para pronosticar la siniestralidad a corto plazo en entornos urbanos complejos. No obstante, también señaló limitaciones importantes: la capacidad predictiva del modelo entrenado en CABA podría no ser directamente transferible a otras ciudades o periodos, dado que los resultados están ajustados a las características específicas de los datos locales. Esto subraya la necesidad de calibrar y validar cuidadosamente los modelos cuando se apliquen en diferentes contextos.

En otros países latinoamericanos encontramos esfuerzos similares. **En Perú**, por ejemplo, se han desarrollado modelos de predicción para tramos carreteros específicos, combinando análisis estadístico y técnicas de minería de datos para identificar los factores que más contribuyen a la siniestralidad en carreteras nacionales. **En Chile**, además del estudio mencionado con método Bayesiano, la Comisión Nacional de Seguridad de Tránsito (CONASET) ha apoyado investigaciones para incorporar predicciones de accidentes en la evaluación de proyectos viales. Incluso **en Cuba** y otros países del Caribe se han reportado aplicaciones de modelos predictivos en zonas urbanas (por ejemplo, la predicción de accidentes de tránsito en Guayaquil, Ecuador, también fue explorada con modelos de pronóstico de series de tiempo como Prophet, obteniendo errores de predicción razonables según un estudio cubano).

En síntesis, América Latina muestra un interés creciente en aprovechar modelos predictivos de siniestros viales. Si bien los casos pioneros suelen ser proyectos académicos o pilotos, están sentando el camino para que instituciones gubernamentales adopten estas metodologías. Un denominador común es la colaboración entre universidades, observatorios de seguridad vial y autoridades de transporte para compartir datos y conocimiento. No hay que obviar que persisten retos importantes en la región,

como la calidad de los datos disponibles y la continuidad de estas iniciativas a escala institucional.

ESTUDIOS INTERNACIONALES DESTACADOS

A nivel internacional (fuera de Latinoamérica), la literatura sobre modelos predictivos de accidentes es amplia y diversa. Podemos destacar varias líneas de trabajo y casos prácticos:

- **Comparativas de algoritmos de ML:** Numerosos estudios han comparado el desempeño de distintos algoritmos de machine learning para pronosticar la severidad o frecuencia de accidentes. Un ejemplo reciente proviene de Tailandia, donde investigadores analizaron un dataset de más de 112.000 accidentes ocurridos en cinco años, enfocándose en casos donde la culpa era del conductor. Evaluaron ocho algoritmos supervisados (entre ellos Árboles de Decisión, SVM, Random Forest, k-NN, Redes Neuronales, Naïve Bayes, Regresión Logística y Gradient Boosting) buscando cuál clasificaba mejor la severidad del accidente. Tras un riguroso preprocesamiento y balanceo de datos, hallaron que el Random Forest ofreció la mejor performance global en la tarea de clasificar accidentes fatales vs. no fatales, con un AUC ~0,77 y accuracy ~0,78. Identificaron además las variables más influyentes en las predicciones: tipo de vía (accidentes en autopistas), la conducta de exceso de velocidad, el momento del día (accidentes diurnos), la ausencia de iluminación nocturna y el género del conductor (conductor masculino) emergieron como factores de peso en la probabilidad de accidentes graves. Un hallazgo importante fue que, si bien el modelo clasificaba bien los casos no fatales, su capacidad para "atrapar" casos de accidentes fatales fue limitada (sensibilidad del ~20% para el caso fatal). Esto evidencia la dificultad de predecir eventos muy raros (las fatalidades son una fracción pequeña del total de accidentes) y sugiere que incluso con ML avanzados, sigue siendo desafiante anticipar los choques más severos debido a la complejidad multicausal de estos eventos. Los autores recomiendan explorar ingeniería de características más avanzada y técnicas de ensamble para mejorar la detección de casos fatales.

- **Modelos híbridos y enfoques innovadores:** Investigaciones en países desarrollados han ido más allá de aplicar algoritmos por separado, proponiendo enfoques híbridos o de múltiples etapas. Por ejemplo, en el Reino Unido se han utilizado sistemas de ensamble (como votación mayoritaria de múltiples modelos) para mejorar la robustez de la predicción de severidad de lesiones. En un estudio, la combinación de varios clasificadores mediante un esquema de voting ensemble logró el desempeño más alto comparado con cualquier modelo individual, aprovechando la fortaleza de cada algoritmo. Asimismo, algunos trabajos recientes incorporan técnicas de aprendizaje profundo: redes neuronales convolucionales aplicadas a imágenes de cámaras de tráfico para detectar condiciones peligrosas, o redes recurrentes para predecir accidentes inminentes en función de series temporales de velocidad y flujo vehicular. Un ejemplo innovador es el uso de datos de vehículos conectados y sensores de ciudad (Internet de las Cosas, IoT) que alimentan plataformas de ciudad inteligente: se analizan frenadas bruscas, activación de ABS, o llamadas telemáticas para identificar en tiempo real zonas y momentos de alto riesgo de accidente.
- **Enfoques en tiempo real y gestión dinámica**: En línea con lo anterior, en lugares como Estados Unidos y Europa se han desarrollado prototipos de sistemas de gestión de seguridad vial en tiempo real. Estos sistemas integran algoritmos predictivos con centros de control de tráfico, de modo que si el modelo indica una probabilidad inusualmente alta de accidente en cierto tramo durante la próxima hora (quizás debido a una tormenta súbita más pico de congestión), se puedan tomar medidas inmediatas: alertar a conductores vía paneles informativos, reducir los límites de velocidad temporalmente, o pre-posicionar equipos de emergencia. Por ejemplo, investigadores en Nueva Zelanda combinaron datos de flujo vehicular, clima y estado de la carretera para predecir la probabilidad horaria de accidentes en una autopista, logrando identificar correctamente muchos de los hotspots y periodos pico de siniestros. En otro caso, en España, se investigó la predicción de la lesividad (gravedad de las lesiones) en accidentes a partir de datos de la Red de Carreteras, utilizando técnicas de ML para estimar si un choque resultaría en heridos leves, graves o fallecidos. Estos estudios internacionales proporcionan un banco de conocimiento muy útil: muestran qué variables suelen ser universales (ej. velocidad, flujo, clima) y cuáles pueden ser

contextuales (p. ej., comportamiento local de conductores, características únicas de ciertas vías).

- **Integración con políticas y seguros:** Fuera del ámbito estrictamente académico, empresas de tecnología y aseguradoras también han incursionado en el uso de modelos predictivos de riesgo vial. Grandes aseguradoras emplean modelos de riesgo para fijar primas, considerando factores históricos de siniestralidad de perfiles de conductores y ubicaciones. Por su parte, compañías de análisis de datos promueven soluciones de Big Data para seguridad vial urbana. Por ejemplo, la empresa PREDIK Data-Driven señala que las técnicas de análisis masivo de datos e IA pueden eliminar la toma de decisiones "instintivas" en la gestión de riesgos, reemplazándolas por decisiones basadas en datos que permiten anticipar problemas y mitigar riesgos antes de que se materialicen. Si bien esa afirmación es general a distintos sectores, aplica perfectamente al campo de los accidentes de tráfico: con suficientes datos, es posible predecir dónde hay mayor probabilidad de choques y actuar preventivamente. Algunos gobiernos locales en países desarrollados han empezado a colaborar con empresas de análisis para desarrollar dashboards predictivos de seguridad vial, integrando datos de tránsito en tiempo real con históricos de accidentalidad para priorizar intervenciones (similar a lo que se busca con la plataforma ViaSegura en desarrollo por la Universidad de los Andes en Colombia, que combina IA para mapear elementos de seguridad vial en las vías y podría en el futuro integrar predicción).

SECCIÓN 4 - EXPLORACIÓN Y ENTENDIMIENTO DEL DATASET DE SINIESTROS VIALES EN BOGOTÁ

4.1. ESTRUCTURA DEL ARCHIVO DE DATOS

El conjunto de datos *"Siniestros Viales Consolidados Bogotá D.C."* se presenta en un archivo .xlsx con múltiples hojas de cálculo interrelacionadas, reflejando un diseño de esquema relacional. La clave primaria que enlaza todas las hojas es el *CODIGO_ACCIDENTE*, identificador único de cada siniestro vial registrado. A continuación, se describen las secciones principales del archivo y su contenido:

Hoja ***"SINIESTROS":*** Tabla principal que resume cada accidente de tránsito con atributos básicos. Incluye campos como *FECHA* (fecha del siniestro), *HORA*, *CLASE* (tipo de accidente), *GRAVEDAD* (severidad del accidente), *CODIGO_LOCALIDAD* (identificador de la localidad de Bogotá donde ocurrió), entre otros datos contextuales del suceso. Esta es la tabla central y cada registro corresponde a un siniestro vial específico, identificado por un *CODIGO_ACCIDENTE*.

Hojas ***"ACTOR_VIAL", "VEHICULOS" y "HIPOTESIS"***: Conjuntos de datos secundarios que detallan distintos aspectos de cada siniestro, vinculados a la tabla principal mediante *CODIGO_ACCIDENTE*. La hoja *ACTOR_VIAL* contiene información de las víctimas o involucrados (ej. peatón, conductor, pasajero) y sus características (edad, condición de lesionado o fallecido, etc.). La hoja *VEHICULOS* lista los vehículos implicados en cada accidente (tipo de vehículo, servicio, modelo, etc.). La hoja *HIPOTESIS* registra las causas probables o factores contribuyentes del accidente según el informe policial (por ejemplo, exceso de velocidad, no respetar señales, embriaguez, fallas mecánicas, etc.). Estos datos relacionales permiten un análisis detallado por actor vial, tipo de vehículo y causa asociada a cada evento.

Hoja ***"DICCIONARIO":*** Glosario de datos que define y codifica las variables categóricas presentes en las otras hojas. Aquí se encuentran las descripciones de códigos para campos como *CLASE* y *GRAVEDAD*, entre otros. Por ejemplo, los códigos numéricos de *CLASE* de accidente se traducen a categorías descriptivas (choque, atropello, volcamiento, etc.), y los códigos de *GRAVEDAD* especifican si el accidente tuvo solo daños materiales, lesionados o

víctimas fatales. Esta hoja de diccionario garantiza la correcta interpretación de los datos categóricos y asegura la consistencia semántica del análisis.

4.2. VARIABLES CLAVE DEL CONJUNTO DE DATOS

Para llevar a cabo un análisis de datos significativo es fundamental identificar las variables clave que describen cada siniestro vial y comprender su codificación. A continuación, se destacan los campos más importantes del conjunto de datos, junto con su significado y relevancia analítica:

- **FECHA / HORA:**
  - Formatos: *FECHA* en dd/mm/aaaa; *HORA* en HH:MM:SS (24 h).
  - Derivables: *AÑO, MES, DIA_SEMANA, HORA_NUM.*
  - Uso analítico: tendencias anuales y estacionales; identificación de horas pico y diferencias laboral vs. fin de semana; posibilidad de cruce con luz diurna/nocturna o clima.
- **CODIGO_LOCALIDAD / DIRECCION / DISEÑO_LUGAR**
  - Ubicación administrativa (20 localidades urbanas + Sumapaz).
  - No incluye coordenadas geográficas; permite análisis por tasas por localidad (al normalizar con población o parque automotor) y caracterización del diseño del lugar (intersección, tramo, etc.).
- **CLASE (tipo de siniestro)**
  - Codificada; mapea a: Choque, Atropello, Caída de ocupante, Volcamiento, Incendio, Autolesión, Otro.
  - Campos asociados: *CHOQUE* (contra vehículo/objeto fijo/semoviente) y *OBJETO_FIJO* (p. ej., muro, poste, árbol, semáforo).
  - Uso analítico: dinámica del evento; segmentación para severidad/causas; priorización de intervenciones por tipo de siniestro.

- **GRAVEDAD (severidad del siniestro)**
    - ❖ Codificada y mutuamente excluyente: 1 = Con muertos, 2 = Con heridos, 3 = Solo daños.
    - ❖ Complementos (si existen): conteos de Muertos y Heridos por accidente.
    - ❖ Uso analítico: proporciones fatal/lesionado/daños; tasa de letalidad por clase, actor, vehículo o causa.
- **ACTOR_VIAL (detalle de personas involucradas)**
    - ❖ Claves: *CODIGO_ACCIDENTE*, *CODIGO_ACCIDENTADO*.
    - ❖ Variables: *CONDICION* (conductor, peatón, pasajero, motociclista, etc.), *ESTADO* (ileso, herido, fallecido), *EDAD*, *SEXO*, vínculo a *VEHICULO*.
    - ❖ Uso analítico: perfil de usuarios vulnerables (peatones, ciclistas, motociclistas); víctimas por accidente; distribución por edad/sexo; severidad por rol.
- **VEHICULOS (detalle de automotores implicados)**
    - ❖ Variables: *VEHICULO* (id), *CLASE* (auto, moto, bus, camión, bici, etc.), *SERVICIO* (público/particular), *MODALIDAD*, *ENFUGA*.
    - ❖ Uso analítico: participación modal; severidad por tipo de vehículo; análisis de fuga; cruces actor–vehículo.
- **HIPOTESIS (causa probable)**
    - ❖ Variable: *CODIGO_CAUSA* → descripciones (p. ej., Exceso de velocidad, No respetar señales, Embriaguez, Falla mecánica, Imprudencia peatonal).
    - ❖ Uso analítico: ranking de factores de riesgo; cruce causa × clase × gravedad × actor/vehículo.
    - ❖ Consideración: es un registro preliminar del informe policial (puede tener subjetividad); útil como indicador de tendencia.

4.3. ESTADÍSTICAS DESCRIPTIVAS INICIALES

A partir de la consolidación de los datos, se procedió a calcular estadísticas básicas y visualizaciones para obtener una visión general de la siniestralidad vial en Bogotá durante el periodo 2015-2020. En total, el dataset contiene del orden de ~180 mil a 200 mil registros de siniestros ocurridos en ese lapso (un promedio de decenas de miles de accidentes por año). A continuación, se resumen los hallazgos principales en cuanto a tipos de accidentes, severidad de los mismos, distribución temporal y geográfica:

- **Tipos de accidentes más frecuentes:** La clase de siniestro con mayor incidencia es el choque entre vehículos. Este tipo representa la mayoría de los accidentes ocurridos. En contraste, los atropellos a peatones, si bien son menos comunes en número absoluto, constituyen una fracción importante, seguidos por las caídas de ocupante (principalmente de motocicletas). Otros tipos como volcamientos, incendios vehiculares y autolesiones son relativamente infrecuentes. En la *Figura 4.1* se observa la distribución de frecuencias por tipo de siniestro, evidenciando el claro predominio de los choques sobre las demás clases.

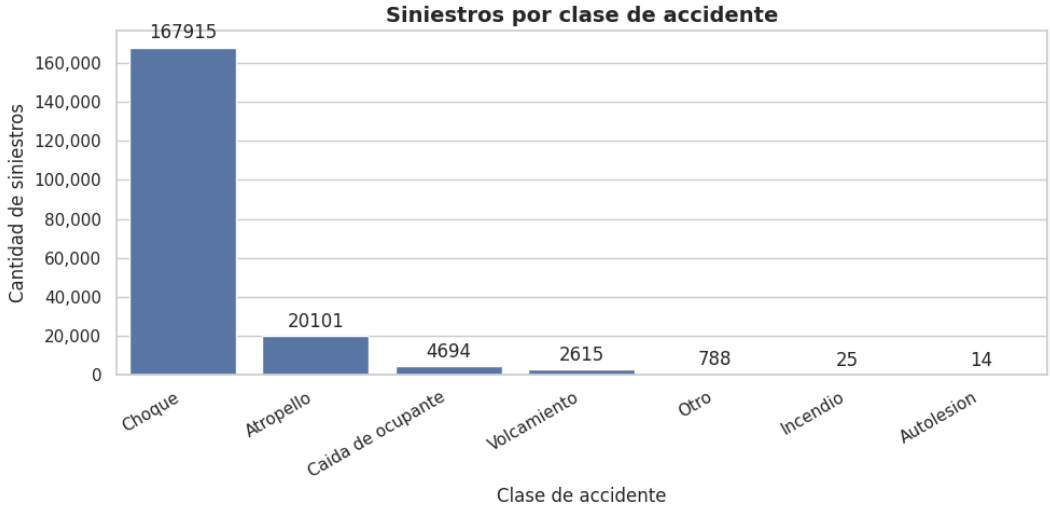

Figura 4.1. Distribución de siniestros por tipo de accidente en Bogotá D.C. (2015–2020).

Se aprecia que la categoría Choque abarca la mayor proporción de eventos reportados, muy por encima del resto. Los Atropellos constituyen el segundo tipo más común, mientras que caídas de ocupante (ejemplo típico: motociclistas que caen), volcamientos y otras clases suman porcentajes menores. Esta predominancia de colisiones vehiculares concuerda con la literatura de siniestralidad: por ejemplo, en

contextos urbanos similares se ha documentado que los choques representan cerca de la mitad o más de los incidentes viales. A pesar de ser menos frecuentes, los atropellos tienden a ser más graves en sus consecuencias, aspecto que se analiza al considerar la severidad.

- **Gravedad y víctimas de los siniestros:** Del total de accidentes registrados, la gran mayoría no involucraron víctimas fatales. En términos de proporción, alrededor de ~1–2% de los siniestros fueron fatales (con al menos un fallecido), aproximadamente 20–30% resultaron con lesionados, y el restante ~70–80% correspondieron a eventos de solo daños materiales (sin heridos) valores estimados coherentes con la información oficial de la ciudad. Esta distribución refleja que, aunque los accidentes fatales son relativamente raros, su reducción es prioritaria por el impacto en vidas humanas. Asimismo, al sumar todas las víctimas, se confirma que los peatones y motociclistas conforman un porcentaje sustancial de los lesionados y fallecidos, consistente con tendencias internacionales donde los usuarios vulnerables de la vía aportan una gran parte de las víctimas. Por ejemplo, en Bogotá se reportó que en 2023 más del 60% de las víctimas mortales eran motociclistas o peatones, lo que al extrapolar al periodo 2015-2020 sugiere una carga similar en esos grupos. En cuanto a las víctimas totales anuales, 2019 presentó uno de los números más altos de lesionados, mientras que 2020 mostró una caída pronunciada debido a las restricciones de movilidad durante la pandemia de COVID-19 (que redujeron el tráfico y, por ende, la siniestralidad).

- **Tendencias anuales:** La evolución temporal anual de los siniestros (ver *Figura 4.2*) muestra una tendencia al aumento en el número de accidentes desde 2015 hasta 2019, seguida por un descenso marcado en 2020. Entre 2015 y 2019 se observa un incremento paulatino (por ejemplo, de ~30 mil accidentes en 2015 a casi 35 mil en 2019, según el dataset). El año 2020 rompe la tendencia con una caída significativa en la siniestralidad total (se registraron muchos menos casos, del orden de ~22 mil) por las medidas de confinamiento y reducción de tránsito vehicular durante varios meses de ese año. Este comportamiento anómalo en 2020 coincide con reportes internacionales y locales sobre la disminución de accidentes durante la pandemia.

Cabe destacar que la reducción se dio principalmente en siniestros leves (colisiones menores), mientras que los patrones de accidentes graves no variaron tan drásticamente en proporción. En años posteriores (fuera del rango de este conjunto de datos) se espera un repunte conforme retornó la movilidad habitual, pero dichos datos no están incluidos en este consolidado.

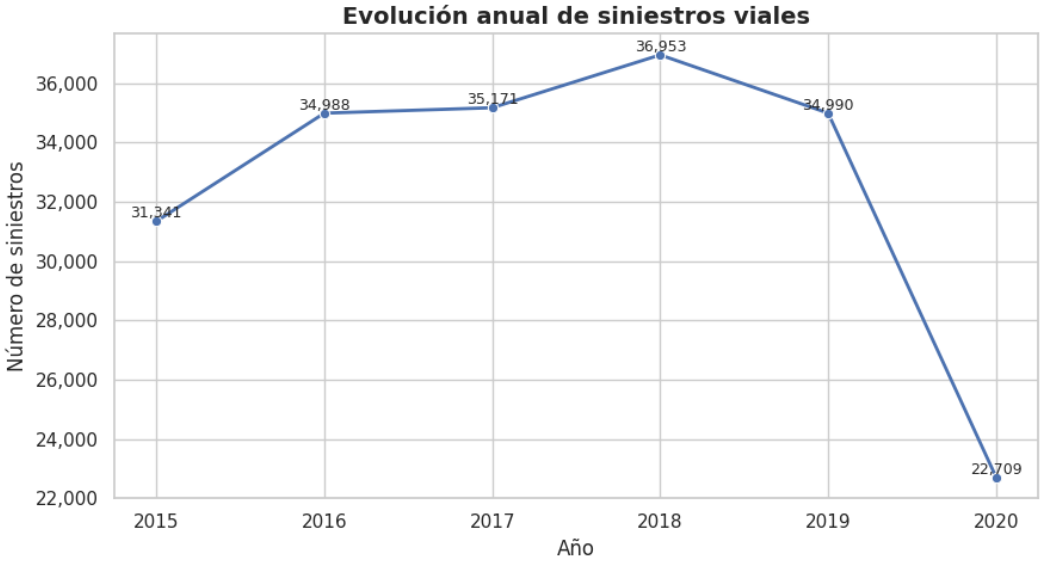

Figura 4.2. Evolución anual del número de siniestros viales en Bogotá D.C. (2015–2020).

La serie temporal evidencia un aumento sostenido de los accidentes año tras año hasta 2018 (línea ascendente en color azul), seguido de una caída abrupta en 2020. Este descenso coincide temporalmente con la emergencia sanitaria por COVID-19, que implicó reducciones en la circulación vehicular y en la exposición al riesgo, resultando en menos accidentes reportados. El comportamiento hasta 2019 sugiere que, en ausencia de intervenciones fuertes, la siniestralidad tendía al alza, por lo que las políticas de *Visión Cero* y otros programas de seguridad vial implementados a finales de la década de 2010 buscaban contrarrestar esta tendencia. Los datos posteriores a 2020 deberán analizarse para comprobar si el decrecimiento se mantuvo o si la siniestralidad retornó a niveles previos una vez normalizada la movilidad.

- **Huella diaria: perfil horario típico (mediana + banda P10–P90):** La curva de mediana confirma un patrón diario muy estable: entre 0:00–4:00 cada hora aporta apenas ≈1–2% del total diario; desde 5:00–7:00 hay un ascenso brusco (arranque de la movilidad), y entre 10:00–16:00 se observa una meseta donde cada hora concentra

≈5–6.5% del día, con máximo alrededor de 14:00. A partir de 19:00 el peso horario desciende de forma sostenida hasta ≈2% a las 23:00. Como las cifras están normalizadas por día, el perfil no está sesgado por días "grandes" o por 2020: lo que vemos es la forma promedio del día en Bogotá, independientemente del volumen.

La banda P10–P90 evidencia variabilidad muy baja de madrugada y más amplia en la meseta diurna, lo que sugiere que los picos de la mañana y la tarde ocurren prácticamente todos los días, pero su magnitud relativa fluctúa según el contexto (clima, congestión, eventos). En términos operativos, esta huella habilita metas realistas: si una intervención afirma "aplanar" picos, debería reducir de forma consistente la mediana (no solo casos extremos) y estrechar el P10–P90; si solo cae la mediana sin cerrar la banda, el sistema se vuelve más impredecible, no más seguro.

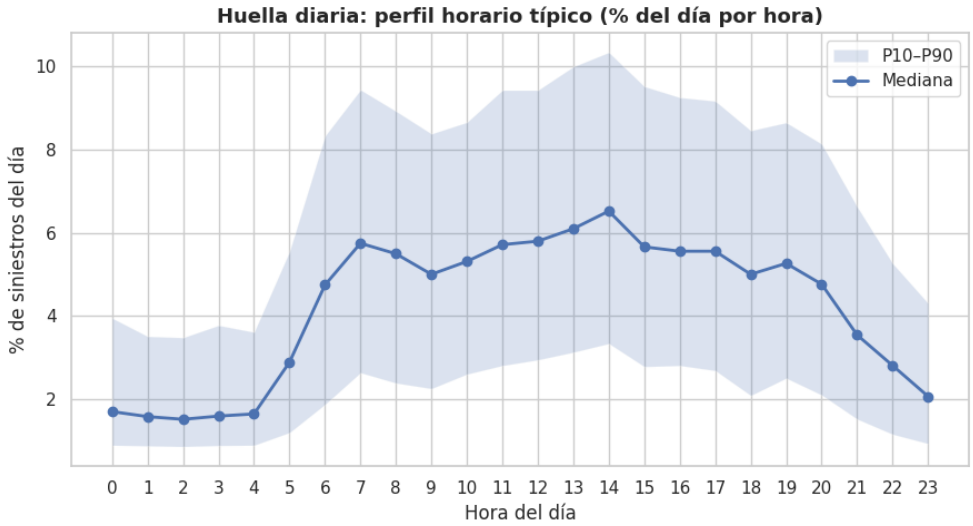

Figura 4.3. Huella diaria de siniestros viales en Bogotá D.C. (2015–2020).

- **Descomposición STL, serie horaria, componente diaria (24h) y tendencia:** La descomposición STL separa la serie en tres piezas: (i) la serie original confirma altos conteos horarios a lo largo del periodo y una ruptura visible en 2020; (ii) la estacionalidad diaria (periodo 24 h) capta el ciclo que se repite todos los días, con forma estable (dos elevaciones diurnas y valle nocturno) y amplitud que se contrae durante los meses de restricciones por COVID-19; (iii) la tendencia muestra un crecimiento 2015–2019 y una caída abrupta en 2020, seguida de recuperación parcial.

Así, el efecto diario recurrente se mantiene, pero el nivel del sistema cambia por choques exógenos.

Analíticamente, STL permite comparar periodos desestacionalizados (p. ej., evaluar políticas sin el ruido del ciclo diario), mejorar pronósticos (modelando tendencia + estacionalidad) y diseñar tableros donde los desvíos se midan sobre el residuo (anomalías reales). Como la serie son conteos, para modelado causal conviene verificar supuestos (p. ej., varianza>media → overdispersion, apta para modelos Poisson/NegBin); no obstante, para diagnóstico descriptivo STL es ideal: demuestra formalmente que el patrón horario es estructural y que la caída de 2020 responde a nivel, no a un cambio de "forma" del día.

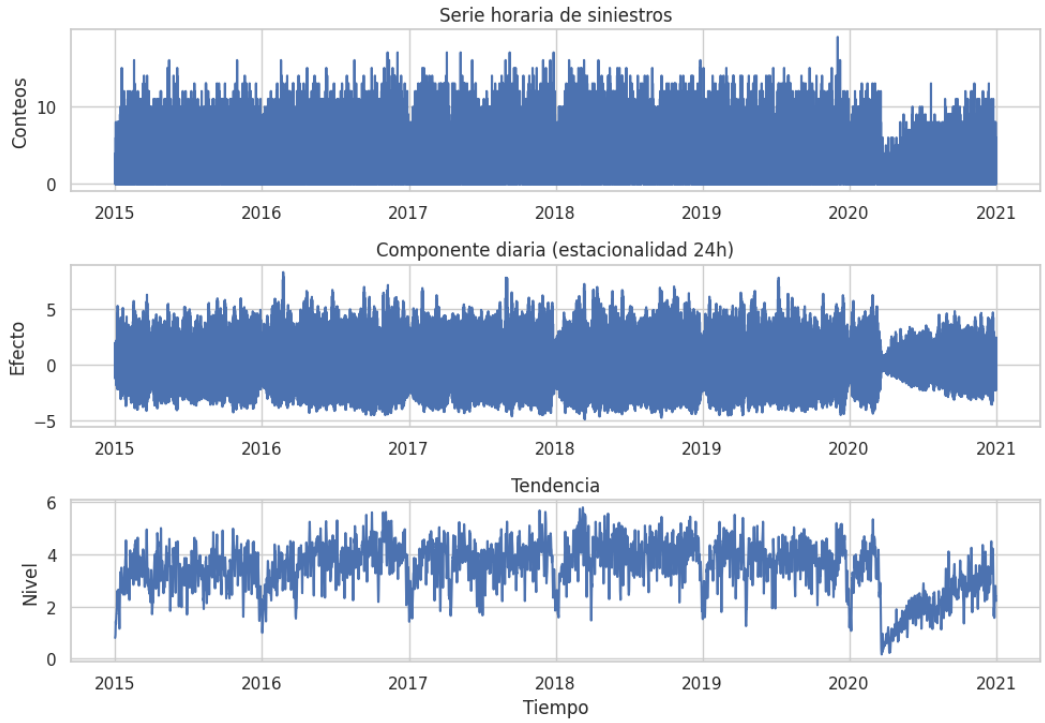

Figura 4.4. Serie de siniestros viales en Bogotá D.C. (2015–2020).

- **"Spaghetti" de días + mediana global:** Las curvas semitransparentes por día, normalizadas como % del día, se agrupan alrededor de la mediana global (línea negra). La nube compacta en la madrugada y su ensanchamiento controlado en la franja 7–19 h evidencian alta repetibilidad del perfil: la mayoría de los días siguen la misma forma con pequeñas variaciones. Los pocos trazos que se apartan marcan

outliers (feriados, eventos climáticos, restricciones puntuales), útiles para monitoreo operativo y alertas.

Este gráfico es excelente para control estadístico: puedes cuantificar la dispersión por hora (IQR/mediana), etiquetar días atípicos (p. ej., si en una hora su % supera P90 histórico) y hasta clusterizar la forma diaria (k-means sobre los 24 puntos normalizados) para separar laborables, sábados y domingos sin usar etiquetas. En síntesis: la figura valida que el patrón horario sí ocurre todos los días, ofrece una línea base robusta (mediana) y muestra explícitamente cuánto varía, que es lo que necesitas para evaluar impacto de intervenciones.

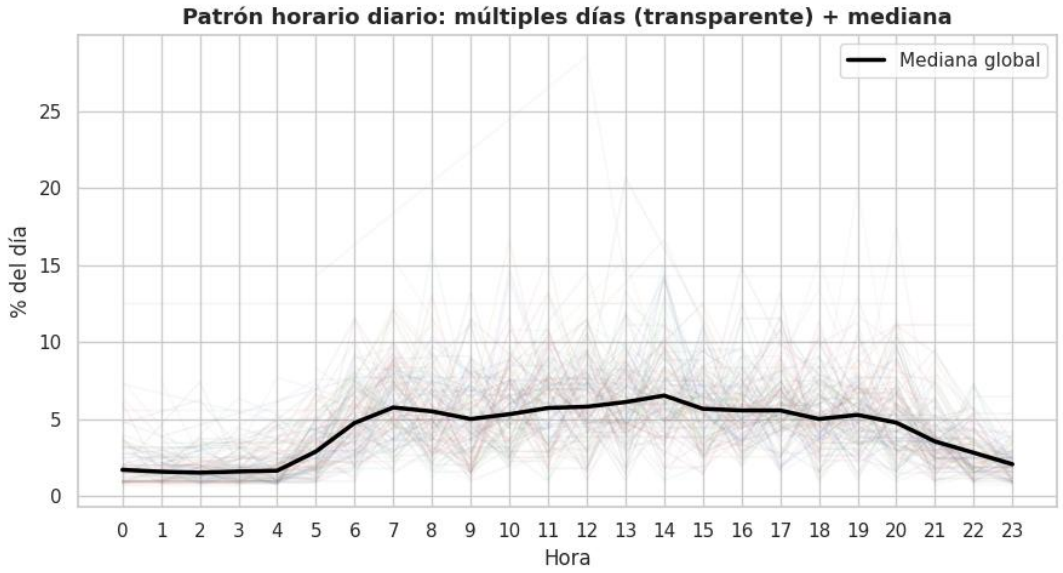

Figura 4.5. Patrón horario diario de siniestros viales en Bogotá D.C. (2015–2020).