

Análisis multivariado y clustering de indicadores de siniestralidad vial en Ecuador: Un estudio 2020 – 2023

Multivariate analysis and clustering of road accident indicators in Ecuador: A 2020 – 2023 study

Jessica Andrea Vivanco Correa¹ , Ciro Diego Radicelli-García² 

¹ Universidad Nacional de Chimborazo, jessica.vivanco@unach.edu.ec, Riobamba – Ecuador

² Universidad Nacional de Chimborazo, cradicelli@unach.edu.ec, Riobamba – Ecuador

Autor para correspondencia: jessica.vivanco@unach.edu.ec

RESUMEN

En Ecuador, los siniestros de tránsito no muestran el mismo comportamiento en todas las provincias. En algunos casos predominan los altos volúmenes de accidentes, población y vehículos; en otros, la frecuencia es menor, pero la mortalidad relativa resulta más alta. Con base en esta diferencia, el estudio analizó patrones de siniestralidad vial durante el período 2020–2023. La información se obtuvo del INEC y de la ANT y se organizó una base de datos de 92 registros provinciales que corresponde a los cuatro años de estudio en 23 provincias del Ecuador, sin considerar a Galápagos por los registros nulos que presentó en los diferentes indicadores de siniestralidad. Se consideraron siete variables de estudio: siniestros, fallecidos, lesionados, vehículos matriculados, población, tasa de mortalidad y siniestros por vehículo. Con dicha información se procedió a estandarizar los datos mediante Z-score y se aplicó el Análisis de Componentes Principales. Los dos primeros componentes explicaron el 86,63 % de la variabilidad total: PC1 representó principalmente el volumen general de siniestralidad y exposición vial, PC2 estuvo asociado con la severidad relativa y PC3 con la intensidad de siniestros respecto al parque vehicular. Posteriormente, se aplicó el algoritmo K-means. El agrupamiento distinguió tres perfiles: alta siniestralidad absoluta, comportamiento intermedio y mayor severidad relativa. Estos resultados muestran que el uso conjunto de PCA y K-means facilita la identificación de perfiles territoriales de riesgo vial.

Palabras clave: Análisis multivariado; Siniestralidad vial; Componentes principales; K-means; Ecuador.

ABSTRACT

In Ecuador, traffic accidents do not follow the same pattern across all provinces. In some cases, high volumes of accidents, population, and vehicles predominate; in others, the frequency is lower, but the relative fatality rate is higher. Based on this difference, the study analyzed road accident patterns during the 2020–2023 period. Data were taken from INEC and ANT records. The final database included 92 province-year observations, corresponding to four years of study and 23 provinces in Ecuador. The Galápagos Islands were excluded because some accident indicators had missing or very low values. Seven variables were analyzed: accidents, fatalities, injuries, registered vehicles, population, mortality rate, and accidents per vehicle. The data were first standardized using Z-scores, and then Principal Component Analysis was applied. The first two components explained 86.63% of the total variability: PC1 mainly represented the overall volume of road crashes and road exposure, PC2 was associated with relative severity, and PC3 with crash intensity relative to the vehicle fleet. Finally, the K-means algorithm was used to group the observations. The clustering identified three profiles: high absolute accident rate, intermediate performance, and higher relative severity. These results show that the combined use of PCA and K-means facilitates the identification of regional road risk profiles.

Keywords: Multivariate analysis; Road traffic accidents; Principal components; K-means; Ecuador.

Derechos de Autor

Los originales publicados en las ediciones electrónicas bajo derechos de primera publicación de la revista son del Instituto Superior Tecnológico Universitario Rumiñahui, por ello, es necesario citar la procedencia en cualquier reproducción parcial o total. Todos los contenidos de la revista electrónica se distribuyen bajo una [licencia de Creative Commons Reconocimiento-NoComercial-4.0 Internacional](#).



Citar como:

Vivanco Correa, J. A., & Radicelli-García, C. D. (2026). Análisis multivariado y clustering de indicadores de siniestralidad vial en Ecuador: Un estudio 2020 – 2023. *CONECTIVIDAD*, 7(2), e517. <https://doi.org/10.37431/conectividad.v7i2.517>

1. INTRODUCCIÓN

La siniestralidad vial en Ecuador es un problema que no puede analizarse únicamente a partir del número total de accidentes. En cada provincia intervienen condiciones distintas, como la cantidad de población, el parque vehicular, la movilidad y las características del territorio. Por esta razón, no todas las observaciones presentan el mismo comportamiento. Las provincias con mayor población suelen concentrar también un mayor parque vehicular y, en consecuencia, registrar un mayor número de siniestros.

En Ecuador existen páginas oficiales donde se registran los datos obtenidos sobre siniestros, fallecidos, lesionados, vehículos matriculados y la población (Agencia Nacional de Tránsito, 2023; Instituto Nacional de Estadística y Censos, 2023). Con esta información se pueden construir diferentes indicadores de frecuencia, de exposición y de severidad. Sin embargo, dichos reportes institucionales permiten visualizar la información mediante gráficas, tablas o comparaciones generales que no siempre permiten identificar las relaciones entre variables y poder realizar una comparación de comportamientos similares.

Hay diferentes estudios que se han realizado utilizando herramientas espaciales, estadísticas y computacionales para analizar la siniestralidad vial y la gravedad de sus consecuencias. Estos enfoques han mostrado utilidad para reconocer patrones territoriales, zonas críticas y factores asociados al riesgo vial (Alam & Tabassum, 2023; Gedamu et al., 2024; Muñoz Muñoz et al., 2024). En este tipo de análisis, la Matemática Aplicada y la Matemática Computacional resultan útiles porque permiten organizar varios indicadores en un mismo modelo y explorar relaciones que no se observan fácilmente con tablas descriptivas.

En los últimos años, la aplicación de técnicas de análisis multivariado y agrupamiento para el estudio de la siniestralidad vial ha adquirido mayor relevancia. Diversas investigaciones han empleado el Análisis de Componentes Principales (PCA) y algoritmos de agrupamiento para identificar patrones de riesgo, reducir la dimensionalidad de los datos y clasificar observaciones con características similares, proporcionando información útil para la planificación de estrategias de seguridad vial. En Ecuador y otros contextos, estudios recientes han utilizado métodos multivariados para analizar la distribución de los siniestros entre provincias, evidenciando la utilidad de estas herramientas para apoyar la toma de decisiones basada en datos. Sin embargo,

aún son limitados los trabajos que integran el análisis conjunto de indicadores de frecuencia, exposición y severidad mediante la combinación de PCA y K-means en una base de datos provincia-año para el período 2020–2023 (Li & Chen, 2024; Muñoz Muñoz et al., 2024).

Para este propósito se empleó el Análisis de Componentes Principales, identificado en adelante como PCA por sus siglas en inglés (Principal Component Analysis). Este método transforma variables posiblemente correlacionadas en nuevos componentes no correlacionados, los cuales concentran una parte importante de la variabilidad de los datos (Jolliffe & Cadima, 2016; Maugeri et al., 2023). En la práctica, esto permite pasar de varias variables originales a pocos componentes que resumen la mayor parte de la información y que pueden graficarse con mayor claridad.

Junto con el PCA, el algoritmo K-means permite agrupar observaciones de acuerdo con su similitud multivariada (MacQueen, 1967; Kaufman & Rousseeuw, 2005; Muñoz Muñoz et al., 2024). Al realizar este estudio en relación provincia-año se logra distinguir observaciones con alto volumen de siniestros e identificar diferencias en los niveles de severidad relativa.

El análisis de la siniestralidad vial requiere considerar no solo la frecuencia de los eventos, sino también indicadores de exposición y severidad. La frecuencia permite cuantificar la ocurrencia de siniestros, la exposición incorpora variables asociadas con la población y el parque vehicular, mientras que la severidad permite valorar las consecuencias de los accidentes mediante indicadores como la mortalidad relativa (World Health Organization, 2023; Instituto Nacional de Estadística y Censos, 2024). Analizar estas dimensiones de forma conjunta permite identificar diferencias territoriales que no serían evidentes mediante el uso aislado de valores absolutos. Asimismo, el período 2020–2023 resulta particularmente relevante debido a los cambios en los patrones de movilidad asociados con la pandemia por COVID-19 y la posterior recuperación de la circulación vehicular, condiciones que pudieron modificar la ocurrencia y distribución de los siniestros de tránsito (World Health Organization, 2023).

Aunque en Ecuador se dispone de información oficial sobre siniestros de tránsito y existen investigaciones recientes sobre el tema, aún son limitados los estudios que integran técnicas de análisis multivariado para evaluar conjuntamente indicadores de frecuencia, exposición y severidad con fines de apoyo a la toma de decisiones en seguridad vial.

Por ende, la presente investigación tiene como objetivo identificar los principales patrones

multivariados de la siniestralidad vial en Ecuador durante el período 2020–2023, determinando los indicadores que estructuran la variación entre las observaciones provincia-año, identificando las observaciones más representativas y caracterizando perfiles territoriales diferenciados de riesgo vial. Para ello, se emplea el Análisis de Componentes Principales para reducir la dimensionalidad de los datos y reconocer las variables que explican la mayor parte de la variabilidad, seguido de la aplicación del algoritmo K-means para clasificar las observaciones según comportamientos similares de frecuencia, exposición y severidad.

2. MATERIALES Y MÉTODOS

El estudio se desarrolló con un enfoque cuantitativo y aplicado, orientado a explorar el comportamiento conjunto de varios indicadores de siniestralidad vial en Ecuador durante el período 2020–2023. El análisis no se centró en una sola variable, sino en la relación entre siniestros, fallecidos, lesionados, población y vehículos matriculados, junto con indicadores derivados de exposición y severidad. La información utilizada se la obtuvo de las estadísticas de siniestros de tránsito publicadas por el Instituto Nacional de Estadística y Censos en el portal Ecuador en Cifras.

Los datos se obtuvieron del portal Ecuador en Cifras del INEC, elaboradas a partir de registros administrativos de la Agencia Nacional de Tránsito, por lo que se utilizaron como fuente oficial para el análisis de la siniestralidad vial en Ecuador (Agencia Nacional de Tránsito, 2023; Instituto Nacional de Estadística y Censos, 2023; Instituto Nacional de Estadística y Censos, 2024). El período 2020–2023 se seleccionó porque para esos años se disponía de datos consolidados y comparables sobre las variables de interés. Además, este intervalo permite observar el comportamiento de los siniestros durante y después del período de restricciones de movilidad asociado a la pandemia por COVID-19.

La unidad de análisis fue cada observación provincia-año. Después de organizar la información, se trabajó con 92 registros provincia-año, correspondientes a 23 provincias del Ecuador continental. Galápagos quedó fuera del análisis porque en varios años presentó valores nulos o muy reducidos, lo que podía afectar la comparación con el resto de provincias, la estimación de los componentes principales y la conformación de los grupos.

Las variables utilizadas en el análisis fueron: el número de siniestros de tránsito, fallecidos,

lesionados, vehículos matriculados, población, tasa de mortalidad por accidentes de tránsito por cada 100.000 habitantes y siniestros de tránsito por cada 1.000 vehículos matriculados. Estas variables permiten observar el fenómeno desde tres dimensiones: frecuencia, exposición y severidad (World Health Organization, 2023; Instituto Nacional de Estadística y Censos, 2024; Toalombo Vargas & Montalvo Márquez, 2024). Las variables provincia y año se mantuvieron únicamente para identificar cada registro de la base, pero no se utilizaron en los cálculos del PCA ni del agrupamiento.

Antes de realizar el análisis multivariado, se revisó la base de datos para comprobar que no existieran valores faltantes y que la información sea la correcta. También se volvió a calcular la tasa de mortalidad y el indicador de siniestros por vehículo, para asegurar que todos los registros fueran trabajados bajo el mismo criterio.

En el caso de la tasa de mortalidad, se relacionó el número de fallecidos con la población de cada provincia y el resultado se expresó por cada 100.000 habitantes. Además, se incluyó un indicador relativo de siniestros frente al número de vehículos matriculados. Para hacerlo comparable entre provincias, el valor se presentó por cada 1.000 vehículos. Con estos indicadores se buscó complementar los valores absolutos con medidas relativas de riesgo y exposición vial (World Health Organization, 2023; Acosta-González et al., 2025).

Como las variables no estaban en la misma escala, se aplicó la estandarización Z-score. Con este procedimiento, todas las variables quedaron comparables y se evitó que los indicadores con valores más grandes tuvieran mayor peso en el análisis. Esto fue necesario antes de aplicar el Análisis de Componentes Principales, debido a que la base combinó conteos, población, vehículos e indicadores relativos (Maugeri et al., 2023; Jolliffe & Cadima, 2016).

Posteriormente, se aplicó el Análisis de Componentes Principales con el propósito de reducir la dimensión de la matriz original y representar la mayor parte de la información en un número menor de componentes (Jolliffe & Cadima, 2016; Maugeri et al., 2023). En el modelo se incluyeron siete variables cuantitativas activas: siniestros, fallecidos, lesionados, vehículos matriculados, población, tasa de mortalidad y siniestros por vehículo. Al trabajar con siete variables cuantitativas, el análisis produjo siete componentes principales. Para los gráficos y el agrupamiento se tomaron únicamente los dos primeros, porque fueron los que resumieron la

mayor parte de la variación observada en la base.

La aplicación del PCA antes del algoritmo K-means permitió reducir la dimensionalidad de los datos, eliminar la redundancia generada por la correlación entre variables y concentrar la mayor parte de la información en un número reducido de componentes principales. Esto favorece la formación de grupos más estables e interpretables, al disminuir el efecto de variables altamente correlacionadas sobre el proceso de agrupamiento.

Luego de obtener los dos primeros componentes principales, se utilizó K-means para agrupar los registros provincia-año con comportamientos similares. El número de clústeres se evaluó mediante el método del codo y las métricas internas Silhouette, Davies-Bouldin y Calinski-Harabasz para valores de k entre 2 y 6. Aunque la solución con $k = 2$ presentó el mejor desempeño global en las métricas internas, se analizó adicionalmente la solución con $k = 3$ debido a que permitió diferenciar un perfil específico asociado con mayor severidad relativa, además de los grupos de alta siniestralidad absoluta y comportamiento intermedio. Por esta razón, $k = 3$ se mantuvo como una solución de carácter descriptivo e interpretativo, considerando conjuntamente los criterios estadísticos y la utilidad sustantiva de los perfiles obtenidos.

Una vez definidos los grupos, se calcularon los promedios de las variables originales dentro de cada clúster. Con estos valores se pudo describir qué características tenía cada clúster y comparar sus diferencias en cuanto a siniestros, exposición vehicular y severidad.

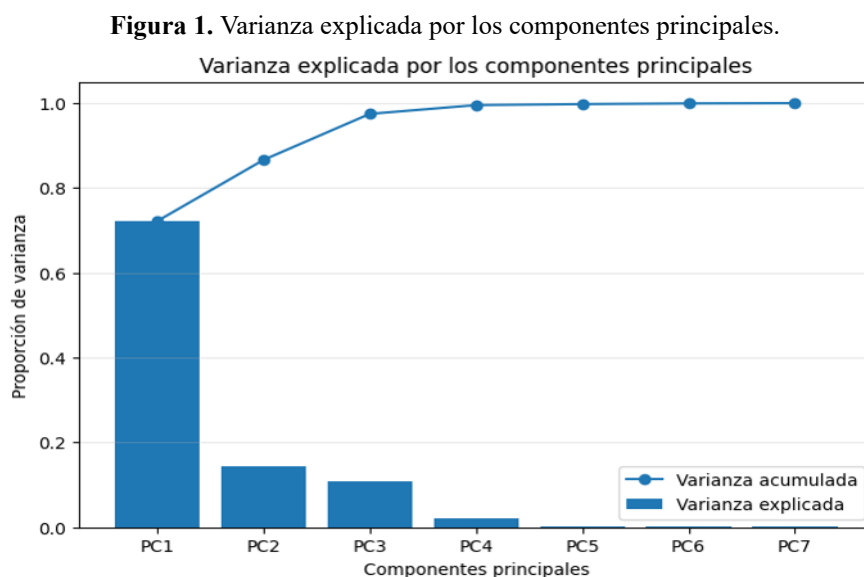
El procesamiento computacional se realizó en el entorno Google Colab empleando las librerías pandas, NumPy, Matplotlib y scikit-learn. La estandarización de las variables se efectuó mediante el método Z-score utilizando StandardScaler. Posteriormente se aplicó el Análisis de Componentes Principales (PCA) y el algoritmo K-means implementados en scikit-learn. Para garantizar la reproducibilidad de los resultados, K-means se ejecutó con una semilla fija ($\text{random_state} = 42$) y diez inicializaciones independientes ($n_init = 10$). La selección del número de grupos se apoyó en el método del codo y en las métricas internas Silhouette, Davies-Bouldin y Calinski-Harabasz. Todas las figuras y tablas presentadas en este trabajo fueron generadas mediante los procedimientos implementados en dicho entorno computacional.

3. RESULTADOS Y DISCUSIÓN

Con el objetivo de identificar patrones multivariados en los indicadores de siniestralidad vial,

se aplicó el Análisis de Componentes Principales sobre las variables cuantitativas previamente estandarizadas, siguiendo criterios habituales del análisis multivariado para reducción de dimensionalidad (Jolliffe & Cadima, 2016; Maugeri et al., 2023).

Al aplicar el PCA, la mayor parte de la información quedó resumida en los primeros componentes. Como se observa en la Figura 1, el primer componente explicó el 72,20% de la varianza total y el segundo alcanzó el 14,43%. Sumados, PC1 y PC2 reunieron el 86,63% de la variabilidad de la base.



Fuente: Elaboración propia a partir de datos del INEC y ANT.

Este porcentaje fue suficiente para trabajar con una representación bidimensional de las observaciones sin perder una parte relevante de la información original. El detalle de la varianza explicada y acumulada se muestra en la Tabla 1.

Tabla 1. Varianza explicada por componente principal.

Componente	Varianza explicada	Varianza acumulada
PC1	0,7220	0,7220
PC2	0,1443	0,8663
PC3	0,1085	0,9748
PC4	0,0206	0,9954
PC5	0,0024	0,9978
PC6	0,0016	0,9994
PC7	0,0006	1,0000

Fuente: Elaboración propia a partir de datos del INEC y ANT.

Los tres primeros componentes reunieron el 97,48% de la variabilidad total. Aun así, para los gráficos y el agrupamiento se trabajó con PC1 y PC2, ya que juntos explicaron más del 86% de la información y permitieron representar las observaciones en un plano bidimensional.

Para interpretar qué representa cada componente, se revisaron las cargas factoriales de las variables originales. La Tabla 2 muestra estos valores para los tres primeros componentes principales.

Tabla 2. Cargas factoriales de los tres primeros componentes principales.

Variable	PC1	PC2	PC3
Siniestros	0,442	0,005	-0,023
Fallecidos	0,433	0,143	-0,153
Lesionados	0,427	0,021	0,006
Vehículos	0,425	-0,003	-0,210
Población	0,437	-0,028	-0,155
Tasa de mortalidad	-0,046	0,987	0,072
Siniestros por vehículo	0,246	-0,057	0,950

Fuente: Elaboración propia a partir de datos del INEC y ANT.

La Tabla 2 muestra que el primer componente principal (PC1) presenta cargas positivas y de magnitud similar para las variables siniestros, fallecidos, lesionados, vehículos matriculados y población, lo que indica que este componente resume principalmente el nivel general de siniestralidad y exposición de las observaciones provincia-año. Por su parte, el segundo componente principal (PC2) está dominado por la tasa de mortalidad, con una carga de 0,987, evidenciando que este componente representa principalmente la severidad relativa de los siniestros. En conjunto, ambos componentes permiten diferenciar observaciones tanto por el volumen de eventos como por la gravedad de sus consecuencias.

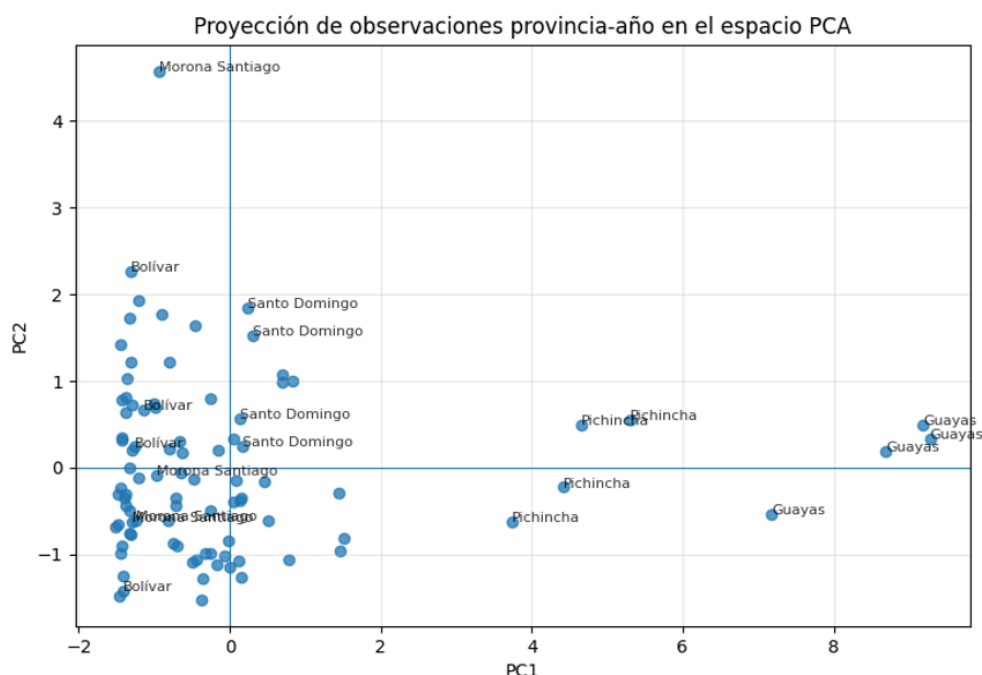
PC1 resume sobre todo las variables relacionadas con el volumen de siniestralidad y la exposición vial: siniestros, fallecidos, lesionados, vehículos matriculados y población. Por tanto, PC1 puede interpretarse como una dimensión de volumen general de siniestralidad vial. El segundo componente principal presenta una carga elevada en la tasa de mortalidad, por lo que puede interpretarse como una dimensión de severidad relativa. Esto significa que PC2 permite diferenciar observaciones que, aunque no necesariamente presentan altos volúmenes absolutos de siniestros, sí muestran mayor mortalidad en relación con su población.

Por su parte, el tercer componente principal está asociado principalmente con el indicador de siniestros por vehículo, lo que sugiere una dimensión relacionada con la intensidad de siniestros respecto al parque vehicular. Aunque PC3 no se incluyó en los gráficos bidimensionales, ayuda a entender otra parte del comportamiento de los indicadores, especialmente la relación entre siniestros y parque vehicular.

3.1. Proyección de las observaciones en el espacio reducido

En la Figura 2 se observa la ubicación de las observaciones provincia-año en el plano formado por PC1 y PC2. Para evitar una gráfica sobrecargada, solo se colocaron etiquetas en algunos casos representativos o extremos. La distribución evidencia una separación entre observaciones asociadas a mayor volumen de siniestralidad y aquellas relacionadas con mayor severidad relativa.

Figura 2. Proyección de observaciones provincia-año en el espacio de los dos primeros componentes principales.



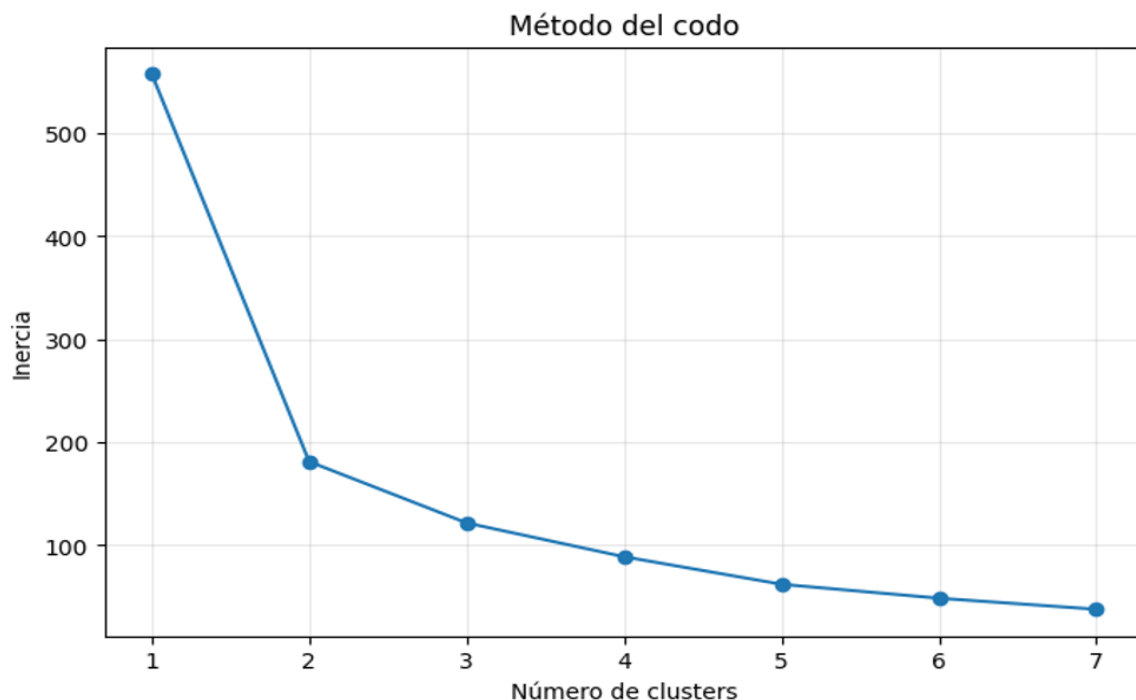
Fuente: Elaboración propia a partir de datos del INEC y ANT.

En el plano PC1–PC2 se observa que las observaciones no se distribuyen de manera uniforme. Algunas aparecen alejadas del centro, principalmente por sus mayores valores en siniestros, fallecidos, lesionados, población y vehículos matriculados. En cambio, la mayoría se ubica en zonas más cercanas al origen, lo que indica valores más moderados en los indicadores analizados. El eje PC1 permite diferenciar principalmente observaciones con mayor o menor volumen general de siniestralidad vial. En cambio, el eje PC2 permite distinguir observaciones con mayor severidad relativa, asociada principalmente a la tasa de mortalidad. Esta representación constituye la base para la aplicación posterior del algoritmo de agrupamiento K-means.

3.2. Determinación del número de clústeres

El número de grupos se revisó con el método del codo, tomando como entrada las coordenadas de PC1 y PC2. La Figura 3 presenta la curva resultante.

Figura 3. Método del codo para la selección del número de clústeres.



Fuente: Elaboración propia a partir de datos del INEC y ANT.

Al revisar la curva de inercia, se observa que el cambio más importante ocurre hasta $k = 3$. Luego de ese valor, la disminución de la inercia es más pequeña, por lo que no se justifica aumentar muchos más grupos. Aunque las métricas internas mostraron un mejor desempeño global para $k = 2$, se mantuvo la solución con $k = 3$ debido a que permitió diferenciar tres perfiles de interés sustantivo: alta siniestralidad absoluta, comportamiento intermedio y mayor severidad relativa. De este modo, la selección final consideró conjuntamente el comportamiento de la inercia, las métricas de validación y la utilidad interpretativa de los perfiles obtenidos.

Para complementar esta decisión, se calcularon métricas internas de validación del agrupamiento. En la Tabla 3 se presentan los valores de los índices Silhouette, Davies-Bouldin y Calinski-Harabasz para distintos números de clústeres, con el fin de comparar la separación y consistencia de las soluciones evaluadas.

Tabla 3. Métricas internas de validación del clustering.

Número de clústeres	Silhouette	Davies-Bouldin	Calinski-Harabasz
2	0,757	0,451	188,06
3	0,417	0,799	159,40
4	0,427	0,676	154,78
5	0,403	0,668	173,50
6	0,433	0,685	180,19

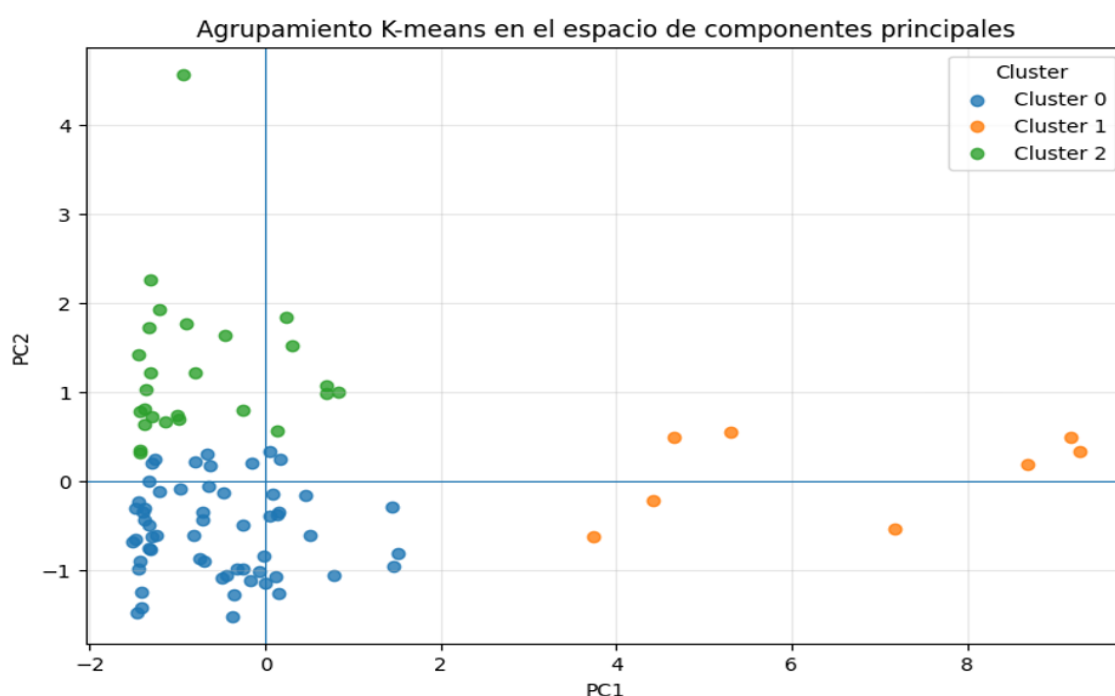
Fuente: Elaboración propia a partir de datos del INEC y ANT.

El índice de silueta muestra que una solución con dos clústeres presenta mayor separación global. Sin embargo, desde el punto de vista interpretativo, la solución con tres clústeres permite diferenciar de mejor manera los perfiles de siniestralidad vial, distinguiendo observaciones de alto volumen, comportamiento intermedio y mayor severidad relativa. Por ello, se seleccionó $k = 3$ como una solución metodológicamente adecuada y sustantivamente interpretable.

3.3. Agrupamiento de observaciones provincia-año mediante K-means

Después de definir $k = 3$, se aplicó K-means con las coordenadas de las observaciones en PC1 y PC2. La Figura 4 muestra cómo quedaron distribuidos los grupos en el plano de componentes principales.

Figura 4. Distribución de los grupos provincia-año obtenidos con K-means en el plano PC1–PC2.



Fuente: Elaboración propia a partir de datos del INEC y ANT.

En la Figura 4 se observan tres grupos claramente diferenciados. El Clúster 1 reúne las observaciones con mayor siniestralidad en valores absolutos y se relaciona, principalmente, con provincias de alta población y mayor parque vehicular. El Clúster 0 agrupa la mayor parte de los registros y representa un comportamiento intermedio. En cambio, el Clúster 2 se distingue por presentar mayor severidad relativa, reflejada en tasas de mortalidad más altas.

Esta separación confirma que las observaciones provincia-año no siguen un mismo patrón. Algunas se diferencian por el volumen de eventos y la exposición vehicular, mientras que otras destacan por la gravedad de sus consecuencias.

3.4. Caracterización de los clústeres

Para caracterizar los grupos obtenidos, se calcularon los promedios de las variables originales en cada clúster. Estos resultados se muestran en la Tabla 4.

Tabla 4. Promedios de indicadores de siniestralidad vial por clúster.

Clúster	N.º observaciones	Siniestros	Fallecidos	Lesionados	Vehículos	Población	Tasa de mortalidad	Siniestros por vehículo
0	58	467,90	50,60	363,83	78.044,12	513.923,38	10,03	5,46
1	8	5.574,62	438,75	4.886,25	587.242,25	3.876.459,88	11,34	9,48
2	26	358,23	71,12	309,19	61.856,38	386.437,42	18,71	4,68

Fuente: Elaboración propia a partir de datos del INEC y ANT.

En el Clúster 1 se observa que el mayor número de siniestros, fallecidos, lesionados, y vehículos matriculados se encuentran en este clúster, el cual corresponde principalmente a las observaciones en las provincias de Pichincha y Guayas, estas provincias tienen una mayor población y parque vehicular, por ello también presentan una mayor exposición vial, lo que se puede relacionar con los altos niveles de siniestralidad registrados en este clúster.

En el Clúster 0 hay el mayor número de observaciones provincia-año, sus valores promedio permiten interpretar como el grupo de comportamiento intermedio, con niveles moderados de siniestros y lesionados.

El Clúster 2 está conformado por 26 observaciones y representa una alta tasa promedio de mortalidad. A pesar de que sus valores de siniestros, lesionados, vehículos y población son menores que los del grupo de alta siniestralidad, registra un mayor nivel de mortalidad relativa. Por esta razón, se puede catalogar como el clúster con mayor severidad.

Los resultados muestran que no es suficiente con revisar únicamente el número total de siniestros. El análisis multivariado permite comparar provincias con mayor cantidad de eventos frente a otras donde la tasa de mortalidad es más elevada. De esta manera, se reconocen perfiles de riesgo que podrían no observarse con claridad mediante un análisis solamente descriptivo.

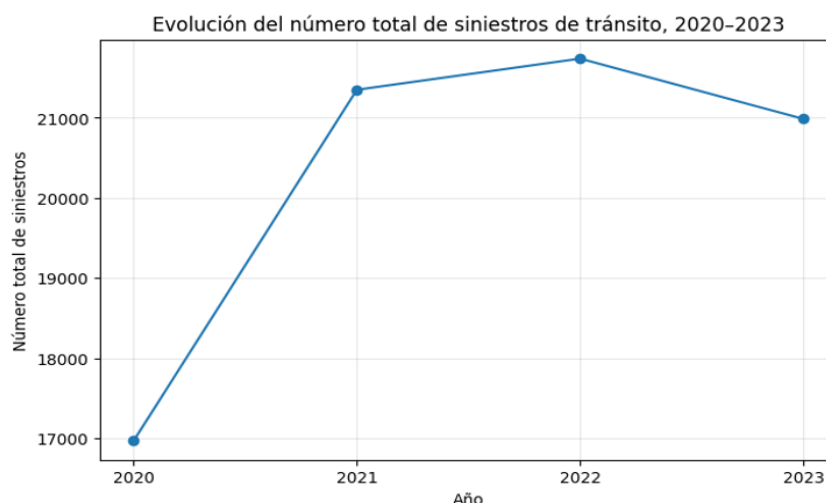
3.5. Análisis temporal complementario

Con el propósito de complementar el análisis multivariado, se examinó la evolución anual del número total de siniestros de tránsito durante el período 2020–2023. Los resultados se presentan en la Figura 5.

Al revisar los datos por año, se observa que el número de siniestros no fue igual durante todo

el período. En el 2020 se registra el valor más bajo, cerca de 17 mil siniestros, lo cual puede relacionarse con las restricciones de movilidad aplicadas durante la pandemia de COVID-19. En el 2021 se observa un incremento importante y en el 2022 se observa el valor más alto del período. Para el 2023, el número de siniestros disminuye ligeramente, aunque se mantiene alrededor de los 21 mil registros.

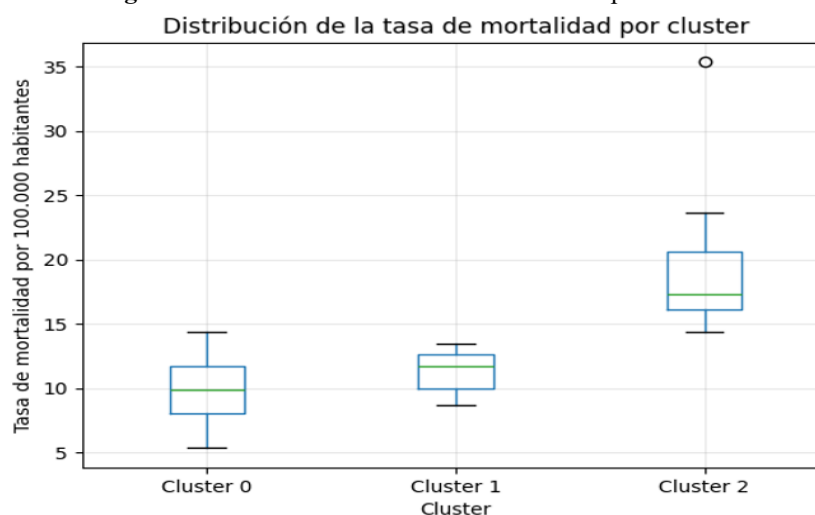
Figura 5. Evolución del número total de siniestros de tránsito en Ecuador, 2020–2023.



Fuente: Elaboración propia a partir de datos del INEC y ANT.

En la Figura 6 se muestra, además, cómo se distribuye la tasa de mortalidad en cada clúster.

Figura 6. Distribución de la tasa de mortalidad por clúster.



Fuente: Elaboración propia a partir de datos del INEC y ANT.

La distribución de la tasa de mortalidad evidencia diferencias entre los grupos identificados. En particular, el clúster asociado a mayor severidad relativa presenta valores superiores de mortalidad, lo que confirma que los grupos no solo se diferencian por el volumen de siniestros, sino también por la gravedad de sus consecuencias.

3.6. Discusión

Los resultados muestran que la siniestralidad vial en Ecuador, durante el período 2020–2023, no tuvo el mismo comportamiento en todas las observaciones analizadas. PC1 se relaciona principalmente con el volumen de siniestros y la exposición vial, mientras que PC2 se asocia con la tasa de mortalidad. Esto muestra que una mayor cantidad de siniestros no siempre implica una mayor mortalidad, por lo que resulta necesario analizar de manera conjunta los indicadores de frecuencia, exposición y severidad. Este comportamiento coincide con lo reportado por la Organización Mundial de la Salud y por estudios recientes, los cuales señalan que la evaluación de la seguridad vial requiere integrar múltiples indicadores para identificar territorios con diferentes perfiles de riesgo y orientar de mejor manera las estrategias de prevención (World Health Organization, 2023; Acosta-González et al., 2025; Toalombo Vargas & Montalvo Márquez, 2024). Al revisar las cargas de los componentes, se observa que PC1 está más relacionado con siniestros, fallecidos, lesionados, vehículos matriculados y población. En cambio, PC2 se relaciona principalmente con la tasa de mortalidad, lo que permite identificar observaciones donde la letalidad es mayor, aunque el número total de siniestros no sea muy alto. Por esta razón, el análisis ayuda a comparar provincias con muchos eventos frente a otras donde las consecuencias son más fuertes en relación con la población.

Aunque el tercer componente principal (PC3) no se utilizó para el proceso de agrupamiento, explicó un 10,85 % de la variabilidad total y presentó una alta asociación con la variable Siniestros por vehículo. Esto sugiere que dicho componente captura un comportamiento complementario relacionado con la intensidad relativa de los siniestros respecto al parque vehicular, constituyendo un indicador potencialmente útil para futuros estudios orientados a profundizar el análisis de la exposición al riesgo vial.

Con K-means se obtuvieron tres perfiles de siniestralidad, el grupo de alta siniestralidad se relaciona principalmente con observaciones de provincias con mayor población y parque vehicular, como Guayas y Pichincha. Otro grupo muestra un comportamiento intermedio, mientras que el tercer grupo presenta mayor mortalidad relativa. Esto confirma que no basta con revisar únicamente el total de siniestros, ya que una provincia puede tener muchos accidentes, pero otra puede presentar una tasa de mortalidad más alta.

Al revisar los años de estudio, 2020 aparece como un caso particular, debido a que registra el menor número de siniestros del período. Esto puede explicarse, al menos en parte, por la menor circulación de personas y vehículos durante las medidas aplicadas frente al COVID-19. En los años posteriores, el número de eventos vuelve a incrementarse, lo que muestra que los cambios en la movilidad también influyeron en la ocurrencia de siniestros.

El estudio también presenta algunas limitaciones. La información utilizada corresponde a datos provinciales y anuales, es por ello que no fue posible revisar diferencias más específicas a nivel cantonal, parroquial o por tipo de vía. Además, no se incluyeron variables como infraestructura vial, clima, horarios de ocurrencia, causas del siniestro o acceso a servicios de emergencia, factores que también pueden influir en la concentración y mortalidad de los siniestros de tránsito (Gedamu et al., 2024; Zhang et al., 2023; Yang & Han, 2025; Zeng et al., 2024).

Aun con estas limitaciones, el uso de PCA y K-means permitió observar patrones relevantes en la base analizada. Este tipo de procedimiento puede servir como apoyo para priorizar territorios y orientar acciones preventivas en seguridad vial (Li & Chen, 2024).

4. CONCLUSIONES

El análisis multivariado de los indicadores de siniestralidad vial en Ecuador durante el período 2020–2023 permitió identificar patrones diferenciados en las observaciones provincia-año. La aplicación del Análisis de Componentes Principales evidenció que PC1 representa principalmente el volumen general de siniestralidad y exposición vial, mientras que PC2 se relaciona con la severidad relativa, principalmente a través de la tasa de mortalidad. Por su parte, PC3 se asocia principalmente con la intensidad de siniestros respecto al parque vehicular. Con la aplicación de PCA y K-means se pudo organizar las observaciones en tres clústeres. El Clúster 1 agrupa los casos con mayor siniestralidad absoluta, principalmente por sus valores altos de siniestros, fallecidos, lesionados, población y vehículos matriculados. El Clúster 0 presenta un comportamiento intermedio, mientras que el Clúster 2 se diferencia por su mayor tasa de mortalidad, lo que indica un perfil de mayor severidad relativa.

Los resultados muestran que la siniestralidad vial no se presenta de la misma forma en todo el Ecuador. Las diferencias entre provincias se relacionan con la cantidad de siniestros, el parque vehicular y también con las consecuencias que dejan los accidentes. Por esta razón, las

provincias con mayor número de siniestros no necesariamente son las que presentan mayor severidad relativa.

El procedimiento aplicado ayudó a organizar la base de datos y a comparar observaciones que no siempre se diferencian fácilmente con tablas o gráficos descriptivos. Al reducir las variables con PCA y después agruparlas con K-means, se pudieron reconocer perfiles de riesgo que complementan el análisis tradicional.

Para próximos estudios se podrían agregar otras variables que no estuvieron disponibles en esta base de datos, por ejemplo, el tipo de vía, el estado de la infraestructura, las causas del siniestro, las condiciones climáticas, las características socioeconómicas y el acceso a servicios de emergencia. También sería importante trabajar con datos a nivel cantonal o parroquial, porque al analizar solo la información provincial se pueden perder diferencias internas entre territorios. En conjunto, los resultados obtenidos demuestran que la integración del Análisis de Componentes Principales y el algoritmo K-means constituye una herramienta útil para apoyar el análisis de la siniestralidad vial desde una perspectiva multivariada. La metodología propuesta permite complementar los análisis descriptivos tradicionales, facilitando la identificación de perfiles territoriales diferenciados que pueden servir como apoyo para la planificación de estrategias de prevención, priorización de recursos y formulación de políticas públicas orientadas a mejorar la seguridad vial en Ecuador.

Contribución de los Autores (CRediT): JAVC: Conceptualización, Curación de datos, Análisis formal, Metodología, Software, Validación, Visualización, Redacción - borrador original. CDRG: Metodología, Supervisión, Validación, Redacción - revisión y edición.

Conflicto de Intereses: Los autores declaran que no existen conflictos de intereses en esta publicación.

Declaración de Uso de Inteligencia Artificial: Los autores declaran que utilizaron herramientas de inteligencia artificial (IA) en la redacción del artículo. La herramienta utilizada para este fin fue: CHATGPT.

REFERENCIAS

- Acosta-González, N., Cahueñas, S., & Pérez, C. (2025). Risk factors for fatal road traffic accidents in Ecuador. *Transportation Research Interdisciplinary Perspectives*, 32, 101515. <https://doi.org/10.1016/j.trip.2025.101515>
- Agencia Nacional de Tránsito. (2023). *Estadísticas de siniestros de tránsito*. <https://www.ant.gob.ec/estadisticas/>
- Alam, M. S., & Tabassum, N. J. (2023). Spatial pattern identification and crash severity analysis

- of road traffic crash hot spots in Ohio. *Heliyon*, 9(5), e16303. <https://doi.org/10.1016/j.heliyon.2023.e16303>
- Gedamu, W. T., Plank-Wiedenbeck, U., & Wodajo, B. T. (2024). A spatial autocorrelation analysis of road traffic crash by severity using Moran's I spatial statistics: A comparative study of Addis Ababa and Berlin cities. *Accident Analysis & Prevention*, 200, 107535. <https://doi.org/10.1016/j.aap.2024.107535>
- Instituto Nacional de Estadística y Censos. (2023). *Estadísticas de siniestros de tránsito*. <https://www.ecuadorencifras.gob.ec/estadisticas-siniestros-de-transito/>
- Instituto Nacional de Estadística y Censos. (2024). *Siniestros de tránsito: cuarto trimestre 2023*. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Estadistica%20de%20Transporte/2023/iv_trimestre/2023_RESULTADOS_SINIESTROS_IVT.pdf
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Wiley. <https://doi.org/10.1002/9780470316801>
- Li, J., & Chen, F. (2024). Gauging road safety advances using a hybrid EWM–PROMETHEE II–DBSCAN model with machine learning. *Frontiers in Public Health*, 12, 1413031. <https://doi.org/10.3389/fpubh.2024.1413031>
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics: 5.1 (pp. 281–298). University of California Press. <https://projecteuclid.org/ebooks/berkeleysymposium-on-mathematical-statistics-andprobability/Proceedings-of-the-Fifth-Berkeley-Symposium-onMathematical-Statistics-and/chapter/Some-methods-forclassification-and-analysis-of-multivariateobservations/bsmsp/1200512992>
- Maugeri, A., Barchitta, M., Favara, G., La Mastra, C., La Rosa, M. C., Magnano San Lio, R., & Agodi, A. (2023). The Application of Clustering on Principal Components for

- Nutritional Epidemiology: A Workflow to Derive Dietary Patterns. *Nutrients*, 15(1), 195. <https://doi.org/10.3390/nu15010195>
- Muñoz Muñoz, E. G., Verduga Alcívar, D. A., Yandri Francinet, G. A., Lapo Palacios, M. A., & Zorrilla Briones, O. (2024). Búsqueda de patrones con machine learning en datos de siniestros de tránsito. *Ciencia Latina Revista Científica Multidisciplinar*, 8(2), 1617–1638. https://doi.org/10.37811/cl_rcm.v8i2.10592
- Toalombo Vargas, V. M., & Montalvo Márquez, F. J. (2024). Severity of traffic accidents in Ecuador. *Tesla Revista Científica*, 4(1), e290. <https://doi.org/10.55204/trc.v4i1.e290>
- World Health Organization. (2023). *Global status report on road safety 2023*. World Health Organization. <https://www.who.int/publications/i/item/9789240086517>
- Yang, W., & Han, S. (2025). Direct and indirect effects of road attributes on traffic safety. *Journal of Safety Research*, 93, 156–169. <https://doi.org/10.1016/j.jsr.2025.02.017>
- Zeng, Y., Qiang, Y., Zhang, N., Yang, X., Zhao, Z., & Wang, X. (2024). An Influencing Factors Analysis of Road Traffic Accidents Based on the Analytic Hierarchy Process and the Minimum Discrimination Information Principle. *Sustainability*, 16(16), 6767. <https://doi.org/10.3390/su16166767>
- Zhang, X., Qi, S., Zheng, A., Luo, Y., & Hao, S. (2023). Data-driven analysis of fatal urban traffic accident characteristics and safety enhancement research. *Sustainability*, 15(4), 3259. <https://doi.org/10.3390/su15043259>