



**udp** UNIVERSIDAD  
DIEGO PORTALES

# **CARACTERIZACIÓN Y CLASIFICACIÓN DE CONDUCTORES BASADA EN CARACTERÍSTICAS SOCIODEMOGRÁFICAS Y MÉTRICAS DE RIESGO**

**CLAUDIA JIMENA AHUMADA MOYA**

Tesis para optar al Título de Ingeniera Civil Industrial

Profesor Guía: Raúl Pezoa Zamorano

**FACULTAD DE INGENIERÍA Y CIENCIAS  
ESCUELA DE INGENIERÍA INDUSTRIAL**

Santiago, Chile — 2025



**udp** UNIVERSIDAD  
DIEGO PORTALES

# **CARACTERIZACIÓN Y CLASIFICACIÓN DE CONDUCTORES BASADA EN CARACTERÍSTICAS SOCIODEMOGRÁFICAS Y MÉTRICAS DE RIESGO**

**CLAUDIA JIMENA AHUMADA MOYA**

Tesis para optar al Título de Ingeniera Civil Industrial

Profesor Guía: Raúl Pezoa Zamorano

Profesor Comisión: Franco Basso Sotz

**FACULTAD DE INGENIERÍA Y CIENCIAS  
ESCUELA DE INGENIERÍA INDUSTRIAL**

Santiago, Chile — 2025

---

## Agradecimientos

Al concluir esta etapa, me invade una profunda emoción y gratitud. Han sido cinco años intensos, llenos de aprendizajes, desafíos, buenos momentos, risas, amistades, exámenes y desvelos. Nada de esto habría sido posible sin el apoyo incondicional de mi familia, mis amigos, mi pareja y los docentes que me acompañaron en este proceso. A todos ellos, deseo expresar mi más sincero agradecimiento.

A mi familia, por ser el pilar fundamental en todo este proceso formativo. A Luis y Jimena, gracias por esperarme cada día después de largas jornadas universitarias, por preocuparse siempre por mí y recordarme la importancia del esfuerzo y la perseverancia. Espero retribuirles cada día por todo su amor y apoyo. A Alejandra, gracias por ser mi hermana, amiga y muchas veces mi segunda madre. Compartir esta etapa universitaria juntas fue una de las mejores experiencias; las noches de estudio, las conversaciones y el apoyo mutuo quedarán siempre en mi memoria. A Daniel, gracias por ser ese hermano incondicional que siempre entendió mis ausencias y apoyó mis esfuerzos con cariño y paciencia.

A mi pareja, Julio, no existen palabras suficientes para expresar lo agradecida que estoy por todo tu apoyo durante estos años. Estuviste ahí en los momentos más difíciles, cuando quise rendirme, y lograste hacer que todo fuera más llevadero. Gracias por facilitarme la vida, por confiar en mí incluso más de lo que yo lo hacía, y por acompañarme con tanto amor durante todo este proceso, especialmente en la tesis.

A mis amigas de la universidad, Paula, Stephanie, Karla y Fabiana, por ser una fuente constante de inspiración y apoyo. Porque una siempre brilla más cuando está rodeada de mentes brillantes y corazones nobles. Conocerlas ha sido uno de los mayores regalos que me dejó la universidad. Este camino no fue fácil, pero sí mucho más ameno al lado de ustedes. Atesoro cada momento compartido y espero que nuestra amistad siga creciendo más allá de esta etapa.

A mis amigas del alma, Javiera, Fernanda, Alondra, Ignacia, Antonia y Constanza, gracias por creer siempre en mí, por acompañarme en los momentos más difíciles —especialmente durante la tesis— y por recordarme todo lo que vendría después de terminar esta etapa. Gracias por su cariño incondicional, por las risas, los consejos, los viajes y los sueños compartidos. Cada una de ustedes ocupa un lugar muy especial en mi corazón, y saber que me acompañan en este cierre es uno de mis mayores orgullos.

A los profesores y profesoras de la Universidad Diego Portales, admiro profundamente la vocación y compromiso con que ejercen su labor, y me llevo de muchos de ellos aprendizajes que van más allá de lo académico. En especial, quiero agradecer a mi profesor guía de tesis, Raúl Pezoa, por su dedicación, empatía y compromiso durante todo este proceso. Gracias por tomarse el tiempo, semana a semana, de guiarme, por su paciencia y por ser un profesional que realmente admiro.

Finalmente, a la Universidad Diego Portales, por todos estos años de formación y por permitirme vivir una de las experiencias más enriquecedoras de mi vida: mi intercambio en la Universidad de Alicante durante el primer semestre de 2024, una vivencia que marcó profundamente mi crecimiento personal y profesional.

---

A todos quienes creyeron en mí, que me acompañaron en cada paso y celebraron cada logro, les dedico este resultado. Este logro también es suyo.

---

## Resumen

El presente estudio analiza los estilos de conducción en la Autopista Central de Santiago y su relación con la ocurrencia de accidentes. A partir de datos reales del sistema de Identificación Automática de Vehículos (AVI) entre marzo y mayo de 2024, complementados con registros de accidentes del Ministerio de Obras Públicas y antecedentes sociodemográficos, se construyeron indicadores de velocidad, *headway*, tiempo hasta la colisión (TTC), nivel de conflicto y uso de carriles. Mediante el algoritmo de agrupamiento *K-Means* se identificaron cuatro perfiles de conducción: flujo estable, conducción conservadora, alta exposición y conducción rápida. Adicionalmente, se realizó una caracterización sociodemográfica de los conductores en cada clúster, vinculando los registros de patentes con bases del Registro Electoral y la Encuesta CASEN (2022), lo que permitió estimar variables como edad, sexo e ingreso promedio comunal. Posteriormente, se analizó si la composición de estos clústeres variaba sistemáticamente en torno a la ocurrencia de accidentes. Los resultados muestran que las diferencias entre las ventanas previas y las condiciones habituales son pequeñas y no estadísticamente significativas, lo que indica una alta estabilidad del flujo vehicular y descarta una relación causal entre la presencia de un tipo de conductor y la probabilidad de accidente. Estos hallazgos sugieren que los perfiles de conducción no son, por sí solos, un predictor suficiente de riesgo vial, destacando la necesidad de integrar variables contextuales —como condiciones del tráfico y meteorología— para mejorar los modelos predictivos de seguridad vial.

Código Junio?

---

## Abstract

This study analyzes driving behavior patterns on Santiago's Autopista Central and their relationship with traffic accidents. Using real vehicle-by-vehicle data from the Automatic Vehicle Identification (AVI) system between March and May 2024, complemented by accident records from the Ministry of Public Works and sociodemographic data, indicators of speed, headway, time-to-collision (TTC), conflict level, and lane usage were constructed. A *K-Means* clustering algorithm identified four driving profiles: stable flow, conservative driving, high exposure, and fast driving. Additionally, a sociodemographic characterization of drivers in each cluster was performed by linking vehicle registration data with the Electoral Register and CAsEN (2022) survey, estimating variables such as age, gender, and average household income by commune. Subsequently, the study examined whether the distribution of these clusters changed systematically around accident events. Results indicate that differences between pre-accident and usual conditions are minimal and statistically insignificant, suggesting a high degree of stability in traffic flow and no causal relationship between specific driver types and accident occurrence. These findings imply that driving profiles alone are insufficient predictors of accident risk, highlighting the importance of incorporating contextual variables—such as traffic density and weather conditions—to enhance predictive road safety models.

---

# Índice

<b>Resumen</b>	<b>3</b>
<b>Abstract</b>	<b>4</b>
<b>Índice de Figuras</b>	<b>6</b>
<b>Índice de Tablas</b>	<b>7</b>
<b>1. Introducción</b>	<b>8</b>
<b>2. Revisión de la literatura</b>	<b>9</b>
<b>3. Descripción de los datos</b>	<b>10</b>
3.1. Datos de tránsito . . . . .	10
3.2. Datos de accidentes . . . . .	13
3.3. Datos de conductores . . . . .	14
<b>4. Metodología</b>	<b>17</b>
4.1. Construcción de variables y ponderación . . . . .	17
4.2. Algoritmo de clusterización <i>K-Means</i> . . . . .	18
4.3. Selección del número de clusters y variables . . . . .	19
<b>5. Resultados</b>	<b>21</b>
5.1. Selección del número de clusters . . . . .	21
5.2. Clusterización y caracterización . . . . .	23
5.3. Relación entre la ocurrencia de accidentes y la composición de clústeres . . .	29
<b>6. Conclusiones</b>	<b>32</b>
<b>Referencias</b>	<b>34</b>

---

# Índice de figuras

1.	Autopista Central, Santiago, Chile . . . . .	11
2.	Distribución de accidentes por hora del día. . . . .	14
3.	Proporción de conductores por sexo. . . . .	15
4.	Distribución de conductores por comuna de residencia. . . . .	15
5.	Distribución de edad estimada de los conductores . . . . .	16
6.	Distribución de conductores por clúster en el espacio PCA . . . . .	23
7.	Proporción promedio de horarios de conducción por clúster. . . . .	28



---

## Índice de Tablas

1.	Variables del sistema de Identificación Automática de Vehículos (AVI). . . .	11
2.	Resumen mensual del flujo vehicular registrado por el sistema AVI. . . . .	12
3.	Variables del registro de accidentes del Ministerio de Obras Públicas (MOP). . . . .	13
4.	Variables consideradas en el análisis de accidentes. . . . .	13
5.	Descripción de las variables utilizadas para la clusterización y caracterización de los conductores. . . . .	17
6.	Estadística descriptiva de las variables utilizadas para la clusterización. . . .	22
7.	Métricas internas de validez para la selección del número de clústeres . . . .	22
8.	Resumen de características promedio por clúster. . . . .	24
9.	Características sociodemográficas promedio por clúster. . . . .	26
10.	Tipos de vehículo predominantes por clúster. . . . .	26
11.	Proporción promedio de horarios de conducción por clúster. . . . .	27
12.	Resultados del test $t$ pareado por clúster y escenario. . . . .	31

---

# 1. Introducción

La seguridad vial continúa siendo una de las principales preocupaciones en las autopistas y carreteras de Chile. A pesar de los esfuerzos por reducir los accidentes, las estadísticas de siniestros muestran una tendencia preocupante. Durante el año 2023 se registraron 78,238 siniestros de tránsito en el país, resultando en 1,635 fallecidos y 45,679 lesionados (Conaset, 2023). Al comparar estas cifras con las del año 2013, donde se reportaron 73,276 siniestros, 1,623 fallecidos y 59,592 lesionados, se evidencia un aumento significativo en el número total de accidentes, aunque se observó una leve reducción en la cantidad de fallecidos (Conaset, 2023).

En términos generales, durante los últimos 10 años, la cantidad de siniestros ha mostrado un crecimiento constante, impulsado principalmente por factores como el aumento del parque vehicular y los problemas persistentes de imprudencia al volante (Lizama Cisternas and Moreno Alveal, 2011). Este contexto subraya la importancia de analizar los comportamientos de los conductores para entender mejor los factores detrás de los accidentes de tránsito. Estudios anteriores muestran que las principales causas de los accidentes son la falta de cumplimiento de normas de tránsito, como no respetar la prioridad de paso, la conducción en sentido contrario y el exceso de velocidad. Estos comportamientos incrementan significativamente el riesgo de accidentes, especialmente en intersecciones y carreteras con alta densidad vehicular (Karacasu et al., 2014). Además, se ha identificado que factores como el estrés, la fatiga y las distracciones temporales también afectan negativamente el desempeño al volante, reduciendo la capacidad de reacción ante contingencias viales y aumentando el riesgo de siniestros (Petridou and Moustaki, 2000).

En esta tesis, se estudiarán los factores sociodemográficos y de comportamiento de los conductores para evaluar cómo inciden en la tasa de accidentabilidad en las autopistas. Utilizando datos de Autopista Central en Santiago de Chile, se realizará un análisis vehículo a vehículo, lo que permitirá identificar patrones y perfiles específicos de conducción y clasificarlos según su nivel de riesgo. Posteriormente estos perfiles serán utilizados como una variable explicativa de ocurrencia de accidente. A partir de los resultados obtenidos, se podrán proponer medidas preventivas más efectivas, enfocadas en los conductores de mayor riesgo, lo que contribuirá a la reducción de accidentes y a la mejora de la seguridad en las carreteras del país.

La literatura sobre seguridad en carreteras ha explorado diversos métodos, como modelos de visión por computadora (Sivaraman and Trivedi, 2013), análisis de datos simulados y naturalistas (Ding et al., 2022), para identificar patrones de riesgo en el comportamiento vehicular. Sin embargo, la mayoría de estos enfoques se basa en datos simulados o instrumentados (Miyajima et al., 2007), lo que introduce sesgos y limita la representatividad de los resultados. En esta investigación, se utilizarán datos reales recopilados durante un periodo amplio, integrando información demográfica y de comportamiento de conducción. Este enfoque no solo proporciona mayor precisión, sino que también aborda de otra manera la identificación de perfiles de conducción y su asociación con niveles de riesgo, un aspecto aún poco estudiado en la literatura.

El resto de este trabajo se organiza de la siguiente forma. En la sección 2 se presenta la revisión de la literatura, donde se analizan estudios previos relacionados con el comportamiento

---

vehicular, factores de riesgo y metodologías aplicadas en la identificación de patrones de conducción. la sección 3 establece los objetivos generales y específicos; la sección 4 presenta la hipótesis, que se complementa con los resultados esperados en la sección 5; finalmente, la sección 6 define los alcances de la investigación y la sección 7 describe la metodología que se empleará.

## 2. Revisión de la literatura

En el ámbito de la seguridad vial, múltiples estudios han explorado metodologías avanzadas para analizar el comportamiento de los conductores y predecir riesgos en carretera. Está investigación busca innovar al analizar perfiles de conducción a partir de las placas patentes de los vehículos que circulan por la autopista, un enfoque poco estudiado hasta ahora.

Sivaraman and Trivedi (2013) analiza el comportamiento vehicular mediante visión por computadora, utilizando datos obtenidos a partir de imágenes y secuencias de video del tráfico vehicular. Este estudio combina técnicas de detección basadas en algoritmos de aprendizaje, como SVM y AdaBoost, junto con modelos espacio-temporales para predecir maniobras como cambios de carril y adelantamientos.

Weng and Meng (2012) analiza factores de riesgo en zonas de trabajo en autopista mediante el modelo de árbol de decisión, clasificando a conductores según variables como el entorno, el vehículo y las características de él. Este estudio fue realizado con una base de datos de accidentes de tráfico del año 2004 al 2008, en China. Los resultados destacan que condiciones adversas, vehículos antiguos y giros incrementan significativamente el riesgo de accidentes.

Ding et al. (2022) identifica patrones en la conducción, utilizando datos simulados con un modelo GMM para identificarlos. La velocidad y la distancia de seguimiento son suficientemente únicos para distinguir entre diferentes conductores. Similarmente, Miyajima et al. (2007) analiza el uso de los pedales del vehículo (acelerador y freno) con resultados similares, además identifica la identidad del conductor.

Monselise et al. (2019) identifica tareas como la manipulación de dispositivos, cambios bruscos de carril, frenados repentinos y aceleraciones rápidas como los factores más predictivos de accidentes. Por otro lado, Barnard et al. (2016) identificó comportamientos de riesgo como distracción, agresividad y negligencia en eventos críticos y patrones cotidianos, este análisis ocupó datos naturalistas a través de vehículos instrumentados, cámaras y sensores al igual que el estudio anterior.

de Winter and Dodou (2010) resalta la influencia de las violaciones intencionales en la probabilidad de accidentes. Por su parte, Kim et al. (2006) analiza cómo las características del conductor (edad, género, consumo de alcohol) y del vehículo (antigüedad, sistemas de seguridad) impactan tanto el riesgo como las tasas de supervivencia. Concluye que los cinturones de seguridad y vehículos más pesados aumentan las tasas de supervivencia, mientras que los adultos mayores, mujeres y pasajeros en asientos delanteros tienen menor probabilidad de sobrevivir. Los accidentes son más frecuentes con conductores jóvenes ( $\leq 24$  años), mayores ( $\geq 65$  años), mujeres, aquellos que consumen alcohol, y con ciertos tipos de vehículos más ligeros.

---

Di Milia et al. (2011) analiza cómo los factores demográficos, como la edad, sexo y estado socioeconómico y las condiciones laborales como turnos, horas extra y traslados) influyen en la fatiga y el riesgo de accidentes. La metodología consiste en una revisión de literatura de estudios previos. Se utilizaron autoinformes, mediciones objetivas y datos de accidentes laborales y viales. Los resultados muestran que la fatiga aumenta con la edad, afecta más a mujeres y trabajadores de bajo nivel socioeconómico, y está asociada a largas jornadas, turnos nocturnos y tiempos de traslado prolongados.

Hussain et al. (2025) analiza cómo factores demográficos (edad, género, experiencia y nivel educativo), tipos de vehículos líderes y límites de velocidad influyen en el comportamiento de seguimiento de vehículos y el riesgo de accidentes mediante simulaciones de conducción con 61 participantes. Los resultados muestran que el riesgo de colisión es mayor a 50 km/h, especialmente al seguir motos, debido a distancias de seguimiento más cortas. Las mujeres y los conductores con menor educación presentan mayor riesgo, mientras que conductores más experimentados reaccionan más rápido pero tienden a seguir vehículos más de cerca.

Aunque se han explorado múltiples enfoques para analizar el comportamiento vehicular, esta investigación se enfocará en utilizar datos reales de un periodo amplio, integrando información demográfica y de comportamiento de conducción de todos los conductores. Este enfoque difiere de lo abordado en revisiones previas, ya que emplearemos datos y métodos que no han sido utilizados con anterioridad, lo que presenta ventajas en términos de precisión. Además, abordaremos uno de los principales desafíos de investigaciones previas, donde el uso de datos o métodos no del todo fiables ha introducido sesgos que pueden afectar la validez de los resultados.

### **3. Descripción de los datos**

#### **3.1. Datos de tránsito**

Para esta investigación se utilizaron registros desagregados vehículo-a-vehículo del sistema de Identificación Automática de Vehículos (AVI) de Autopista Central, esta tiene una longitud de 60.5 km y cruza la ciudad de norte a sur a través de los ejes Ruta 5 y General Velásquez, mostrados en la Figura 1. El período de análisis corresponde a los meses de marzo, abril y mayo de 2024 e incluye 31 pórticos de control. Cada observación contiene la fecha-hora de paso por pórtico, la matrícula del vehículo, el identificador del pórtico, el carril de circulación, la categoría vehicular y la velocidad instantánea (km/h).

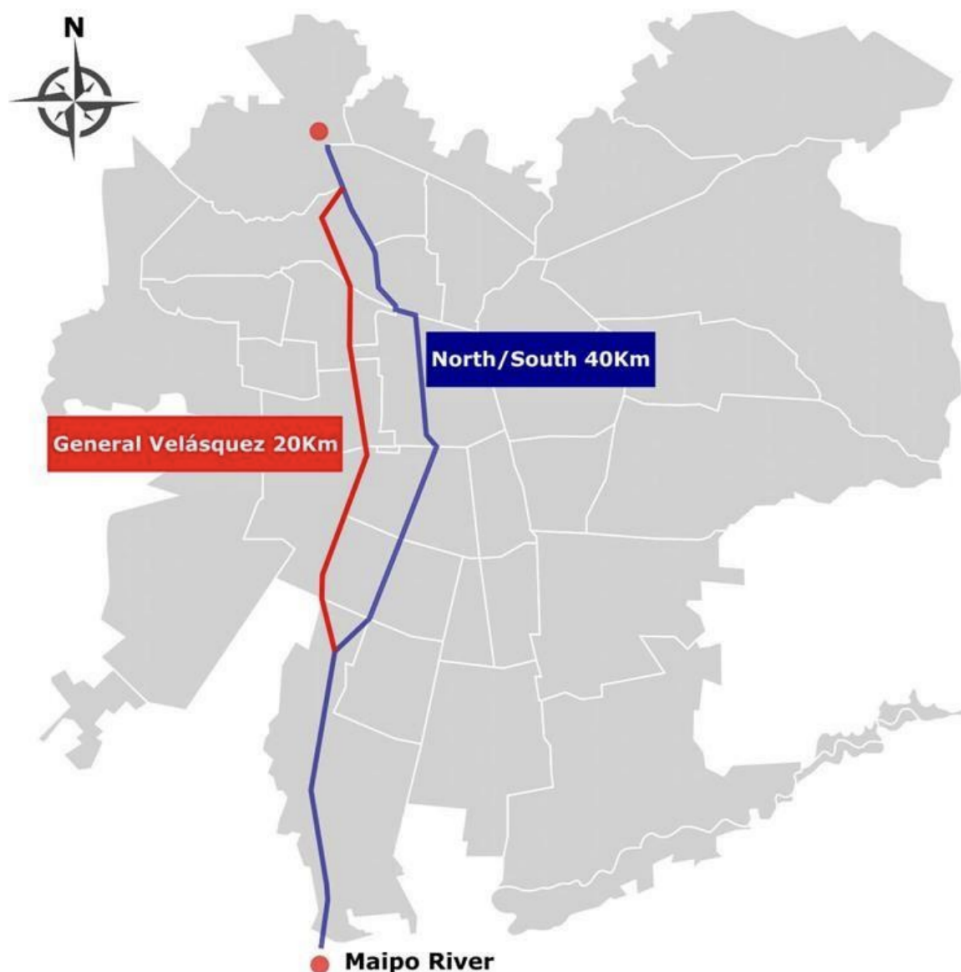


Figura 1: Autopista Central, Santiago, Chile

*+ outliers según velocidad*

Con el objetivo de asegurar consistencia en los datos, se aplicó una primera limpieza que excluye los carriles 0 y 4. Tras este filtro se conserva aproximadamente el 95 % de las observaciones mensuales. La Tabla 1 describe las variables de la base AVI, mientras que la Tabla 2 presenta un resumen mensual del flujo vehicular total, la velocidad promedio y la distribución del tránsito por carril.

Tabla 1: Variables del sistema de Identificación Automática de Vehículos (AVI).

Variable	Descripción
Fecha	Fecha y hora en que el vehículo pasó por el pórtico.
Matrícula	Patente del vehículo.
Pórtico	Identificador del pórtico.
Categoría	Tipo de vehículo (ligero, pesado, motocicleta).
Carril	Carril de paso (derecho, central o izquierdo).
Velocidad	Velocidad instantánea (km/h).

*- Filtro por tamaño matrícula (5-6)  
- velocidades > 0 y head > 0*

Tabla 2: Resumen mensual del flujo vehicular registrado por el sistema AVI.

Mes	Flujo vehicular	Vel. prom. (km/h)	Carril der.	Carril cent.	Carril izq.
Marzo	46,271,155	85.614	12,494,477	17,348,108	16,428,570
Abril	45,634,285	84.250	12,428,674	16,890,326	16,315,285
Mayo	44,001,485	84.010	12,133,673	16,397,994	15,469,818
<b>Promedio</b>	<b>45,302,308</b>	<b>84.625</b>	<b>12,352,275</b>	<b>16,878,809</b>	<b>16,071,224</b>

Como se observa en la Tabla 2, el flujo vehicular presenta un comportamiento relativamente estable entre los meses analizados, la velocidad promedio también se mantiene constante, en torno a los 84 km/h, lo que refleja una circulación fluida y sin anomalías notorias durante el período.

Respecto a la distribución por carril, se aprecia que la mayor parte del flujo circula por el carril central (aproximadamente 37 % del total), seguido del carril izquierdo y finalmente el derecho. Esta tendencia es coherente con la dinámica típica de autopistas urbanas, donde los carriles centrales concentran el tránsito de paso y los laterales se utilizan para adelantamientos o incorporaciones.

---

### 3.2. Datos de accidentes

De manera complementaria, se utilizaron registros de accidentes ocurridos en las vías, provistos por el Ministerio de Obras Públicas (MOP) para el año 2024. Estos registros incluyen información detallada sobre la fecha y hora de inicio, el tramo, la calzada, el kilómetro y el tipo de evento.

Tabla 3: Variables del registro de accidentes del Ministerio de Obras Públicas (MOP).

Variable	Descripción
Fechas Inicio	Fecha de ocurrencia del accidente.
Hora Inicio	Hora de ocurrencia del accidente.
Tipo	Clasificación del evento (se retienen solo accidentes).
Tramo	Tramo de la autopista donde ocurrió el evento.
Calzada	Sentido de circulación asociado al evento.
Km.	Kilómetro de referencia del accidente.

Para este estudio se consideraron exclusivamente los eventos clasificados como accidentes, descartando cualquier otro tipo de incidencia vial. Cada accidente fue analizado considerando los pórticos ubicados aguas arriba (AA1–AA2) y aguas abajo (AB1–AB2) del punto de ocurrencia, con el fin de observar el comportamiento del flujo vehicular en torno al siniestro.

Tabla 4: Variables consideradas en el análisis de accidentes.

Variable	Descripción
Tramo	Identificador del tramo vial.
Calzada	Sentido del tramo (oriente/poniente).
Km	Kilómetro de referencia del tramo.
Aguas Arriba 1–2	Pórticos más cercanos en sentido ascendente.
Aguas Abajo 1–2	Pórticos más cercanos en sentido descendente.

Para caracterizar temporalmente los siniestros, la Figura 2 muestra la distribución de accidentes por horas del día.

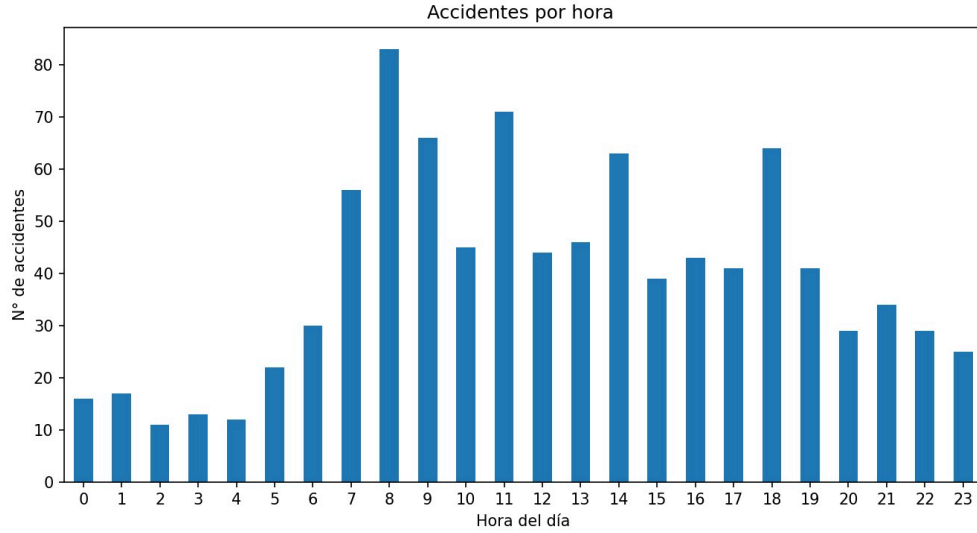


Figura 2: Distribución de accidentes por hora del día.

La Figura 1 muestra que la mayor concentración de accidentes ocurre durante los periodos de mayor congestión vehicular, particularmente entre las 7:00 y las 9:00 horas, y nuevamente entre las 17:00 y las 19:00 horas. Estos intervalos coinciden con los horarios de mayor flujo vehicular en las autopistas de Santiago de Chile.

### 3.3. Datos de conductores

Además de las bases mencionadas anteriormente, se complementó el análisis con una base socio-demográfica de los conductores que incorpora RUT, nombre, sexo, domicilio electoral, región y comuna. Se trabajó bajo el supuesto de que el propietario corresponde al conductor habitual, se caracterizaron las distribuciones de sexo y comuna de residencia.

Del total de 1.048.575 registros originales, se conservaron 1.026.421 observaciones válidas (97,9%). Entre los registros válidos, solo 827.071 casos presentaron sexo identificado (masculino/femenino). Aproximadamente, 62,1% corresponden a hombres y 37,9% a mujeres representado en la Figura 3. La Figura 4 muestra la distribución por comuna, se observa que las comunas de Maipú, Puente Alto y San Bernardo concentran la mayor cantidad de registros, seguidas por Las Condes, La Florida y Santiago, lo que refleja la relación entre densidad poblacional, nivel socioeconómico y tenencia vehicular en la Región Metropolitana.



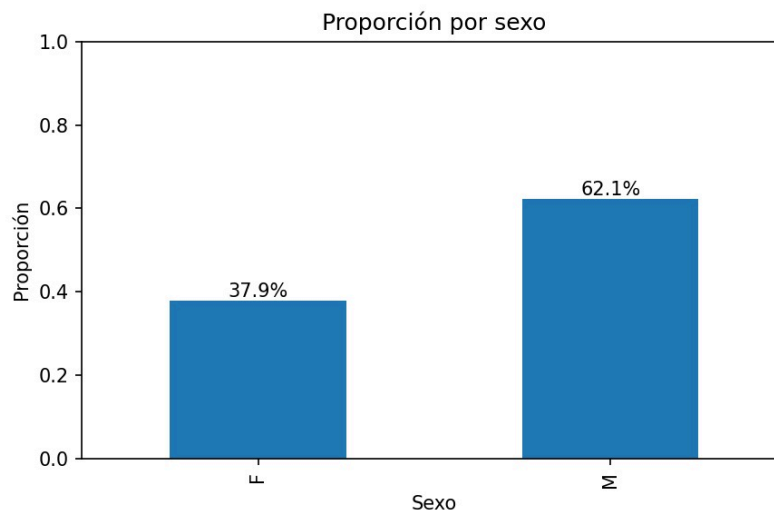


Figura 3: Proporción de conductores por sexo.

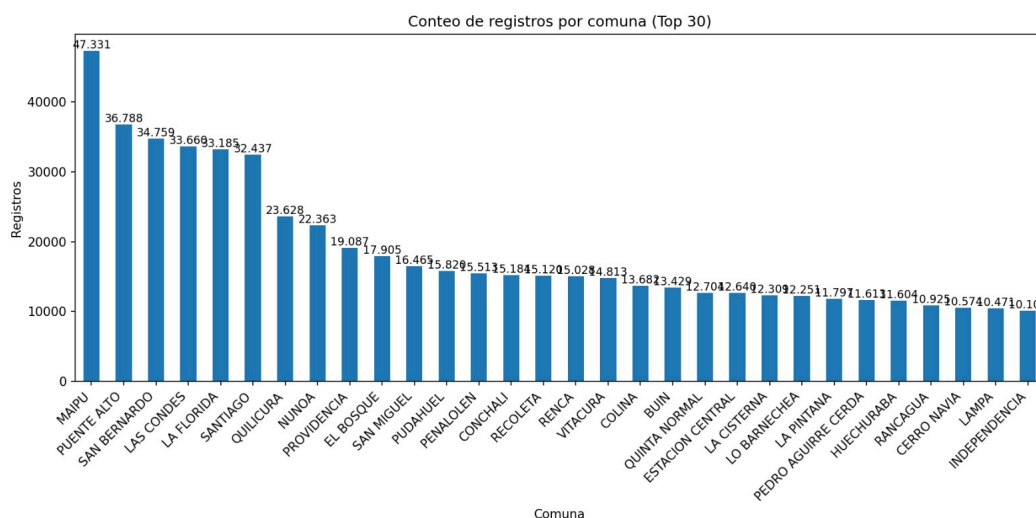


Figura 4: Distribución de conductores por comuna de residencia.

Adicionalmente, se estimó la distribución etaria de los conductores a partir del número de RUT, esto permitió obtener una edad promedio cercana a los 50 años, con un rango entre 18 y 89 años.

Se identificó que aproximadamente un 3,1 % de los registros presentaba edades inferiores a 18 años, lo cual se asocia principalmente a RUT extranjeros o formatos no válidos, estos casos fueron excluidos del análisis.

La Figura 5 muestra la distribución etaria estimada, donde se observa una mayor concentración entre los 35 y 55 años, evidenciando que la mayoría de los propietarios corresponde a grupos etarios laboralmente activos.

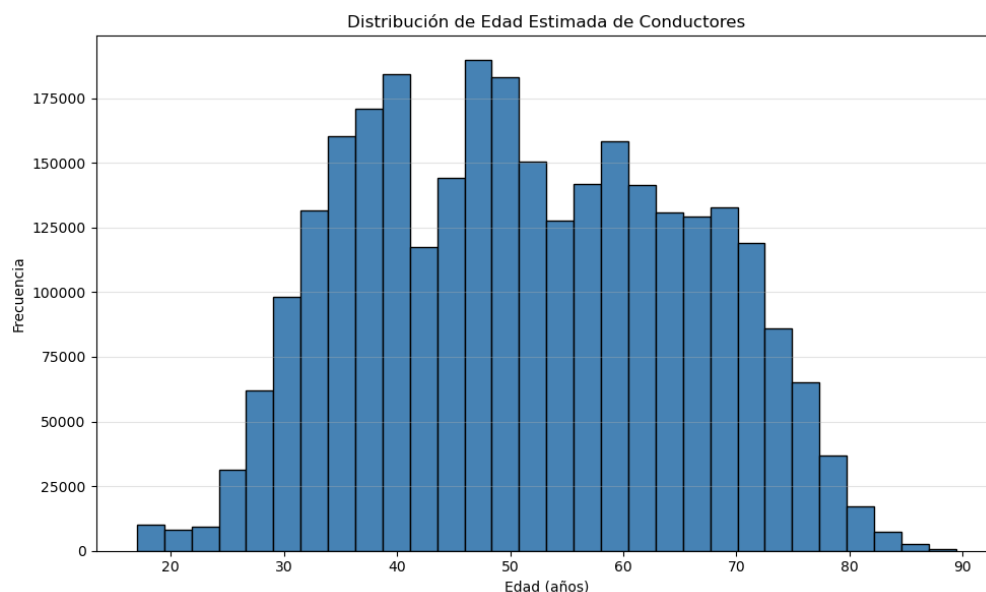


Figura 5: Distribución de edad estimada de los conductores

En conjunto, las variables de sexo, comuna y edad permiten caracterizar el perfil de los conductores en la Región Metropolitana, lo que será relevante al momento de analizar patrones de circulación y el comportamiento vial.

La integración de las tres fuentes —AVI, accidentes y socio-demografía— permite construir un dataset consolidado que enlaza, para cada matrícula y pórtico, las condiciones de circulación (velocidad, carril, paso) con el contexto de accidentabilidad y atributos del conductor. Esta base es el insumo para la construcción de variables agregadas por matrícula y la posterior clusterización.

## 4. Metodología

En esta sección se presenta la metodología utilizada para el análisis de los datos. Primero, en la Sección 4.1, se detallan las variables construidas por matrícula a partir del flujo vehicular y, al final de esa misma subsección, se explica la ponderación utilizada para hacer comparables y estables los indicadores. Luego, en la Sección 4.2, se describe el algoritmo de clusterización *K-Means* y las consideraciones prácticas de su aplicación. Finalmente, en la Sección 4.3, se expone el criterio para definir el número de clusters.

### 4.1. Construcción de variables y ponderación

A partir de la base descrita en la Sección 3.1, se construyó un conjunto de variables que busca representar el comportamiento de conducción de cada conductor. Las variables utilizadas se presentan en la Tabla 5, la cual muestra el nombre de la variable y su descripción.

Tabla 5: Descripción de las variables utilizadas para la clusterización y caracterización de los conductores.

Variable	Descripción
Velocidad promedio	Desplazamiento de un objeto dividido entre el tiempo transcurrido; en la práctica, corresponde a la media de la velocidad instantánea observada al cruzar los pórticos (km/h).
Headway promedio	Diferencia temporal entre dos vehículos consecutivos que pasan por un mismo pórtico y carril; se promedia por matrícula para representar su separación típica (s).
Time to Collision (TTC)	Tiempo estimado que demoraría el vehículo seguidor en alcanzar al líder si ambos mantuvieran sus velocidades (s).
Proporción de conflicto	Proporción de pasadas en que se observa una condición riesgosa definida por un umbral de TTC ( $\leq \tau$ ).
Proporciones de carril	Proporción de pasadas en cada carril (izquierdo, central y derecho), que refleja preferencias de circulación.
Distribución horaria	Porcentaje de pasadas en franjas del día (Mañana, Tarde, Noche y madrugada).

Velocidad relativa!

A continuación se explican algunas variables:

**Time To Collision (TTC).** Para cada pasada  $p$  correspondiente a la matrícula  $i$ , se define  $v_{ip}$  como la velocidad del vehículo observado, y  $v_{ip,a}$  como la velocidad del vehículo anterior que circuló justo antes por el mismo pórtico y carril. Sea además  $HW_{ip}$  el headway temporal observado entre ambos vehículos, expresado en segundos. Bajo el supuesto de que las velocidades se mantienen constantes, el Time To Collision se calcula como:

$$TTC_{ip} = \frac{HW_{ip} v_{ip}}{v_{ip} - v_{ip,a}}.$$

*Handwritten calculation:*  $\frac{5 \cdot \frac{4}{h}}{\frac{4}{h}} = 5$

Si  $v_{ip} \leq v_{ip,a}$ , el vehículo  $i$  no se aproxima al que le antecede y, por lo tanto, se considera  $\text{TTC}_{ip} = +\infty$ . En el análisis se emplea el promedio por matrícula,

$$\overline{\text{TTC}}_i = \frac{1}{N_i} \sum_{p=1}^{N_i} \text{TTC}_{ip},$$

que representa el nivel de exposición promedio del conductor  $i$  al riesgo de alcance con el vehículo precedente.

**Proporción de conflicto.** Se define un *evento de conflicto* cuando el TTC cae por debajo de un umbral  $\tau$  (en segundos):

$$\text{Conflicto}_{ip} = \begin{cases} 1, & \text{si } \text{TTC}_{ip} \leq \tau, \\ 0, & \text{si } \text{TTC}_{ip} > \tau \end{cases} \quad \overline{C}_i = \frac{1}{N_i} \sum_{p=1}^{N_i} \text{Conflicto}_{ip}.$$

ttc - max ?

Es importante mencionar que el umbral  $\tau$  responde a la lógica de medidas sustitutas de seguridad, donde se considera que existe conflicto cuando un indicador supera un valor crítico; en particular, cuando el TTC es menor o igual a  $\tau$ , se asume una condición riesgosa

Las variables anteriores se calculan por matrícula. Para evitar que conductores con pocas observaciones distorsionen la caracterización global y la clusterización, se utiliza una ponderación por número de viajes (verificar si lo cambie a numero de viajes o numero de pasadas) Si  $w_i = N_i$  es el peso de la matrícula  $i$ , entonces cualquier promedio global  $\tilde{x}$  se obtiene como

$$\tilde{x} = \frac{\sum_i w_i \bar{x}_i}{\sum_i w_i}.$$

En la etapa de clusterización, esta decisión equivale a dar mayor estabilidad a las observaciones con mayor evidencia empírica, manteniendo la interpretabilidad de los centroides.

## 4.2. Algoritmo de clusterización *K-Means*

Para identificar perfiles de conducción se aplicó el algoritmo de agrupamiento no supervisado *K-Means*. Su objetivo es particionar un conjunto de  $n$  observaciones en  $k$  grupos, de modo que los elementos de cada grupo sean lo más similares posible entre sí y diferentes respecto de los otros grupos. Esto se logra minimizando la suma de las distancias cuadráticas entre cada observación y el centroide del cluster correspondiente:

$$\min_{C, \mu_1, \dots, \mu_k} J(C, \mu) = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|_2^2, \quad (1)$$

donde  $\mu_j$  es el centroide del cluster  $C_j$  y  $\|\cdot\|_2$  corresponde a la distancia euclidiana.

El algoritmo funciona de manera iterativa, alternando entre dos pasos:

---

(i) **Asignación.** Cada observación se asigna al cluster con el centroide más cercano:

$$\mathbf{x}_i \in C_j \iff \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2 \leq \|\mathbf{x}_i - \boldsymbol{\mu}_h\|_2^2, \quad \forall h \in \{1, \dots, k\}. \quad (2)$$

(ii) **Actualización.** Los centroides se recalculan como el promedio de los puntos asignados:

$$\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i, \quad j = 1, \dots, k. \quad (3)$$

El proceso se repite hasta que las asignaciones dejan de cambiar o la reducción de la función objetivo es menor que un umbral predefinido. K-means es un algoritmo eficiente para grandes volúmenes de datos. Además, su salida es fácilmente interpretable, ya que cada cluster puede describirse en función de su centroide, lo que facilita la caracterización de perfiles de conducción.

### 4.3. Selección del número de clusters y variables

La definición del número de variables a considerar y del número de clusters a utilizar en el algoritmo K-Means constituye una decisión central en el análisis. Dado que no existe un valor único “correcto” para estos parámetros, se optó por un enfoque comparativo sustentado en métricas internas, complementado con criterios de interpretabilidad de los resultados.

En la etapa exploratoria se realizaron pruebas eliminando variables de manera individual. Sin embargo, estas pruebas no mostraron mejoras relevantes en las métricas de agrupamiento, tales como Silhouette, Calinski–Harabasz y Davies–Bouldin. Por este motivo, se mantuvo el conjunto completo de variables calculadas, apoyando esta decisión en cuatro razones principales. En primer lugar, cada variable aporta una dimensión distinta del estilo de conducción. En segundo lugar, la exclusión de variables no se tradujo en mejoras consistentes en las métricas. En tercer lugar, las correlaciones observadas entre variables fueron moderadas y no reflejan la misma mecánica subyacente, sino fenómenos distintos. Finalmente, el número total de variables resultó acotado y manejable para K-Means, lo que no compromete la interpretación de los resultados.

La única excepción fue la exclusión de redundancias lineales evidentes, con el fin de evitar colinealidad estricta. En este sentido, la tercera proporción de carril fue omitida, ya que su valor es determinístico al cumplirse que la suma de las tres proporciones es igual a uno.

El análisis se apoyó principalmente en variables dinámicas como la velocidad, el headway, el tiempo hasta la colisión (TTC) y la ocurrencia de conflictos. A estas se añadieron las proporciones de carril, las variables horarias y la velocidad promedio, lo que permitió contextualizar de manera más completa la forma de conducción.

Finalmente, la selección del número de clusters  $k$  se realizó evaluando un rango de valores  $k \in \{2, \dots, K_{\text{máx}}\}$  a través de métricas internas estándar. El criterio de decisión buscó un equilibrio entre parsimonia e interpretabilidad, priorizando soluciones que mostraran consistencia en las métricas y que a la vez fueran comprensibles desde la perspectiva del comportamiento de conducción.

---

La elección de  $k$  se realizó comparando métricas internas estándar en un rango  $k \in \{2, \dots, K_{\text{máx}}\}$  y privilegiando la parsimonia e interpretabilidad. A continuación se describen las métricas y sus fórmulas

**(1) Coeficiente de Silhouette.** Para cada observación  $i$ , sea  $a(i)$  la distancia media de  $i$  a los puntos de su propio cluster y  $b(i)$  la mínima distancia media a los puntos de cualquier otro cluster. El índice individual es:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1],$$

y el valor reportado es el promedio  $\bar{s} = \frac{1}{n} \sum_{i=1}^n s(i)$ . Valores cercanos a 1 indican clusters compactos y bien separados.

**(2) Índice de Davies–Bouldin (DB).** Sea  $C_j$  el cluster  $j$  con  $n_j$  puntos y centroide  $\mu_j$ . Definimos la *dispersión* intra-cluster como

$$S_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|_2,$$

y la distancia entre centroides  $M_{jh} = \|\mu_j - \mu_h\|_2$ . Entonces,

$$DB = \frac{1}{k} \sum_{j=1}^k \max_{h \neq j} \frac{S_j + S_h}{M_{jh}}.$$

Valores más bajos indican mejor separación relativa.

**(3) Índice de Calinski–Harabasz (CH).** Sea  $\mu$  el centroide global,  $B_k$  la *dispersión entre clusters* y  $W_k$  la *dispersión intra-cluster*. Usando trazas:

$$\text{tr}(W_k) = \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|_2^2, \quad \text{tr}(B_k) = \sum_{j=1}^k n_j \|\mu_j - \mu\|_2^2.$$

El índice es

$$CH = \frac{\text{tr}(B_k)/(k-1)}{\text{tr}(W_k)/(n-k)}.$$

Valores altos reflejan mayor separación entre clusters relativa a su compactación interna.

El número de clusters  $k$  se determinó evaluando en conjunto las métricas internas: se privilegió aquel  $k$  que, de forma simultánea, mostrara un coeficiente de Silhouette promedio  $\bar{s}$  y un índice de Calinski–Harabasz (CH) elevados, un índice de Davies–Bouldin (DB) bajo. Cuando las métricas entregaron señales levemente dispares, se eligió la alternativa más parsimoniosa que preserva una separación nítida entre grupos y favorece su interpretabilidad.

---

## 5. Resultados

En esta sección se presentan los principales hallazgos obtenidos a partir del proceso de clusterización de conductores. Luego de definir el conjunto de variables y seleccionar el número de clústeres más adecuado mediante el análisis comparativo de métricas, se procedió a caracterizar los grupos resultantes y a evaluar sus diferencias en función de las variables de conducción y los antecedentes socio-demográficos.

El análisis busca identificar patrones de conducción presentes en la autopista y determinar si ciertos grupos de conductores tienden a estar más asociados a la ocurrencia de accidentes. Para ello, se consideraron ocho variables representativas del comportamiento en autopista: número total de viajes, velocidad promedio, velocidad relativa, headway promedio, nivel de conflicto, proporción de uso de los tres carriles y proporción de cambios de pista. A partir de estas dimensiones se buscó obtener una segmentación coherente que permitiera, comparar la distribución de los grupos en condiciones normales de flujo y en tramos cercanos a accidentes.

### 5.1. Selección del número de clusters

Las variables seleccionadas permiten observar el comportamiento de los conductores desde diferentes perspectivas. Por un lado, se consideró el nivel de exposición de cada vehículo a través del número total de viajes realizados. También se incluyeron medidas que reflejan cómo se desplazan longitudinalmente, como la velocidad promedio, la velocidad relativa y el headway entre vehículos. A esto se suman indicadores que capturan la interacción con el entorno —como el nivel promedio de conflicto— y variables que representan el comportamiento lateral, tales como el uso de carriles y los cambios de pista. En conjunto, estas dimensiones permiten caracterizar de forma integral los distintos estilos de conducción presentes en la autopista.

La Tabla 6 presenta la estadística descriptiva de las variables utilizadas para la clusterización, calculada a partir de la base consolidada del periodo marzo–mayo. A grandes rasgos, los resultados muestran una alta heterogeneidad entre conductores, especialmente en la cantidad total de viajes y en su headway, lo que evidencia la coexistencia de distintos niveles de exposición y estilos de manejo. Las velocidades promedio se concentran entre 75 y 93 km/h, en línea con los valores esperados en autopistas urbanas, mientras que la velocidad relativa promedio (0.98) indica un comportamiento general acorde al flujo del entorno. En contraste, variables como la proporción de cambios de pista y el uso de carriles presentan mayor dispersión, lo que sugiere la presencia de conductores con patrones más activos o agresivos frente a otros más estables. En conjunto, estas características confirman que las variables seleccionadas capturan adecuadamente la diversidad del flujo vehicular, otorgando una base sólida para la clusterización.

Tabla 6: Estadística descriptiva de las variables utilizadas para la clusterización.

Variable	Media	Desv. Est.	Mín.	P25	P50	P75	Máx.
Número de viajes	20.281	37.325	1.0	2.0	6.0	20.0	1457.0
Velocidad promedio	83.538	14.016	1.0	75.2	84.2	92.8	187.0
<u>Velocidad relativa promedio</u>	0.980	0.117	0.018	0.914	0.988	1.055	3.873
Headway promedio	5.279	9.626	0.0	3.0	4.1	5.746	5279.5
Conflicto promedio	0.072	0.125	0.0	0.0	0.024	0.100	1.0
Proporción carril 1	0.318	0.265	0.0	0.125	0.250	0.447	1.0
Proporción carril 2	0.371	0.221	0.0	0.234	0.355	0.500	1.0
Proporción carril 3	0.311	0.251	0.0	0.089	0.286	0.500	1.0
Proporción cambio de pista	0.489	0.230	0.0	0.357	0.500	0.600	1.0

Antes de clusterizar se determinó el número óptimo de grupos ( $K$ ), para esto se evaluaron diferentes combinaciones mediante tres métricas internas de validación: el coeficiente de Silhouette, el índice de Calinski–Harabasz (CH) y el índice de Davies–Bouldin (DB). En este análisis se utilizaron  $N = 7$  variables, excluyendo la proporción del carril 3 por su colinealidad con las otras dos proporciones. La Tabla 7 muestra los valores obtenidos para cada métrica según el número de clusters considerado.

Tabla 7: Métricas internas de validez para la selección del número de clústeres

K	Silueta $\uparrow$	DB $\downarrow$	CH $\uparrow$
2	0.2135	1.6114	<b>15029.17</b>
3	0.2191	1.7787	12784.84
4	0.2332	<b>1.2697</b>	14776.83
5	<b>0.2334</b>	1.3100	14222.33
6	0.2063	1.4365	12156.51
7	0.2048	1.2974	12279.04
8	0.1685	1.4528	11307.72
9	0.1639	1.3951	11051.88
10	0.1669	1.3346	10642.50

En el gráfico anterior, los mejores valores de cada métrica se muestran en negrita. Las tres métricas no maximizan/minimizan en el mismo  $K$ , por lo que la selección debe basarse en el equilibrio entre cohesión intra-cluster y separación inter-cluster, junto con la parsimonia del modelo. En particular: (i)  $K = 5$  obtiene el máximo Silhouette (0.2334), lo que sugiere una ligera mejora en la compacidad relativa respecto de  $K = 4$ ; sin embargo, esa ganancia es marginal y viene acompañada de un DB mayor (1.31) y un CH menor (14,222), señal de clusters algo menos definidos. (ii)  $K = 2$  presenta el máximo CH (15,029), consistente con grandes separaciones entre dos grupos muy agregados, pero lo hace con una Silhouette más baja (0.2135) y un DB alto (1.61), lo que indica heterogeneidad interna elevada dentro de cada conglomerado y, por ende, menor utilidad descriptiva. (iii)  $K = 4$  no lidera todas las métricas de forma individual, pero minimiza DB (1.2697) y mantiene valores competitivos en Silhouette (0.2332), prácticamente igual a  $K = 5$  y CH (14,777). Este patrón —DB mínimo



---

con Silhouette alta y CH elevado— sugiere una estructura con buena compacidad interna y fronteras definidas entre grupos, sin fragmentar en exceso el espacio de características.

Además, al aumentar  $K$  por encima de 5 se observa una degradación sistemática: Silhouette desciende y DB tiende a aumentar, mientras que CH disminuye, lo que es consistente con particiones más pequeñas y menos separadas que no aportan interpretabilidad adicional. En síntesis,  $K = 4$  ofrece el mejor compromiso entre las tres métricas y la complejidad del modelo, proporcionando una partición parsimoniosa que distingue perfiles de conducción con suficiente claridad sin generar grupos residuales o excesivamente específicos.

De esta forma, se seleccionaron cuatro clusters como el número más representativo de los distintos estilos de conducción. En los apartados siguientes se profundiza en la descripción de cada grupo, en su relación con la frecuencia de accidentes y las características sociodemográficas de los conductores.

## 5.2. Clusterización y caracterización

Definido el número de grupos óptimo, se aplicó *K-Means* sobre el conjunto de siete variables seleccionadas. Para fines descriptivos, la asignación de grupos se proyectó en dos componentes principales (PCA). Como es habitual, las componentes principales no poseen una interpretación física directa, por lo que la lectura sustantiva se realiza sobre las variables originales mediante los datos por clúster.

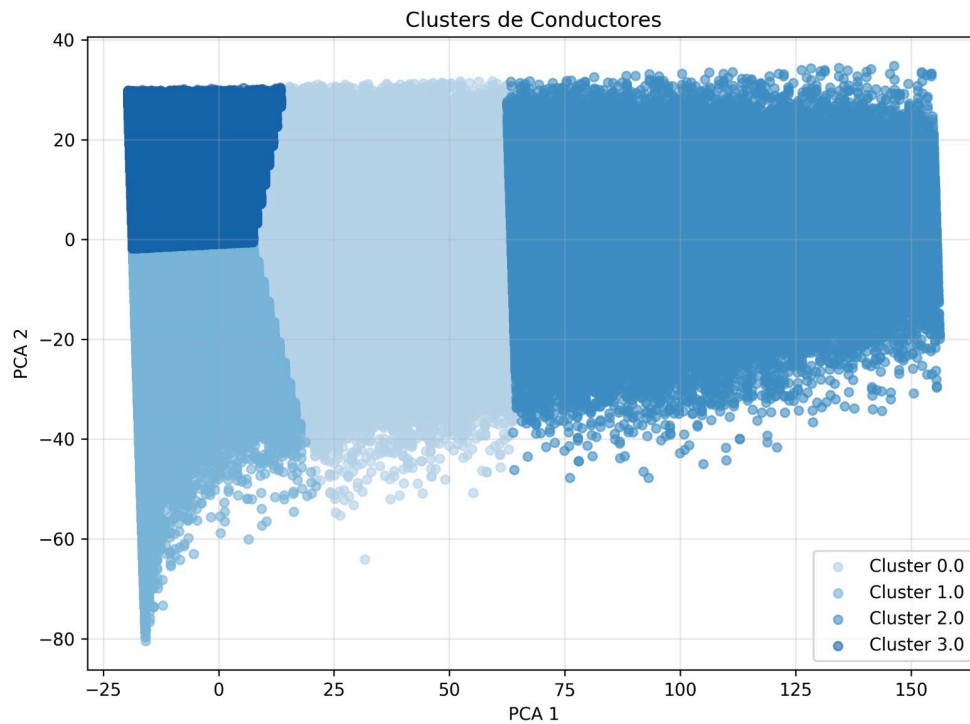


Figura 6: Distribución de conductores por clúster en el espacio PCA

La Figura 6 muestra que los clústeres ocupan regiones diferenciadas del plano, con una

superposición acotada. Sin embargo, dado que el PCA combina linealmente las variables y no preserva su significado original, la caracterización sustantiva se apoya en los promedios por grupo de la siguiente tabla

Tabla 8: Resumen de características promedio por clúster.

Clúster	Tamaño	%	Vel.	Vel. Rel.	Headway	Conflicto	Carril 1	Carril 2	Carril 3
0	270,938	14.38	82.74	0.990	5.06	0.074	0.314	0.365	0.321
1	651,839	34.60	72.62	0.916	4.46	0.049	0.428	0.387	0.185
2	104,364	5.54	81.72	0.994	5.01	0.078	0.317	0.351	0.332
3	856,954	45.48	93.57	1.037	5.06	0.068	0.200	0.377	0.423

Con respecto a la Tabla 8, al analizar conjuntamente las variables que caracterizan cada grupo, se identifican cuatro perfiles de conducción diferenciados en función de su velocidad, nivel de exposición, relación con el flujo vehicular y ocupación de carriles. A continuación, se describen en detalle.

**Clúster 0 — Conductores de flujo estable.** Este grupo representa un 14.4 % de la muestra y se caracteriza por una velocidad promedio intermedia (82.7 km/h) y una velocidad relativa cercana a la unidad (0.99), lo que sugiere una conducción alineada con el ritmo general del tránsito. El headway (5.06 s) se encuentra en un rango medio, indicando una distancia prudente, mientras que la distribución entre carriles es equilibrada (Carril 1: 0.314; Carril 2: 0.365; Carril 3: 0.321), sin preferencias marcadas por ninguna pista. El nivel de conflicto (0.074) es moderado, consistente con un patrón de conducción regular y sin comportamientos extremos. Este clúster puede asociarse a un estilo estable y representativo del flujo promedio de la autopista, constituyendo una referencia base frente a los demás clusters.

**Clúster 1 — Conductores de baja velocidad y perfil conservador.** Con una participación del 34.6 %, este es el segundo grupo más numeroso. Sus miembros presentan las velocidades promedio más bajas (72.6 km/h) y una velocidad relativa inferior a 1 (0.916), lo que refleja desplazamientos sistemáticamente más lentos que el flujo circundante. A su vez, el headway (4.46 s) es el más corto del conjunto, sugiriendo circulación en entornos de mayor densidad. El rasgo más distintivo de este grupo es su alta permanencia en el Carril 1 (0.428) y baja utilización del Carril 3 (0.185), patrón característico de conductores que priorizan estabilidad y menor exposición a adelantamientos. El conflicto promedio (0.049) es el más bajo de todos los clústeres, lo que refuerza la idea de una conducción precavida y poco interactiva con el entorno, asociada a perfiles de bajo riesgo. (asociarlo a horas de conducción de mayor frecuencia y el promedio de número de viajes entre cond del cluster)

**Clúster 2 — Conductores de alta exposición y conflicto elevado.** Este grupo minoritario (5.5 %) combina niveles de velocidad promedio similares al flujo (81.7 km/h; vel. rel. 0.994) con una frecuencia de viajes superior al promedio, evidenciando una mayor exposición en el sistema. El headway (5.01 s) es similar al de Cluster 0, pero su nivel de conflicto promedio (0.078) es el más alto de todos los grupos, lo que sugiere una conducción con mayor

---

interacción o tensión con el entorno, posiblemente derivada del uso intensivo de la red. Su distribución por carriles es equilibrada, aunque con ligera inclinación hacia el Carril 3 (0.332), coherente con desplazamientos de carácter más dinámico. En conjunto, este clúster puede interpretarse como un grupo de usuarios frecuentes con conducción activa y potencialmente más riesgosa.

**Clúster 3 — Conductores de alta velocidad en carriles rápidos.** Con un 45.5 % del total, constituye el grupo mayoritario. Presenta las mayores velocidades promedio (93.6 km/h) y una velocidad relativa superior a 1 (1.037), lo que refleja una tendencia a circular por sobre la velocidad promedio del flujo. El headway (5.06 s) no difiere significativamente del resto, pero la ocupación de carriles muestra una clara preferencia por el Carril 3 (0.423), a diferencia del Carril 1 (0.200), evidenciando un comportamiento de adelantamiento o tránsito sostenido en la pista rápida. El nivel de conflicto (0.068) se mantiene en valores moderados, lo que podría asociarse a conductores con mayor dominio o confianza en condiciones de velocidad elevada. Este patrón describe un perfil de conducción ágil, orientado al sobrepaso y al aprovechamiento de la vía rápida.

Los resultados evidencian la coexistencia de distintos estilos de conducción en la autopista Central de Chile, desde los más conservadores y de baja interacción (C1), hasta los de alta velocidad en carriles rápidos (C3), pasando por perfiles estables (C0) y usuarios de alta exposición con mayores niveles de conflicto (C2). La presencia de estos grupos es coherente con la dinámica heterogénea del tráfico urbano. Particularmente, el conflicto promedio emerge como una variable relevante para el análisis posterior de seguridad: los valores más elevados en C2 y, en menor medida, en C3, sugieren una posible asociación entre frecuencia de interacción y riesgo de accidente. En la sección siguiente se examina esta hipótesis al comparar la composición de clústeres en ventanas previas a accidentes frente a condiciones habituales, evaluando si la presencia de ciertos perfiles de conducción tiende a incrementarse antes de un accidente.

Los perfiles anteriores permiten anticipar que podría existir una relación entre los estilos de conducción y el riesgo operativo: (i) C3 podría incrementar la probabilidad de maniobras a alta velocidad en la pista rápida; (ii) C1, por el contrario, exhibe baja conflictividad; (iii) C2 combina exposición y conflicto sistemáticamente mayores. Esto sugiere la hipótesis de que la proporción de ciertos conductores tendería a aumentar en ventanas previas a accidentes respecto de condiciones habituales. La siguiente sección evalúa esta hipótesis mediante ventanas balanceadas PRE y USUAL en pódicos aguas arriba/abajo, evaluando si la composición de clústeres varía significativamente antes de un siniestro.

Con el fin de complementar el análisis de estilos de conducción, se realizó una caracterización sociodemográfica de los conductores asociados a cada clúster identificado. Para ello, se integraron los registros de patentes con información proveniente de la base de datos de conductores y las encuestas Casen del año 2022, lo que permitió estimar variables como sexo, nivel de ingreso promedio por comuna y prioridad social (IPS 2022).

Al vincular la información de domicilio electoral con las bases de ingresos promedio comunales de la Encuesta CASEN (2022), se observaron diferencias en la distribución territorial y

socioeconómica de los conductores. Esta integración de fuentes permitió identificar patrones de residencia asociados a cada clúster, los cuales se presentan a continuación.

Tabla 9: Características sociodemográficas promedio por clúster.

Clúster	% Hom- bres	% Muje- res	Ingreso Promedio Comunal (CLP)	IPS medio	Nº Con- ductores	Año Pro- medio del Vehículo
0	61.8	38.2	1,388,250	61.69	66,471	2014
1	59.8	40.2	1,519,260	58.23	139,710	2013
2	61.5	38.5	1,368,910	62.16	22,889	2015
3	61.2	38.8	1,569,656	56.87	179,490	2014

La Tabla 9 muestra que la proporción masculina se mantiene relativamente estable entre clústeres, cercana al 60 %, lo que sugiere que el género no constituye un factor diferenciador relevante en los estilos de conducción. En contraste, las diferencias socioeconómicas y territoriales sí resultan significativas: los clústeres 1 y 3 agrupan conductores residentes en comunas con mayores ingresos promedio comunales, tales como Las Condes, Vitacura o Providencia, mientras que los clústeres 0 y 2 concentran conductores provenientes de comunas con ingresos medios o bajos, como San Bernardo, Maipú y Puente Alto.

Estas diferencias sugieren que el nivel socioeconómico y la localización geográfica podrían estar relacionados con los patrones de conducción observados. Conductores de comunas periféricas tienden a recorrer distancias mayores y a enfrentarse a mayores niveles de congestión, mientras que quienes residen en comunas de mayores ingresos suelen tener trayectos más cortos y mejor infraestructura vial. En este sentido, la caracterización socioeconómica y territorial de los clústeres entrega un contexto esencial para comprender las diferencias en el comportamiento vehicular y su posible vinculación con la ocurrencia de accidentes, aspecto que se aborda en la siguiente sección.

Tabla 10: Tipos de vehículo predominantes por clúster.

Clúster	Tipo 1	Tipo 2	Tipo 3	Año promedio del vehículo
0	Automóvil (39.1 %)	Station Wagon (24.6 %)	Camioneta (15.5 %)	2014
1	Automóvil (43.8 %)	Station Wagon (27.5 %)	Camioneta (13.1 %)	2013
2	Automóvil (36.0 %)	Station Wagon (21.3 %)	Camioneta (16.6 %)	2015
3	Automóvil (44.4 %)	Station Wagon (35.6 %)	Camioneta (13.5 %)	2014

A partir de la composición por tipo de vehículo, se observa que en todos los clústeres predomi-

na el uso de automóviles y station wagon, lo que refleja el carácter eminentemente particular del flujo vehicular analizado. Sin embargo, existen matices que permiten distinguir el perfil de cada grupo.

El Clúster 0 combina una proporción relevante de station wagon (24,6 %) y camionetas (15,5 %), lo que sugiere una mayor presencia de vehículos de uso mixto, posiblemente asociados a desplazamientos familiares y laborales. Este grupo presenta un año promedio del vehículo de 2014, lo que indica una antigüedad intermedia del parque vehicular.

El Clúster 1 concentra la mayor proporción de automóviles (43,8 %) y (27,5 %), representando un grupo claramente vinculado a conductores particulares, con un parque vehicular algo más antiguo (año promedio 2013).

En contraste, el Clúster 2 presenta una distribución más heterogénea, con una menor proporción de automóviles (36,0 %) pero un peso superior de camionetas (16,6 %) y furgones (10,0 %), lo que sugiere un uso más asociado a labores de transporte o trabajo. Este grupo posee el parque vehicular más reciente (año promedio 2015), lo que podría relacionarse con un recambio más frecuente de unidades operativas.

Finalmente, el Clúster 3 destaca por combinar la mayor proporción de automóviles (44,4 %) y station wagon (35,6 %), además de contar con un parque vehicular relativamente moderno (año promedio 2014). Este grupo podría corresponder a conductores con mayor nivel adquisitivo tal como fue mencionado anteriormente.

Tabla 11: Proporción promedio de horarios de conducción por clúster.

Clúster	%AM	%PM	%NOCHE	%MADRUGADA
0	0.374	0.436	0.121	0.069
1	0.362	0.486	0.103	0.049
2	0.414	0.419	0.103	0.064
3	0.303	0.469	0.142	0.086

Al analizar la distribución temporal de los desplazamientos, se observa que los conductores de los distintos clústeres presentan patrones diferenciados de circulación a lo largo del día.

El Clúster 0 mantiene una distribución relativamente equilibrada entre los horarios de mañana (37.4 %) y tarde (43.6 %), con menor presencia nocturna (12.1 %) y en la madrugada (6.9 %). Este comportamiento corresponde al perfil de *flujo estable* identificado previamente, caracterizado por desplazamientos regulares y sin una concentración marcada en franjas de alta congestión.

El Clúster 1, asociado a conducción conservadora y de baja velocidad, concentra la mayor proporción de viajes en horario PM (48.6 %), lo que sugiere desplazamientos principalmente vespertinos o de retorno laboral. Su menor participación en las franjas nocturnas y de madrugada refuerza su perfil de conducción prudente y planificada, propia de usuarios que circulan en contextos de tránsito controlado y estable.

El Clúster 2, caracterizado por alta exposición y mayor conflicto, presenta la proporción más altas en horario AM (41.3 %) y un (41.9 %) en horario PM, lo que evidencia una presencia sostenida durante las horas de mayor flujo vehicular. Esto sugiere que se trata de conductores altamente expuestos que circulan principalmente en los periodos de mayor congestión, probablemente por motivos laborales o de transporte regular. Su patrón de conducción en franjas de alto tráfico refuerza su condición de grupo intensivo en uso de la autopista y explica los niveles relativamente mayores de conflicto observados en este clúster.

El Clúster 3, vinculado a conducción rápida y uso preferente de carriles de alta velocidad, presenta una proporción destacada de viajes nocturnos (14.2 %) y en la madrugada (8.6 %), evidenciando una mayor presencia en horarios de baja congestión y condiciones de flujo más libre. Este patrón es coherente con un estilo de conducción más ágil y orientado al aprovechamiento de la vía rápida, en contextos donde la densidad vehicular es reducida.

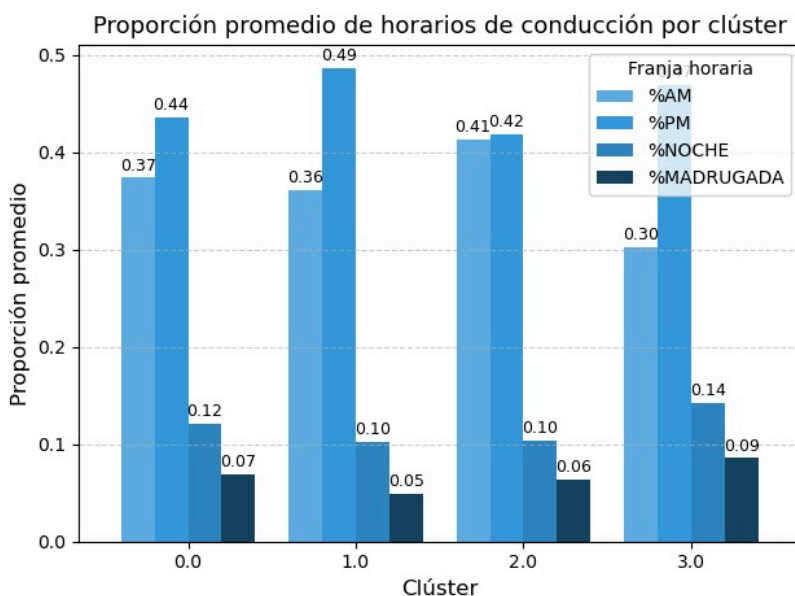


Figura 7: Proporción promedio de horarios de conducción por clúster.

La Figura 7 muestra la distribución promedio de los horarios de conducción para cada clúster. El gráfico permite visualizar de forma comparativa los momentos del día en que cada grupo concentra su actividad, evidenciando que los estilos de conducción no solo difieren en velocidad o uso de carriles, sino también en sus patrones temporales de desplazamiento. Esta dimensión temporal aporta una perspectiva complementaria a la caracterización de los conductores, mostrando que los hábitos de circulación —como preferir horas de alta o baja congestión— forman parte integral de los distintos perfiles identificados.

---

### 5.3. Relación entre la ocurrencia de accidentes y la composición de clústeres

El objetivo de este análisis es evaluar si la composición de clústeres en el flujo vehicular cambia de forma sistemática alrededor de un accidente, en comparación con condiciones habituales en los mismos pódicos y horarios. Para ello, se analizaron los accidentes reales ocurridos entre marzo y mayo, considerando los pódicos asociados tanto aguas arriba (AA1, AA2) como aguas abajo (AB1, AB2), utilizando ventanas temporales simétricas de  $\pm 5$  minutos en torno al instante del accidente.

Sea  $RATIO_{ijk}$  la proporción del clúster  $k$  observada en el pódico  $i$  bajo la condición  $j$ , donde:

- $i \in \{1, 2, 3, 4\}$  representa el pódico: AA2 (1), AA1 (2), AB1 (3) y AB2 (4).
- $j \in \{0, 1\}$  indica la posición temporal: 0 para los cinco minutos antes y 1 para los cinco minutos después del evento.
- $k \in \{0, 1\}$  distingue la condición del flujo: 0 para *no accidente* (USUAL) y 1 para *accidente*.

El contraste estadístico se realiza sobre las diferencias  $\Delta_k = RATIO_{ij1k} - RATIO_{ij0k}$ , es decir, la variación promedio entre la composición durante la ventana de accidente y su equivalente habitual. Con el fin de identificar posibles patrones anticipatorios o reactivos en la dinámica del tráfico, se consideraron tres escenarios analíticos:

- **AA, PRE:** pódicos aguas arriba en los cinco minutos previos al accidente, donde podrían manifestarse señales de conducción anómala antes del siniestro.
- **AB, PRE:** pódicos aguas abajo observados antes del accidente, utilizados como referencia o control, ya que no deberían mostrar alteraciones asociadas al evento.
- **AA, POST:** pódicos aguas arriba en los cinco minutos posteriores al accidente, utilizados también como punto de control, ya que no deberían reflejar alteraciones predictivas asociadas a los accidentes.

#### Composición de clústeres en pódicos aguas arriba (AA1 y AA2) durante la ventana previa al accidente ( $j = 0$ )

El análisis de los pódicos ubicados aguas arriba del accidente permite observar el comportamiento del flujo inmediatamente antes del evento. Como se aprecia en la Tabla 12, las diferencias promedio entre la condición de accidente y la habitual son muy pequeñas —en torno a décimas de punto porcentual—, lo que sugiere una alta estabilidad en la composición de clústeres.

En los pódicos que están inmediatamente antes del accidente (AA1), las proporciones se mantienen prácticamente constantes entre las ventanas de accidente y las habituales, sin evidencias de un cambio sistemático en la conducción previa. Ahora en los pódico que se enuncian antes de los pódicos de aguas arriba 1 (AA2), se observa una ligera disminución del clúster asociado a conducción rápida (C3) y un leve aumento en los grupos intermedios (C0 y C1). Esta diferencia alcanza un nivel de significancia marginal ( $p \approx 0,049$ ), pero su

---

magnitud es reducida ( $|\Delta| < 1\%$ ), lo que sugiere un efecto estadísticamente débil y sin relevancia práctica.

En primera instancia podría interpretarse que una menor proporción del clúster C3 antes del accidente refleja una reducción del número de conductores rápidos, lo cual podría asociarse a condiciones más densas o de riesgo potencial. Sin embargo, para comprobar si esta diferencia —aunque débil— corresponde realmente a un efecto particular de los accidentes o simplemente a la variabilidad habitual del tráfico, se decidió contrastar estos resultados con escenarios de control. Estos puntos de referencia corresponden a condiciones en las que no deberían observarse alteraciones marcadas, como los pódicos aguas abajo o las ventanas posteriores al evento. Si las diferencias se repiten también en dichos contextos, puede inferirse que las fluctuaciones observadas responden al comportamiento natural del flujo y no a un patrón causal asociado a los accidentes.

### **Análisis de escenarios de control**

Los escenarios de control permiten validar si los pequeños cambios observados aguas arriba responden a un patrón real o simplemente a fluctuaciones aleatorias del tráfico. En los pódicos aguas abajo, cinco minutos antes del accidente (AB PRE), las diferencias promedio también son inferiores al 1% y no presentan una dirección consistente entre clústeres. Incluso aquí, donde no debería existir efecto del accidente, el clúster C3 muestra una reducción significativa ( $p = 0,002$ ), lo que confirma que estas fluctuaciones pueden ocurrir por azar.

De forma similar, en la ventana de los pódicos aguas arriba, cinco minutos después del accidente (AA POST) se observan pequeñas variaciones en C1 y C3 que alcanzan significancia estadística  $p=0.002$  y  $p=0.011$ , respectivamente, pero con magnitudes igualmente pequeñas. Dado que el patrón se repite en condiciones donde el accidente no debiera tener impacto, se concluye que no existe evidencia sólida de una relación causal entre la composición de clústeres y la ocurrencia de accidentes., sino que más bien es un comportamiento esperado del tráfico

En conjunto, los análisis de AA1, AA2, AB y AA POST muestran un comportamiento coherente:

- Las variaciones en la composición de clústeres son pequeñas, generalmente menores al 1%.
- El único caso con significancia marginal, el clusters 3 en los pódicos aguas arriba 2 cinco minutos antes del accidente, carece de respaldo cuando se observan los escenarios de control.
- Los p-valores cercanos a 0.05 indican que las diferencias están en el límite de la significancia estadística, lo que reduce su confiabilidad.

Por lo tanto, no se observan evidencias de que la presencia de un tipo específico de conducción aumente o anticipe la probabilidad de accidente. Las diferencias identificadas parecen deberse a la variabilidad inherente del tráfico y no a un cambio estructural previo al siniestro.



Tabla 12: Resultados del test  $t$  pareado por clúster y escenario.

Escenario	Clúster	Mean_Acc	Mean_Usual	$\Delta$	$t$	$p$ -valor
AA1 PRE	C0	0.357	0.355	+0.0019	0.58	0.559
	C1	0.128	0.125	+0.0025	0.87	0.386
	C2	0.347	0.349	-0.0017	-0.47	0.638
	C3	0.169	0.171	-0.0027	-0.68	0.496
AA2 PRE	C0	0.360	0.355	+0.0046	1.27	0.204
	C1	0.124	0.121	+0.0024	1.01	0.315
	C2	0.349	0.349	-0.0001	-0.04	0.968
	C3	0.168	0.175	-0.0069	-1.98	<b>0.049</b>
AB PRE	C0	0.356	0.354	+0.0020	0.74	0.461
	C1	0.126	0.124	+0.0021	1.22	0.225
	C2	0.364	0.361	+0.0035	1.24	0.217
	C3	0.153	0.161	-0.0076	-3.16	<b>0.002</b>
AA POST	C0	0.355	0.357	-0.0019	-0.68	0.494
	C1	0.128	0.123	+0.0054	3.10	<b>0.002</b>
	C2	0.350	0.346	+0.0035	1.12	0.265
	C3	0.168	0.175	-0.0070	-2.56	<b>0.011</b>

A partir de los resultados mostrados en la Tabla 12, se observa que las diferencias en la composición de clústeres entre las ventanas de accidente y las habituales son pequeñas y no sistemáticamente significativas. Aunque algunos contrastes puntuales presentan valores de  $p$  cercanos o inferiores a 0.05 —como el clúster C3 en AA2 PRE o C1 y C3 en AA POST—, las magnitudes de cambio son reducidas (menores al 1 %), lo que limita su relevancia práctica.

En particular, la leve disminución del clúster C3, asociado a conductores de tipo rápido, en AA2 PRE podría sugerir inicialmente un cambio en la composición del flujo antes de los accidentes. Sin embargo, al comparar con los escenarios de control (AB PRE y AA POST), se advierte que reducciones de magnitud similar también aparecen en contextos donde el accidente no influye directamente. Esto indica que las variaciones detectadas son parte de la variabilidad natural del tráfico y no evidencian un patrón sistemático previo al siniestro.

Por lo tanto, la composición del flujo vehicular previa a los accidentes no muestra alteraciones significativas respecto a condiciones habituales. Esto sugiere que la ocurrencia de un siniestro no está asociada a un cambio estructural en los estilos de conducción presentes en el tramo. En otras palabras, la presencia de un tipo particular de conductor —como los pertenecientes al clúster asociado a conducción rápida o riesgosa— no modifica de forma consistente la distribución del tráfico antes de un evento.

Los resultados observados respaldan la hipótesis de estabilidad del flujo en torno a los accidentes, indicando que las pequeñas variaciones detectadas corresponden principalmente a fluctuaciones naturales y no a un patrón causal. De este modo, se descarta que la incorporación de los perfiles de conducción aporte una mejora significativa en la capacidad predictiva

---

de los modelos de accidentes. Por consiguiente, se rechaza la hipótesis H3, que postulaba que la inclusión de estos perfiles mejoraría de manera significativa el rendimiento de los modelos de predicción de accidentes considerados en este estudio.

## 6. Conclusiones

El presente estudio tuvo como propósito caracterizar los estilos de conducción presentes en la Autopista Central y evaluar si ciertos perfiles de conductores —identificados mediante clústeres de comportamiento— se asocian a una mayor probabilidad de ocurrencia de accidentes. A partir de datos reales del sistema AVI (marzo–mayo de 2024), complementados con registros del Ministerio de Obras Públicas y antecedentes sociodemográficos, se construyeron indicadores de velocidad, headway, tiempo hasta la colisión, nivel de conflicto y uso de carriles. Mediante la aplicación del algoritmo K-Means se identificaron cuatro grupos de conducción claramente diferenciados: un clúster de flujo estable (C0), un perfil conservador y de baja velocidad (C1), un grupo de alta exposición y conflicto (C2), y un clúster rápido en carriles de alta velocidad (C3). De esta manera, se cumplieron los dos primeros objetivos específicos del estudio: desarrollar métricas de riesgo y generar perfiles de conducción a partir de conductores con características similares.

Además, la caracterización permitió identificar diferencias significativas entre los grupos, tanto en su comportamiento dinámico como en sus atributos sociodemográficos y temporales. Por ejemplo, los conductores de los clústeres 1 y 3 se concentran en comunas de mayores ingresos promedio y presentan vehículos más antiguos, mientras que los del clúster 2 —asociado a alta exposición y mayor nivel de conflicto— corresponden a usuarios frecuentes que circulan principalmente en horas punta, lo que los expone a contextos de tráfico más denso. Estos hallazgos muestran que los estilos de conducción no solo difieren en su desempeño vehicular, sino también en sus condiciones sociales, territoriales y horarias, configurando un panorama integral del comportamiento en la autopista.

Posteriormente, se puso a prueba la hipótesis que proponía que la incorporación de perfiles de conducción mejoraría significativamente el rendimiento de los modelos de predicción de accidentes. El análisis de ventanas temporales antes y después de los siniestros, junto con escenarios de control, demostró que la composición del flujo vehicular permanece prácticamente constante entre condiciones normales y de accidente, sin variaciones sistemáticas ni estadísticamente significativas. Por lo tanto, se concluye que los perfiles de conducción no son, por sí solos, un predictor suficiente de riesgo vial, lo que sugiere que la ocurrencia de accidentes responde principalmente a factores contextuales o situacionales del entorno vial.

En contraste, los resultados confirman la existencia de estilos de conducción claramente diferenciados y consistentes en el tiempo, lo que valida la metodología aplicada y resalta el potencial de este tipo de análisis para el diseño de políticas de seguridad vial y gestión del tráfico. En particular, la identificación de conductores altamente expuestos o propensos al conflicto puede servir como base para estrategias focalizadas de educación, control o gestión preventiva.

Por otra parte, este trabajo presenta algunas limitaciones propias de los datos y del diseño

---

metodológico. La asociación entre matrícula y conductor se establece bajo el supuesto de que el propietario del vehículo es quien habitualmente lo conduce, lo que introduce cierta incertidumbre en la caracterización sociodemográfica. Asimismo, los registros AVI no contemplan información contextual relevante —como condiciones meteorológicas, densidad del tráfico o incidentes menores— que podrían influir en las métricas instantáneas de conducción. Finalmente, el uso de ventanas temporales de cinco minutos otorga estabilidad a las mediciones, pero puede suavizar eventos de corta duración o comportamientos transitorios inmediatamente previos a un siniestro.

Los hallazgos de este estudio sugieren que, en las condiciones actuales de observación, la composición de estilos de conducción no constituye un indicador suficiente para anticipar la ocurrencia de accidentes en tiempo real. Esto implica que los esfuerzos de gestión vial y de seguridad deberían concentrarse en fortalecer la calidad y resolución de los datos operacionales, integrando información de tráfico, condiciones meteorológicas y reportes automáticos de incidentes, antes de implementar modelos predictivos basados en comportamiento. A su vez, las políticas de prevención podrían orientarse hacia la gestión de puntos críticos de alta densidad vehicular o en horarios donde se concentren mayores niveles de conflicto, en lugar de basarse exclusivamente en la tipología de conductores.

A partir de los resultados obtenidos, se abren diversas líneas de investigación a futuro. Se propone incorporar nuevas variables contextuales —como el clima, la densidad del tránsito o la geometría de la vía— con el fin de mejorar la capacidad explicativa de los modelos de riesgo. Asimismo, realizar un análisis longitudinal de los conductores involucrados en accidentes permitiría determinar si presentan una mayor probabilidad de reincidencia o si su comportamiento previo puede anticipar futuros siniestros. Por último, el uso de metodologías avanzadas, como modelos de aprendizaje profundo o redes bayesianas dinámicas, podría capturar relaciones no lineales entre las variables de conducción y las condiciones del entorno, avanzando hacia sistemas predictivos más realistas y adaptativos.

---

## Referencias

- Barnard, Y., Utesch, F., van Nes, N., Eenink, R., and Baumann, M. (2016). The study design of udrive: the naturalistic driving study across europe for cars, trucks and scooters. *European Transport Research Review*, 8:1–10.
- Conaset (2023). Tres de cada cinco chilenos observa mucha agresividad y violencia en conductores y peatones. Accedido el 9 de diciembre de 2024.
- de Winter, J. C. and Dodou, D. (2010). The driver behaviour questionnaire as a predictor of accidents: A meta-analysis. *Journal of safety research*, 41(6):463–470.
- Di Milia, L., Smolensky, M. H., Costa, G., Howarth, H. D., Ohayon, M. M., and Philip, P. (2011). Demographic factors, fatigue, and driving accidents: An examination of the published literature. *Accident Analysis & Prevention*, 43(2):516–532.
- Ding, Z., Xu, D., Tu, C., Zhao, H., Moze, M., Aioun, F., and Guillemard, F. (2022). Driver identification through heterogeneity modeling in car-following sequences. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):17143–17156.
- Hussain, Z., Mohammed, S. S., Dias, C., Hussain, Q., and Alhajyaseen, W. K. (2025). Empirical analysis of car-following behavior: Impacts of driver demographics, leading vehicle types, and speed limits on driver behavior and safety. *Transportation Research Part F: Traffic Psychology and Behaviour*, 108:188–205.
- Karacasu, M., Ergül, B., and Altin Yavuz, A. (2014). Estimating the causes of traffic accidents using logistic regression and discriminant analysis. *International journal of injury control and safety promotion*, 21(4):305–313.
- Kim, H. S., Kim, H. J., and Son, B. (2006). Factors associated with automobile accidents and survival. *Accident Analysis & Prevention*, 38(5):981–987.
- Lizama Cisternas, L. and Moreno Alveal, J. (2011). Estudio exploratorio de la industria del transporte y accidentes de tránsito en chile. Technical report, Universidad de Chile. Disponible en repositorio de la Universidad de Chile.
- Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., and Itakura, F. (2007). Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, 95(2):427–437.
- Monselise, M., Liang, O. S., and Yang, C. C. (2019). Identifying important risk factors associated with vehicle injuries using driving behavior data and predictive analytics. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–10. IEEE.
- Petridou, E. and Moustaki, M. (2000). Human factors in the causation of road traffic crashes. *European journal of epidemiology*, 16:819–826.
- Sivaraman, S. and Trivedi, M. M. (2013). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 14(4):1773–1795.

---

Weng, J. and Meng, Q. (2012). Effects of environment, vehicle and driver characteristics on risky driving behavior at work zones. *Safety science*, 50(4):1034–1042.