Anonimización y Análisis de Datos Vehiculares

Un Estudio sobre Permisos de Circulación en La Serena (2011-2024)

Autor: Cristian Patricio Méndez Fuenzalida
Programa: Magíster en Estadística
Profesora guía: María José González Clares
Ayudante: Daniel Alejandro Saavedra Morales

Fecha: 21-11-2025

1. Resumen

El presente proyecto de tesis aborda el desafío de equilibrar la privacidad de los datos personales con su utilidad estadística en el ámbito municipal. Utilizando un conjunto de datos longitudinal de 1.35 millones de permisos de circulación de la comuna de La Serena (2011-2024), se evalúa la implementación de técnicas de protección de datos. La metodología incluye la aplicación de *hashing* criptográfico (HMAC) para la seudonimización de identificadores directos y algoritmos de control de revelación estadística (**k-anonimidad** y **l-diversidad**) para proteger cuasi-identificadores.

Los resultados demuestran que es posible eliminar el riesgo de reidentificación por singularidad (alcanzando 0 % de violaciones de k-anonimidad con $k = 5$), aunque persisten desafíos de homogeneidad en atributos sensibles (l-diversidad mínima de 1.0). Se concluye que la combinación de estas técnicas permite publicar datos abiertos seguros, manteniendo la capacidad de realizar análisis de tendencias históricas del parque automotriz sin comprometer la privacidad de los contribuyentes.

Palabras clave: Anonimización, Seudonimización, Privacidad de Datos, Gestión Municipal, R.

2. Objetivos

2.1. Objetivo General

Evaluar el impacto de diversas técnicas de anonimización y seudonimización en la calidad analítica y el riesgo de reidentificación de un conjunto de datos longitudinales de permisos de circulación vehicular de la comuna de La Serena (2011-2024), con el fin de proponer un marco metodológico que equilibre la protección de la privacidad individual con la utilidad de los datos para análisis estadístico en contextos municipales.

2.2. Objetivos Específicos

1. **Preparar y Caracterizar el Conjunto de Datos:**
   - Consolidar y limpiar la base de datos longitudinal de permisos de circulación vehicular de La Serena (2011-2024).
   - Realizar un análisis descriptivo exhaustivo del conjunto de datos para identificar variables clave, cuasi-identificadores potenciales y características temporales relevantes.
2. **Seleccionar e Implementar Técnicas de Anonimización/Seudonimización:**
   - Investigar y seleccionar un conjunto de técnicas de anonimización (ej. k-anonimidad, l-diversidad, t-cercanía, generalización, supresión) y seudonimización (ej. hashing, encriptación con clave) aplicables a los datos de permisos de circulación, con especial énfasis en las placas patente.
   - Implementar las técnicas seleccionadas sobre el conjunto de datos preparado.
3. **Definir y Aplicar Métricas de Privacidad y Utilidad:**
   - Establecer métricas cuantitativas para evaluar el nivel de protección de la privacidad ofrecido por cada técnica implementada (ej. riesgo de reidentificación, k-anonimidad alcanzado).
   - Definir métricas para cuantificar la utilidad de los datos anonimizados/seudonimizados para análisis estadísticos específicos (ej. precisión de estimaciones de tendencias, capacidad para realizar consultas relevantes para la gestión municipal).
4. **Evaluar el Equilibrio Privacidad-Utilidad:**
   - Analizar comparativamente el rendimiento de las diferentes técnicas en función del equilibrio logrado entre el nivel de privacidad y la preservación de la utilidad de los datos.
   - Identificar cómo diferentes parámetros dentro de cada técnica afectan este equilibrio (*trade-off*).
5. **Analizar el Impacto Específico en Análisis Estadísticos Clave:**
   - Determinar cómo las técnicas de anonimización/seudonimización más prometedoras impactan la capacidad de realizar análisis longitudinales relevantes sobre el parque automotriz (ej. evolución de tipos de vehículos, antigüedad promedio, tendencias de marcas).

3. Marco Teórico

En la última década, la disponibilidad de registros administrativos en formato digital ha generado oportunidades significativas para el desarrollo de estudios estadísticos aplicados en el ámbito público. En Chile, este fenómeno ha cobrado fuerza especialmente a nivel municipal, donde bases de datos como los permisos de circulación vehicular constituyen fuentes valiosas de información sobre el parque automotriz, las tendencias de motorización y ciertos patrones de comportamiento territorial.

No obstante, estos datos también plantean dilemas importantes en torno a la privacidad. Aunque no contienen nombres ni identificadores explícitos, variables como la placa patente, el tipo de vehículo, el color, el año de fabricación y la forma de pago, entre otras, pueden funcionar como cuasi-identificadores. Esto implica que, si no se aplican técnicas adecuadas de anonimización, es posible reconstruir trayectorias individuales o deducir información sensible, vulnerando derechos fundamentales. En este contexto se sitúa el presente estudio, que tiene como eje la aplicación y evaluación de distintas técnicas de anonimización y seudonimización en los datos de permisos de circulación de la comuna de La Serena, en la Región de Coquimbo, entre los años 2011 y 2024.

Diversos enfoques han sido desarrollados para enfrentar estos desafíos. La **k-anonimidad** (*k-anonymity*) (Samarati, 2001), por ejemplo, propone que cada registro sea indistinguible de al menos otros k-1 respecto de ciertos atributos clave. Sin embargo, se ha demostrado que este método puede ser insuficiente ante ataques más complejos, lo que ha llevado a la formulación de propuestas más exigentes como la **l-diversidad** (*l-diversity*) (Machanavajjhala et al., 2007) y la **t-cercanía** (*t-closeness*) (Li et al., 2007), que incorporan diversidad y proximidad en la distribución de atributos sensibles.

Un enfoque alternativo y más reciente es el de la *privacidad diferencial* (Dwork, 2006), que ofrece una garantía matemática robusta contra intentos de reidentificación, incluso en presencia de información externa. Este modelo ha sido especialmente valorado en contextos donde se requiere realizar análisis agregados sin comprometer datos individuales, como ocurre en estudios longitudinales o de planificación urbana.

Ahora bien, toda técnica de anonimización implica necesariamente un grado de pérdida de información. Por ello, una parte esencial de esta investigación consiste en evaluar el **equilibrio** entre protección de la privacidad y preservación de la utilidad de los datos. Diversas métricas han sido propuestas para este propósito, como el riesgo de reidentificación, la precisión en consultas específicas o la reproducibilidad estadística (Abu-Sinna et al., 2024; Pasquini et al., 2025). En el caso particular de La Serena, donde existe un interés municipal por analizar el comportamiento histórico del parque vehicular, la utilidad de los datos se traduce en la capacidad de estimar tendencias por tipo de vehículo, marca, motorización y antigüedad, entre otros indicadores relevantes.

Por último, este trabajo se sustenta en marcos éticos y normativos que, si bien muchas veces se desarrollan en contextos internacionales como el Reglamento General de Protección de Datos (GDPR) europeo, encuentran eco en las obligaciones de confidencialidad que rigen a los organismos públicos en Chile. Durante muchos años, la Ley N°19.628 sobre Protección de la Vida Privada sirvió como marco legal en esta materia, estableciendo principios básicos de tratamiento y seguridad de los datos personales. Sin embargo, con la reciente promulgación de la Ley N°21.719 en 2024, el país da un paso decisivo hacia un estándar más moderno, incorporando principios de responsabilidad proactiva, consentimiento explícito y la creación de una Agencia de Protección de Datos Personales. Estas normativas no solo refuerzan la necesidad de aplicar medidas de anonimización robustas, sino que también establecen un marco de referencia legal y ético que este estudio considera fundamental.

4. Metodología

Este estudio adopta una aproximación cuantitativa aplicada, orientada al análisis y transformación de datos administrativos longitudinales, con el propósito de evaluar el impacto de diversas técnicas de anonimización sobre la utilidad estadística y la protección de la privacidad en contextos municipales. El eje de trabajo será la base de datos histórica de permisos de circulación vehicular de la comuna de La Serena, correspondiente al período 2011–2024.

4.1. Fundamentos Técnicos de las Técnicas

Para cumplir con los objetivos del estudio, se implementará un conjunto de técnicas de seudonimización y anonimización. A continuación, se detalla el fundamento técnico de cada una de las principales técnicas a emplear.

Técnicas de Seudonimización

La seudonimización es el proceso de reemplazar un identificador directo (como la placa patente) por un seudónimo, con el fin de desvincular los datos de la identidad del individuo.

Hashing Criptográfico (SHA-256)

El hashing es una función matemática que toma una entrada de cualquier tamaño (la placa) y la convierte en una cadena de caracteres alfanuméricos de longitud fija (el "hash").

1. **Propiedades Clave:**
   a. **Determinista:** La misma placa siempre producirá el mismo hash, permitiendo el análisis longitudinal (seguir un vehículo a través de los años).
   b. **Irreversible (One-Way):** Es computacionalmente inviable obtener la placa original a partir del hash.
   c. **Resistente a Colisiones:** Es extremadamente improbable que dos placas diferentes generen el mismo hash.
2. **Aplicación:** Se utilizará para crear el `placa_hash` como un identificador anónimo persistente.

Hashing con Clave (HMAC)

HMAC (Hash-based Message Authentication Code) es una forma de hashing que incorpora una "clave secreta" en el cálculo.

1. **Propiedades Clave:**
   a. Mantiene todas las propiedades del hashing estándar, pero añade una capa de seguridad.
   b. El mismo hash solo puede ser reproducido por alguien que posea la clave secreta.
2. **Aplicación:** Se utilizará para crear el `placa_hmac` como un seudónimo alternativo, demostrando una técnica más robusta contra ataques de diccionario o tablas precalculadas.

Técnicas de Anonimización (Control de Revelación Estadística)

La anonimización se enfoca en proteger la identidad contra la reidentificación indirecta a través de la combinación de **cuasi-identificadores (QIs)**. En este estudio, los QIs son variables como `ano_fabricacion`, `marca_limpia`, `vehiculo_limpio`, etc.

k-Anonimidad (k-Anonymity)

Es una propiedad que garantiza que cualquier registro en el conjunto de datos no pueda ser distinguido de al menos `k-1` otros registros basándose en sus cuasi-identificadores.

1. **Objetivo:** Lograr que cada combinación de QIs aparezca al menos `k` veces en la base de datos.
2. **Implementación Técnica:** Para alcanzar `k=5`, el algoritmo `kAnon` (de la librería `sdcMicro`) modifica los datos originales usando dos métodos principales:

   a. **Generalización:** Reemplaza valores específicos por valores más generales (ej. `ano_fabricacion = 2019` se convierte en `ano_fabricacion = "2015-2020"`).
   b. **Supresión:** Elimina los valores que son demasiado únicos y no pueden ser generalizados (los convierte en `NA`).
3. **Debilidad:** Es vulnerable al **ataque de homogeneidad**, que ocurre si todos los `k` registros de un grupo comparten el mismo valor en un atributo sensible (ej. todos tienen el mismo `tipo_pago`).

l-Diversidad (l-Diversity)

Es una extensión de la k-anonimidad diseñada para solucionar el ataque de homogeneidad.

1. **Objetivo:** Asegura que dentro de cada grupo k-anónimo, exista un mínimo de `l` valores "bien representados" para el atributo sensible.
2. **Implementación Técnica:** Se define un atributo sensible (en nuestro caso, `tipo_pago`) y se establece un objetivo de `l=2`. El algoritmo `ldiversity` evalúa si cada grupo k-anónimo tiene al menos 2 tipos de pago diferentes.

4.2. Diseño del Estudio

La investigación se estructura en cinco fases principales:

1. **Preparación y limpieza de datos:** Integración de archivos CSV anuales (con delimitador `;`) provenientes del registro de permisos de circulación. Se contempla una etapa de estandarización de columnas. Se identificará y consolidará una estructura común para facilitar la comparabilidad interanual.
2. **Caracterización descriptiva del conjunto de datos:** Se realizará un análisis exploratorio de variables tales como tipo de vehículo, año de fabricación, cilindrada (CC), tipo de combustible, forma de pago y valor del permiso. Asimismo, se identificarán variables potencialmente sensibles o cuasi-identificadoras, como la **placa patente**, que será el foco principal de las estrategias de anonimización.
3. **Implementación de técnicas de anonimización y seudonimización:** En función de los fundamentos técnicos descritos, se aplicarán los métodos de *hashing* criptográfico y *HMAC* a la placa patente. Posteriormente, se aplicarán **k-anonimidad** y se evaluará la **l-diversidad** sobre el conjunto de cuasi-identificadores.
4. **Evaluación del impacto sobre privacidad y utilidad:** Se definirán indicadores cuantitativos como el riesgo estimado de reidentificación, el nivel de indistinguibilidad (k alcanzado), y métricas de distorsión o pérdida de información relevante. Paralelamente, se medirán indicadores de utilidad estadística mediante comparaciones entre análisis realizados con la base original y con versiones anonimizadas: distribución de antigüedad vehicular, frecuencia por marca o tipo, evolución del valor de permisos, entre otros.
5. **Análisis del equilibrio privacidad–utilidad:** Finalmente, se compararán las distintas técnicas aplicadas en función del balance alcanzado. Se evaluará su aplicabilidad en entornos municipales, donde la transparencia y la protección de datos deben coexistir, y se propondrá un marco metodológico replicable para otras comunas u organismos públicos.

4.3. Herramientas y Software

Todo el desarrollo técnico y estadístico del estudio se realizó en **lenguaje R**, utilizando el entorno **Quarto** para la documentación reproducible y la generación de reportes. Se emplearon librerías especializadas como `tidyverse` para limpieza y análisis exploratorio; `digest` para funciones de hashing; y `sdcMicro` para la anonimización estadística basada en *k-anonimidad*, *l-diversidad* y supresión.

Para garantizar la transparencia y reproducibilidad de la investigación, el código fuente completo, los scripts de procesamiento y este informe dinámico se encuentran disponibles públicamente en el siguiente repositorio de GitHub:

https://github.com/crimen-cl/Proyecto-Magister

5. Resultados

5.1. Análisis Exploratorio de Datos

En este capítulo se presentan los resultados obtenidos a partir del procesamiento y análisis del conjunto de datos de permisos de circulación de la comuna de La Serena para el período 2011-2024. La primera sección detalla la preparación y limpieza de los datos, mientras que la segunda sección se enfoca en el análisis exploratorio de datos (EDA), sentando las bases para la posterior evaluación de técnicas de anonimización.

Preparación y Limpieza del Conjunto de Datos

El conjunto de datos inicial consistió en 14 archivos anuales en formato CSV, correspondientes al registro de permisos de circulación entre los años 2011 y 2024. El primer paso consistió en la consolidación de estos archivos en una única base de datos longitudinal (N = 1,356,351 registros).

Durante este proceso, se realizó una exhaustiva limpieza y estandarización. Se corrigieron los tipos de datos de columnas numéricas clave como `ano_fabricacion`, `tasacion` y `valor_permiso`. Adicionalmente, se llevó a cabo una homologación de los nombres de todas las variables a un formato consistente (`snake_case`).

Un paso crítico fue la inspección y limpieza de las variables categóricas, donde se detectaron diversas inconsistencias. Se aplicaron las siguientes correcciones:

1. **Vehículo:** Se unificaron categorías sinónimas (ej. “MOTO”, “MOTOCICLETA”) bajo la etiqueta única “MOTOCICLETA”.
2. **Marca:** Se estandarizaron nombres para corregir inconsistencias (ej. “KIA MOTORS” a “KIA”).
3. **Combustible y Transmisión:** Se agruparon múltiples abreviaciones y variaciones (ej. “Benc”, “BENC” a “BENCINA”; “Dies”, “Diés” a “DIESEL”).
4. **Color:** Dada la alta cardinalidad, se decidió mantener esta variable en su estado original para esta fase.

El resultado de esta etapa es un `dataframe` limpio, consolidado y estandarizado, denominado `datos_limpios`, que constituye la base para todos los análisis subsecuentes.

Análisis Exploratorio de Datos (EDA)

Evolución General del Parque Vehicular

Para comprender la magnitud y el crecimiento del parque automotriz en La Serena, se cuantificó el número de vehículos únicos por año. El cálculo se basó en el conteo de placas patentes distintas para cada período, para evitar el conteo duplicado por pago en cuotas.

La Figura 1 muestra la evolución de la cantidad de vehículos únicos desde 2011 hasta 2024.

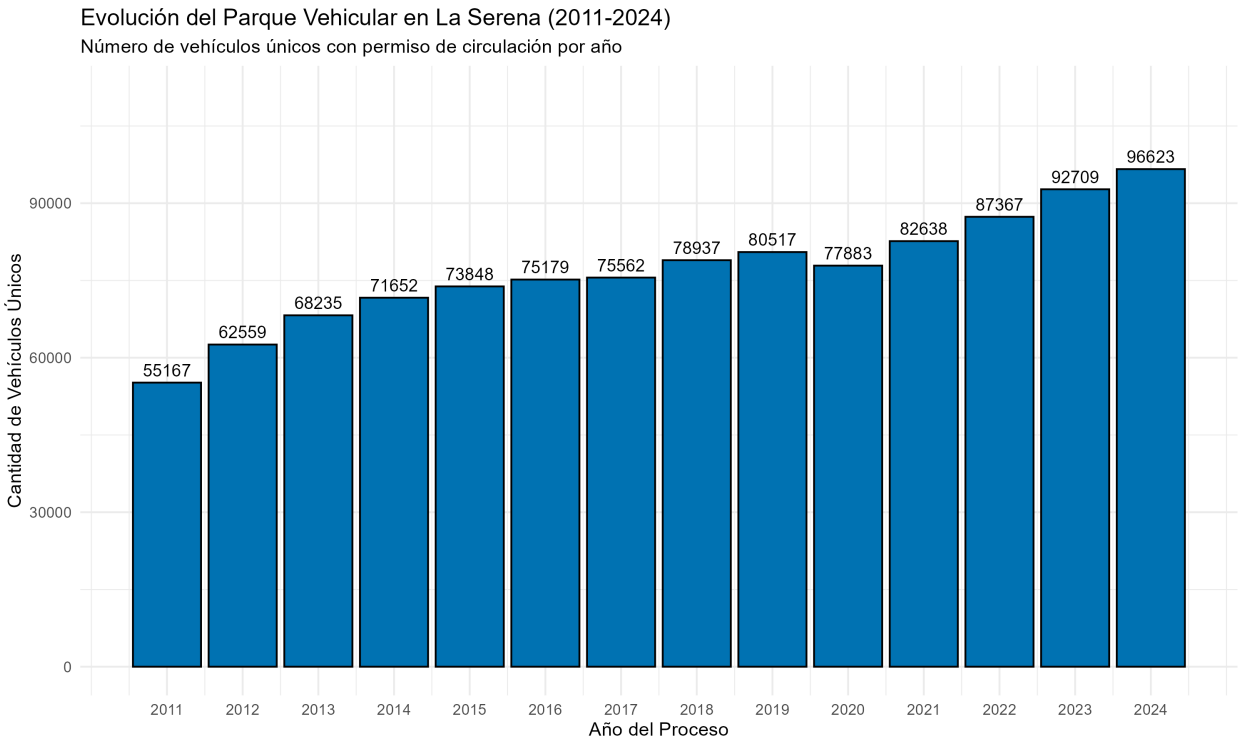

Figura 1: Evolución del Parque Vehicular en La Serena (2011-2024).

Como se puede apreciar en la Figura 1, existe una tendencia general de crecimiento sostenido a lo largo del período, pasando de **55,167** vehículos únicos en 2011 a **96,623** en 2024. Esto representa un aumento del **75.2 %** en 14 años. Se destaca una notoria caída en el año 2020 (77,883 registros), interrumpiendo la tendencia alcista vista hasta 2019 (80,517 registros), un fenómeno que puede atribuirse al impacto socioeconómico de la pandemia de COVID-19. A partir del año 2021, la tendencia de crecimiento se retoma con fuerza.

Composición del Parque Vehicular en 2024

Para obtener una caracterización detallada del estado actual del parque vehicular, se analizó la composición de los datos del año más reciente, 2024. La Figura 2 sintetiza esta composición en tres paneles.

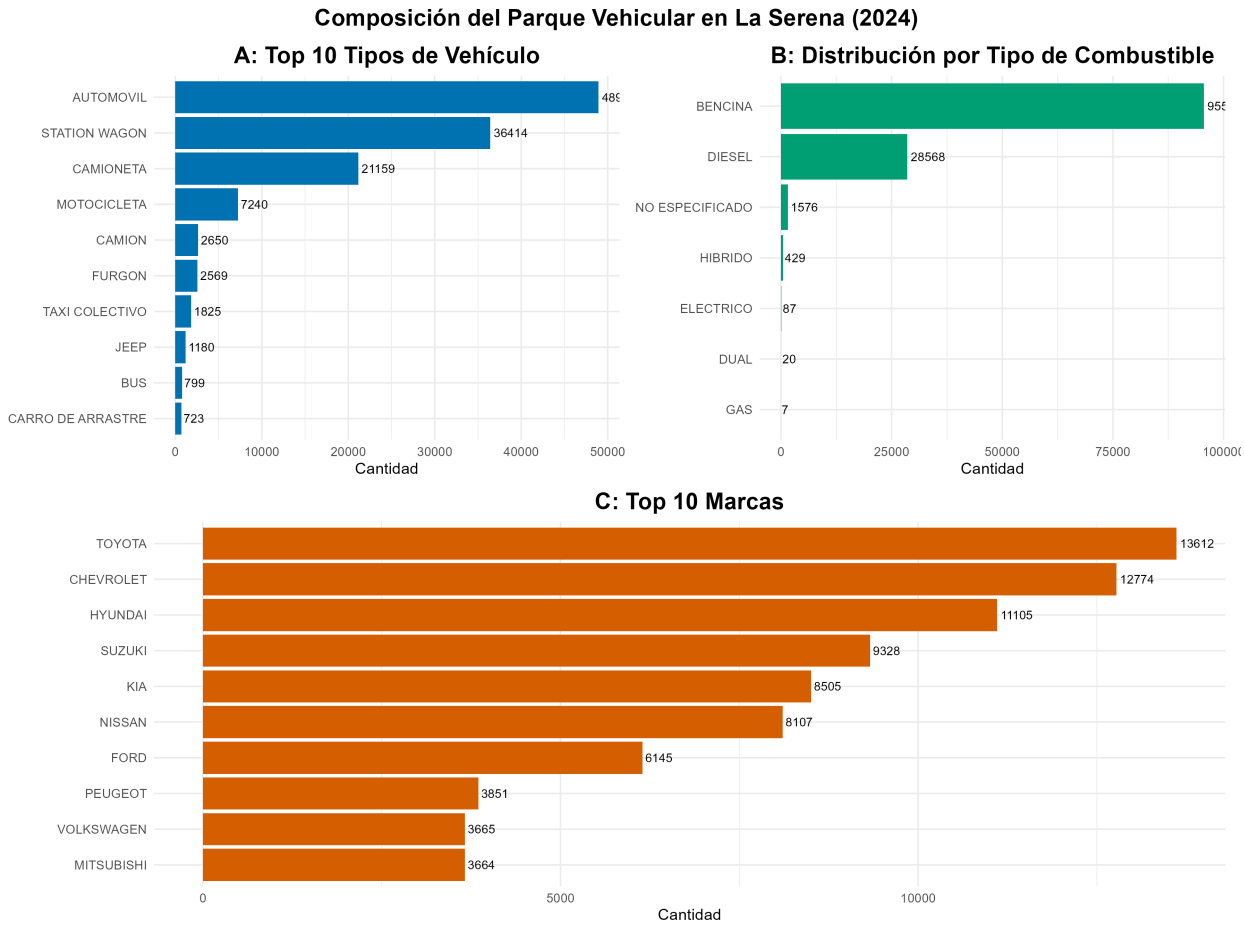

Figura 2: Composición del Parque Vehicular en La Serena (2024).

El análisis de la composición revela varios puntos clave:

- **Panel A (Tipo de Vehículo):** Las categorías "AUTOMOVIL", "CAMIONETA" y "STATION WAGON" representan la mayoría del parque, reflejando las preferencias de los conductores en la comuna.
- **Panel B (Combustible):** Existe un dominio absoluto de la bencina, que representa la gran mayoría de los vehículos. El diésel se posiciona como la segunda opción más relevante, mientras que las tecnologías híbridas y eléctricas mantienen una participación minoritaria.
- **Panel C (Marcas):** Se observa una alta concentración en un número reducido de fabricantes. Chevrolet, Hyundai y Toyota lideran el mercado, seguidas de cerca por otras marcas de origen asiático.

Tendencias en la Composición de Combustibles (2011-2024)

Finalmente, para añadir una perspectiva temporal al análisis de composición, se examinó la evolución de la proporción de cada tipo de combustible a lo largo de todo el período de estudio. La Figura 3 muestra estas tendencias.

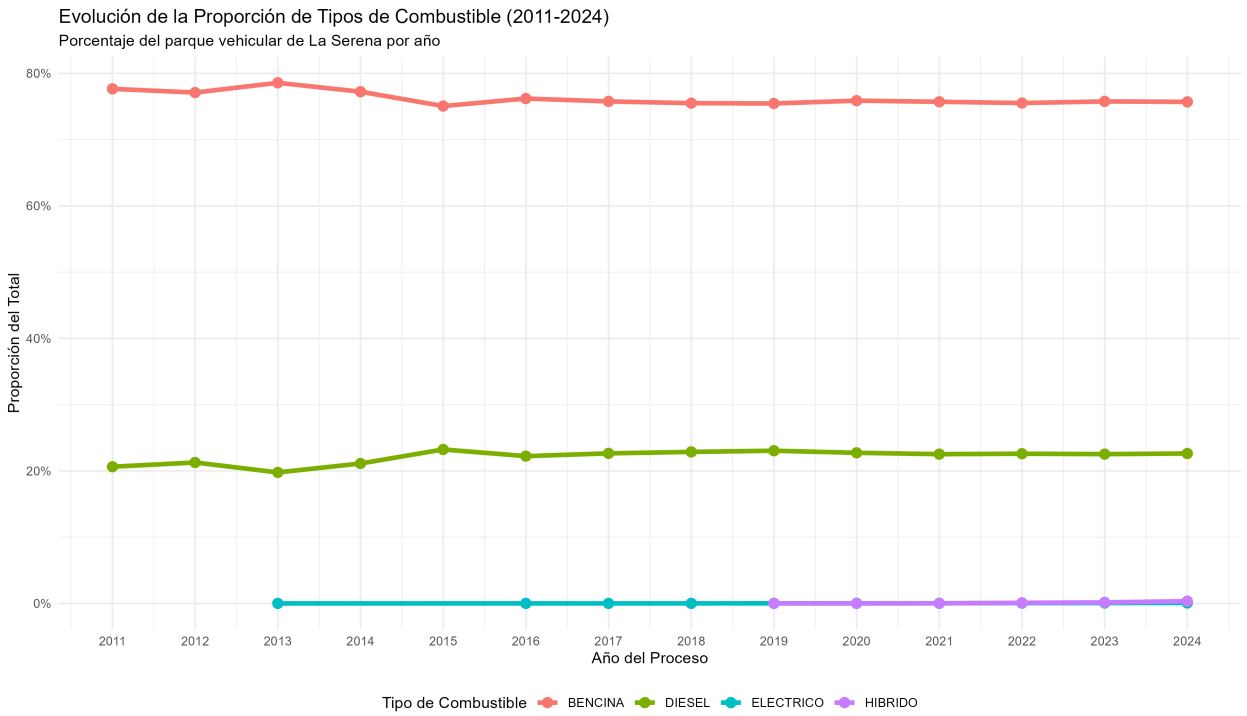

Figura 3: Evolución de la Proporción de Tipos de Combustible (2011-2024).

Este análisis longitudinal confirma varias tendencias importantes. Si bien la bencina sigue siendo mayoritaria, su participación ha disminuido ligeramente en el tiempo. Por el contrario, el diésel ha experimentado un crecimiento sostenido y significativo, ganando cuota de mercado año tras año. Por su parte, los vehículos híbridos y eléctricos, aunque marginales, muestran una clara aparición y crecimiento en los últimos años del período, señalando el inicio de una posible diversificación en la matriz energética del transporte local.

5.2. Desarrollo de Objetivos Específicos: Aplicación de Técnicas de Anonimización

Una vez caracterizado el conjunto de datos, el siguiente paso es abordar el Objetivo Específico 2: "Seleccionar e Implementar Técnicas de Anonimización/Seudonimización". En esta sección se presenta una implementación de cuatro técnicas distintas: dos de seudonimización para proteger los identificadores directos y dos de anonimización para mitigar el riesgo de reidentificación a través de los cuasi-identificadores.

Técnicas de Seudonimización de Identificadores Directos

La variable `placa` es el identificador directo más sensible en el conjunto de datos. Para protegerla, se implementaron dos técnicas de seudonimización basadas en hashing. El objetivo es reemplazar la placa original por un seudónimo consistente (la misma placa siempre genera el mismo seudónimo), lo que permite mantener la capacidad de realizar análisis longitudinales sin exponer el dato real.

A continuación, se describe cada técnica y sus resultados.

Técnica 1: Hashing Criptográfico (SHA-256)

El hashing criptográfico es un método de un solo sentido, considerado computacionalmente irreversible. Es una técnica estándar para crear seudónimos cuando no existe la necesidad de revertir el proceso para obtener el valor original. La seguridad del método se basa en la robustez del algoritmo criptográfico utilizado. La **Tabla 1** muestra el resultado de aplicar SHA-256 a un conjunto de placas de ejemplo.

Tabla 1: Resultado de la Seudonimización con Hashing.

Placa Original	Seudónimo Hash (SHA-256)
AA-1600	`d74bf534effd86d31ee72063394edc6b5c2e293dfa34571df5db86f631a41242`
AA-1890	`b0a1b8327293f123839167b1d782a3397440977274cb9af45518807ff779cd3d`
AA-2647	`b708fa3fb87cfd6bdddb7d575787332b879ae94bece86deb2a5c44d7486d4fee`
AA-2729	`85fcc352bd9f8be472f21107544f258c6c8b821d8f74f3a5640b40284d3f9009`

Técnica 2: Hashing con Clave (HMAC-SHA256)

El Hashing con Clave (HMAC) es una variante que incorpora una "clave secreta" en el proceso de hashing. El seudónimo solo puede ser generado si se conoce tanto el dato original como la clave, lo que añade una capa extra de seguridad contra ataques de diccionario o tablas precalculadas (*rainbow tables*). Al igual que el hashing estándar, es un proceso consistente e irreversible. La **Tabla 2** muestra los resultados de esta técnica.

Tabla 2: Resultado de la Seudonimización con HMAC.

Placa Original	Seudónimo HMAC (con clave)
AA-1600	`c771491033f56e0cedd248fcec552ebcd4f3f33e91818bdf29b925dc56a1da7b`
AA-1890	`fe1be29760f0d978d1310b5d3ec64bf578146c37be0570c13bb392fa8e594b05`
AA-2647	`820388221bcb9b02b8dc44aa0329dc55ef41bd64850a2f9b525be62eaccb968a`
AA-2729	`cc6e9ca54a4ab5a04423f3547877ae81521328db6588fdcd2d42607540df229e`

Ambas técnicas cumplen eficazmente el objetivo de crear un identificador sustituto para el análisis. La elección entre una y otra dependería de los requerimientos de seguridad específicos del contexto de aplicación.

Técnicas de Anonimización

Técnica 1: k-Anonimidad

Para la anonimización de cuasi-identificadores se aplicó la técnica de **k-anonimidad** sobre el conjunto de datos completo, que consta de más de 1.3 millones de registros. El objetivo fue asegurar que cada registro

fuera indistinguible de al menos otros cuatro (un valor de **k=5**), utilizando como cuasi-identificadores las variables `ano_fabricacion`, `marca_limpia`, `vehiculo_limpio` y `combustible_limpio`.

El análisis de riesgo inicial sobre el `conjunto de datos` completo reveló la presencia de registros vulnerables:

- **Antes de la anonimización:** Un total de **9,432 registros (el 0.695 % del total)** no cumplían con la regla de 5-anonimidad. Aunque el porcentaje es bajo, este número absoluto representa miles de registros que pertenecían a grupos pequeños y, por tanto, con un riesgo de reidentificación más elevado.

Tras la aplicación del algoritmo `kAnon`, el resultado fue el siguiente:

- **Después de la anonimización:** El número de registros que violan la condición de 5-anonimidad se redujo a **cero (0.000 %)**.

El resultado demuestra que el proceso fue exitoso al aplicar la regla de k=5 sobre la totalidad de los datos. El conjunto de datos ahora cumple con el criterio de privacidad establecido, habiéndose logrado a través de las modificaciones (generalización o supresión de valores) que el algoritmo aplicó internamente para eliminar los grupos de riesgo.

Técnica 2: l-Diversidad

Finalmente, se exploró la técnica de **l-diversidad** para fortalecer la protección contra ataques de homogeneidad, que ocurren cuando todos los valores de un atributo sensible dentro de un grupo son idénticos. Se utilizó el atributo `tipo_pago` como variable sensible, con un objetivo de **l=2**, lo que significa que cada grupo de registros indistinguibles debe contener al menos dos tipos de pago distintos.

El análisis se aplicó sobre el conjunto de datos completo, utilizando los mismos cuasi-identificadores que en la k-anonimidad. El resultado del análisis de diversidad se resume en la siguiente estadística:

- **Mínimo de l-diversidad encontrado:** 1.0
- **Media de l-diversidad:** 2.923
- **Máximo de l-diversidad:** 3.0

El hallazgo más relevante es que el **valor mínimo de diversidad (`Min.`) en los grupos de equivalencia es de 1.0**. Este valor indica que, a pesar de que los datos pueden cumplir con la regla de k-anonimidad, todavía existen grupos de registros que no satisfacen la condición de l=2. En estos grupos específicos, todos los individuos comparten el mismo valor para el atributo sensible `tipo_pago`, lo que representa una vulnerabilidad ante un ataque de homogeneidad.

Por lo tanto, la l-diversidad sirve como una métrica de riesgo más estricta que la k-anonimidad, permitiendo identificar vulnerabilidades que esta última no considera por sí sola, y demostrando la necesidad de aplicar múltiples capas de protección dependiendo de la sensibilidad de los datos.

5.3. Evaluación de Utilidad y Equilibrio

Una vez aplicadas las técnicas de anonimización, es imperativo medir el impacto que estas han tenido sobre la utilidad analítica de los datos. En un proceso ideal, se busca maximizar la protección de la privacidad minimizando la distorsión o pérdida de información.

Para esta evaluación, se comparó la distribución temporal del parque vehicular entre el conjunto de datos original y el conjunto anonimizado final. La comparación permite visualizar la cantidad de registros que debieron ser suprimidos (eliminados) por el algoritmo para cumplir estrictamente con la regla de **k-anonimidad** ($k = 5$).

La **Figura 4** presenta los resultados de esta comparación.

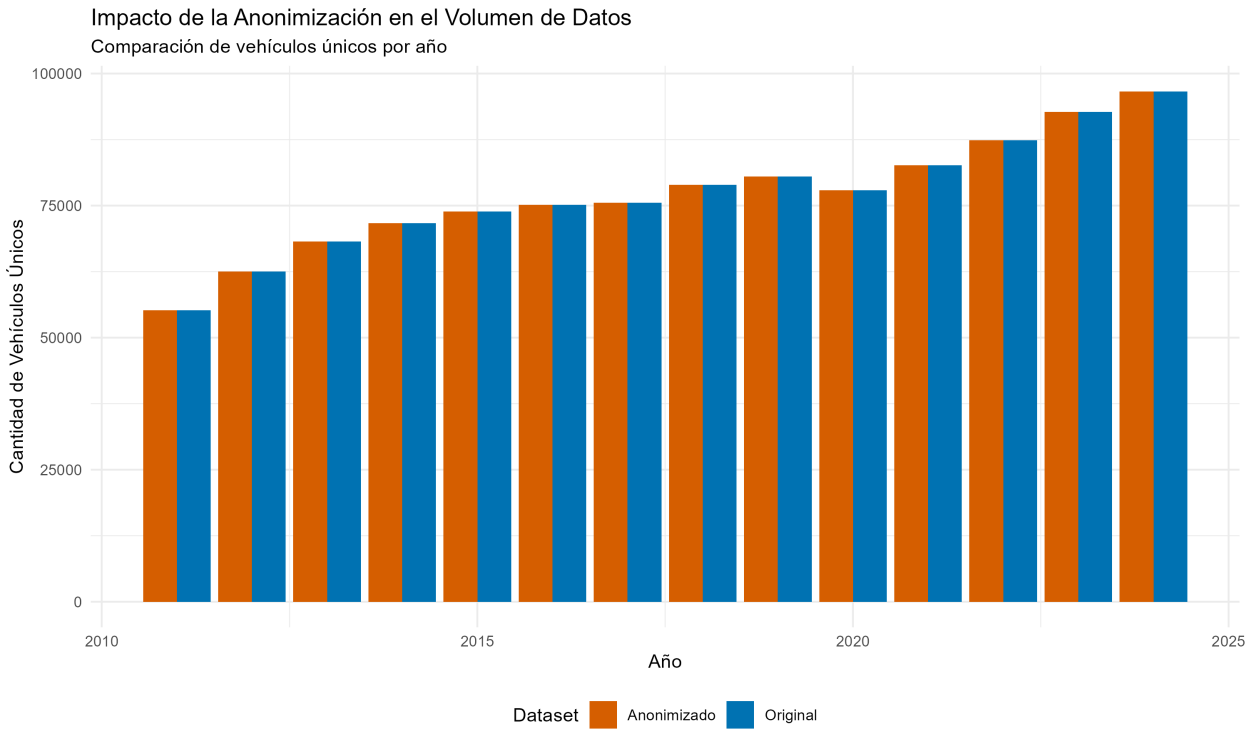

Figura 4: Comparación de la Evolución del Parque Vehicular: Original vs. Anonimizado.

Análisis de Conservación de Información

Como se evidencia en la figura anterior, la aplicación combinada de *HMAC* (para identificadores) y **k-anonimidad** (para cuasi-identificadores) tuvo un impacto marginal en la estructura macroscópica de los datos.

Al contrastar las barras azules (Original) con las naranjas (Anonimizado), la diferencia visual es prácticamente imperceptible. Esto indica que la supresión de registros afectó a una porción insignificante de la población (correspondiente a combinaciones de atributos extremadamente atípicas o únicas que no alcanzaban el umbral de $k = 5$).

Conclusión de Utilidad: El conjunto de datos protegido conserva intactas las tendencias históricas y el volumen general de información, validando su utilidad para la toma de decisiones municipales y el análisis estadístico, cumpliendo así con el objetivo de equilibrar privacidad y utilidad.

6. Conclusiones

Este estudio ha permitido evaluar la viabilidad técnica de implementar estándares avanzados de privacidad en registros administrativos municipales.

6.1. Discusión de Resultados

La evaluación del **equilibrio** entre privacidad y utilidad arrojó resultados positivos.

1. **Eficacia de la Privacidad:** La aplicación de k-anonimidad ($k = 5$) fue exitosa en eliminar registros únicos vulnerables. Sin embargo, el análisis de l-diversidad reveló que la estructura de los datos de pago es altamente homogénea, lo que sugiere que para ciertas variables se requiere una generalización más agresiva.
2. **Preservación de la Utilidad:** Como se evidenció en la comparación gráfica final, la distorsión introducida por la supresión de datos fue mínima. Las tendencias históricas de crecimiento del parque vehicular y la transición de combustibles se mantienen intactas en el set de datos protegido.

6.2. Conclusiones Finales

El proyecto aporta un marco metodológico replicable para que la Municipalidad de La Serena y otros organismos públicos cumplan con los nuevos estándares de la Ley de Protección de Datos Personales. Se concluye que la seudonimización mediante HMAC es la herramienta más robusta para permitir el seguimiento longitudinal de vehículos, mientras que la anonimización estadística debe ser un proceso iterativo que no se base en una sola métrica.

7. Bibliografía

1. Aggarwal, C. C., & Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-Preserving Data Mining: Models and Algorithms* (pp. 11–52). Springer US.
2. Samarati, P. (2001). Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010-1027.
3. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 3.
4. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106-115.
5. Dwork, C. (2006). Differential Privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, Languages and Programming* (pp. 1–12). Springer Berlin Heidelberg.
6. El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association, 15*(5), 627-637.
7. Garfinkel, S. L. (2015). De-identification of personal information. *NIST Internal Report 8053*. National Institute of Standards and Technology.
8. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., & de Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons.
9. Abu-Sinna, M. A., Negm, A. M., Abd El-Salam, M. F., & El-Sayed, A. (2024). Evaluating Privacy-Level Metrics in Privacy-Preserving Data Mining. *International Journal of Intelligent Computing and Information Sciences, 24*(4), 43-59.
10. Pasquini, R., Gkoulalas-Divanis, A., & Fadda, E. (2025). Comparison of anonymization techniques regarding statistical reproducibility. *BMC Medical Informatics and Decision Making, 25*. (Referencia a verificar y completar según PMC11790161 mencionado en búsqueda).
11. Zhang, Q., Yang, L. T., Chen, Z., Li, P., & Bu, F. (2018). Anonymizing Spatiotemporal Data for Privacy-Preserving Data Publishing: A Survey. *IEEE Transactions on Big Data, 4*(3), 300-317.
12. Das, S., Kumar, A., & Chakraborty, D. (2024). A Survey on Differential Privacy for SpatioTemporal Data in Transportation Research. *arXiv preprint arXiv:2407.15868*.
13. Sun, X., Wang, H., Li, J., & Wang, L. (2019). Vehicle License Plate Anonymization Based on Generative Adversarial Networks. *2019 IEEE International Conference on Image Processing (ICIP)*, 3606-3610.
14. Information Commissioner's Office (UK). (n.d.). *Anonymisation*. Recuperado de https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/anonymisation-and-pseudonymisation/anonymisation/
15. Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR), 42*(4), 1-53.

8. Referencias

https://datos.gob.cl/organization/municipalidad_de_la_serena

9. Anexo

Tabla 1:

A General Survey of Privacy-Preserving Data Mining Models and Algorithms
Aggarwal, C. C., & Yu, P. S. 2008
Ofrece una visión general de los modelos y algoritmos existentes en la minería de datos con preservación de privacidad (PPDM).
Proporciona un marco conceptual sobre PPDM, útil para entender el contexto general de las técnicas de anonimización.

Tabla 2:

Protecting Respondents' Identities in Microdata Release
Samarati, P. 2001
Presenta métodos para proteger las identidades en la publicación de microdatos, siendo un trabajo pionero en k-anonymity.
Fundamental para comprender el k-anonymity y las técnicas de generalización y supresión, claves para la anonimización de datos vehiculares.

Tabla 3:

L-diversity: Privacy beyond k-anonymity
Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. 2007
Introduce el concepto de l-diversity para abordar limitaciones del k-anonymity respecto a la protección de atributos sensibles.
Permite explorar un modelo de privacidad más robusto que k-anonymity, aplicable a la diversidad de datos en permisos de circulación.

Tabla 4:

t-Closeness: Privacy Beyond k-Anonymity and l-Diversity
Li, N., Li, T., & Venkatasubramanian, S. 2007
Propone t-closeness, un modelo de privacidad que busca que la distribución de atributos sensibles en una clase sea cercana a su distribución general.
Ofrece otro modelo de privacidad avanzado para comparar, buscando un mejor control sobre la revelación de información sensible en los datos.

Tabla 5:

Differential Privacy
Dwork, C. 2006
Introduce formalmente el concepto de privacidad diferencial, una garantía de privacidad robusta contra diversos ataques.
Provee una base teórica para el estándar de privacidad más fuerte; podría aplicarse a consultas agregadas sobre los datos de permisos.

Tabla 6:

Protecting privacy using k-anonymity
El Emam, K., & Dankar, F. K. 2008
Discute la aplicación práctica y los desafíos del k-anonymity, ilustrado con datos de salud.
Aporta una perspectiva sobre la implementación de k-anonymity y la evaluación de riesgos, transferible al dominio de datos vehiculares.

Tabla 7:

De-identification of personal information
Garfinkel, S. L. 2015
Proporciona una guía conceptual y práctica sobre la de-identificación de información personal según el NIST.
Sirve como referencia para el proceso de de-identificación de las placas patente y otros cuasi-identificadores en los datos.

Tabla 8:

Statistical disclosure control
Hundepool, A., Domingo-Ferrer, J., et al. 2012
Compendio de técnicas y metodologías para el control de la revelación estadística en la publicación de datos.
Fundamental para el enfoque estadístico de la tesis, cubriendo la teoría y práctica del SDC, relevante para datos de permisos.

Tabla 9:

Evaluating Privacy-Level Metrics in Privacy-Preserving Data Mining
Abu-Sinna, M. A., Negm, A. M., et al. 2024
Se enfoca en la evaluación de métricas que cuantifican el nivel de privacidad alcanzado por técnicas de PPDM.
Ayuda a seleccionar y justificar las métricas de privacidad que se utilizarán para evaluar las técnicas de anonimización implementadas.

Tabla 10:

Comparison of anonymization techniques regarding statistical reproducibility
Pasquini, R., Gkoulalas-Divanis, A., & Fadda, E. 2025
Compara el impacto de diversas técnicas de anonimización en la reproducibilidad de resultados y la utilidad estadística de los datos.
Directamente aplicable para medir cómo las técnicas de anonimización afectan la utilidad de los datos vehiculares para análisis estadísticos posteriores.

Tabla 11:

Anonymizing Spatiotemporal Data for Privacy-Preserving Data Publishing: A Survey
Zhang, Q., Yang, L. T., et al. 2018
Revisa las técnicas de anonimización específicamente diseñadas para datos espaciotemporales.
Relevante para el carácter longitudinal de los datos (2011-2024) y si se considera alguna dimensión espacial implícita del seguimiento vehicular.

Tabla 12:

A Survey on Differential Privacy for SpatioTemporal Data in Transportation Research
Das, S., Kumar, A., & Chakraborty, D. 2024
Presenta una revisión de la aplicación de privacidad diferencial a datos espaciotemporales en el contexto de la investigación en transporte.
Conecta la privacidad diferencial con el dominio de datos de transporte, ofreciendo enfoques específicos para datos con características similares a los tuyos.

Tabla 13:

Vehicle License Plate Anonymization Based on Generative Adversarial Networks
Sun, X., Wang, H., et al. 2019
Propone un método de anonimización de placas patente vehiculares utilizando Redes Generativas Antagónicas (GANs) para datos de imágenes.
Aunque enfocado en imágenes, puede inspirar ideas sobre cómo transformar o seudonimizar las placas patente de forma que preserve alguna utilidad.

Tabla 14:

Anonymisation
Information Commissioner's Office (UK) n.d.
Proporciona guías y mejores prácticas sobre anonimización desde la perspectiva del GDPR y la protección de datos.
Ofrece un marco ético y legal, así como buenas prácticas en anonimización que son importantes para la rigurosidad de la tesis.

Tabla 15:

Privacy-preserving data publishing: A survey of recent developments
Fung, B. C. M., Wang, K., et al. 2010
Ofrece una revisión exhaustiva de los desarrollos en la publicación de datos que preservan la privacidad, modelos y desafíos.
Referencia clave para un entendimiento profundo del campo de PPDM, sus conceptos fundamentales y la evolución de las técnicas.