Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

Uso de geometría y topología para ciencia de datos - MA2007B.602

Monterrey, Nuevo León. Fecha, 10 de Junio de 2025.

Entrega Final Reto

Autores:

Gerardo Samuel Reyes Castro A01571141

Marcos Aquino Garcia A00835576

Karla Sofía Cantú Zendejas A01285550

Eduardo Mitrani Krouham A01783220

Profesores:

Lilia Alanís López

Salvador Mancilla Hernández

Resumen

Este trabajo aborda la problemática del alto consumo de combustible en la flota vehicular de SLB México, empresa del sector energético. Mediante un enfoque basado en Análisis Topológico de Datos (TDA), se analiza una base de más de 80 mil transacciones de combustible registradas a través de la plataforma Edenred entre 2021 y 2024. A través de Mapper y homología persistente, se buscaron patrones de consumo, identificando vehículos con rendimientos anómalos respecto a su desempeño esperado. Se utilizaron variables clave como el rendimiento real, recorrido y cantidad de mercancía, y se aplicó similitud coseno para construir complejos topológicos. Finalmente, se propuso un modelo predictivo con XGBoost para clasificar combinaciones con bajo rendimiento esperado. Los hallazgos contribuyen a la toma de decisiones orientadas a mejorar la eficiencia energética de la flota y reducir emisiones de CO, alineándose con los Objetivos de Desarrollo Sostenible.

1. Introducción

La industria del transporte enfrenta actualmente retos significativos en torno a la eficiencia del consumo de combustible y su impacto ambiental. En el caso de SLB México, el departamento de sustentabilidad ha planteado como prioridad la mejora del desempeño energético de su flota vehicular, tanto por razones económicas como por su compromiso con los Objetivos de Desarrollo Sostenible (ODS).

De acuerdo con el CEMDA, el sector transporte en México es responsable del 25 % de las emisiones de gases de efecto invernadero. Este contexto motiva la búsqueda de soluciones innovadoras basadas en datos. Entre ellas, el Análisis Topológico de Datos (TDA) surge como una herramienta prometedora, al permitir descubrir estructuras complejas en grandes volúmenes de información.

Este estudio aplica técnicas de TDA —como Mapper y homología persistente— para analizar el consumo de combustible a partir de las transacciones registradas en Edenred. Se busca responder a las siguientes preguntas:

- ¿Cómo identificar vehículos o estaciones con patrones de consumo anómalos?
- ¿Qué variables influyen en la eficiencia del rendimiento?
- ¿Es posible anticipar, mediante modelos predictivos, combinaciones propensas a bajo rendimiento?

2. Motivación

El consumo de combustible en flotas de transporte representa una problemática crítica tanto por sus implicaciones económicas como por su impacto ambiental. En promedio, una flota pequeña de transporte de carga en México puede consumir más de 240 litros de diésel al día, considerando un rendimiento aproximado de 1.2 km/L y distancias diarias de 200 km por unidad (Polilla Studio, 2018). Esta cifra, multiplicada por cientos de vehículos, implica costos operativos sustanciales y una huella de carbono considerable.

De acuerdo con el Instituto para la Diversificación y Ahorro de la Energía, cada litro de diésel emite aproximadamente 2.64 kg de CO_2 (IDAE, 2016). Así, una sola unidad de transporte puede generar más de 630 kg de CO_2 diarios, contribuyendo a un problema ambiental de gran escala. En el caso particular de SLB México, la identificación de vehículos con rendimientos inferiores al esperado representa una oportunidad tangible para reducir tanto emisiones como gastos innecesarios.

Frente a esta realidad, es fundamental desarrollar estrategias de análisis que permitan detectar comportamientos anómalos, identificar oportunidades de mejora y anticipar escenarios de bajo rendimiento antes de que se traduzcan en pérdidas operativas. La información disponible en las bases de datos de Edenred —que documentan de forma precisa el consumo por unidad, estación y transacción— ofrece un contexto

idóneo para aplicar técnicas de análisis avanzado como el Análisis Topológico de Datos (TDA).

Este enfoque no solo permite visualizar la complejidad estructural del comportamiento de la flota, sino también agrupar vehículos con patrones similares y localizar aquellos cuyo desempeño energético se desvía del comportamiento general. A partir de estas observaciones, se puede diseñar un modelo predictivo que anticipe combinaciones de factores propensas a bajo rendimiento, facilitando una toma de decisiones más eficiente, proactiva y alineada con los principios de sostenibilidad.

3. Marco Teórico

3.1. Análisis Topológico de Datos (TDA)

El Análisis Topológico de Datos (TDA) es un enfoque matemático para estudiar la forma de los datos a través de herramientas de la topología algebraica. A diferencia de los métodos estadísticos tradicionales que se enfocan en relaciones lineales o locales, TDA permite descubrir propiedades globales como componentes conexas, ciclos y vacíos de alta dimensión (Carlsson, 2009).

Una de sus principales aplicaciones es la reducción de la complejidad en espacios de alta dimensión mediante representaciones estructuradas, como complejos simpliciales o grafos, que capturan la geometría y conectividad de los datos. En este trabajo se emplearon dos herramientas clave del TDA: el algoritmo Mapper y la homología persistente.

3.2. Mapper

El algoritmo Mapper, propuesto por Singh, Mémoli y Carlsson (2007), transforma conjuntos de datos complejos en una representación tipo grafo que preserva la topología general del conjunto. Su ejecución consiste en:

1. Aplicar una función *lens* (proyección) que reduce la dimensión o transforma el espacio.
2. Dividir el rango del lens en múltiples intervalos con traslape (definidos por parámetros como `n_cubes` y `overlap`).
3. Aplicar un algoritmo de agrupamiento local (como DBSCAN) en cada subgrupo.
4. Construir un grafo en el que cada nodo representa un clúster, y se conectan si comparten datos en común.

El resultado es un grafo que permite visualizar la estructura global del conjunto de datos, detectar grupos de comportamiento similar o encontrar regiones con propiedades atípicas.

3.3. Filtro de Densidad basado en Similitud Coseno

En este trabajo no se utilizó una proyección clásica como PCA o UMAP, sino un *filtro de densidad* personalizado basado en la **similitud coseno**. Este filtro tiene como objetivo asignar a cada observación un valor que refleje la densidad relativa de su vecindad en el espacio angular, destacando regiones densas con comportamientos similares.

El valor del filtro para un punto x_i se definió como:

$$f_\varepsilon(x_i) = \text{MinMaxScale}\left(\sum_j \exp\left(-\frac{\text{cosine}(x_i, x_j)^2}{\varepsilon}\right)\right) \tag{1}$$

donde $\text{cosine}(x_i, x_j)$ es la similitud coseno entre las observaciones x_i y x_j, y ε es un parámetro de suavizado. Finalmente, los valores se normalizaron al intervalo $[0, 1]$ usando escalamiento MinMax, para ser utilizados como función *lens* en el algoritmo Mapper.

Este enfoque permitió resaltar agrupaciones sutiles en el espacio de datos, priorizando regiones de alta densidad con patrones de consumo energético similares.

3.4. Homología Persistente

La homología es una herramienta matemática que permite cuantificar características topológicas como componentes conexas (H_0), ciclos (H_1) y vacíos (H_2) dentro de un espacio. La *homología persistente* permite seguir la evolución de estas características a lo largo de una *filtración*, es decir, una secuencia de complejos crecientes construidos sobre los datos.

El resultado se representa en un **diagrama de persistencia** o **código de barras**, donde cada barra representa la "vida" de una característica topológica: su aparición y desaparición en la filtración. Barras largas suelen indicar estructuras significativas, mientras que las más cortas pueden representar ruido.

3.5. Complejo de Vietoris–Rips

Para computar la homología sobre datos discretos, se utilizó el complejo de Vietoris–Rips. Dados puntos x_i en un espacio métrico y un umbral ε, se construyen símplices conectando subconjuntos de puntos donde cada par está a una distancia menor o igual que ε.

Este enfoque es ampliamente utilizado por su simplicidad y eficiencia computacional, permitiendo identificar patrones topológicos relevantes incluso en espacios de alta dimensión.

4. Metodología

4.1. Descripción de los datos

El análisis se llevó a cabo utilizando datos históricos de consumo de combustible proporcionados por la plataforma Edenred, correspondientes a los años 2021 a 2024. Cada archivo contenía transacciones detalladas con información como el volumen cargado, el precio unitario, la unidad vehicular, la estación de servicio, el tipo de mercancía y el rendimiento reportado.

Para preparar los datos, se aplicó una función personalizada de limpieza que:

- Homogeneizó los nombres de columnas en minúsculas con guiones bajos.
- Identificó y unificó la columna de fecha.
- Eliminó registros duplicados y columnas completamente nulas.
- Excluyó conceptos como bonificaciones, ajustes administrativos o transacciones atípicas.

Los archivos anuales fueron concatenados en un único `DataFrame`. También se construyó una nueva variable `Unidad`, a partir del identificador del vehículo, para facilitar el análisis por unidad específica.

4.2. Preprocesamiento y selección de variables

Las columnas clave fueron transformadas a formato numérico, incluyendo `rendimiento_real` (rendimiento observado) y `recorrido`, eliminando registros con valores no realistas (como rendimientos negativos).

Se seleccionaron las siguientes variables como características del análisis:

- `recorrido`: distancia recorrida en la transacción.
- `precio_unitario`: costo por litro.
- `cantidad_mercancía`: volumen de combustible cargado.
- `rendimiento`: rendimiento estimado o teórico.
- `rendimiento_real`: rendimiento efectivamente observado.

Estas variables fueron escaladas con `StandardScaler` para asegurar una magnitud comparable entre dimensiones.

4.3. Aplicación del algoritmo Mapper

Se utilizó la biblioteca `KeplerMapper` para implementar el algoritmo Mapper. Como función de proyección (*lens*) se empleó un filtro de densidad basado en la similitud coseno, el cual fue descrito previamente en el Marco Teórico. Esta elección permitió resaltar regiones densas de comportamiento similar en el espacio de variables seleccionadas.

La configuración del Mapper incluyó:

- Número de cubos: 10
- Traslape: 0.5
- Agrupamiento local: `DBSCAN` con $eps = 0{,}5$ y $min\_samples = 5$

Durante el proceso exploratorio, también se evaluaron otras proyecciones, como *UMAP*, sin embargo, se optó por la función de densidad por similitud coseno debido a su mayor estabilidad y su capacidad para reflejar agrupamientos sutiles en la estructura de consumo. Esta proyección generó grafos más legibles y estructuralmente interpretables, facilitando el análisis posterior.

Se probaron distintas configuraciones de cobertura, variando los valores de `n_cubes` (5, 10, 15) y `perc_overlap` (0.1, 0.3, 0.5), con el objetivo de encontrar un balance adecuado entre granularidad, conectividad y claridad estructural. La configuración final elegida fue de `n_cubes = 10` y `overlap = 0.5`, al ser la que mostró una estructura informativa y suficientemente segmentada para identificar patrones relevantes.

El algoritmo de agrupamiento utilizado dentro de cada cubo fue `DBSCAN`, configurado con $eps = 0{,}5$ y $min\_samples = 5$. Esta técnica fue seleccionada por su capacidad de detectar clústers de forma arbitraria y su robustez frente al ruido. Los parámetros fueron ajustados empíricamente considerando la densidad del espacio escalado.

La estructura de Mapper resultante permitió visualizar agrupamientos de unidades vehiculares con patrones similares en términos de rendimiento, recorrido y consumo. Además, facilitó la identificación de nodos con comportamientos atípicos o ineficientes, los cuales fueron estudiados en el análisis posterior.

4.4. Análisis de persistencia y homología

A partir de los agrupamientos generados, se implementó un analisis topológico adicional utilizando herramientas de homología persistente. Este enfoque nos permitió ver la complejidad y variabilidad de los patrones de cada unidad en diferentes escalas.

De primero, se calcularon los diagramas de persistencia para las unidades con mayor cantidad de registros, con variables de entrada el recorrido, el precio unitario y la cantidad transportada. A partir de estos diagramas, se extrajeron metricas (H_0), tales como el número de componentes conexas, la suma total de persistencias y la máxima persistencia. Estas métricas fueron posteriormente correlacionadas con el rendimiento real de cada unidad asi identificando asociaciones entre la dispersión topológica de los datos y la eficiencia de la operacion.

Luego, se implemento el análisis temporal mediante la segmentación de los datos por mes. Esta descomposición mensual nos permitió observar la evolución dinámica de las métricas topológicas y su relación con el rendimiento de cada vehículo en el tiempo.

Por ultimo, se extendió el análisis al dominio de series de tiempo, utilizando técnicas de embedding en las trayectorias temporales de rendimiento y emisiones de CO_2. Sobre estos embeddings, se calcularon diagramas de persistencia en (H_1) y se evaluaron las distancias Wasserstein permitiendo detectar cambios en la dinámica operativa.

Este análisis topológico complementó la interpretación de los agrupamientos obtenidos, proporcionando una caracterización de patrones estables, anomalías y cambios en el desempeño de las unidades vehiculares

4.5. Metodología del Modelo de Clasificación

Durante esta etapa al ver los resultados topologicos, se decidió desarrollar y evaluar un modelo de clasificacion, para predecir la variable objetivo del rendimiento real.

Se definieron las siguientes variables predictoras:

- **Variables numéricas**: `conductor_score`, `vehiculo_score`, `rend_cond_mean`, `rend_veh_mean`.
- **Variables categóricas**: `division`, `bl`, `mercancía`, `no_estacion_pemex`.

Una vez con esto se dividio en entrenamiento y en test, siendo estos test 25 porciento de los datos y el entrenamiento el 75 restante.

Luego se implementó un *pipeline* que integra preprocesamiento y modelo en el cual usamos el clasificador XGBoost. Definimos una serie de parametros los cuales usando GridSearch se encontraron los parametros optimos para nuestro objetivo, los cuales fueron:

- Número de combinaciones: 20.
- Validación cruzada: 5 pliegues.
- Métrica: `f1`.
- Espacio de búsqueda:
  - `n_estimators`: [50, 100, 200, 500]
  - `max_depth`: [3, 5, 7]
  - `learning_rate`: [0.01, 0.1, 0.2]
  - `subsample`: [0.8, 1.0]
  - `colsample_bytree`: [0.8, 1.0]
  - `gamma`: [0, 0.1, 0.2]

Se ajustaron los mejores parametros obtenidos, y a partor de eso se entrenó el mejor estimador con todo el conjunto de entrenamiento y se evaluó en el conjunto de prueba en donde se calcularon las siguientes metricas:

- Exactitud (Accuracy)
- Precisión, Recall y F1-Score por clase
- Matriz de Confusión

Resultados principales: `Accuracy global: 0.85`
`Clase 0 -- Precisión: 0.87, Recall: 0.92, F1-score: 0.90`
`Clase 1 -- Precisión: 0.77, Recall: 0.65, F1-score: 0.70`

Con esto, se logro crear un modelo que va a ayudar a SLB Mexico poder predecir si su rendimiento en proximos viajes va a ser bueno o malo, dependiendo de las variables propuestas anteriormente con una precision del 85 porciento.

5. Conclusiones

5.1. Síntesis de hallazgos

Este proyecto aplicó herramientas de *Topological Data Analysis* (TDA) para explorar estructuras subyacentes en datos operativos relacionados con el rendimiento vehicular (km/L) y las emisiones de CO_2 ($kgCO_2$). Entre los principales hallazgos se destacan:

- La presencia de ciclos topológicos persistentes en H_1, que revelan trayectorias operativas repetitivas o condiciones estructuradas de funcionamiento.
- Una conexión rápida entre componentes en H_0, lo cual indica regiones de alta densidad y agrupamientos bien definidos.
- La ausencia de cavidades topológicas en H_2, consistente con la naturaleza y dimensionalidad de los datos.
- Variaciones significativas en los diagramas de persistencia y landscapes entre diferentes subconjuntos de datos, revelando estructuras no evidentes mediante técnicas tradicionales.

Estos resultados demuestran que el enfoque topológico aporta una perspectiva novedosa y complementaria al análisis clásico de datos operativos, permitiendo detectar patrones profundos sin imponer supuestos métricos o lineales.

5.2. Implicaciones teóricas y prácticas

Desde el punto de vista teórico, el uso de TDA en este proyecto permitió:

- Capturar propiedades globales del espacio de datos, como agrupamientos (H_0) y ciclos (H_1), sin supuestos de linealidad o normalidad.
- Representar la información en formas geométricas persistentes, facilitando una comprensión estructural de los datos.
- Complementar métricas numéricas tradicionales con una interpretación cualitativa de la organización interna de los datos.
- Extender la noción de distancia mediante distancias Wasserstein entre diagramas, evidenciando variaciones estacionales.

En cuanto a sus aplicaciones prácticas:

- Se identificaron agrupamientos operativos relevantes no detectados por PCA ni clustering convencional.
- Se revelaron trayectorias y condiciones de operación repetitivas, clave para estrategias de diagnóstico o mantenimiento.
- Se caracterizaron zonas con distintas densidades, facilitando la segmentación operativa de unidades, estaciones y conductores.
- Se generó información de apoyo para decisiones logísticas y de sostenibilidad con base en la estructura topológica de los datos.

5.3. Limitaciones del estudio

El trabajo presenta varias restricciones que deben considerarse:

- El análisis se basó en un solo dataset consolidado, sin comparación contra fuentes externas.
- No se aplicaron técnicas de reducción de ruido o umbrales de persistencia.
- No se realizó validación estadística formal mediante *bootstrap* o permutaciones.
- Se utilizó exclusivamente el complejo Vietoris–Rips, sin comparar otras opciones como Čech o Alpha complexes.

Estas limitaciones fueron determinadas por las condiciones computacionales y temporales del proyecto, pero abren la puerta a mejoras metodológicas en el futuro.

5.4. Investigación futura

Entre las extensiones posibles para fortalecer y ampliar este trabajo se proponen:

- Incluir otros complejos simpliciales (Alpha, Čech) para comparar estructuras topológicas.
- Validar la robustez de ciclos mediante pruebas estadísticas como *bootstrap* o simulaciones Monte Carlo.
- Aplicar técnicas de reducción de dimensionalidad (UMAP, Isomap) antes del TDA.
- Usar *persistence images* o *landscapes* como *features* en modelos predictivos.
- Explorar diferentes funciones de filtración y su impacto en los diagramas obtenidos.

5.5. Aplicaciones potenciales

El enfoque utilizado puede trasladarse a otros contextos operativos o industriales. Algunas aplicaciones potenciales incluyen:

- Monitoreo de eficiencia y consumo en sistemas vehiculares o industriales.
- Detección de anomalías operativas sin definir umbrales preestablecidos.
- Agrupamiento estructural de agentes, estaciones o trayectos sin restricciones métricas.
- Análisis de estabilidad o redundancia en sistemas complejos.

TDA permite descubrir relaciones implícitas y estructuras profundas en los datos, lo que lo convierte en una herramienta valiosa para sectores que buscan eficiencia, estabilidad y diagnóstico en contextos no lineales.

5.6. Resumen ejecutivo

- Se aplicó homología persistente sobre datos reales de rendimiento vehicular y emisiones de CO_2, utilizando complejos Vietoris–Rips.
- Se interpretaron los resultados con base en componentes H_0 y H_1, descartando H_2 por la naturaleza de los datos.
- Se identificaron ciclos persistentes y zonas de alta densidad no visibles por técnicas tradicionales.
- Se visualizaron estructuras mediante diagramas de persistencia, *barcodes* y *landscapes*.
- Se analizaron variaciones temporales mediante distancias Wasserstein.
- Se propusieron mejoras metodológicas y aplicaciones futuras en contexto industrial.

Estas contribuciones reafirman el valor de TDA como una herramienta prometedora para el análisis estructural de datos operativos complejos, particularmente en el ámbito de la eficiencia vehicular y el control de emisiones.

Referencias

- Polilla Studio. (2018). Control de combustible para flotas: errores que afectan empresas en México. *Tecnomotum*. https://tecnomotum.com.mx/post/control-de-combustible-para-flotas-errores-que-afectan-empresas-en-mexico
- IDAE. (2016). Consumo de carburante y emisiones. *IDAE*. Recuperado de: https://coches.idae.es/consumo-de-carburante-y-emisiones#:~:text=Por%20cada%20litro%20de%20gasolina,64%20kg%20de%20CO2
- Carlsson, G. (2009). *Topology and data*. Bulletin of the American Mathematical Society, 46(2), 255–308.
- Singh, G., Mémoli, F., & Carlsson, G. (2007). *Topological methods for the analysis of high dimensional data sets and 3D object recognition*. Eurographics Symposium on Point-Based Graphics.