UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Trabajo de Fin de Máster

Detección de anomalías en la monitorización de una flota de autobuses

Curso 2021-2022

Autor: **Diana Vanessa Silva Arando**
Tutores: **Eva Onaindía de la Rivaherrera**
Bernardo V. Tormos Martínez
Máster: **Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital**
Departamento: **Sistemas Informáticos y Computación**

Agradecimientos

Quisiera agradecer en primer lugar a mi directora de proyecto Eva Onaindia por su constante apoyo además de sus conocimientos, paciencia y fabulosa actitud en el desarrollo de este trabajo. Sus constantes esfuerzos y guías permitieron la realización de este trabajo.

Así también, dar gracias al equipo de Motores Térmicos de la UPV: Bernardo Tormos codirector del proyecto, Santiago Ballester, Vicente Macián y a los miembros restantes del equipo, que apoyaron el proyecto con sus conocimientos en el área automotriz, son un gran equipo, fue muy motivador trabajar con ustedes.

Muchas gracias a la Empresa Municipal de Transporte de Valencia (EMT) y en especial a Josep Chiner Palmi y Vicente Buendía Ramón que fueron los principales responsables de proporcionar los datos para desarrollar el proyecto y por orientarnos en el intrincado mundo del transporte urbano.

Una mención especial a la Fundación Carolina que me brindó la oportunidad de estudiar en España y en la UPV y me proporcionó los medios para cursar este Máster, sin ustedes no hubiera podido conocer el otro lado del mundo de esta forma y conocer tantos lugares y personas increíbles, gracias infinitas.

Con mucho cariño agradezco a mis padres por todo el amor que me dieron toda la vida. Gracias mamá por creer en mi e impulsarme a mejorar, por su comprensión y consejos. Gracias papá por siempre dar todo por mí, quisiera que siguieras con nosotras para ver esto, un abrazo y un beso hasta el cielo, te extraño mucho.

Y, por último, pero no menos importante quiero agradecer a mi amado novio Brayan, gracias por las risas, la compañía, la ayuda, la paciencia y todo el amor durante este proceso, te amo, se vienen más aventuras juntos.

Resumen

El objetivo de este TFM consiste en desarrollar una aplicación para el análisis de datos recogidos mediante el protocolo FMS (Fleet Management System) de una flota de autobuses con el fin de detectar patrones de comportamiento en el funcionamiento de los vehículos y detectar posibles anomalías en aquellos que se salgan del patrón. Dicho análisis proporcionará una información útil para el mantenimiento predictivo de la flota y el seguimiento del estado del vehículo. El trabajo se centra en la detección de anomalías en series temporales de datos FMS de la flota de autobuses recogidos durante un período de tiempo. Al no disponer de un dataset de entrenamiento etiquetado con secuencias anómalas, la propuesta se centra en desarrollar un modelo no supervisado de detección de anomalías. Con este fin, se proporciona un conjunto de datos con secuencias temporales que se consideran normales, a partir del cual se entrena un modelo que aprende a analizar diferencias con otras secuencias y a clasificarlas como anomalías o no. Se implementarán dos técnicas de detección de anomalías, una basada en los k-vecinos más cercanos y otra basada en técnicas de agrupamiento (clustering) que se aplicarán en función de la tipología de los datos FMS.

Palabras clave: autobús, anomalía, trayecto, señales, modelo.

Abstract

The objective of this TFM is to develop an application for the analysis of data collected through the FMS (Fleet Management System) protocol of a fleet of buses in order to detect behavior patterns in the operation of vehicles and detect possible anomalies in those That they get out of the pattern. This analysis will provide useful information for predictive fleet maintenance and vehicle health monitoring. The work focuses on the detection of anomalies in time series of FMS data of the bus fleet collected over a period of time. In the absence of a training dataset labeled with anomalous sequences, the proposal focuses on developing an unsupervised model of anomaly detection. To this end, a set of data with temporal sequences that are considered normal is provided, from which a model is trained that learns to analyze differences with other sequences and to classify them as anomalies or not. Two anomaly detection techniques will be implemented, one based on k-closest neighbors and the other based on clustering techniques that will be applied based on the type of FMS data.

Keywords: autobus, anomaly, route, signals, model.

Resum

L'objectiu d'aquest TFM consisteix a desenvolupar una aplicació per a l'anàlisi de dades recollides mitjançant el protocol FMS (Fleet Management System) d'una flota d'autobusos amb la finalitat de detectar patrons de comportament en el funcionament dels vehicles i detectar possibles anomalies en aquells que s'isquen del patró. Aquesta anàlisi proporcionarà una informació útil per al manteniment predictiu de la flota i el seguiment de l'estat del vehicle. El treball se centra en la detecció d'anomalies en sèries temporals de dades FMS de la flota d'autobusos recollits durant un període de temps. Al no disposar d'un dataset d'entrenament etiquetat amb seqüències anòmales, la proposta se centra en desenvolupar un model no supervisat de detecció d'anomalies. A aquest efecte, es proporciona un conjunt de dades amb seqüències temporals que es consideren normals, a partir del qual s'entrena un model que aprén a analitzar diferències amb altres seqüències i a classificar-les com a anomalies o no. S'implementaran dues tècniques de detecció d'anomalies, una basada en els k-veïns més pròxims i una altra basada en tècniques d'agrupament (clustering) que s'aplicaran en funció de la tipologia de les dades FMS.

Paraules clau: autobús, anomalia, trajecte, senyals, model.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 Introducción

El transporte urbano en las ciudades es muy importante para ciudades como Valencia, cada día cientos de vehículos de transporte público circulan por la ciudad conectando la ciudad, tal magnitud de movimiento y fuerza motriz requiere mantenimiento regular por lo que cada entidad encargada del transporte metropolitano tiene la necesidad de desarrollar procesos en que los que se determina qué componentes de la flota de autobuses deben ser revisados y enviados a mantenimiento para aprovechar al máximo su tiempo de vida útil a la vez que se garantiza la seguridad de los operadores y usuarios. La aproximación clásica es el llamado mantenimiento preventivo, donde el fabricante ha definido periodos de intervención o sustitución de piezas independientemente de su condición real, de cara a asegurar una baja probabilidad de aparición de fallo en servicio. Una primera evolución es el llamado mantenimiento predictivo, donde esas intervenciones fijas se sustituyen, al menos una parte importante, por revisiones que permiten determinar la condición o estado de cada pieza o sistema y tomar la decisión de si hay que sustituir o puede continuar en servicio. En un caso hay una pérdida de vida útil del componente ya que se sustituye cuando podría aún disponer de mayor vida útil y en el otro caso se evita este problema con la determinación del estado, pero a expensas de un tiempo de indisponibilidad para permitir realizar la inspección sobre el sistema.

Es aquí cuando entra en juego el estudio de la detección de anomalías que pueden ser enfocadas a este problema en particular ya que con el avance de la tecnología se tiene a mano gran cantidad de información sobre el funcionamiento de un autobús, los cuales pueden obtenerse en intervalos de tiempo pequeños que dan seguimiento bastante minucioso a los componentes de un autobús (principalmente el motor). Recurriendo a un análisis detallado de la información disponible es posible construir un modelo capaz de identificar un patrón en el funcionamiento de los autobuses y por lo tanto hacer identificable cuando uno requiera ser revisado por un experto al presentar datos que no se ajusten a lo que el modelo indica como patrón normal. Esta aproximación se asemeja al llamado predictivo mencionado anteriormente sin la desventaja de la indisponibilidad para la realización de las medidas de la condición ya que se aprovecha información que se está obteniendo en el propio servicio del vehículo.

1.1 Objetivos y motivación

A través de la recolección de señales del funcionamiento de un vehículo se tiene una variedad de datos por cada autobús, los cuales indican el valor de una variable en un instante de tiempo. El objetivo es determinar la posibilidad de realizar un diagnóstico a partir de los datos recopilados por lo que surge la siguiente pregunta:

¿Es posible detectar de forma temprana que un autobús requiera mantenimiento basándonos en su comportamiento con la información disponible?

Lo cual lleva a formular una serie de objetivos juntos con sus objetivos secundarios.

Objetivo 1. ***Realizar una recopilación de información sobre el análisis y mantenimiento de vehículos, detección de anomalías y modelos para detección de anomalías.***

- Recopilar textos y conocimiento de expertos sobre las consideraciones que se deben tener para el mantenimiento de motores y vehículos en general.
- Sintetizar la información sobre la detección de anomalías para identificar los tipos, problemas comunes y soluciones propuestas
- Seleccionar y explicar las técnicas y/o modelos más empleados en la detección de anomalías.

Objetivo 2. ***Analizar y determinar la funcionalidad de los datos proporcionados por el sistema de recopilación de datos.***

- Identificar el propósito y concepto de los datos de monitorización.
- Identificar las relaciones y correspondencias entre los distintos archivos producto de la monitorización.
- Clasificar los tipos de datos y ordenarlos por orden de relevancia en la detección de anomalías.
- Realizar una limpieza de los datos para ser procesados.

Objetivo 3. ***Proponer un protocolo para la detección de anomalías en autobuses***.

- Seleccionar una o más técnicas y desarrollar un modelo de *clustering*.
- Entrenar uno a más modelos con los datos limpios producto del objetivo anterior.
- Probar variaciones de las parametrizaciones del modelo.
- En base al mejor modelo, analizar y comparar resultados.
- Tras la comparación de resultados desarrollar y proponer un protocolo general para aplicar en otros casos.
- Proponer recomendaciones y futuras mejoras del protocolo.

1.2 Estructura del trabajo

Una vez definidos los objetivos de este trabajo se desarrollarán una serie de apartados para poder alcanzarlos, los cuales irán en el siguiente orden:

- Primeramente, se abarcará el estado del arte en el cual se describen los conceptos relacionados con el tema de este trabajo, iniciando con un detalle de los sistemas inteligentes en flotas de autobuses y su uso en el día a día; se continúa con una síntesis sobe la investigación actual sobre la detección de anomalías en la cual se parte desde la definición, tipos y clasificaciones y finaliza mencionando las técnicas que suelen emplearse.
- En segundo lugar, se explicará el procesamiento de la información disponible, la recopilación y el análisis de datos disponibles sobre la monitorización de autobuses partiendo en describir el sistema bus CAN, pasando a identificar las unidades básicas de información y la forma en que se conectan entre sí, hasta describir las señales disponibles para la búsqueda de anomalías.

- En cuarto lugar y un punto importante a abarcar es el análisis del comportamiento de las series temporales que dará un panorama de las técnicas necesarias para lograr los objetivos de este trabajo.
- Como quinto paso, se hablará sobre el modelo empleado para realizar el reconocimiento de anomalías junto con las consideraciones, los resultados obtenidos y la discusión de estos. En base a los análisis hechos se propondrá un protocolo.
- Finalmente, se hablará sobre unas consideraciones y trabajos futuros en base al punto anterior.

2 Estado del arte

El estado del arte de este trabajo se divide en varias secciones. Por un lado, se describe el uso de sistemas inteligentes en flotas de autobuses y algunos casos de éxito. Posteriormente se hace una introducción al concepto de serie temporal. Las dos secciones restantes se dedican a explicar los tipos de detección de anomalías y las principales técnicas utilizadas en dicha detección.

2.1 Sistemas inteligentes en flotas de autobuses

El mantenimiento de una flota de vehículos de transporte, como es el caso de autobuses urbanos, consiste en realizar una serie de tareas adaptadas a la tipología del vehículo con el fin de maximizar el rendimiento de este, minimizar su coste por kilómetro, garantizar la seguridad de los pasajeros y certificar que el vehículo cumple con la normativa vigente.

El procedimiento operativo más rudimentario consiste en realizar un *mantenimiento correctivo*, es decir, llevar el vehículo a taller para ser reparado cuando se detecta un fallo de funcionamiento. El *mantenimiento preventivo* consiste en aquellas intervenciones periódicas realizadas sobre el vehículo en aras a reducir la probabilidad de aparición del fallo en servicio (como el típico cambio de aceite del motor). El *mantenimiento predictivo*, por otro lado, se fundamenta en detectar posibles fallos en etapas tempranas con el fin de evitar que estos se manifiesten en averías importantes durante el funcionamiento del vehículo.

Hoy en día, y con el avance de la tecnología, los vehículos modernos llevan dispositivos capaces de emitir cientos de señales provenientes de diferentes partes del vehículo como el motor, la transmisión, frenos y una gran variedad de otros controladores en el vehículo. Además, los dispositivos telemáticos son capaces de transmitir esta información junto con la posición GPS de los vehículos y otros datos, para ser analizados.

Cada vez es más frecuente la necesidad de sistemas inteligentes para la prevención de anomalías o averías en flotas de autobuses. El objetivo es predecir potenciales averías y reducir así el costo de mantenimiento de la flota. A continuación, se describen algunos ejemplos de sistemas inteligentes en flotas de autobuses y particularmente para mantenimiento predictivo.

El proyecto de I+D+i **Predicbus** es un sistema inteligente para prevenir y anticiparse a posibles averías y, además, optimizar la conducción de los autobuses, con un consecuente ahorro energético y de emisiones contaminantes [PREDICBUS21]. Predicbus desarrolla un sistema inteligente capaz de monitorizar en tiempo real los parámetros de funcionamiento del autobús y aplica técnicas de *data mining* y reconocimiento de patrones para el análisis y la toma de decisiones. Mientras que muchos sistemas de mantenimiento predictivo se limitan a la sustitución de piezas en función del tiempo o la distancia recorrida, Predicbus infiere la necesidad de sustitución de piezas para evitar averías no previstas mediante el análisis de otros parámetros más precisos como la temperatura del aceite y del refrigerante, presión del sistema de aire comprimido, o carga del alternador.

Al igual que en el trabajo realizado en este proyecto, muchos sistemas de mantenimiento predictivo utilizan el bus CAN (sistema de comunicación desarrollado para intercambiar información entre las unidades de control electrónicas de un automóvil) con el fin de extraer información del vehículo (ver más detalles en sección 3.1). Como ejemplo, el trabajo publicado en [Massaro19] describe una unidad de control eléctrico (UCE) para la monitorización de una flota de autobuses utilizando los estándares SAE J1962 y SAE J1939 para el diseño y desarrollo de un sistema de IoT (Internet of Things) que recupera parámetros a partir de la UCE. El sistema de IoT está conectado a un motor de Inteligencia Artificial que implementa un Perceptrón Multicapa capaz de realizar un mantenimiento predictivo de cada vehículo analizando el comportamiento del conductor. El comportamiento se clasifica en categorías como conducción segura, conducción ecológica y otros estilos dependiendo del cumplimiento de límites de velocidad por parte del conductor. Para estimar los comportamientos se utilizan técnicas de *clustering* que agrupa a los conductores por posicionamiento del acelerador, revoluciones del motor, velocidad indicada por el GPS, etc.

2.2 Series temporales

Debido a la naturaleza y funcionamiento de los sistemas de monitoreo en flotas de autobuses, es de esperar que, al recoger datos en tiempo real registrados durante el funcionamiento de los vehículos a lo largo del tiempo, estos tengan un gran volumen y el tiempo sea la mayor característica de los datos recogidos. Es por eso que se requiere hablar sobre las series temporales ya que es la forma en la que se presentan los datos de monitoreo para este trabajo.

Una serie temporal es un conjunto o colección de observaciones sobre una característica, univariable o multivariable, a lo largo de un periodo de tiempo [ALONSO20]. Una serie temporal suele tener un tamaño considerable y, por tanto, tener un costo computacional mayor que otros tipos de datos [Laptev15].

Las series temporales son muy usadas para hacer predicciones de comportamientos o detectar anomalías en la característica observada. Tienden a presentar tres componentes, los cuales son:

- **Tendencia**: Se refiere a los cambios que puedes ser notables a lo largo del tiempo de la observación, pueden ser crecientes, decrecientes o estables.
- **Estacionalidad**: Se refiere a la observación de ciclos o comportamiento regulares que se repiten en el tiempo de observación.
- **Aleatoriedad**: Se refiere a observaciones que no se ajustan a la evolución común de la serie y no parecen tener un comportamiento explicable.

Las series temporales pueden ser estudiadas u organizadas en:

- **Series temporales completas**. Es cuando las series temporales son discretas e individuales. Esta es la forma más común de atacar problemas relacionados con series temporales.

- **Series temporales secuenciadas**. Las series están subdivididas en secuencias generadas por una ventana deslizante sobre la misma. Esta forma de organizar los datos es cuestionada, los resultados de los algoritmos son esencialmente aleatorios.
- **Puntos temporales**. A cada dato de la serie se asigna las distancias a los puntos más próximos además de una medida de similitud con los mismos

El objetivo del estudio de las series temporales es el conocimiento del comportamiento de una variable a través del tiempo para poder realizar predicciones, es decir, determinar qué valor tomará la variable objeto de estudio en uno o más períodos de tiempo situados en el futuro. O para detectar datos anómalos en la serie para la toma decisiones (por ejemplo: mantenimiento de vehículos). Estos objetivos se consiguen mediante la aplicación de un determinado modelo al que es sometido una serie temporal o varias series temporales dependiendo del caso [Parra19].

2.3 Detección de anomalías

La detección de anomalías es el reconocimiento de datos que se encuentran fuera de lo que suele denominarse un "comportamiento normal", es decir, se centra en identificar y reconocer valores inusuales o atípicos en un entorno determinado [Chandola09, Lopez-Avila19]. Se definen tres tipos principales de anomalías:

- Puntuales: Son anomalías de datos individuales que pueden ser etiquetadas como anómalo o no anómalo, sin importar el contexto del problema o la naturaleza de los datos**.**
- Contextuales: Son anomalías que dependen del contexto en que se evalúe cierta instancia de datos ya que las condiciones suelen cambiar el concepto de anomalía. Por lo que, si bajo un conjunto de condiciones los datos observados se consideran como anomalía, al cambiar estas varios o todos los elementos dejan de ser considerados anomalía.
- Colectivas: Son anomalías que considera un conjunto de datos en lugar de instancias individuales. Puede que un elemento observado de forma individual no se considere anomalía, pero al ser observado en conjunto con otros ese conjunto si se considera anómalo.

A la hora de realizar un estudio de anomalías en un conjunto de datos, un aspecto relevante es la disponibilidad de datos etiquetados como "normales" o "anómalos". En función de esta disponibilidad, las técnicas de detección de anomalías pueden clasificar en uno de los siguientes tipos:

- Supervisadas: Son las más costosas ya que para poder entrenar un modelo de detección de anomalías supervisado se requiere la intervención de un experto humano en el área que pueda identificar un dato o conjunto de datos como una anomalía o un comportamiento normal. Esto se traduce en una gran cantidad de tiempo y recursos para esta tarea, eso sin considerar que el concepto de anomalía puede variar dependiendo las condiciones del entorno siendo necesario un re etiquetado de los datos.

- Semi supervisadas: Estas técnicas requieren que solo se etiqueten los datos considerados “normales”, permitiendo que la clasificación de datos sea menos costosa.
- No supervisadas: Estas técnicas no requieren datos de entrenamiento etiquetados, se apoya en otro tipo de técnicas de clasificación como k-means o árboles.

2.4 Técnicas de detección de anomalías

En esta sección se describen brevemente las principales técnicas computacionales para la detección de anomalías. En líneas generales, las técnicas de detección de anomalías son diversas y se clasifican en función del tipo de algoritmo que emplean [Chandola09, Thudumu20]:

- técnicas basadas en clasificación
- técnicas basadas en el vecino más próximo
- técnicas estadísticas
- técnicas de agrupamiento o *clustering*

Dado que en este caso de estudio los datos no están etiquetados y no se dispone de prototipos de muestras que indiquen un comportamiento normal o anómalo, la propuesta para abordar este problema será utilizar técnicas de *clustering*. Por dicho motivo, se explica brevemente el resto de las técnicas en esta sección, y dedicamos la siguiente sección a explicar en detalle las técnicas de agrupamiento.

2.4.1 Técnicas basadas en clasificación

Estos métodos son los que requieren del etiquetado de datos para poder entrenar un modelo que distinguirá entre datos anómalos y no anómalos. Es aplicable tanto a una variable de una sola clase como multiclase. Entre las opciones aplicables para estos tipos de variable se puede considerar:

- Recurrir a las redes neuronales con anomalías uniclase.
- Con anomalías multiclase las redes bayesianas ingenuas pueden ser aplicadas con eficiencia, aunque siguen siendo temas de estudio para mejorar los modelos obtenidos.
- Para anomalías uniclase más complejas como las series temporales es recomendable el uso de clasificadores basados en SVM al concentrarse en regiones aprendidas más que en variables individuales.
- En entornos limitados que no consideren muchas variables y que se disponga de expertos en el tema se puede considerar la clasificación basada en reglas que dictaminan lo que se considera normal y etiqueta automáticamente como anomalía aquello que no entra en las reglas establecidas.

2.4.2 Técnicas basadas en el vecino más cercano

En el área de detección de anomalías es común e incluso supuesto que los acontecimientos normales ocurren en vecindarios densos y los anómalos lejos de sus vecinos más cercanos, el detalle es determinar cuándo una instancia de datos es vecina de otra, ahora bien, se proponen dos formas de hacerlo, con distancia, normalmente euclidiana y otra con densidad relativa.

Cuando se opta por el uso de distancias normalmente se emplea la distancia euclídea para calcular la distancia o como otras veces es llamada "similitud", este enfoque es mayormente usado para datos continuos, aunque existen autores que proponen otros criterios de evaluación para aplicar este enfoque en otros tipos de datos. En cambio, si opta por el uso de la densidad relativa se toma en cuenta la densidad de un vecindario, si la concentración de vecinos respecto a una instancia de datos es baja entonces se considera una anomalía, claro que depende de la distancia que se proponga como límite de vecindario.

Se disponen de los métodos de *clustering,* muy usados para aprendizaje no supervisado y semi supervisado que puede ser a su vez enfocado en buscar anomalías en base a su pertenencia o no a grupos, o en la cercanía de su centroide o la combinación de ambos; si bien no tiene mejores resultados que otras técnicas es muy útil debido a su versatilidad a la hora de aplicarlo a datos no etiquetados.

2.4.3 Técnicas estadísticas

Estás técnicas se basan en establecer modelos estocásticos en los que se pueden o no parametrizar los datos del modelo. La base se centra en que cada instancia de datos tiene una probabilidad de ocurrir en el entorno estudiado y las anomalías son aquellas que tienen asociadas una probabilidad baja tras ser evaluadas con el modelo. Este tipo de técnicas suelen tener un buen desempeño en la detección de anomalías, pero como se trabaja bajo el supuesto de que todos los datos son generados por una distribución específica esto no siempre es cierto, sobre todo cuando se tienen datos bastante dispersos que es lo que ocurre en varios casos de estudio.

2.5 Técnicas de agrupamiento o *clustering*

En esta sección se explica en detalle el tipo de algoritmos que se utilizarán para analizar los datos de los autobuses, los cuales vendrán representados en forma de series temporales.

El objetivo del agrupamiento (***clustering***) de series temporales consiste en analizar tendencias de movilidad en las series con el fin de entender cómo evolucionan los datos y la razón de dicha evolución, descubrir patrones de comportamiento en las series, identificar series que muestran patrones similares, etc. Las aproximaciones basadas en agrupamiento también se utilizan para la detección de anomalías o detección de *outliers* (elementos atípicos en el conjunto). La utilización de técnicas de agrupamiento de datos para la detección de anomalías se basa en la idea de que los elementos atípicos no pertenecen a ningún conjunto (clúster) o forman su propio clúster.

Como se ha visto en la sección anterior, además de las técnicas de agrupamiento, existen también otras aproximaciones basadas en proximidad, árboles o reducción de la dimensionalidad para estudiar detección de anomalías. Algunas de estas técnicas, sin embargo, requieren la existencia de muestras etiquetadas como "anomalía" mientras que nuestro conjunto de datos no dispone de dicha información.

En esta sección presentamos dos de los algoritmos de *clustering* más conocidos y que, como se explicará más adelante, son dos técnicas que se ajustan bien a la tipología de los datos disponibles de los autobuses.

2.5.1 Algoritmo K-means

El algoritmo K-means es un algoritmo de clasificación no supervisada y es uno de los más utilizados. Consiste en separar o clasificar una colección que tiene *n* elementos en *k* grupos. Estos grupos son formados en base a un **centroide**, un centroide es un punto o conjunto de puntos que denotan el centro de cada clúster. Esta separación o asignación a los grupos se hace de forma iterativa para llegar a una separación óptima.

El algoritmo sigue una serie de pasos para lograr una clasificación en k elementos:

1. Se crea de forma aleatoria las coordenadas de los centroides de cada grupo.
2. Para cada elemento se calcula la distancia a cada centroide y aquel que esté a menor distancia será el clúster al que se le asigne. Esta distancia puede ser calculada con varios tipos de métricas: distancia euclídea, distancia euclídea al cuadrado, manhattan, entre otros, métrica debe seleccionarse en base a un análisis de la naturaleza de los datos a evaluar.
3. Una vez asignados los elementos a los grupos se reajustan las coordenadas de los centroides de modo que estos estén a la misma distancia de cada elemento al que fue asignado.
4. Los pasos 2 y 3 se repiten hasta que se cumpla uno de los siguientes criterios: Hasta que se supere un umbral de tolerancia previamente definido o hasta que se cumpla un número de iteraciones predefinido.

Al iterar *n* veces estos pasos se busca encontrar la agrupación óptima, esto es, elegir aquellos centroides que **minimicen la suma de distancias al cuadrado**.

En este algoritmo uno de los retos es poder elegir adecuadamente el valor de *k*, es decir elegir en cuantos grupos se desea clasificar lo elementos que se estudian. Existen varias técnicas para poder hacerlo, se mencionan las dos más relevantes:

- **El método del codo (*elbow*).** Es un método empírico para encontrar el número óptimo de clústeres para un conjunto de datos. En este método, se elige un rango de valores candidatos del número de separaciones posibles, posteriormente se aplica el algoritmo de K-means usando cada uno de los valores de k. En cada resultado se calcula la distancia intra-cluster e inter-cluster. Lo que se busca es encontrar un valor de *k* en el que se maximice la distancia inter-cluster y se minimice la distancia inter-cluster.

- **Método de la silueta (*silhouette*).** Este método mide qué tan cerca se encuentra un punto de sus puntos vecinos más cercanos, en todos los grupos. Proporciona información sobre la calidad del agrupamiento que se puede utilizar para determinar si se debe realizar un refinamiento adicional en el agrupamiento actual

Los métodos Elbow y Silhouette se utilizan para encontrar el número óptimo de grupos. Surge ambigüedad en el método del codo para elegir el valor de k. El análisis de silueta se puede utilizar para estudiar la distancia de separación entre los grupos resultantes y se puede considerar un mejor método en comparación con el método Elbow [Kumar21].

2.5.2 Algoritmo de *clustering* jerárquico

Es una técnica de aprendizaje no supervisado que, a diferencia del algoritmo K-means, no dispone de centroides sobre los cuales calcular las distancias, sino que las distancias se calculan entre cada uno de los elementos a clasificar, lo cual simplifica el proceso de agrupamiento. Este tipo de algoritmo es bastante recomendable cuando se tienen pocos datos.

Existen dos tipos de *clustering* jerárquico:

- El **aglomerativo** comienza asignando un clúster a cada elemento y fusiona aquellos más cercanos entre sí hasta llegar a un solo clúster.
- El **divisivo** trabaja de forma inversa, asumiendo un único clúster al inicio el cual se va dividiendo en otros más pequeños.

En la aplicación de este algoritmo es necesario definir la forma en que se calcularán las distancias entre los clústeres, a esta medida se la llama medida de vinculación, se pueden mencionar cuatro importantes:

- **Conexión completa:** La distancia se mide entre los dos puntos más lejanos de cada clúster
- **Conexión simple:** Es la opuesta a la conexión completa. Toma la distancia mínima entre dos puntos de cada clúster.
- **Distancia entre medias:** La distancia entre dos clústeres se calcula como la distancia entre las medias de cada uno.
- **La distancia promedio entre pares:** Es el promedio entre todas las distancias que podemos obtener entre todos los pares de puntos.

Por otra parte, también es necesario definir qué medidas de distancia se implementarán para aplicarlas en alguna de las métricas anteriormente mencionadas. Las métricas más comunes son las siguientes:

- **Distancia Euclidiana**. Es la distancia ordinaria en línea recta entre dos puntos en el espacio euclidiano. Una variación de esta distancia es la distancia euclídea cuadrada.
- **Distancia Manhattan**. Similar a la euclidiana, pero la distancia se calcula sumando el valor absoluto de la diferencia entre las dimensiones.

- **Distancia del coseno**. Esta métrica toma en cuenta la forma de las variables, más que sus valores, esto tiende a reducir el ruido. Tiende a asociar observaciones que tienen las mismas variables máxima y mínima, independientemente de su valor efectivo.

Los resultados del algoritmo se representan gráficamente mediante un dendrograma. Un dendrograma suele tener este aspecto:

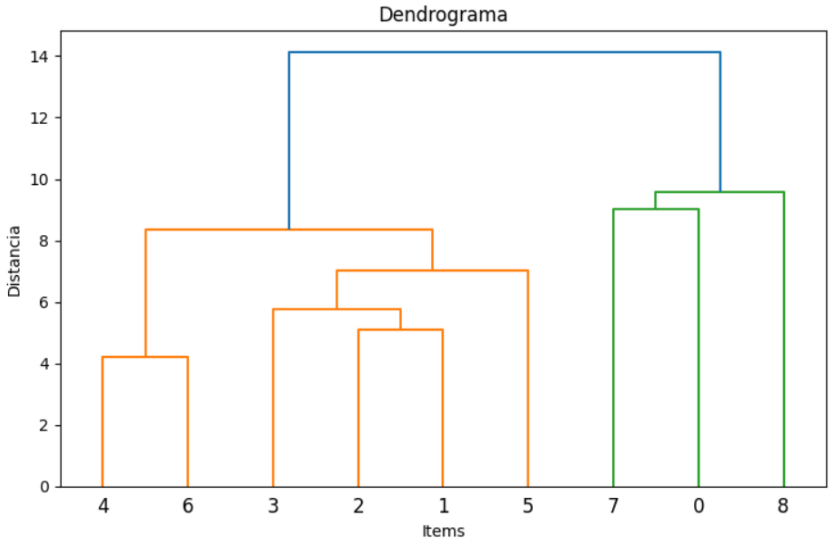

Figura 2.1 Ejemplo de dendrograma

La Figura 2.1 muestra un ejemplo de dendrograma donde el eje X representa las observaciones u objetos del estudio, y el eje Y representa la medida de distancia que se haya empleado en el algoritmo de *clustering*. Las líneas horizontales representan fusión de clústeres. En la interpretación de un dendrograma se debe analizar la altura de las líneas verticales de modo que cuanto menor sea la altura, más similares son los objetos que aparecen agrupados. En la Figura 2.1, los elementos 4 y 6 son los más similares entre sí ya que la altura de la línea que une estos dos objetos es la menor de todas (una altura ligeramente superior a 4). Los siguientes elementos más similares entre sí son los objetos 1 y 2; a continuación, se formaría el grupo de los elementos 1-2 con elemento 3, y así sucesivamente. La Figura 2.1 muestra una partición de dos clústeres, diferenciados por el color verde y naranja, que tiene lugar a una distancia próxima a 10 (número de líneas verticales a partir del corte horizontal a la distancia especificada). Si se cortara el dendrograma más abajo, por ejemplo, a una distancia 7, la distancia sería menor pero el número de clústeres o agrupamientos sería mayor. En este caso se obtendrían 6 clústeres, uno con los elementos 4 y 6; otro clúster con los elementos 3, 2 y 1; y luego cuatro clústeres que contendrían cada uno un único elemento, con los datos 5, 7, 0 y 8, y cada uno de estos clústeres aparecería de un color diferente.

Una forma de calcular el número de clústeres es elegir un punto de corte en el eje Y, a ese punto de corte se le traza una línea horizontal paralela al eje X, y se cuenta el número de líneas verticales que corta dicha línea horizontal. A ese punto de corte se lo denomina *threshold*. Por ejemplo: en la figura 5.1 si se elige un punto de corte en el punto 10 del eje Y, y se traza una línea horizontal al eje Y en ese punto, la línea cortaría 2 líneas verticales, por lo tanto, se tendría dos clústeres. Esta línea horizontal puede escogerse manualmente para analizar el

dendrograma, pero es más recomendable usar otras métricas como la silueta o el método del codo para seleccionar el número óptimo de clústeres.

Entre las desventajas del algoritmo de *clustering* jerárquico se puede mencionar:

- No es recomendable su uso en conjuntos de datos muy grandes ya que tiene un costo computacional alto.
- Es sensible al ruido o a *outliers*

3 Procesamiento de los datos

Este capítulo se centra en explicar los mecanismos de recopilación de los datos relativos a cada autobús a través del protocolo **bus CAN**, y la organización y ordenación de dichos datos en estructuras adecuadas para construir la información disponible sobre el funcionamiento y logística de la flota de autobuses. Se realizará un análisis detallado de las señales recopiladas a través del protocolo bus CAN para determinar posteriormente las variables relevantes para el diagnóstico.

3.1 El sistema bus CAN

El **bus CAN** (*Controller Area Network*) es un protocolo de comunicación de mensajes desarrollado por Bosch que permite que los microcontroladores y componentes electrónicos que hay en un vehículo puedan comunicarse entre sí sin requerir de una computadora para ello (utilizaremos el término BUS-CAN de aquí en adelante para referirnos a dicho protocolo). Los vehículos modernos disponen de múltiples subsistemas electrónicos cada uno de los cuales tiene su propia unidad de control electrónico. A partir del sistema BUS-CAN, un vehículo es capaz de administrar y monitorizar diferentes funciones: sistema de frenado, ventanillas eléctricas, control de crucero, GPS, consumo de combustible, etc.

Debido a la naturaleza del BUS-CAN, la transmisión de información es considerable ya que cada nodo (componente del BUS-CAN que consta de una CPU, un controlador y un transceptor) puede enviar y recibir información en el autobús de datos [Smith21].

Los vehículos de la Empresa Municipal de Transportes de Valencia (EMT) utiliza el BUS-CAN según el FMS (Fleet Management System) basado en el protocolo SAE (Sociedad de Ingenieros Automotrices) J1939. El protocolo J1939 es una norma de la SAE para el envío de datos por un BUS-CAN en sistemas de automoción y es utilizado por muchas marcas como Hyundai, Volvo, Renault, Scania, etc. A través del protocolo J1939 se recopila señales del funcionamiento de los autobuses a una frecuencia determinada de registro. El FMS para autobuses es una interfaz común basada en el estándar FMS para camiones. La información disponible a través de la interfaz Bus-FMS-Standard depende del fabricante.

En este trabajo se ha utilizado el documento FMS-Standard Interface Document vers.04 (Bus and Truck) (13/10/2017) que indica las señales (obligatorias u opcionales) que deben poner los fabricantes a disposición de los usuarios [HDEI17]. En el resto de las secciones de este capítulo se definirán las señales a evaluar que, a priori, se consideran más relevantes y se establecerán las frecuencias de adquisición. El objetivo es registrar estas señales durante un período de tiempo y posteriormente analizarlas para establecer un grupo de señales y una frecuencia de registro que permitan obtener una información útil para el mantenimiento predictivo de la flota, el seguimiento del estado del vehículo y la potencial detección de anomalías.

3.2 Estructura de la información

Este apartado tiene por finalidad describir los componentes principales del flujo de información de la EMT de Valencia, así como la comunicación entre los mismos. Primeramente, describimos las unidades principales que conforman la estructura de los datos y, posteriormente, se detallará las señales capturadas de los autobuses a través del sistema BUS-CAN.

3.2.1 Unidades principales de información de la red de autobuses

Las unidades principales de información son aquellos componentes sobre los que se asientan las bases del flujo de información de la monitorización de autobuses, se identifican cuatro

Autobús: es la unidad principal del análisis ya que el diagnóstico se realiza a nivel de autobús. Se define como autobús aquella unidad sobre las cual se tienen datos diarios (excepto los días que el vehículo no circula) y a partir de la cual se hace análisis de datos a mayor escala. La forma en la que se identifica un autobús en el entorno de análisis es por un código único e irrepetible asignado en el momento de su ingreso en el inventario de la flota de autobuses.

Línea de autobús: una línea hace referencia a una ruta diseñada por la EMT que consta de un número de paradas que se ubican en puntos estratégicos en la ciudad de Valencia. Una línea consta de los siguientes datos para poder identificar la ruta que sigue:

1. Un código de identificación único. Por ejemplo: las líneas 10, 70 o 90, que suelen pasar por el centro de la ciudad de Valencia.
2. Un sub-identificador para diferenciar si se trata de una línea en sentido de ida o de vuelta. Por ejemplo, de las líneas mencionadas anteriormente se tienen los sub-identificadores 10-1, 10-2, 70-1, 70-2, 90-1 y 90-2 respectivamente.
3. Unos identificadores de las paradas de origen y destino para ambos sentidos de la línea, así como lo identificadores de todas las paradas intermedias junto con el orden en que deben ser visitadas.

LÍNEA 10					
Trayecto de vuelta (10-1)			Trayecto de ida (10-2)		
Orden	Parada	Nombre de parada	Orden	Parada	Nombre de parada
0	1440	Dr. Tomás Sala - Sant Marceli	0	1240	Lliri Blau - Mestre Ventura Pascual
1	976	Carters (impar) - Primer de Maig	1	965	Germans Villalonga - Enric Navarro
2	979	Carters - Mossén Febrer	2	966	Germans Villalonga - Dr. Vicent Zaragozá
3	1859	Carters - Pl. Santiago Suárez	3	967	Ramón Asensio - Guardia Civil
4	1860	Uruguai - Marques de Bellet	4	188	Primat Reig - Campus Universitari
5	1547	Uruguai - Crta. Escrivá	5	1239	Primat Reig - Xabia
6	1783	Uruguai (impar) - Jeroni Munyos	6	2213	Aragó (impar) - Blasco Ibáñez

7	742	Albacete - Mestre Sosa	7	1083	Aragó - Amadeu de Savoia
8	738	Sant Vicent Mártir - Dr. Gil i Morte	8	1035	Aragó - Finlandia
9	815	Espanya - Julio Antonio	9	219	Aragó - Passeig Albereda
10	2210	Sant Vicent Martir - Ánimes	10	764	América - Pont d'Aragó
11	2314	Periodista Azzati - Pl. de l'Ajuntament	11	2251	Porta de la Mar (impar) - Navarro Reverter
12	2280	Barques - Pl. de l'Ajuntament	12	2258	Porta de la Mar - Colón
13	2281	Pintor Sorolla - A. MagnÃ nim	13	2259	Colón - els Pinazo
14	2250	Porta de la Mar (par) - Navarro Reverter	14	2260	Colón - Pascual i Genís
15	2206	Pl. América - Navarro Reverter	15	2277	Xátiva - Marqués de Sotelo
16	1054	Aragó - Saragossa	16	814	Jesús - Pare Jofré
17	1085	Aragó - Xile	17	739	Jesús - Sant Francesc de Borja
18	1055	Aragó - Ernest Ferrer	18	741	Jesés - Pintor Segrelles
19	1057	Aragó - Blasco Ibañez	19	1784	Carcaixent - Giorgeta
20	157	Blasco Ibañez - Gascó Oliag	20	1785	Uruguai (par) - Jeroni Munyós
21	192	Primat Reig - Camí Vera	21	972	Uruguai - Misser Rabassa
22	1032	Tirig - Guardia Civil	22	1036	Uruguai - Venezuela
23	985	Jaume Esteve Cubells - Dr. Vicent Zaragozá	23	974	Carters - Llanera de Ranes
24	987	Mistral - Murta	24	975	Carters - Calvo Acacio
25	1240	Lliri Blau - Mestre Ventura Pascual	25	978	Carters (par) - Primer de Maig
			26	1443	Dr. Tomás Sala - Carters
			27	1444	Dr. Tomás Sala - Jerónima Galés
			28	563	Gaspar Aguilar -Crematorio
			29	564	Sant Doménec de Guzman – Cementeri
			30	1474	Camí vell Picassent - Acc. Cementeri
			31	1475	Camí vell Picassent - Tanatori Municipal
			32	1472	Pius IX - Salvador Perles
			33	1473	Pius IX - Cabanilles
			34	1440	Dr. Tomás Sala - Sant Marcelí

Tabla 3.1 Ejemplo de paradas de la línea 10

En los lineamientos de la EMT se espera que para cada línea exista un conjunto de autobuses predominantes los cuales se esperan que realicen una línea casi siempre, y otro conjunto de autobuses (reservas) que actúan como apoyo a los anteriores, es decir, llenan los huecos de

servicio que podrían presentarse en caso de desperfecto u otras circunstancias de un autobús designado como predominante en la línea, es por esta razón que un autobús puede realizar el recorrido de más de una línea en un mismo día.

Parada: se entiende por parada a aquel sitio en el que se detiene un autobús para dejar o recoger pasajeros y dispone de una ubicación geográfica estática. Cualquier parada en la ciudad de Valencia dispone de una marquesina o un poste en el que se puede identificar el nombre de la parada, el código de parada y las líneas que pasan por esa parada (ver Figura 3.1). En el Figura se puede ver la marquesina de la parada código 1035 que según la Tabla 3.1 forma parte de la ruta vuelta de la línea 10 (trayecto 10-2) y como se puede apreciar indica que la línea 10 se detendrá en esa parada.

Figura 3.1 Ejemplo de marquesina de parada. Fuente: EMT de Valencia

Trayecto: el sub-identificador del sentido de ida o vuelta de una línea que se ha mencionado anteriormente, se denomina trayecto. Así, por ejemplo, el sub-identificador 10-1 indica el trayecto de ida de la línea 10 y el sub-identificador 10-2 indica el trayecto de vuelta de la misma línea. Ambos trayectos son diferentes ya que a pesar de empezar y terminar en las mismas paradas no necesariamente comparten mismas paradas intermedias, ni son el mismo número de paradas. En la Tabla 1 se muestra el detalle de la ruta 10. La línea 10 tiene un recorrido que inicia (en el trayecto de ida 10-1) en Dr. Tomás Sala - Sant Marceli (parada 1440), y finaliza en Lliri Blau (parada 1240) entre ambos extremos la línea tiene 23 paradas. En el caso inverso, el trayecto de vuelta 10-2 empieza en la parada 1240 y termina en la parada 1440, sin embargo, tiene 32 paradas intermedias, esto sucede en otras líneas ya que muchas calles en la ciudad de Valencia no son de doble sentido, por lo que al realizar un trayecto inverso se debe circular por otras calles y por lo tanto detenerse en otras paradas.

Visto desde un mapa puede apreciarse el recorrido de la línea 10 en el trayecto 10-2 (ver Figura 3.2).

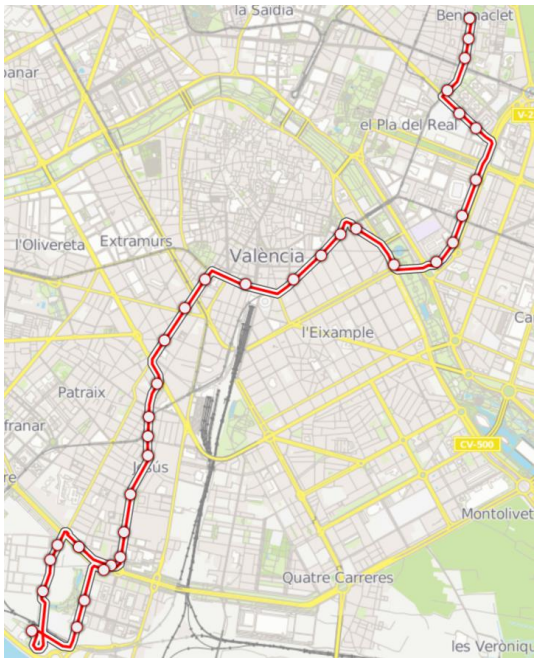

Figura 3.2 Vista en el mapa del trayecto 10-2. Fuente: EMT de Valencia

3.2.2 Señales disponibles

Desde el momento en que un autobús inicia su funcionamiento en la línea y trayecto asignado, el BUS-CAN captura señales. Dichas señales son capturadas cada 10 segundos y cada señal dispone de la siguiente información (ver Tabla 3.2).

Nombre campo	**Descripción**
CATEGORIA	Identificador del tipo de señal
BUS CÓDIGO	Código del autobús al cual pertenece la señal
POSICION X	Valor del eje X en la posición global
POSICION Y	Valor del eje Y en la posición global
VALOR	Valor de la señal

Tabla 3.2 Descripción de la información de una señal capturada por el BUS-CAN

El dato CATEGORÍA es un número que permite identificar a qué tipo de señal hace referencia el registro capturado por el BUS-CAN. EMT Valencia proporciona 15 señales que en realidad son pocas comparadas con las posibles (alrededor de 40) pero debido a la Configuración actual del sistema BUS-CAN estás se redujeron a lo que el sistema INDRA pueda podía proporcionar, por ello este trabajo además se orienta a valorar la viabilidad de estudio con este número limitado de señales (ver tabla 3.3).

Código	Descripción	Detalle	Unidades	Nombre unidades	Tipo
1	Combustible en el depósito	Nivel de combustible	%	Porcentaje	Entero

2	Consumo de combustible total	Combustible total usado del motor	l	Litros	Entero
3	Kilometraje del vehículo	Distancia total del vehículo de alta resolución	km	Kilómetros	Entero
4	Velocidad del vehículo	Velocidad del vehículo del tacógrafo	km/h	Kilómetros por hora	Decimal
5	Revoluciones de motor	Velocidad del motor	rpm	Revoluciones por minuto	Entero
6	Posición del acelerador	Posición 1 del pedal del acelerador	%	Porcentaje	Entero
7	Marcha del vehículo	Marcha seleccionada	Número	Número de marcha	Entero
8	Presión del circuito de frenado 1	Circuito de presión de aire del freno de servicio n. ° 1	kPa	Kilo pascales	Entero
9	Presión del circuito de frenado 2	Circuito de presión de aire del freno de servicio n. ° 2	kPa	Kilo pascales	Entero
10	Frenos accionados	Posición del pedal freno	0/1	no/si	Booleano
11	Estado del freno de estacionamiento	Interruptor del freno de estacionamiento	0/1	no/si	Booleano
12	Posición del elevador de rampas	Estado de la rampa	0/1	no/si	Booleano
13	Puertas abiertas	Estado de las puertas	0/1	no/si	Booleano
14	Consumo puntual de combustible (l/h)	Tasa de combustible	l/h	Litros por hora	Decimal
15	Consumo puntual de combustible (km/l)	Economía de combustible instantánea	km/l	Kilómetros por litro	Decimal

Tabla 3.3 Detalle de los tipos de señales disponibles

Entre estas 15 señales pueden identificarse 4 grupos de datos que brindan información sobre una misma tipología de señal:

- Las señales 1, 2, 14 y 15 brindan información sobre el combustible.

- Las señales 3, 4, 5, 6, y 7 están relacionadas con la velocidad del vehículo y el kilometraje.
- Las señales 8, 9, 10 y 11 están relacionadas con el sistema de frenado.

Las señales 12 y 13 no se consideran parte de algún grupo por que se considera que no son relevantes para el estudio de anomalías. Todas las señales binarias no cumplen con la regla de una señal cada 10 segundos, sino que se capturan en el momento en que el componente correspondiente se activa o desactiva.

No todos los autobuses tienen registros de las 15 señales todos los días, por diversos factores pueden no tenerse registros de algunas señales. Esto se debe usualmente a fallos en la lectura de los sensores o que el BUS-CAN no se ha encendido correctamente.

3.3 Recopilación de datos

Una vez se ha visto la información de la que dispone la EMT, en esta sección se describe la organización de los datos que recoge diariamente la EMT a través del sistema BUS-CAN. Además, se describirá el proceso de limpieza de los datos recogidos.

Las unidades de información mencionadas previamente en la sección 3.2 convergen en un componente fundamental en la monitorización de un autobús, el concepto de **viaje**.

Viaje: es el recorrido de un trayecto que realiza un autobús en un día específico durante un cierto horario previamente asignado por el área de logística de la EMT de Valencia. Un viaje aúna todas las unidades previamente mencionadas para detallar las actividades de un autobús a lo largo de su vida útil. Cada viaje tiene los datos necesarios para poder identificar el ciclo de funcionamiento de un autobús (ver Tabla 3.4).

Nombre campo	**Descripción**
HORA SALIDA	Hora en que el un autobús inició una ruta
CODIGO LINEA	Identificador del código de línea (por ejemplo: línea 10)
BUS CÓDIGO	Código del autobús
TRAYECTO	Identificador de la línea que identifica si el viaje fue en la ruta de ida o de vuelta
PARADA ORIGEN	Código de la parada que marca el inicio de la ruta
PARADA DESTINO	Código de la parada que marca el fin de la ruta

Tabla 3.4 Detalle de la información de un viaje

Retomando el ejemplo de la línea 10, en la Tabla 3.5 se muestra los viajes realizados por un autobús en una fecha y horarios específicos. En este caso se trata del autobús 6470 que realizó la ruta 10 en fecha 25 de mayo de 2021.

Viajes del autobús 6470 en fecha 25-05-2021			
HORA SALIDA	**TRAYECTO**	**PARADA ORIGEN**	**PARADA DESTINO**
08:13	010-2	1440	1240
09:02	010-1	1240	1440
10:00	010-2	1440	1240
10:46	010-1	1240	1440
11:42	010-2	1440	1240
12:31	010-1	1240	1440
13:29	010-2	1440	1240
14:18	010-1	1240	1440
15:13	010-2	1440	1240
16:01	010-1	1240	1440
16:59	010-2	1440	1240
17:46	010-1	1240	1440
18:41	010-2	1440	1240
19:30	010-1	1240	1440
20:27	010-2	1440	1240
21:16	010-1	1240	1440
22:02	010-2	1440	1240

Tabla 3.5 Viajes hechos por el autobús 6470 en fecha 25-05-2021

En la Tabla 5 puede verse que el autobús 6470 inició su actividad a las 08:13 de la mañana del 25 de mayo en el trayecto de vuelta de la ruta 10 (trayecto 010-2). A este respecto, cabe destacar que no es imperativo que un autobús inicie en el trayecto de ida de una línea. Asimismo, el último viaje realizado por el autobús comenzó a las 22:02, recorriendo el trayecto 010-2. De este modo, el día 25 de mayo el autobús 6470 recorrió la ruta de la línea 10 varias veces, 8 veces en el trayecto de ida y 9 veces en el trayecto de vuelta.

Una vez identificados los viajes y los periodos de tiempo entre ellos es posible obtener las señales capturadas por el BUS-CAN durante cada viaje. A partir de la información disponible sobre las señales se tienen algunos datos para poder hacer un cruce de información, sobre todo para obtener las señales obtenidas en ese periodo de tiempo (ver Tabla 3.6).

Nombre campo	**Descripción**
NRO MUESTRA	Identificador único de la señal registrada
CATEGORIA	Identificador del tipo de señal
BUS CÓDIGO	Código del autobús al cual pertenece la señal
FECHA Y HORA	Día, mes, año, hora, minuto y segundo de registro de la señal
POSICION X	Valor del eje X en la posición global
POSICION Y	Valor del eje Y en la posición global
VALOR	Valor de la señal

Tabla 3.6 Detalle de la información de una señal

Con estos datos si se desea saber qué señales se capturaron en el primer viaje del autobús 6470 entonces se deben analizar todas las señales capturadas en el periodo de tiempo comprendido entre 08:13 y 09:02; es decir, se tienen en cuenta todos los registros menores al inicio del siguiente viaje. Estos registros son identificables a través del campo FECHA Y HORA. El resumen de las señales registradas durante el primer viaje realizado por el autobús 6470 el día 25 de mayo se muestra en la Tabla 3.7.

Dato	**Descripción**
Número total de registros	2479
Señales faltantes	7, 11, 12
Señales con fallo	2, marcador en cero
Hora último registro	08:50
Tiempo sin señales hasta siguiente tramo	12 minutos

Tabla 3.7 Resumen de información autobús 6470

- Entre el periodo de tiempo desde el inicio del trayecto hasta antes del inicio del siguiente se tienen 2479 registradas globalmente sin distinguir la categoría.
- De las 15 señales no se tiene registro de las señales 7, 11 y 12.
- Los registros de la señal 2 marcan cero lo cual indica fallo de sensor.
- El último registro antes del inicio del siguiente tramo fue registrado a las 08:50, lo que deja unos 12 minutos sin registros, esto se debe a que un tramo no inicia justamente después de otro, hay un pequeño periodo de espera entre tramos antes de que el conductor active el sistema de recojo de señales.

3.3.1 Organización de los datos

Los datos que se registran diariamente la EMT de Valencia, desde el enfoque de volumen de información y la organización de los archivos reúne las siguientes características:

1. En la flota de la EMT de Valencia no todos los autobuses disponen de BUS-CAN por la antigüedad de varios de ellos, debido a esto hay un recorte de información por lo cual el diagnóstico solo puede hacerse en 214 autobuses que sí disponen del BUS-CAN.
2. El formato en que los datos obtenidos por cada bus es en ficheros CVS y vienen presentados de la siguiente forma: un fichero con los autobuses de la flota, un fichero de las líneas, y por cada día se generan dos ficheros, el que registra las señales de cada bus y el que registra los horarios de cada viaje que realizó cada bus en un día.
3. En promedio se puede tener información de hasta 180 autobuses por día.
4. En promedio un bus puede realizar 15 viajes cada día y no necesariamente en la misma línea.
5. Las señales que no son de tipo binario en el fichero de señales se registran cada 10 segundos y son registradas en un mismo fichero.
6. Al tener un solo fichero de señales de todos los buses por cada día, el volumen de ese fichero es considerable, alrededor de 350 Mb.
7. La duración de los trayectos de los autobuses dependerá de la longitud de la línea y de la hora en la que el autobús realiza el trayecto, ya que existen franjas horarias con un

mayor volumen de tráfico. Con todo ello, el volumen de señales que se recoge diariamente se mueve aproximadamente alrededor de los 7 millones de registros.

A partir de estas características se tienen en cuenta las siguientes consideraciones para poder analizar los datos de forma ordenada y, sobre todo, poder realizar el diagnóstico de cada autobús:

1. Se depura el archivo de los autobuses para que solo queden aquellos que entrarán en el análisis, es decir, los que disponen de BUS-CAN.
2. El archivo diario de señales se particiona para cada autobús, obteniendo un fichero independiente para cada uno de ellos.
3. A su vez, se divide el archivo de datos de un autobús en un fichero independiente para cada viaje diario realizado por el autobús.
4. Con los viajes identificados se puede realizar una segunda partición a los ficheros de señales en un fichero por cada viaje.

Para este proceso se ha utilizado la herramienta Pandas de Python por la facilidad para manejar este tipo de ficheros en Dataframes. Como nota adicional, comentar que no se recurre al uso de base de datos para tal cantidad de datos ya que, a pesar de poder facilitar las cosas, el problema se centra en migrar cada día los 7 millones de datos que demora alrededor de 7 horas, lo cual no es factible de aplicar en este caso.

3.4 Análisis de las señales

En esta sección se explica el comportamiento de las 15 señales disponibles, para ello se toma de referencia el autobús y las lecturas de las señales a lo largo de un día

Señal 1 (combustible en el depósito). Esta señal tiene un comportamiento descendente por indicar el combustible disponible en depósito, al ser una variable de tipo porcentaje los valores posibles están entre los límites de 100 y 0 (Ver Figura 3.3).

A pesar de ser descendente se aprecia irregularidades en las lecturas, esto se debe a la naturaleza del sensor que es de tipo flotador y por lo tanto no es preciso al representar solo porcentajes en valores enteros de nivel de combustible.

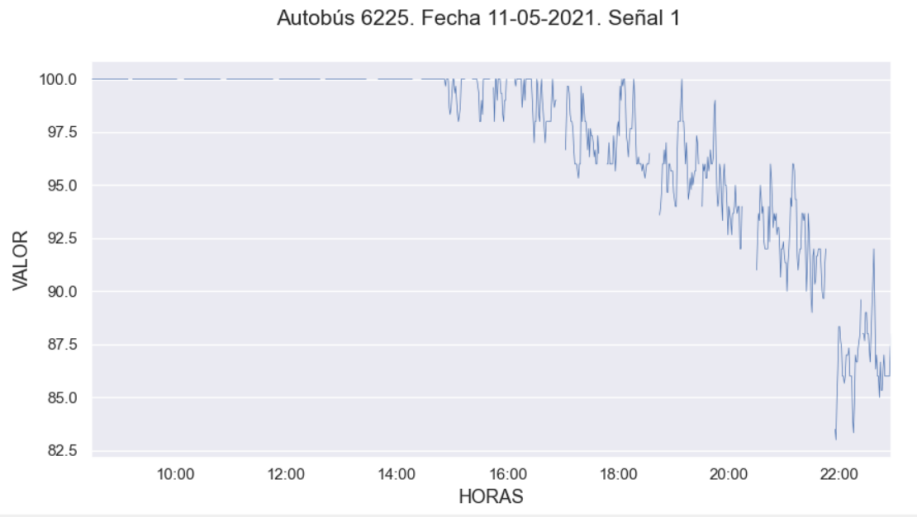

Figura 3.3 Evolución de la señal 1 para el autobús 6225

Señal 2 (consumo de combustible total). Esta señal tiene una naturaleza histórica ya que marca los litros consumidos por el autobús durante toda su vida útil, por lo tanto, tiene un comportamiento relativamente lineal ascendente (ver Figura 3.4).

Para el análisis mensual, diario, por trayecto, entre otros, es necesario obtener la diferencia entre la primera y última lectura del intervalo de tiempo que se desea analizar para saber cuántos litros se consumieron.

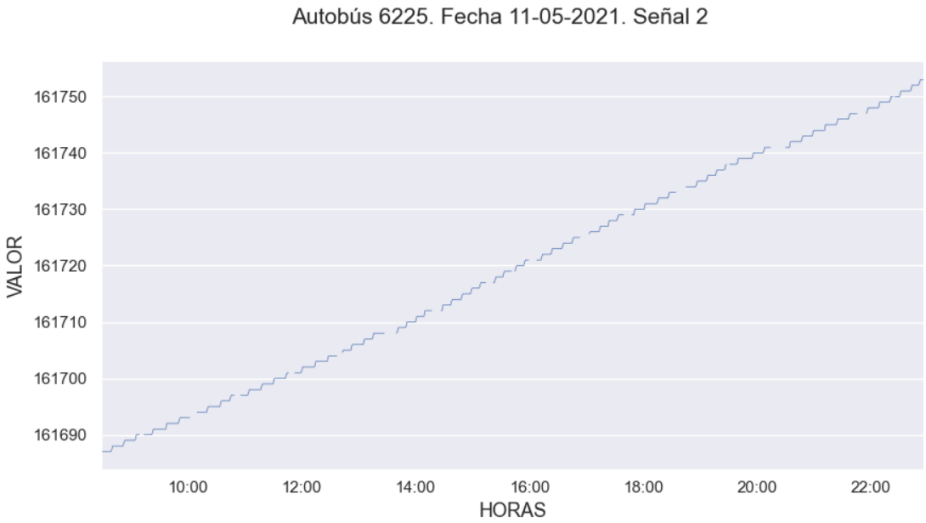

Figura 3.4 Evolución de la señal 2 para el autobús 6225

Señal 3 (kilometraje). Esta señal también tiene una naturaleza histórica y se aplica el mismo comportamiento y análisis que la señal 2 (ver Figura 3.5).

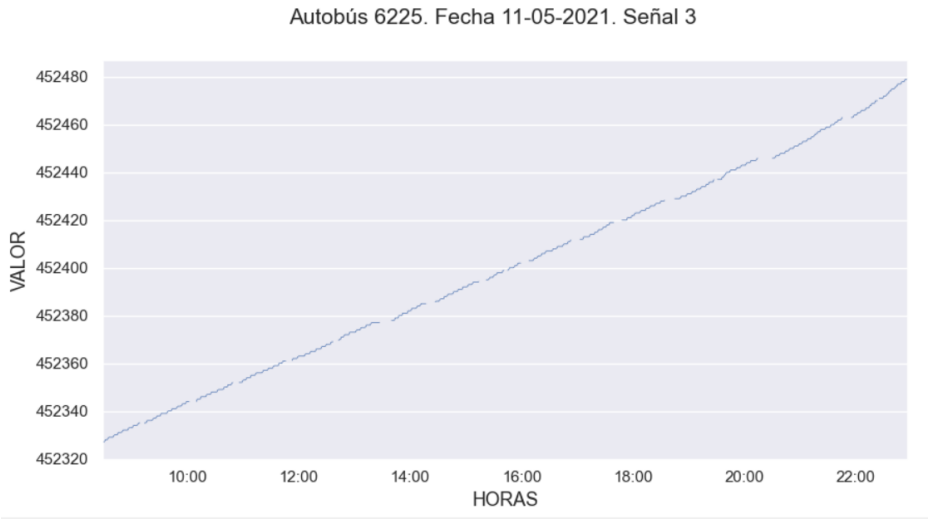

Figura 3.5 Evolución de la señal 3 para el autobús 6225

Señal 4 (velocidad). Es una señal bastante variable incluso en periodos más pequeños de tiempo ya que un autobús va registrando diferentes velocidades en base al tráfico vehicular de ese momento, aunque se puede llegar a establecer límites aceptables de velocidad máxima de circulación (ver Figura 3.6).

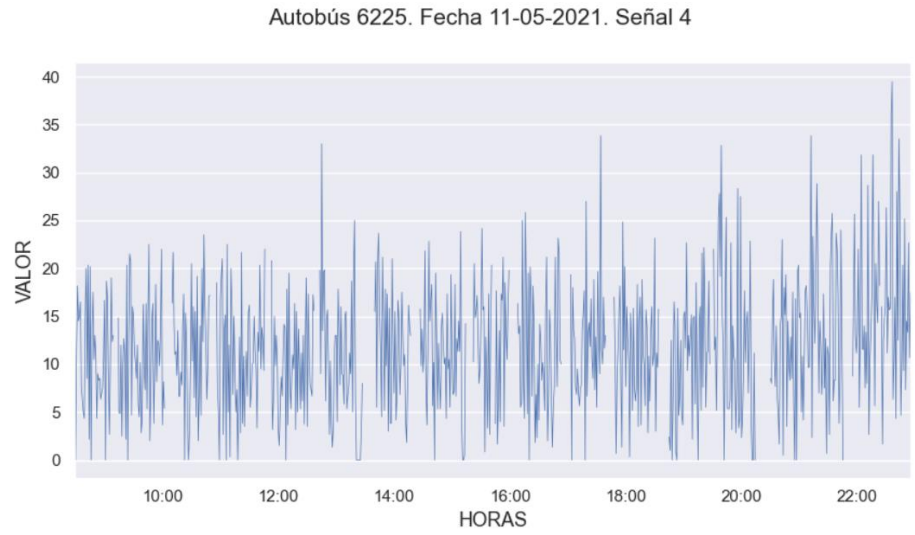

Figura 3.6 Evolución de la señal 4 para el autobús 6225

Señal 5 (revoluciones). Esta señal comparte características similares a la señal 4 al marcar las revoluciones del motor lo cual también es bastante variable, ambas pueden usarse de forma cruzada para validarse entre sí (ver Figura 3.7).

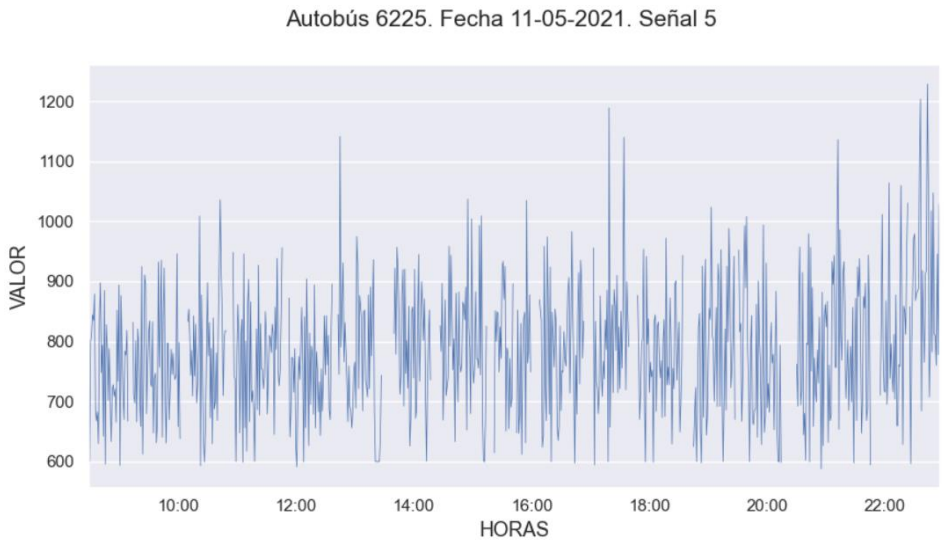

Figura 3.7 Evolución de la señal 5 para el autobús 6225

Señal 6 (posición del acelerador). La posición del acelerador presenta las mismas limitaciones que la señal 1 al solo poder marcar un porcentaje, en este caso presenta un comportamiento bastante variable a lo largo del tiempo (ver Figura 3.8).

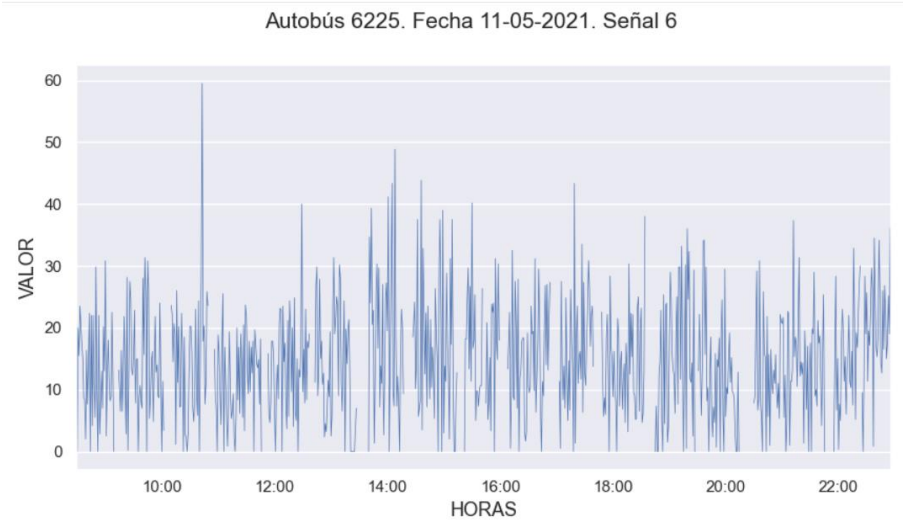

Figura 3.8 Evolución de la señal 6 para el autobús 6225

Señal 7 (marcha). La posición de la marcha puede usarse para cruzar información con la velocidad del autobús, se puede apreciar que el bus del ejemplo no sobrepasa la marcha 4 lo cual es entendible por el tamaño del vehículo (ver Figura 3.9).

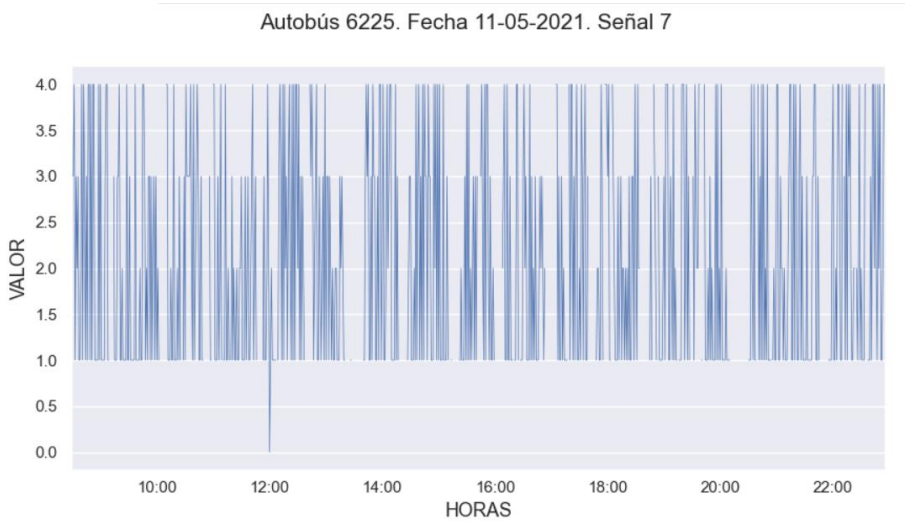

Figura 3.9 Evolución de la señal 7 para el autobús 6225

Señales 8 y 9 (circuito de frenado). Como ambas señales dan razón de los componentes de frenado comparten las mismas características, describen la presión en el circuito de frenado y son igual de variantes que las señales anteriores (ver Figura 3.10).

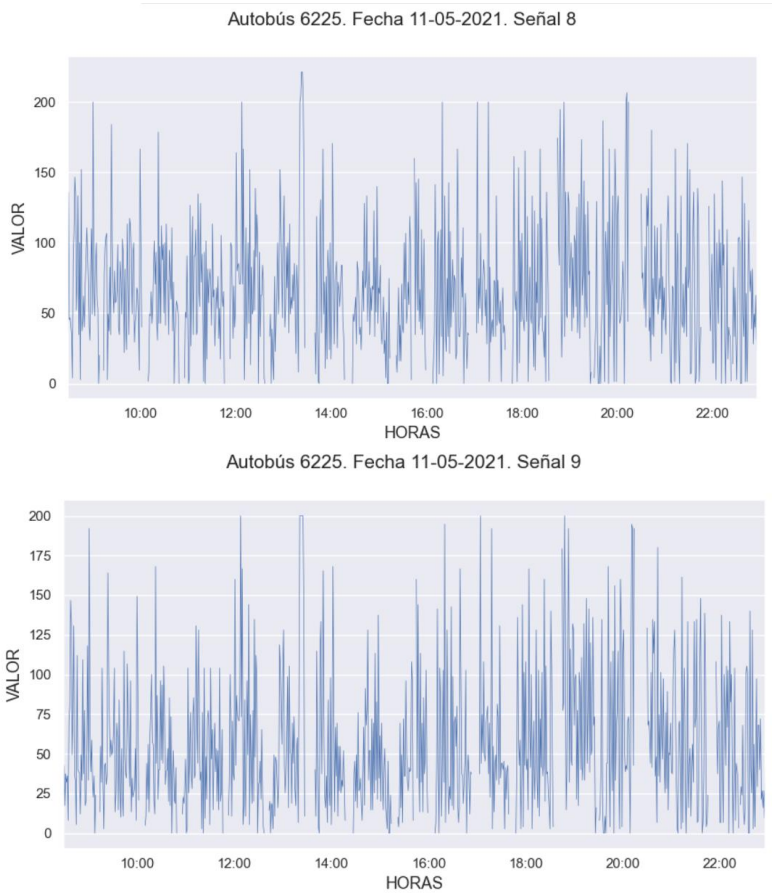

Figura 3.10 Evolución de la señal 8 y 9 para el autobús 6225

Señales 10, 11, 12 (frenos y rampas). Son de tipo binario y no presentan una regularidad de funcionamiento el cual puede ser demasiado variable, las puertas del vehículo se abren y

cierran únicamente en cada parada, y el mecanismo de la rampa puede o no accionarse en un día, por lo que estás señales no serán tomadas en cuenta para la detección de anomalías.

Señal 13 (puertas). La señal de apertura y cierre de puertas del autobús es un buen indicador del lugar dónde es muy probable que se encuentre una parada ya que es donde se recogen y dejan pasajeros.

Señales 14 y 15 (consumo puntual de combustible). Las señales de litros por hora y km por litro no son capturadas por sensores, sino que son calculadas en base a otras señales como el consumo de combustible y el kilometraje, por lo que serán descartadas en el análisis.

3.4.1 Análisis espacial

El análisis espacial se centra en analizar a cada autobús según las rutas y trayectos que realice, lo cual hace que la comparación de datos entre autobuses sea uniforme porque, al tener una ruta fija demarcada por líneas y trayectos asignados en el día a día, ya se considera un comportamiento regular, por lo tanto, se plantea que es posible establecer patrones de comportamiento y obtener información cruzada al considerar que las asignaciones de líneas rotan por cada autobús.

Este análisis toma como base las coordenadas registradas por cada señal disponible y la información geográfica sobre las paradas de cada línea y trayecto. Dada la frecuencia de recogida de señales es posible determinar el recorrido de un autobús y su nivel de fidelidad al recorrido oficial determinado por la EMT así también se puede analizar su comportamiento a lo largo de un trayecto con las señales recogidas durante un viaje.

Por ejemplo, se tiene el viaje de la línea 70 trayecto 70-2 hecho por 3 autobuses diferentes en un mismo día, en este caso 24-05-2021 (ver Figura 3.11).

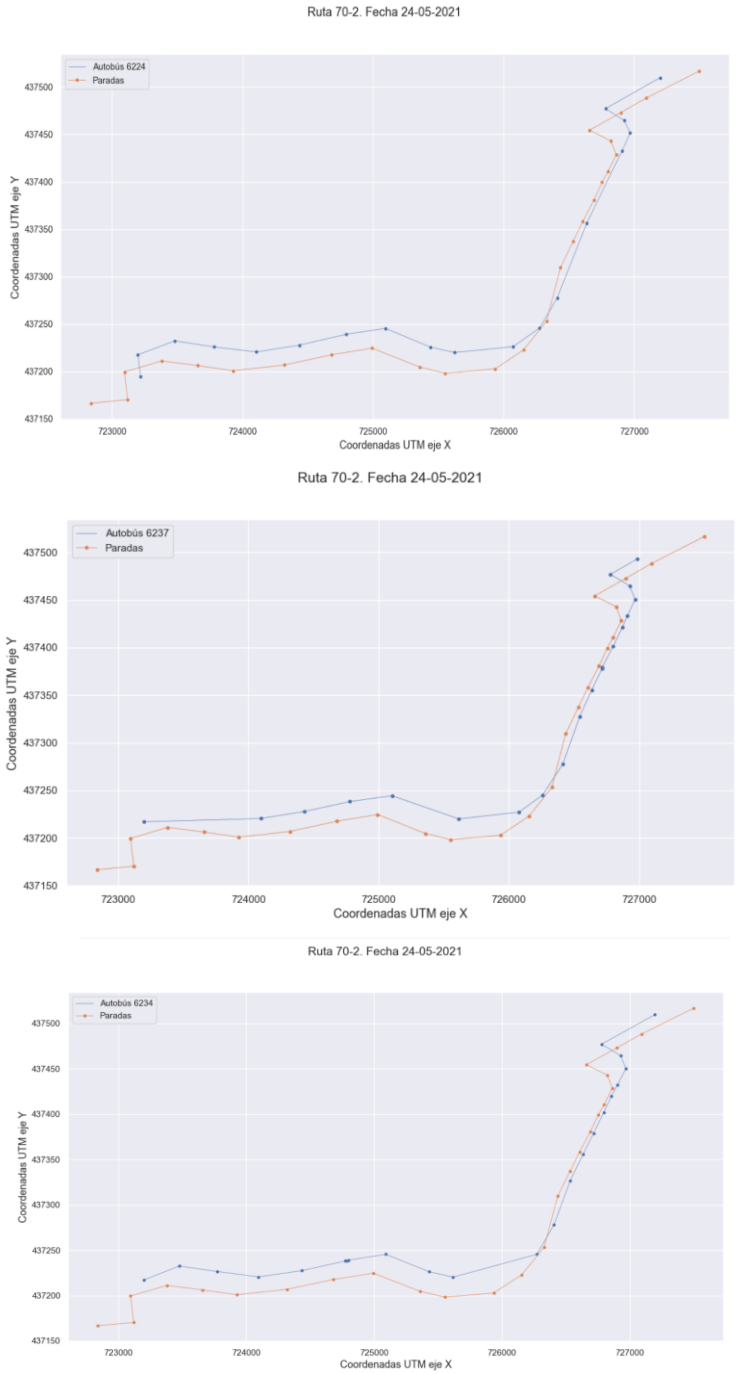

Figura 3.11 Comparación línea 70, trayecto 70-2

La línea roja en los tres planos detalla las posiciones de las paradas del trayecto 70-2 y las otras representan el viaje de los autobuses marcando los puntos en los que se abrieron y cerraron las puertas de acceso al autobús. El indicador se perta abierta o cerrada es dado por la señal 13 donde 1 es abierto y 0 es cerrado.

Las figuras muestran que el patrón del trayecto de los tres autobuses es muy similar al patrón del trayecto diseñado por la EMT, con ligeras diferencias de la ubicación geográfica de las paradas del trayecto. Asimismo, se ha analizado una señal importante en el análisis espacial, los kilómetros recorridos según el registro de la señal 3 (ver Tabla 3.8).

Bus	6224	6237	6234
Trayecto	70-2	70-2	70-2
Fecha y hora inicio	24/5/2021 06:58	24/5/2021 08:28	24/5/2021 08:50
Fecha y hora fin	24/5/2021 07:27	24/5/2021 09:13	24/5/2021 09:33
Kilometraje registrado (señal 3)	8	8	8
Kilómetros recorridos (según ubicación geográfica)	7.71	7.88	7.91

Tabla 3.8 Comparativa de 3 autobuses en el trayecto 70-2

Según la Tabla 3.8 los autobuses de este ejemplo realizaron el viaje en horarios diferentes y según la variable 3 referente al kilometraje del autobús los tres realizaron 8 kilómetros en ese viaje. Al ser la señal de kilometraje una variable de tipo entera el margen de error no permite apreciar las pequeñas diferencias en el recorrido realizado por los autobuses.

En este punto se recurrió a la ubicación geográfica de los autobuses mediante las coordenadas UTM y se calculó los kilómetros por los autobuses con un mayor nivel de precisión (última fila de la Tabla 3.8). Al disponer de la ubicación geográfica cada 10 segundos de funcionamiento es posible conocer la distancia recorrida entre dos señales mediante la **Distancia Euclídea**. Al tener coordenadas en formato UTM este cálculo es correcto al ser un formato de representación en un plano 2D. Entre cada señal registrada durante el viaje el autobús se mueve unos cuantos metros, y sumando los desplazamientos entre las diferentes lecturas de señales durante un viaje se obtiene la distancia total recorrida. Si bien las lecturas obtenidas con las coordenadas y el uso de la distancia euclídea no llegan a los 8 kilómetros marcados por la señal 3, se puede observar que estos valores son relativamente cercanos al valor de la señal y a la vez relativamente uniformes entre sí.

Este análisis se aplicó a distintas líneas, y siempre se pudo observar diferencias entre la ruta oficial y la marcada por la EMT. Esto es debido a la precisión de las coordenadas, que puede tener rangos de confianza de hasta 79 metros a la redonda [GVA21].

Un autobús puede ser asignado a varias líneas en un mes, esto indicaría un comportamiento bastante variable si se estudiara el comportamiento de la flota de autobuses en base a la línea que recorrió en un tiempo de observación. Por este motivo el análisis espacial no fue considerado como principal en este trabajo; sin embargo, ha sido de mucha utilidad para determinar la mejor forma de cuantificar los kilómetros recorridos en cada viaje. Por tanto, se considera que para tratar con la variable de distancia recorrida esta será obtenida mediante la distancia euclídea entre cada lectura en cada viaje, de esta forma se prescinde de la variable 3 y se descarta el problema de precisión de la misma al carecer esta de decimales que puedan hacer la lectura más exacta.

3.5 Creación del dataset final

Una vez analizadas las 15 señales que se registran a través del sistema BUS-CAN se procede a describir en esta sección el conjunto de datos que constituyen el dataset final con el que se trabajará.

A la hora de estudiar el comportamiento de un conjunto de objetos, autobuses en este caso, es necesario determinar la variable o variables que se utilizarán para estudiar dicho comportamiento. Estas variables tienen que ser representativas del funcionamiento de un autobús y, al mismo tiempo, establecer un criterio significativo de comparación entre autobuses.

De las 15 variables recogidas de los autobuses a través del sistema BUS-CAN (ver Tabla 3.3), se puede descartar las señales referentes al sistema de frenado, rampas y puertas ya que estas señales no ofrecen un perfil de funcionamiento significativo para el estudio de posibles anomalías. Respecto a las señales 14 y 15 ya se indicó que son variables calculadas a partir del consumo de combustible y distancia recorrida. Todo ello deja el análisis con tres parámetros relevantes:

1. Distancia recorrida en un viaje
2. Velocidad del autobús a lo largo del viaje
3. Consumo de combustible en el viaje

Debe tenerse en cuenta que no todos los autobuses de la flota tienen instalado el sistema BUS-CAN de registro de señales y, en algunos de los autobuses que sí lo tienen instalado, la señal de combustible no siempre devuelve un valor correcto. Esto se debe a dos posibles razones: (1) en el BUS-CAN se definen variables obligatorias y variables opcionales y puede suceder que la señal consumo de combustible no esté definida para un modelo de autobús determinado; (2) la variable consumo de combustible sí está definida, pero se produce un fallo en el sistema de transmisión de datos. Por la primera razón expuesta, cuando el BUS-CAN devuelve valores nulos del consumo de combustible suele producirse de manera sistemática en todos los vehículos de un modelo de autobús. De este modo, se eliminan del estudio aquellos autobuses para los que o bien no se dispone del dato del consumo o bien se produce un error en la práctica totalidad de las lecturas obtenidas.

Analizando los datos disponibles de **mayo de 2021 a octubre de 2021**, y tras un filtrado de aquellos autobuses de los que no se dispone de lecturas de consumo de combustible o bien son nulos, el dataset final con el que se realizará el estudio de anomalías se resumen en la Tabla 3.9.

La Tabla 3.9 muestra la flota de autobuses operativa para el análisis, la cual está formada por autobuses de cinco modelos diferentes, así como el número de viajes disponibles para cada modelo de autobús. En el siguiente capítulo se describirá la formación de series temporales y su posterior tratamiento a partir del dataset que se resume en la Tabla 3.9.

Modelo de autobús	Número de autobuses	Número de datos
43 SCANIA N250 E6 ZF	9	487 lecturas de viajes en total
38 CITARO E5 VOITH	2	150 lecturas de viajes en total
42 CITARO E6 VOITH	5	469 lecturas de viajes en total
49 IVECO HEULIEZ GX 337	6	555 lecturas de viajes en total
50 IVECO HEULIEZ GX 437 ART	11	843 lecturas de viajes en total

Tabla 3.9 Resumen de datos operativos para el análisis

4 Análisis de los datos

El objetivo de este capítulo es analizar los datos recogidos de los autobuses a lo largo del período mayo a octubre 2021 así como las propiedades de las series temporales que pueden formarse a partir de dichos datos. En este capítulo abordaremos un estudio detallado de la granularidad, periodicidad y estacionalidad de los datos antes de realizar el análisis de detección de anomalías en el siguiente capítulo.

4.1 Series temporales

Como se ha comentado en la sección 3.5, las tres señales más relevantes para determinar el comportamiento de un autobús son:

1. Distancia recorrida en un viaje
2. Velocidad del autobús a lo largo del viaje
3. Consumo de combustible en el viaje

Se considerará la variable ***tasa de consumo de combustible (litros) por distancia recorrida (km)*** para analizar el comportamiento de los autobuses ya que es la variable más representativa del funcionamiento y rendimiento de un autobús.

Así, el objetivo es generar **series temporales** (conjunto de observaciones generadas secuencialmente en el tiempo) que reflejen la tasa de consumo de combustible de los autobuses a lo largo del tiempo. Para ello, y como se ha visto en el capítulo anterior, debemos considerar los siguientes aspectos para un autobús determinado:

1. Datos recogidos diariamente a lo largo de varios **meses**.
2. Un número indeterminado de **lecturas diarias**, cada una correspondiente a un viaje realizado por el autobús
3. Cada viaje de un autobús está asociado a una **franja horaria**, la cual dependerá del horario en el que el autobús ha realizado el viaje.
4. Por cada viaje, se recogen señales con un período de **frecuencia de 10 segundos**.

Además, se debe tener en cuenta los siguientes factores:

1. El consumo de combustible de un autobús determinado para una misma ruta dependerá de la franja horaria en la que se realice el viaje.
2. Un autobús no circula obligatoriamente todos los días ya que puede haber períodos de descanso en los que el autobús está en cocheras o bien en el taller. Esto significa que no se dispone de lecturas de todos los días de un mes, pero las lecturas disponibles diarias son correlativas.
3. Para un día concreto, es posible que el autobús no haya circulado en todas las franjas horarias. Esto deja un número variable de lecturas por día, es decir un número variable de viajes para cada autobús. Los periodos de inactividad de un autobús tendrán valores nulos y serán ignorados en la formación de las series temporales. Esto significa que

una serie temporal de un autobús recogerá datos en instantes de tiempo consecutivos pero las series temporales de los diferentes autobuses tendrán diferentes tamaños.

El objetivo es realizar primeramente una inspección visual de las observaciones de la variable tasa de consumo de combustible/distancia, y posteriormente realizar un análisis estadístico de los datos. Para ello, se decidió agregar los datos de los viajes de los autobuses en franjas horarias, separados en espacios de una hora desde las 6.00 de la mañana hasta las 22.00 de la noche, y calcular el valor medio del consumo de cada franja horaria. Algunas consideraciones a la formación de estas series temporales son:

1. En el caso de que el viaje de un autobús solape con dos franjas horarias, los datos de dicho viaje se sitúan en la franja horaria en la que ocupe mayor duración.
2. En líneas generales, y teniendo en cuenta que la duración media de una es alrededor de 40 minutos, el número de viajes que se situarán en una franja horaria determinada será de uno o a lo sumo dos viajes para todos los autobuses.

En la formación de series temporales debe tenerse en cuenta el aspecto de los **datos faltantes** o **ausencia de datos**. En primer lugar, los autobuses con los que se realizará el análisis, y que aparecen en la Tabla 3.17, son todos aquellos autobuses para los que se dispone de lecturas correctas del consumo de combustible a lo largo del período mayo a octubre 2021. Asimismo, cabe señalar que en algunos casos puntuales el sistema BUS-CAN no registra valores correctos de la señal de GPS por lo que se procedió a subsanar este error tomando como referencia la distancia estándar del trayecto de la línea que estuviera realizando el autobús.

Como se verá en el capítulo siguiente, para el estudio de detección de anomalías se empleará una medida de similitud de series temporales que permite comparar series que no están perfectamente sincronizadas en el tiempo y tienen diferente longitud.

4.2 Inspección visual de los datos

Mediante una representación figura se pueden observar las características más sobresalientes de las series temporales como el movimiento a largo plazo, detección de valores atípicos o amplitud de las oscilaciones. El objetivo es analizar visualmente la forma, tendencia, estacionalidad y estacionariedad de las series de consumo de combustible por franja horaria. Para ello, se escoge un autobús de cada uno de los modelos presentados en la Tabla 3.17, siendo el análisis realizado para estos autobuses extensible al resto de la flota de autobuses.

La Figura 4.1 muestra la serie temporal formada con los valores de tasa de consumo de combustible por franjas horarias desde mayo a octubre de 2021 del autobús 8301 del modelo 50 IVECO. La línea de color naranja representa la media diaria de los valores de consumo.

Analizando visualmente la figura de la Figura 4.1 se tiene que:

1. Se observa una ligera tendencia ascendente de la variable en la primera mitad del conjunto de datos que luego tiende a descender también muy ligeramente, aunque es apenas apreciable.
2. No se observa estacionalidad en la serie, es decir, no se aprecia periodicidad o variación de la variable en cierto período, por ejemplo, mensual o semanal. En este sentido cabe mencionar que al no disponer de series espaciadas regularmente en el tiempo será difícil apreciar un componente estacional.

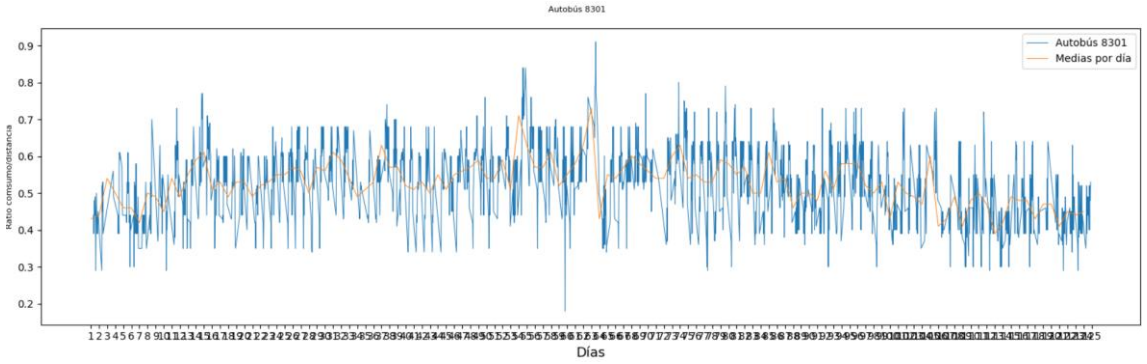

Figura 4.1 Tasa de consumo de combustible del autobús 8301 del modelo 50 IVECO

La Figura 4.2 muestra la serie temporal del autobús 6236 del modelo CITARO. Se observa una serie irregular con una cierta tendencia decreciente en el último período de tiempo, al igual que el autobús 8301. Los picos que se ven en la figura se corresponden con franjas horarias de mayor tráfico, las cuales suelen estar asociadas al período matinal de 8:30 a 9:30, una franja a mediodía que oscila entre las 14.00 y las 16.00, y otra vespertina que generalmente está alrededor de las 19.00. Por otro lado, cuando se observa un período de varios días donde la media del consumo es menor significa que el autobús ha realizado una ruta más larga durante dichos días, generalmente rutas de más de 15km que son trayectos que recorren los extrarradios de la ciudad.

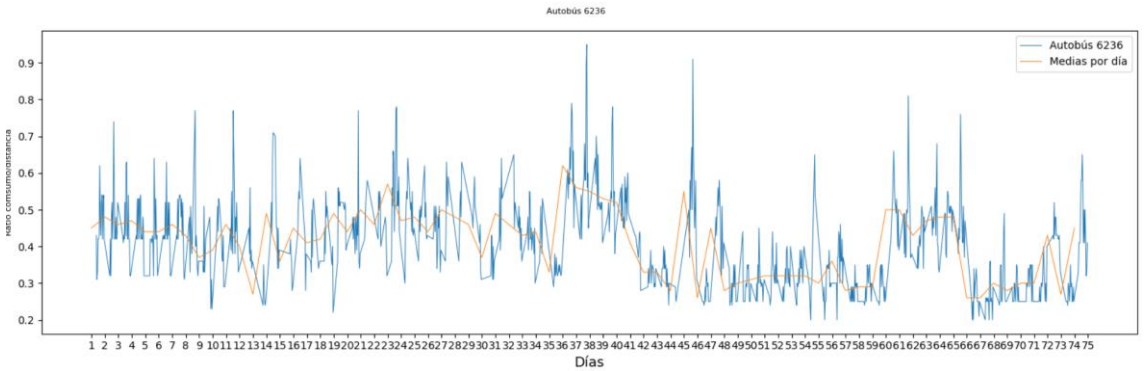

Figura 4.2 Tasa de consumo de combustible del autobús 6236

El comportamiento del autobús 6236 puede ser explicado en base a las rutas asignadas a dicho autobús durante el período de mayo a octubre. Los datos confirman que el autobús 6236 realizó hasta 13 líneas diferentes en los meses de junio y septiembre, y 8 rutas diferentes en los meses de mayo y octubre, entre las cuales hay tanto rutas de larga duración como rutas de corta duración. Así, el autobús ha realizado en un mismo mes rutas de 7km, 10km y 12 km.

Esta variabilidad no es usual en los autobuses de la flota, lo que se refleja en las tasas de consumo tan variables.

Finalmente, se muestran las series temporales de dos autobuses más, el 7114 y 9308 (ver Figuras 4.3 y 4.4), que pertenecen a dos modelos de autobuses distintos a los modelos de los autobuses analizados en las Figuras 4.2 y 4.3. En líneas generales, puede decirse que las series temporales de estos dos autobuses muestran una mayor regularidad que los dos autobuses anteriores donde se observa igualmente una ausencia de efecto estacional. Por otro lado, la amplitud de la señal del autobús 9308 es más variable que en el autobús 7114. Esto se debe a que el autobús 7114 ha realizado mayoritariamente la ruta Campanar-La Malva-rosa-Campanar, que tiene una longitud entre 10,2 y 11 km., mientras que la ruta predominante del autobús 9308 ha sido la ruta Estación del Norte- Torrefiel – Estación del Norte, con una longitud entre 4,8 y 6 km. Esta significativa diferencia en la longitud de las rutas, así como el trazado de estas, determina que la velocidad media real de la ruta del autobús 7114 está entre [13,61-15,44] km/h mientras que en el caso del autobús 9308 está entre [12,19-13,78] km/h. En consecuencia, rutas más largas donde los autobuses pueden tomar velocidades más altas tienden a mostrar un comportamiento más estable a lo largo del tiempo.

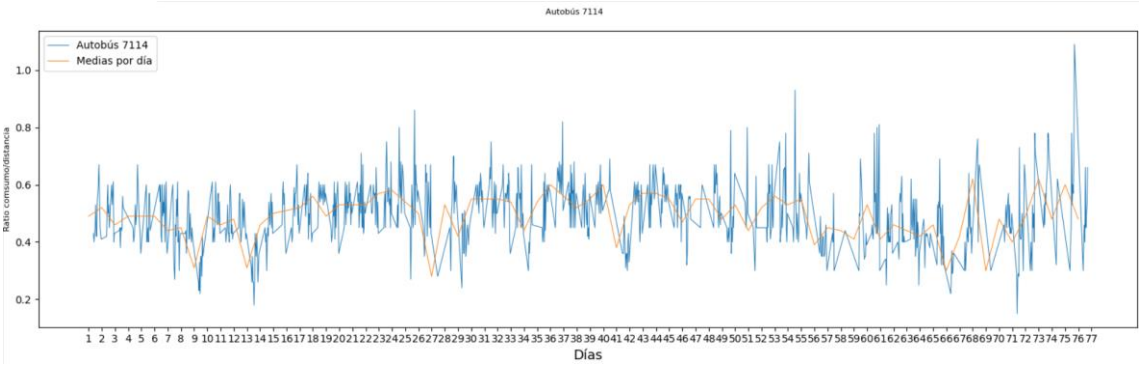

Figura 4.3 Tasa de consumo de combustible del autobús 7114

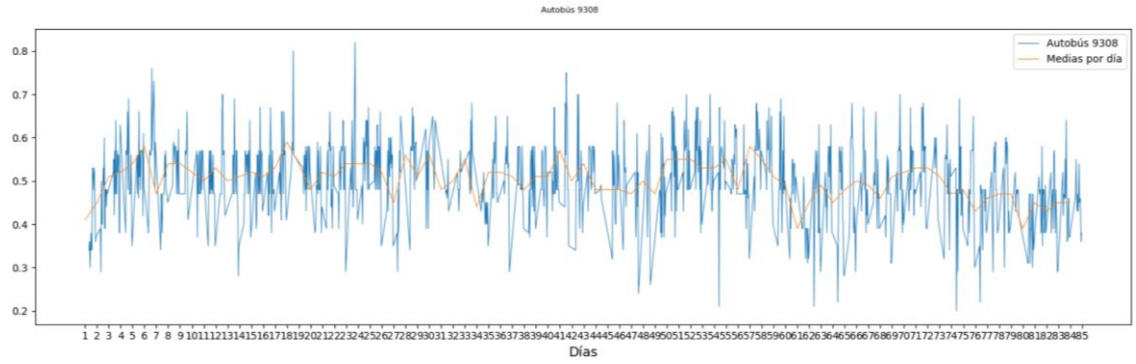

Figura 4.4 Tasa de consumo de combustible del autobús 9308

Salvo en el caso del autobús 6236, en líneas generales puede decirse que las gráficas muestran series temporales que mantienen un comportamiento más o menos estable a lo largo del tiempo. Respecto a la fluctuación de la amplitud de las oscilaciones se observa en todos los casos la existencia de desviaciones grandes (pequeñas) con respecto a la media, aunque no se aprecia que dichas desviaciones presenten algún tipo de inercia (creciente, decreciente o alternando períodos de desviaciones grandes/pequeños con pequeños/grandes de una forma sistemática). La combinación de períodos de alta volatilidad (alta varianza) con períodos de

baja volatilidad (baja varianza) suele ser habitual en series temporales de alta frecuencia, es decir, en series con datos diarios u horarios, como es este el caso. Además, las series pueden venir también condicionadas por otros factores como la ruta que han realizado los autobuses y el uso del aire acondicionado en los meses de verano.

Una de las conclusiones de la inspección visual es que apenas se observan tendencias ascendentes o descendentes en las series temporales. Por otro lado, al tratarse de series de tiempo irregulares por los períodos de inactividad de los autobuses, esto es, al no registrarse los datos en períodos de tiempo espaciados regularmente, no resulta fácil observar la existencia de un componente estacional.

Un aspecto relevante a la hora de trabajar con series temporales es la **estacionariedad**, lo que indicaría que la media y varianza de las series son constantes en el tiempo y que las series son estables a lo largo del tiempo. Si una serie muestra una tendencia creciente o decreciente en el tiempo entonces no es un proceso estacionario en la media. En caso de observar una alta desigualdad en la amplitud de las oscilaciones de los datos sería un indicativo de que el proceso no es estacionario en la varianza. Como se ha comentado, a tenor de las gráficas anteriores, se puede concluir que no existe tendencia apreciable (ascendente o descendente) en las series. Por otro lado, las gráficas muestran cierta desigualdad en la amplitud de las señales, aunque no se trata de fluctuaciones que evidencien una periodicidad reconocible. Esta desigualdad, que es habitual en series temporales de alta frecuencia como en el caso de lecturas de sensores, puede ser indicativo de que la varianza de la serie no es constante a lo largo del tiempo.

4.3 Análisis estadístico de los datos

La mayoría de los modelos para predicción del comportamiento de series de tiempo trabajan bajo el supuesto de que las series son estacionarias. Esto es así porque las series estacionarias son más fáciles de predecir y los modelos que se basan en dicho tipo de series son más fáciles de implementar. En este caso, aunque el objetivo se centra en el estudio de anomalías mediante la aplicación de algoritmos de *clustering*, trabajar con series estacionarias es importante porque muchas herramientas de análisis estadístico para series temporales se basan en este principio. Adicionalmente, el estudio de la estacionariedad permite adentrarse en las características y comportamiento de las series de tiempo.

Las gráficas de la sección anterior no muestran una confirmación visual clara de que las series sean estacionarias por lo que procedemos a realizar una prueba estadística. Para determinar la estacionariedad de las series se ha utilizado la prueba de **Dickey-Fuller**. Este es un método que permite detectar estadísticamente la presencia de una conducta estocástica en las variables de las series temporales mediante un contraste de hipótesis. La formulación de la hipótesis nula establece que la serie temporal se puede representar por una raíz unitaria que no es estacionaria, es decir, que tiene alguna estructura dependiente del tiempo. Por el contrario, si se acepta la hipótesis alternativa, que rechaza la hipótesis nula, indicaría que la serie es estacionaria.

Para ejecutar la prueba Dickey-Fuller se utilizó la librería de Python llamada `statsmodel` [StastsModel21]. La prueba devuelve varios valores estadísticos:

1. El valor estadístico Dickey-Fuller aumentado (ADF - Augmented Dickey Fuller) indica que cuanto más positivo es el valor, más posibilidades de aceptar la hipótesis nula y por tanto de que las series no sean estacionarias. Cuando el valor es negativo se rechaza la hipótesis nula y por tanto la serie temporal es estacionaria.
2. Un p-valor (p-value) de la prueba ADF inferior a 0.05 suele ser un indicativo de que la hipótesis nula se rechaza.
3. Los valores críticos (critical values) indican los intervalos de confianza para rechazar o no la hipótesis nula. Si el valor estadístico ADF es inferior al valor crítico correspondiente entonces se rechaza la hipótesis nula con una confianza del 99% (valor crítico del 1%), con una confianza del 95% (valor crítico del 5%) o una confianza del 90% (valor crítico del 10%).

	8301	**6236**	**7114**	**9308**
ADF Statistic	-4.741723	-4.026574	-7.827402	-4.7972000
p-value	0.000070	0.001278	0.000000	0.000055
Critical value 1%	-3.434	-3.437	-3.437	-3.436
Critical value 5%	-2.863	-2.865	-2.865	-2.864
Critical value 10%	-2.568	-2.568	-2.568	-2.568

Tabla 4.1 Resultados de estacionariedad de las series mediante el test de Dickey-Fuller

A tenor de los valores de la Tabla 4.1 podemos concluir que las series temporales de los cuatro autobuses son estacionarias. Los resultados estadísticos obtenidos son consistentes con las conclusiones de la inspección visual de las gráficas en la sección 4.2. La observación de la Figura 4.2 del autobús 6236 indicaba que este era el autobús que mostraba una mayor variabilidad en los datos y una menor estacionariedad visualmente. Esto se confirma con los datos de la Tabla 4.1 al ser el autobús que ha obtenido un menor valor ADF y un mayor valor de p-valor, aun siendo este inferior a 0.05. Por otro lado, también se puede concluir que el autobús 7114 es el que muestra un comportamiento más estable, al ser el autobús para el que se obtiene el menor valor de ADF de los cuatro y el menor p-valor, lo que asimismo es consistente con la menor variabilidad observada en la amplitud de la señal de este autobús (ver Figura 4.3 y comentarios).

4.4 Análisis de otras frecuencias

La construcción de las series temporales con frecuencia horaria resulta en uno o dos viajes por franja horaria y es la frecuencia que representa más fielmente el registro de datos del sistema BUS-CAN y el funcionamiento de los autobuses. Además, como se ha visto en las secciones anteriores, las series de frecuencia horaria son estacionarias.

En esta sección estudiamos la construcción de las series con frecuencia diaria con el fin de determinar si se obtienen series más estables. Dado que nuestras series temporales son irregulares por los períodos de inactividad de los autobuses, las series de frecuencia diaria se

generaron manualmente calculando la media del consumo de combustible de todas las franjas horarias comprendidas en un día. Al realizar la agregación diaria, se elimina el efecto de la hora del día en el consumo de combustible, lo que significa que consumos más altos de horas de más tráfico o de días de mayor uso de aire acondicionado se agregarían con consumos más bajos y se pierde el efecto 'horario', es decir, se eliminaría una potencial estacionalidad diaria. En cualquier caso, como se ha comentado anteriormente, resulta difícil observar una componente estacional diaria al no disponer de datos equiespaciados en el tiempo.

El objetivo, por tanto, es estudiar series de frecuencia diaria y comprobar tendencias y estacionariedad. La Figura 4.5 muestra las series diarias de los cuatro autobuses estudiados anteriormente.

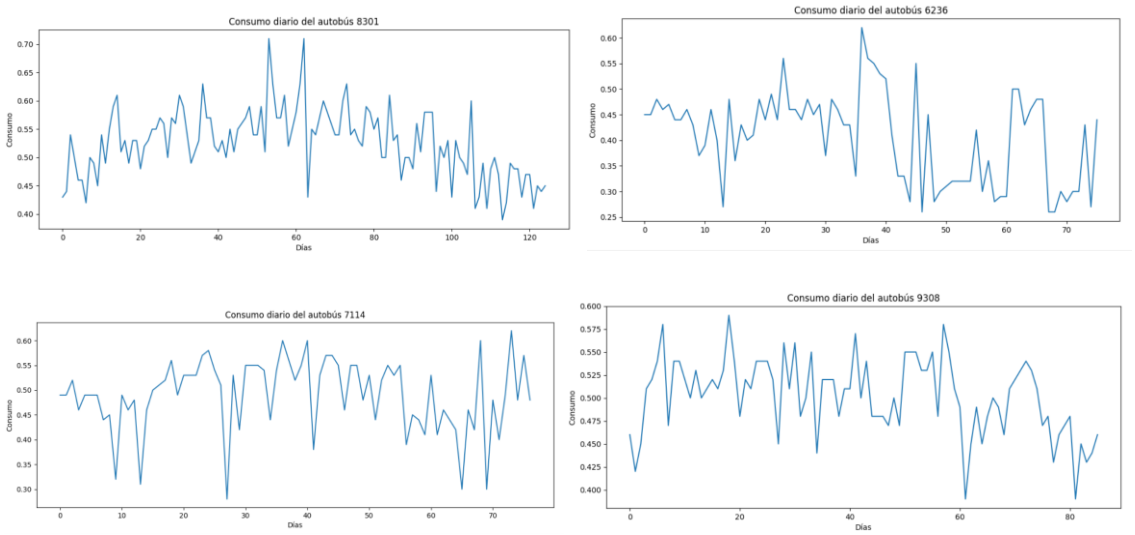

Figura 4.5 Series temporales diaria (autobús 8301 arriba izquierda, autobús 6236 arriba derecha, autobús 7714 abajo izquierda, autobús 9308 abajo derecha).

Visualmente puede observarse que las series muestran una menor estabilidad. Esto es especialmente notable en el autobús 7114, donde la serie temporal horaria mostraba una alta estacionariedad. La tabla 4.2 muestra los resultados de aplicar el test Dickey-Fuller a las series de tiempo diarias.

A tenor de los datos, podemos extraer las siguientes conclusiones:

1. Los datos de los autobuses 7114 y 9308 siguen concluyendo series estacionarias pero el valor ADF se ha incrementado, así como el p-valor (aunque sigue siendo menor de 0.05).
2. En el caso del autobús 6236, la confianza de la prueba Dickey-Fuller ha bajado al 95%.
3. Resulta llamativo el caso del autobús 8301, donde ahora la serie diaria resulta no estacionaria. Esto se debe a que la ligera tendencia ascendente y descendente que se había observado en la serie horaria se ha acentuado en la serie diaria.

En resumen, podemos concluir que las series temporales con frecuencia diaria muestran un menor grado de estacionariedad y, en algunos casos, devuelven series no estacionarias. Esto se debe fundamentalmente a que la media de la ratio de consumo diario de combustible ya

no es estable a lo largo del tiempo, mientras que las observaciones horarias mantienen el efecto del consumo por franja horaria.

	8301	**6236**	**7114**	**9308**
ADF Statistic	-1.379197	-3.122337	-4.246615	-4.021697
p-value	0.592208	0.024944	0.000549	0.001301
Critical value 1%	-3.487	-3.522	-3.521	-3.511
Critical value 5%	-2.886	-2.901	-2.901	-2.897
Critical value 10%	-2.580	-2.588	-2.588	-2.585

Tabla 4.2 Resultados de estacionariedad de las series diarias mediante

4.5 Construcción y procesamiento de las series temporales

Una vez analizada la naturaleza y comportamiento de los datos de estudio de los autobuses, se procede a hacer un proceso de limpieza, construcción y procesamiento de estos para obtener las series temporales que serán los datos de entrada en los algoritmos que utilizamos. Para esto se recurrió a la librería Pandas, la cual es muy usada en el tratamiento y manipulación de series temporales. Los pasos que se siguieron fueron los siguientes:

1. Cada autobús se representa en una estructura de datos `DataFrame` de Pandas compuesto por la fecha, hora de inicial y final de un viaje, combustible consumido en ese viaje y la distancia recorrida.
2. Se calcula la tasa de combustible por viaje, esto es, combustible consumido en un viaje dividido entre la distancia recorrida. La tasa de combustible pasa a reemplazar a las columnas de combustible consumido y kilómetros recorridos.
3. Se ubica la franja horaria de cada viaje, una franja entre las 6 horas hasta las 22 horas. Esto se realiza mediante las opciones de filtrado de datos para datraframes de Pandas. Cuando un viaje sucede en dos franjas diferentes, se considera que pertenece a la franja en la que más tiempo haya circulado.
4. Una vez determinados los viajes que pertenecen a cada franja horaria se procede a uniformar los datos, es decir, para cada franja la tasa de combustible es igual al promedio de todos los viajes asignados a dicha franja.
5. Se obtiene un `DataFrame` final en el que se caracterizan las columnas fecha, franja y tasa combustible. Esto para cada autobús que se vaya a ser incluido en el análisis.
6. Se generan las series con las columnas de tasa de combustible para cada autobús, para lo cual se aísla la columna referente a la tasa de combustible (las columnas en pandas son consideradas series).
7. Una vez obtenidas las series individuales estas pasan a formar parte de una matriz en la que cada fila representa un autobús.
8. Para poder someter las series al análisis es necesario normalizarlas, en este caso se usó una normalización entre 0 y 1 tomando como bases el elemento máximo y el elemento mínimos de todas las series temporales construidas previamente.

Con estos pasoso mencionados se procedió a someter a las series temporales a las evaluaciones de los algoritmos expuestos en la siguiente sección.

5 Detección de anomalías

En este capítulo realizamos el estudio de detección de anomalías de la flota de autobuses con el fin de determinar si existen patrones inusuales en el comportamiento de los vehículos, esto es, si se observa una discordancia en algún vehículo frente al comportamiento predominante del conjunto de objetos del estudio. Este capítulo se estructura del siguiente modo. En la primera sección se explica el criterio de similitud que se ha utilizado para comparar las series temporales, se detallan las especificaciones de los dos algoritmos de *clustering* utilizados, así como la métrica para evaluar la calidad de las particiones resultantes. En las siguientes secciones analizamos los resultados de aplicar los algoritmos de *clustering* a cada modelo de autobuses, indagando en la vida de aquellos vehículos en los que se observe un comportamiento discrepante.

5.1 Agrupamiento de series temporales

En esta sección se detallan los elementos comunes que se utilizarán en la aplicación de los dos algoritmos de agrupamiento de series temporales que se aplican a nuestro problema, y cuyos resultados se muestran en las siguientes secciones.

5.1.1 Medida de similitud

Una de las primeras decisiones a la hora de hacer agrupamiento de series temporales es la medida de similitud o proximidad que se va a utilizar para comparar las series. La medida más utilizada en algoritmos de *clustering* estándar es la distancia Euclídea, que depende exclusivamente de la proximidad que existe entre los valores observados en los correspondientes instantes de tiempo. El problema de la distancia Euclídea es que produce resultados de similitud imprecisos cuando existe una distorsión en el eje del tiempo; es decir, cuando dos series están fuertemente correlacionadas y una de ellas está desplazada incluso un solo instante de tiempo respecto a la otra, la distancia Euclídea devolverá que no son series similares. Por ejemplo, si en las dos series se produce un punto máximo en los datos, pero este pico no ocurre en el mismo instante de tiempo en las dos series, la distancia Euclídea no permitirá alienar los puntos máximos de las dos series. Además, tampoco permite comparar series temporales que no sean de la misma longitud. Y este es precisamente el caso de nuestros datos.

Para tratar con este problema y poder comparar nuestras series temporales se ha utilizado la medida de distancia denominada 'tiempo dinámico de deformación' o *Dynamic Time Warping* (DTW) por sus siglas en inglés. El DTW es una técnica para medir la similitud de dos secuencias temporales que no están exactamente alineadas en el tiempo o no tienen la misma longitud. DTW encuentra el alineamiento no-lineal óptimo entre dos series temporales. Un ejemplo de esto puede verse en la figura 5.1.

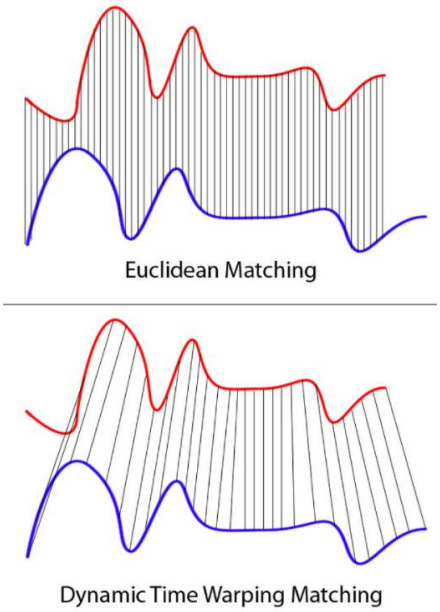

Figura 5.1 Comparación de distancia Euclidea y DTW. Fuente: https://databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html

Dadas dos series temporales, $x = (x_1, .., x_n)$ e $y = (y_1, .., y_m)$, la distancia entre las dos series medida por DTW se formula como

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}$$

siendo π un camino formado por *K* pares de índices (un índice por cada serie) que marca el alineamiento de las dos series temporales. La medida DTW entre la posición i de la serie x y la posición j de y se formula como un problema de optimización que se representa por la siguiente función recursiva:

$$D(i,j) = |x(i) - y(j)| + \min \begin{Bmatrix} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{Bmatrix}$$

donde la expresión $|x(i) - y(j)|$ indica la distancia o coste entre el elemento x_i y el elemento y_j. Existen varios tipos de distancia que pueden aplicarse para medir la distancia entre un valor observado de la serie x y otro de la serie y, es decir para medir $d(x_i , y_j)$, como, por ejemplo, la distancia Euclídea, la más básica; la distancia Hamming, que se aplica principalmente a variables categóricas; distancia Euclídea al cuadrado, aplicada como una variante de la distancia Euclídea normal, por mencionar algunos. La expresión de minimización representa la parte recursiva de la función donde se busca hallar la menor distancia hacia otros elementos vecinos de las series.

El cálculo de DTW se centra en construir una matriz de longitud $n\ x\ m$ para comparar cada elemento de la serie x con cada elemento de la serie y ($n\ x\ m$ comparaciones). Para ello, se utiliza la condición inicial $D(1,1) = |x(1) - y(1)|$, es decir, se comienza con el cálculo de distancia entre los primeros elementos de las series, que es aquí cuando entra en juego el tipo de métrica a emplear. Posteriormente, se calculan las distancias entre los demás elementos avanzando en la construcción de la matriz. Básicamente, la fórmula calcula la distancia entre cada elemento de x y su punto más cercano en y.

El algoritmo se rige bajo ciertas restricciones y reglas que se exponen a continuación donde índice se refiere a una posición en una serie temporal:

- Cada índice de la primera serie debe coincidir con uno o más índices de la otra secuencia, y viceversa.
- El primer índice de la primera serie debe coincidir con el primer índice de la otra serie (pero no tiene que ser su única coincidencia)
- El último índice de la primera serie debe coincidir con el último índice de la otra serie (pero no tiene que ser su única coincidencia)
- El mapeo de los índices de la primera serie a los índices de la otra serie debe ser monótonamente creciente y viceversa, es decir, si *j > i* son índices de la primera serie, entonces no debe haber dos índices *l > k* en la otra secuencia, de modo que el índice *I* coincide con el índice *l*, y el índice *j* coincide con el índice *k*, y viceversa.

El algoritmo, diseñado para calcular la similitud DTW entre dos series temporales, puede extenderse al caso de varias series temporales; en dicho caso, en lugar de hacer comparación de dos series se manejan varias series representados en una matriz y se aplica el algoritmo "todos contra todos", dando como resultado una matriz, llamada **matriz de distancias**. Si se tienen *n* individuos con sus respectivas series temporales entonces la matriz de distancias será cuadrática de longitud *n.*

En lo sucesivo, se utilizará siempre la distancia DTW para los experimentos, los cuales se realizarán con la librería **scikit-learn** (https://scikit-learn.org/).

5.1.2 Aplicación del algoritmo K-means

Para la aplicación del algoritmo K-means se ha utilizado la librería `tslearn` de Python con los siguientes parámetros:

- `n_clusters`: se refiere al número de clústeres o grupos para hacer la partición; en este proyecto se ha trabajado con particiones de dos y tres clústeres.
- `n_init`: número de veces que se ejecuta el algoritmo con semillas de centroides diferentes; para cada una de las pruebas realizadas se ejecutó el algoritmo K-means cinco veces, es decir, se utilizó `n_init=5`.
- `metric`: métrica de evaluación de distancias; se utilizó DTW.
- `max_iter`: máximo número de iteraciones del algoritmo K-means para una ejecución; se utilizó un valor máximo de 50 iteraciones.

- `max_iter_barycenter`: número de iteraciones para el cálculo del baricentro; este parámetro es necesario cuando se emplea la métrica DTW; se utilizó el valor 100.
- `random_state`: determina la generación de un número aleatorio para la inicialización del centroide; se utilizó semillas diferentes para cada una de las ejecuciones del algoritmo K-means.
- `init`: método de inicialización; se utilizó el método kmeans++ el cual selecciona los centroides de los clústeres iniciales tal que se acelera la convergencia del algoritmo.

En la aplicación del algoritmo K-means con DTW, el centro o centroide de los clústeres también se denomina bari-centro (barycenter) y se calcula respecto a la distancia DTW. El bari-centro es la secuencia media de un grupo de series temporales en el espacio DTW. Cuando se aplica la métrica DTW, el algoritmo más utilizado para calcular el bari-centro es el DTW Barycenter Averaging (DBA), el cual minimiza la suma de la distancia DTW al cuadrado entre el bari-centro y las series del clúster [Petitjean11].

Con las definiciones del modelo se procede a entrenarlo con la matriz de distancias DTW en la que cada fila representa las distancias DTW de cada autobús con los demás, de este modo se asegura que el agrupamiento no depende de si las series están alineadas o no.

5.1.3 Aplicación de *clustering* jerárquico

En este trabajo se ha utilizado la técnica de *clustering* jerárquico aglomerativo. Por un lado, es la técnica más utilizada en el agrupamiento jerárquico; por otra parte, al ser los autobuses independientes entre sí, esto permite considerar inicialmente cada autobús como un clúster independiente y buscar comportamientos similares desde abajo hasta lograr agruparlos para poder encontrar anomalías de comportamiento.

La librería que se ha utilizado para este trabajo es `scipy.cluster.hierarchy` de Python, la cual dispone de la función `linkage` que realiza el *clustering* alglomerativo generando una matriz de enlaces que se representará gráficamente como un dendrograma mediante la utilización de la función `dendrogram` perteneciente a dicha librería. La matriz de enlaces es el dendrograma en sí mismo, representado computacionalmente como una matriz.

La función `linkage` se instanció con los siguientes parámetros:

- la matriz de distancias, la cual contiene todas las distancias DTW que hay entre los autobuses, es decir tantas filas y columnas como autobuses porque el cálculo de distancias es “todos contra todos”.
- el método para generar la matriz de enlaces; este es el método que se utiliza para determinar la distancia o similitud entre los individuos y/o agrupaciones. Existen varios métodos con distintos criterios para calcular las distancias: el método de enlace simple que calcula las distancias entre los elementos más cercanos de cada agrupación y/o individuo; el método de enlace completo que calcula la distancia entre los elementos más alejados de cada agrupación y/o individuo; y el método de distancia media que calcula la distancia media inter-clúster. Nosotros optamos por el método de distancia medias, identificado mediante la palabra **`average`**,

porque los otros métodos tienden a distorsionar la métrica inicial lo cual no sería un problema si se dispusiera de más datos [Everitt09, UV22]

En líneas generales, el objetivo que se persigue es identificar un gran salto en la distancia si nuestro propósito es poder explicar la existencia de un cierto número de clústeres. La selección del valor de corte o *threshold* es preferible hacerlo manualmente ya que esto nos permitirá analizar los datos en detalle y detectar casos extremos o valores atípicos.

5.1.4 Métrica de evaluación

Como se ha mencionado en la sección 2, existen varios métodos para determinar el mejor número de clústeres, siendo el método del codo (Elbow method) y el coeficiente de *silhouette* de las particiones dos de los métodos más conocidos. El método del codo no funciona bien si los datos no están muy agrupados o no se dispone de muchos datos por lo que se decidió utilizar la métrica de *silhouette* en este trabajo. Por otro lado, como las pruebas se aplicaron a dos y tres clústeres, el cálculo del *silhouette* no implica un gran costo.

El ***silhouette score*** es una medida de la similitud media de los elementos dentro de un clúster y su distancia a los demás objetos de los otros clústeres. Esta distancia puede ser calculada con diferentes métricas, como ser distancia Euclídea, Manhattan, coseno, entre otros; como en este trabajo se usa la matriz de distancias DTW eso quiere decir que las distancias ya están calculadas. El cálculo del *silhouette score* admite trabajar bajo esas condiciones, por lo que se puede usar la métrica *precomputed*. El valor del *silhouette* se mide con la siguiente fórmula:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

, donde *b(i)* representa la distancia media del punto *i* a todos los puntos del clúster más cercano (distancia inter-cluster o separación), y *a(i)* representa la distancia media del punto *i* a todos los puntos dentro de su clúster (distancia intra-cluster o cohesión). El valor del *silhouette* varía entre +1 y -1, donde:

1. Un valor +1 indica que los clústeres están bien separados unos de otros y claramente diferenciados, es decir, que los elementos de un clúster están bien emparejados en su grupo y mal emparejados con los grupos vecinos
2. Un valor de 0 indica que la distancia entre los clústeres no es significativa
3. Un valor negativo indica que los objetos se han asignado al clúster equivocado

Por tanto, cuanto mayor sea el valor del *silhouette*, más representativa es la agrupación de los datos. El método *silhouette* es el método más utilizado para encontrar el número óptimo de clústeres en el algoritmo K-means.

5.2 Modelo 50 IVECO HEULIEZ GX 437 ART

La primera familia de autobuses que se analizará es la del modelo 50 IVECO HEULIEZ GX 437 ART. Esta es la familia de la que se dispone de más autobuses con datos de tasa de consumo

de combustible. Se realizará primero un análisis aplicando el algoritmo K-means, seguido de la aplicación del algoritmo *clustering* jerárquico y luego se procederá a una comparación y discusión de los resultados.

5.2.1 Aplicación del algoritmo K-means

En la Tabla 5.1 se muestran los resultados de la aplicación del algoritmo K-Means en la generación de particiones de 2 clústeres (C2) y 3 clústeres (C3):

BUS	C2	C3
8301	1	1
8303	1	2
8305	1	2
8307	1	1
8308	1	2
8309	0	0
8310	1	2
8311	1	2
8312	1	1
8313	1	1
8314	1	1

Tabla 5.1 Clustering de los autobuses del modelo 50 IVECO HEULIEZ GX 437 ART

Los resultados de aplicar la métrica del *silhouette* son los siguientes:

- Para dos clústeres, el valor medio del *silhouette* es 0.57
- Para tres clústeres, el valor medio del *silhouette* es 0.33

Sobre la partición de dos clústeres el valor del *silhouette* se considera alto e indica que los clústeres están bien diferenciados. Además, como se puede apreciar en la Tabla 5.1, el autobús 8309 aparece siempre en un clúster diferenciado.

La Figura 5.2 muestra el resultado para la partición de los datos en dos clústeres (C2). El eje X muestra el número de autobús y el eje Y muestra el valor de la distancia DTW. El gráfico de la derecha de la Figura 5.2 muestra los autobuses del clúster 1. Cada una de las líneas azules se corresponde con un autobús perteneciente a dicho clúster, y representan la distancia DTW del autobús correspondiente respecto a los demás autobuses. Se identifica el autobús al que pertenece cada línea cuando la distancia es 0, lo cual representa la distancia consigo mismo. Cuanto más alejada esté la línea azul de un autobús respecto al eje X, más diferencias habrá entre los dos autobuses, es decir, entre el autobús que representa la línea azul con el autobús del eje X. De este modo, se puede ver claramente en la figura de la derecha que todos los autobuses se alejan del autobús 8309, el cual pertenece al otro clúster (clúster 0).

En la Figura 5.2 se puede apreciar también una línea roja predominante que representa el centroide o bari-centro del clúster. Dado que el clúster 0 solo contiene al autobús 8309, la línea roja se solapa totalmente con la línea azul representativa del autobús, siendo el mismo autobús su propio centroide (se puede observar que la distancia 0 se posiciona en la etiqueta

8309 en el eje X). En el caso del clúster 1 se puede ver que tanto el centroide de color rojo como las líneas azules se alejan del marcador del autobús 8309 llegando a estar en las distancias más grandes del gráfico, indicador de la gran diferencia que existe entre los autobuses del clúster 1 con el autobús del clúster 0. Si analizamos el bari-centro del clúster 1 (línea de color rojo), la cual representa la forma promedio de las series que constituyen el clúster 1, podemos observar que la mayor distancia del centroide se produce con respecto al autobús 8309 (que pertenece al otro clúster) y, en segundo lugar, con respecto al autobús 8308 (que pertenece al mismo clúster 1).

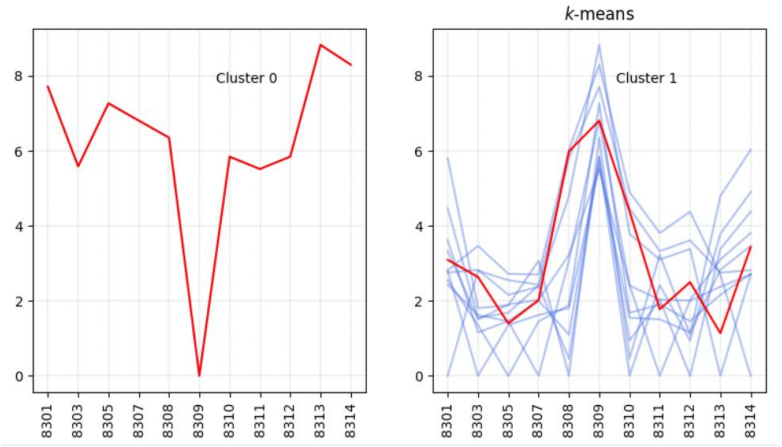

Figura 5.2 Configuración de dos clústeres para los autobuses del modelo 50 IVECO

La Figura 5.3 muestra el resultado de agrupar los autobuses en tres clústeres. Lo que se puede observar en esta figura es que, al igual que en el caso de dos clústeres, el autobús 8309 constituye él solo un clúster, confirmando así un comportamiento discrepante con respecto al resto de autobuses del modelo. Además, la aplicación de la métrica del *silhouette* sobre tres clústeres arrojó un valor de 0.33, más bajo que el obtenido con dos clústeres, por lo que se considera que la partición óptima para la familia de los 50 IVECO es dos clústeres.

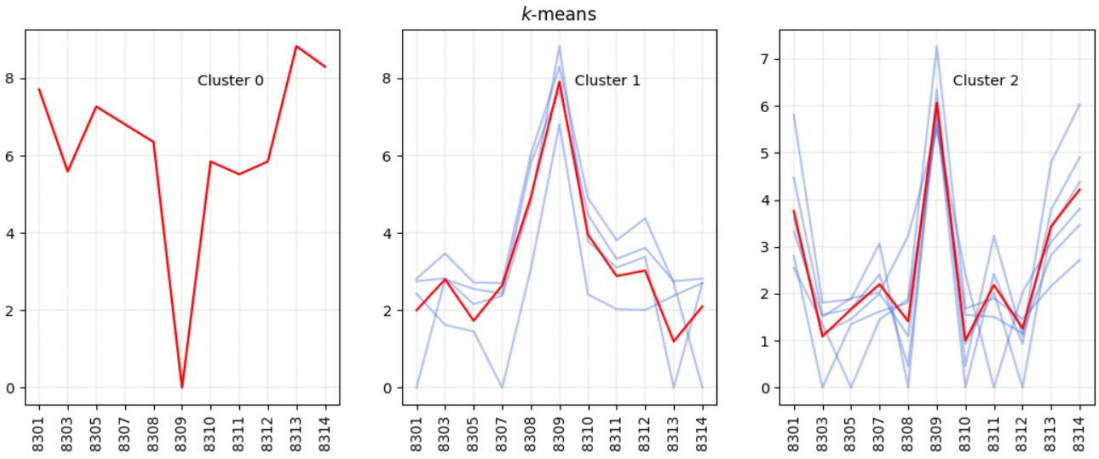

Figura 5.3 Configuración de tres clústeres para los autobuses del modelo 50 IVECO

Para tratar de analizar el comportamiento desigual del autobús 8309 del resto de autobuses, se analizó las rutas seguidas por dicho autobús en el período mayo-octubre, no encontrándose diferencias significativas. El análisis de las rutas no desveló ninguna información relevante ya que otros autobuses pertenecientes al clúster 1 también realizaron las mismas rutas. Hay que

tener en cuenta además que rutas diferentes pueden ser similares en lo que respecta a la longitud de estas, y/o el tiempo medio de recorrido.

Posteriormente se analizó la cantidad de lecturas disponibles para cada autobús, cuantificando las franjas horarias con lecturas de cada autobús por mes, es decir, la longitud de las series temporales de esta familia (ver Tabla 5.2).

BUS	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	**TOTAL**
8301	61	282	374	159	371	365	1612
8303	75	167	89	122	72	4	529
8305	98	111	106	239	138	183	875
8307	143	14	224	195	215	230	1021
8308	130	15	0	0	0	0	145
8309	189	33	89	121	42	31	505
8310	252	8	0	0	49	22	331
8311	117	143	20	65	197	56	598
8312	36	0	0	54	86	203	379
8313	233	333	310	37	321	244	1478
8314	227	257	320	333	308	238	1683

Tabla 5.2 Distribución mensual de franjas horarias con lecturas de tasa de combustible por autobús de 50 IVECO

La tabla 5.2 contiene un resumen de las franjas horarias contabilizadas mensualmente durante los meses de estudio. Los valores de cada mes indican el número de franjas horarias en las que trabajó cada autobús, y la última columna es la sumatoria de todos los meses para cuantificar la cantidad total de franjas horarias trabajadas durante los 6 meses, la cual sería la longitud de la serie temporal de cada autobús.

Basándonos en la cuantificación de la tabla 5.2 se puede ver que se dispone de datos del autobús 8309 para todos los meses, lo cual no explicaría un comportamiento anormal. Sin embargo, sí se puede apreciar que el número de datos del autobús 8308 es muy inferior al del resto de vehículos, es más, no hay datos disponibles desde julio. Esta gran diferencia de datos puede llegar a afectar la clasificación del algoritmo ya que estaría intentando alinear datos demasiado lejanos en el tiempo. Considerando este detalle se procedió a excluir el autobús 8308, cuyos resultados se muestran en la Tabla 5.3.

Los resultados del análisis del *silhouette* son los siguientes:

- Para dos clústeres, el valor medio del *silhouette* es 0.23
- Para tres clústeres, el valor medio del *silhouette* es 0.4

Con la exclusión del autobús 8308 se pueden apreciar cambios significativos en los resultados. En primer lugar, el valor del *silhouette* para dos clústeres es 0.23, mucho más bajo que en el caso de la Figura 5.2. Esto mismo puede apreciarse en la Figura 5.4 la cual muestra que el clúster 0 está considerando al autobús 8309 como parte de él, pero la forma del resto de autobuses del clúster 0 no se corresponde con la forma de dicho autobús. En cambio, el clúster 1 sí tiene cierta armonía con los elementos que lo conforman.

BUS	C2	C3
8301	1	1
8303	1	0
8305	1	0
8307	1	0
8309	0	2
8310	1	0
8311	1	0
8312	1	0
8313	1	1
8314	1	1

Tabla 5.3 *Clustering* de los autobuses del modelo 50 IVECO excluyendo al 8308

En cambio, en la partición de tres clústeres el valor de la métrica de *silhouette* es mayor que para dos clústeres (valor de 0.4 frente a 0.23) y el autobús 8309 continúa aislándose como único elemento de un clúster, tal y como se muestra en la Figura 5.5.

De los experimentos realizados con el algoritmo K-means podemos concluir que cuando se ha obtenido un valor alto de *silhouette*, un valor superior a 0.3, el autobús 8309 siempre ha formado un clúster independiente. Por otro lado, es importante analizar la forma del bari-centro y ver como se separa la secuencia media que está representando de las series de los autobuses. En líneas generales se puede observar que las distancias más significativas de los centroides se producen con respecto a vehículos que no forman parte del clúster, como es de esperar, pero también se observa en ocasiones que el bari-centro se aleja de algún autobús que forma parte del propio clúster, como ocurre en la Figura 5.2. con el autobús 8308. Una vez excluido dicho autobús por la falta de datos, obtenemos la partición de la Figura 5.5, que es la que tiene un mayor valor de *silhouette*, donde las mayores distancias de cada centroide se producen siempre con autobuses que no forman parte de su clúster.

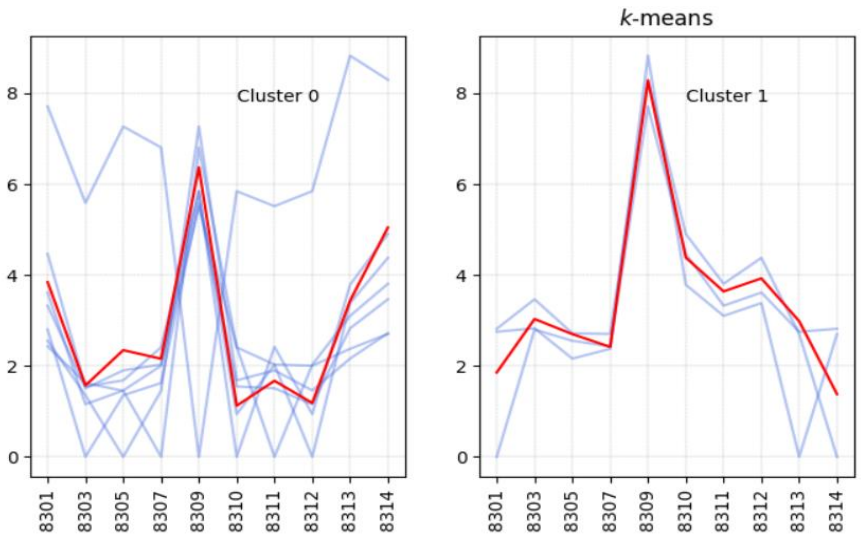

Figura 5.4 Configuración de dos clústeres para los autobuses del modelo 50 IVECO excluyendo al 8308

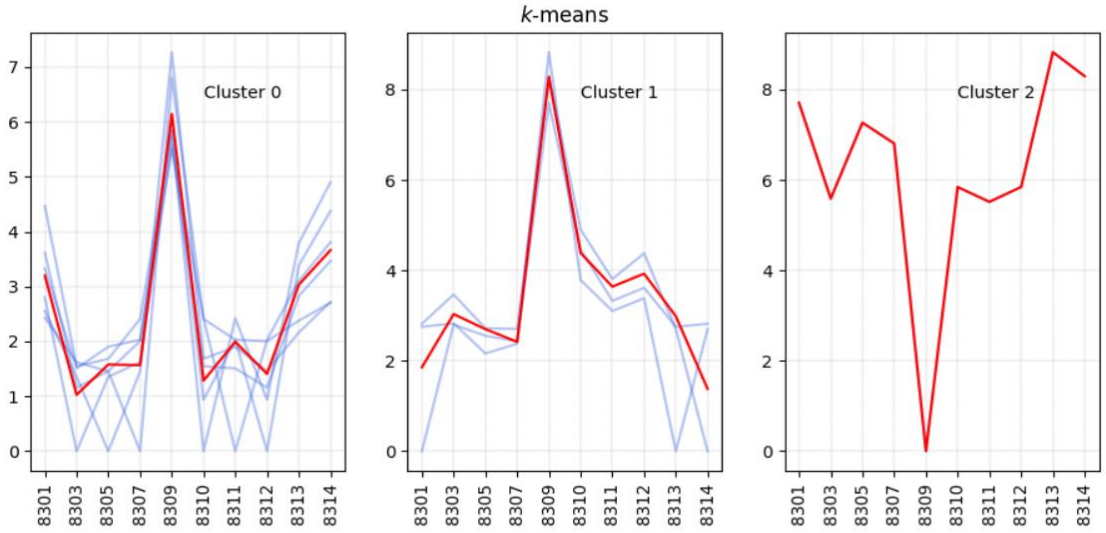

Figura 5.5 Configuración de tres clústeres para los autobuses del modelo 50 IVECO excluyendo al 8308.

5.2.2 Aplicación del algoritmo *clustering* jerárquico

Primeramente, aplicamos el algoritmo de *clustering* jerárquico a los 11 autobuses del modelo 50 IVECO, y obtenemos un dendrograma para un valor de *threshold* = 9, que es el máximo valor de DTW que devuelve el algoritmo del K-means en la sección anterior. Esto resulta en un dendrograma donde no se diferencian clústeres por colores. A partir de este dendrograma seleccionamos a continuación un valor de *threhold* que corte la línea vertical que representa la distancia más grande. El valor seleccionado es *threshold*=5 y el resultado puede verse en la Figura 5.6.

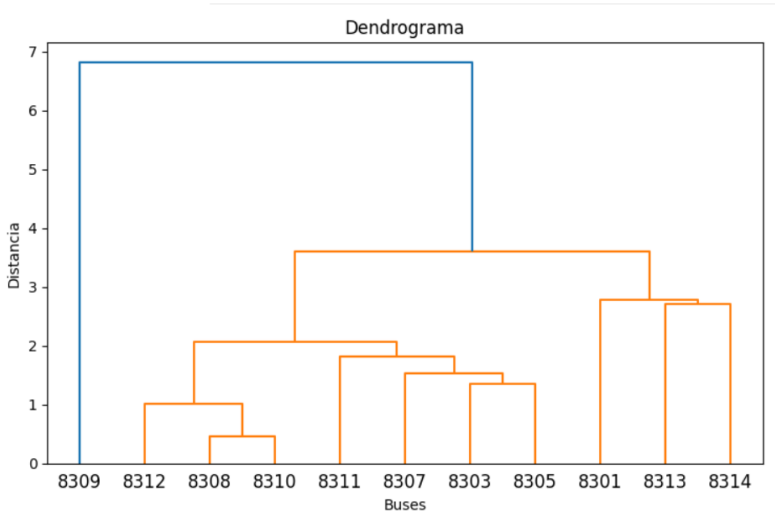

Figura 5.6 *Clustering* jerárquico para los autobuses del modelo 50 IVECO

En la Figura 5.6 se puede apreciar el aislamiento del autobús 8309, lo cual concuerda con los resultados obtenidos con el algoritmo K-means. La Figura 5.6 refleja la formación de dos agrupaciones con un *threshold* de valor 4, una agrupación contiene el autobús 8309 y la otra agrupación contiene el resto de los vehículos. Para un *threshold*=3, tendríamos una partición de tres clústeres: un clúster formado por el autobús 8309, un segundo clúster formado por los

autobuses 8312, 8308, 8310, 8311, 8307,8303 y 8305, y un tercer clúster formado por los autobuses 8301, 8313 y 8314. Se puede observar asimismo que la altura de la línea vertical de color azul que separa el autobús 8309 del resto de autobuses se la que tiene mayor altura, indicativo de la separación que existe entre este autobús con el resto.

Una segunda prueba, similar a la aplicada en el apartado anterior, es decir excluyendo el autobús 8308, y con un *threshold* de 5, brinda los resultados ilustrados en la Figura 5.7.

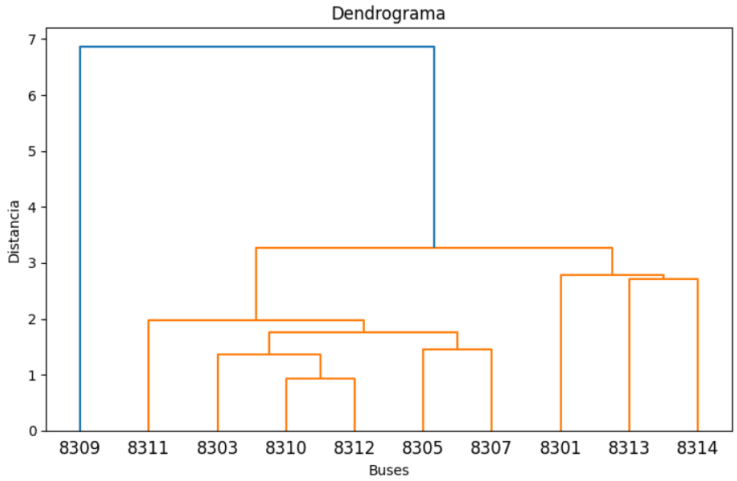

Figura 5.7 *Clustering* jerárquico para los autobuses del modelo 50 IVECO excluyendo el 8308

El dendrograma muestra la diferencia del autobús 8309 respecto a los otros autobuses representados en un clúster de color naranja. Se puede observar igualmente que la línea vertical azul que une el clúster formado por el 8309 con el clúster formado por todos los demás autobuses es la que tiene mayor altura.

Resulta interesante destacar que la agrupación formada por los autobuses 8301, 8313 y 8314 también aparece diferenciada en las figuras 5.4, 5.5 y 5.6, e igualmente en el clúster 1 de la Figura 5.3 junto con el autobús 8307. Esto es un indicativo de un comportamiento similar de estos tres autobuses y diferenciable del resto de autobuses. A menor valor de la distancia DTW, obtendremos agrupaciones más pequeñas de autobuses que denotan una mayor similitud entre sus componentes.

5.2.3 Discusión de resultados

Tras la realización de las dos pruebas de evaluación primero se necesitaba saber si la acción de excluir el autobús 8308 del análisis era correcta por lo que se consultó el motivo por el que dejó de presentar lecturas tan abruptamente, y las personas encargadas de proporcionar los datos indicaron que este autobús no había estado funcionando desde julio por diversos problemas que forzaron llevarlo al taller. Esta información sustenta las acciones de ser retirado del análisis en una segunda prueba al descartarse por falta de datos, al menos por el lapso del tiempo en el que se ejecutaron las pruebas y análisis de esta familia

Con esa aclaración y las pruebas realizadas se tuvo el supuesto de que el autobús 8309 necesitaba ser revisado y entonces se procedió a consultar con las personas encargadas de

generar los datos. La revisión de los registros de fallos durante el servicio y la revisión de los volúmenes de combustible consumidos en el intervalo de tiempo estudiado confirmaron que el autobús 8309:

- Los consumos de combustible se mantuvieron en volúmenes normales.
- Presentaba fallos en el sistema eléctrico del motor, el cual es el que provee información sobre los datos de este estudio.

El autobús fue llevado a mantenimiento en varias ocasiones desde mayo a septiembre entre comprobaciones de rutina, pruebas y aplicación de soluciones al problema, llegando incluso a reiniciar el sistema eléctrico. A pesar de estas acciones se reportaron nuevos problemas de malfuncionamiento en el autobús 8309. Podemos decir, por tanto, que tanto el algoritmo K-means como el de *clustering* jerárquico han sido capaces de detectar un comportamiento anómalo en dicho autobús.

En las Figuras 5.6 y 5.7 se puede ver que los autobuses 8301, 8313 y 8314 formarían un clúster para un *threshold* >= 3 donde la distancia entre los tres autobuses es muy similar. Estos autobuses se agrupan de este modo por dos razones principales:

1. Estos tres autobuses son los que más datos de franjas horarias tienen de todo el grupo de la familia de los 50 IVECO, habiendo más de1400 lecturas registradas para cada uno.
2. nalizando las líneas recorridas por los tres autobuses, estos recorrieron predominantemente la línea 92; además, los autobuses 8313 y 8314 tienen una segunda línea en común, la línea 95. Ambas líneas mencionadas son bastante largas, haciendo un recorrido que va desde la playa hasta el otro extremo de la ciudad, cerca del hospital 9 d'Octubre y ambas tienen 39 paradas. Estas similitudes son las que hacen que se agrupen de forma tan cercana los tres autobuses.

5.3 Análisis por modelos: 43 SCANIA N250 E6 ZF

Esta familia de autobuses dispone de 9 autobuses con datos, por lo que al no tener gran cantidad de datos las agrupaciones pueden ser menos precisas.

5.3.1 Aplicación del algoritmo K-means

Primeramente, se analizan los resultados de aplicar el K-means en agrupaciones de 2 (C2) y 3 (C3) clústeres, cuyos resultados se muestran en la Tabla 5.4.

Los resultados del análisis del *silhouette* son los siguientes:

- Para dos clústeres, el valor medio del *silhouette* es 0.43
- Para tres clústeres, el valor medio del *silhouette* es 0.14

Las gráficas correspondientes a las dos particiones se muestran en la Figura 5.8 y Figura 5.9, respectivamente. En ambos casos el autobús 7117 se presenta en un clúster independiente y la mayor distancia para el resto de los autobuses del modelo se produce precisamente con el

autobús 7117, excepto en algún caso donde también se observa una distancia alta con respecto al autobús 7114. Cabe destacar los siguientes aspectos:

1. Al igual que en el caso del modelo 50 IVECO, el algoritmo K-means aísla un autobús determinado cuando se fuerza la generación de 2 o 3 agrupaciones. El objetivo es, por tanto, determinar si el aislamiento del autobús 7117 responde a algún tipo de anomalía del vehículo o se debe a otras razones relacionadas con la formación de los datos. Hay que tener en cuenta que el parámetro `n_clusters` fuerza al algoritmo K-means a encontrar una partición con dicho número de clústeres.

2. Cabe destacar que en el caso del modelo anterior 50 IVECO, las distancias más grandes respecto al autobús 8309 toman un valor de 9 mientras que en el caso del modelo SCANIA las distancias más altas son algo superiores a 3. Esto indica que los autobuses del modelo SCANIA muestran un comportamiento más similar entre sí.

BUS	C2	C3
7118	0	2
7119	0	2
7121	0	2
7122	0	2
7107	0	0
7109	0	0
7111	0	2
7114	0	0
7117	1	1

Tabla 5.4 *Clustering* de los autobuses del modelo 43 SCANIA N250 E6 ZF

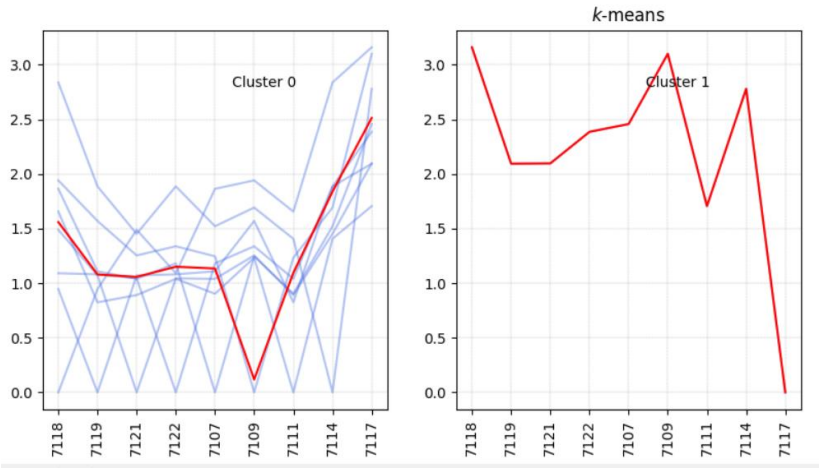

Figura 5.8 Partición de los autobuses del modelo SCANIA en dos clústeres

Figura 5.9 Partición de los autobuses del modelo SCANIA en tres *clusters*

Nos centramos en el análisis de dos clústeres que es el que ha devuelto un mayor valor del *silhouette* (0.43). Se procede a continuación a analizar el número de lecturas en franjas horarias disponibles para cada autobús mensualmente (ver Tabla 5.5), es decir, la longitud de las series temporales.

BUS	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	TOTAL
7118	182	144	132	29	112	334	933
7119	0	17	109	73	0	89	288
7121	83	184	149	80	79	77	652
7122	145	265	118	115	96	18	757
7107	84	94	228	58	72	26	562
7109	147	153	258	42	42	93	735
7111	46	178	112	10	67	15	428
7114	190	264	296	85	22	91	948
7117	20	0	74	14	7	0	115

Tabla 5.5 Distribución mensual de franjas horarias con lecturas de tasa de combustible 43 SCANIA

Con la cuantificación de la extensión de las series temporales expuestos en la tabla 5.5, se puede observar que el autobús 7117 es el que tiene el menor número de datos, de hecho, está en el caso extremo de la familia. Como ya se ha comentado antes, una gran diferencia en las longitudes de las series temporales afecta a los resultados de la clasificación reconociendo elemento de longitud más diferente del resto como diferente.

En este caso, el autobús que aparece diferenciado es precisamente el correspondiente a la serie más corta por lo que este podría ser el motivo de dicha diferenciación. Se procedió entonces a eliminar el autobús 7117 para ver los resultados que se obtenían, los cuales se muestran en la Tabla 5.6.

Los resultados del análisis del *silhouette* son los siguientes:

- Para dos *clusters*, el valor medio del *silhouette* es 0.3
- Para tres *clusters*, el valor medio del *silhouette* es 0.21

BUS	C2	C3
7118	0	2
7119	0	0
7121	0	0
7122	0	0
7107	0	0
7109	0	0
7111	0	0
7114	1	1

Tabla 5.6 *Clustering* de los autobuses del modelo 43 SCANIA N250 E6 ZF excluyendo el autobús 7117

Al forzar dos clústeres, el algoritmo K-means separa ahora el autobús 7114. Los valores del *silhouette* en esta prueba son bastante bajos como para ser considerados una buena separación, lo que indica que el autobús 7114 está mejor emparejado con el resto de los autobuses que formando un clúster independiente. Por lo tanto, a tenor de los resultados, no podemos extraer información concluyente que permita inferir un comportamiento anómalo. Un posible motivo por el que se diferencia el autobús 7117 es por la ausencia de datos de dicho autobús.

Por otro lado, cabe destacar que la mayor distancia del centroide del clúster 0 de la Figura 5.8 se produce con respecto al autobús 7117, que pertenece al otro clúster. Pero también se observa una distancia grande respecto al autobús 7114, que pertenece al mismo clúster 0. En la Tabla 5.5 se muestra que la serie más larga es precisamente la del autobús 7114, siendo la más corta la del autobús 7117.

5.3.2 Aplicación del algoritmo *clustering* jerárquico

Ahora se procede a replicar las pruebas hechas en el apartado anterior con el algoritmo de *clustering* jerárquico aglomerativo. El dendrograma resultante se muestra en la Figura 5.10. En este caso el dendrograma se ha realizado con distancia de 2 (eje Y), razón por la cual todos aquellos elementos que estén por debajo de esa distancia son considerados en un clúster único. Por encima del marcador el dendrograma ha sido señalado con un color diferente porque la distancia a los demás elementos está por encima del marcador 2 agrupándolo en un clúster diferente. Se puede notar que la línea azul es la que tiene la separación más grande respecto al grupo de color naranja, indicador de diferencias en el comportamiento del bus 7117. Puede observarse asimismo que la distancia que separa la formación de los diferentes clústeres es bastante pequeña, siendo la mayor distancia la que separa al autobús 7117 del clúster formado por el resto de los autobuses, pero aun así no es comparativamente muy grande. Si lo medimos con respecto al modelo 50 IVECO, el autobús 7117 está más cerca del resto de autobuses de su familia que el autobús 8309 de los autobuses de su mismo modelo.

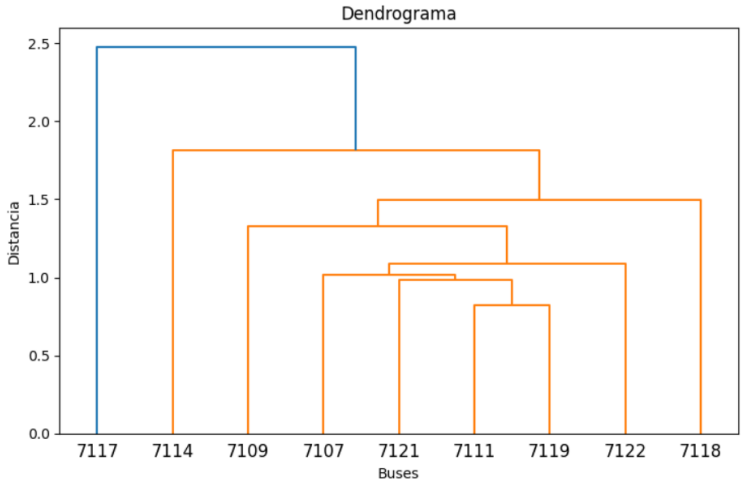

Figura 5.10 *Clustering* jerárquico para los autobuses del modelo 43 SCANIA

Al igual que con el algoritmo K-means, aplicamos el *clustering* jerárquico tras excluir el autobús 7117 por la falta de datos, lo que produce una serie temporal muy irregular, y el resultado se muestra en la Figura 5.11. Puede observarse que, efectivamente, existe una gran similitud entre todos los autobuses. La distancia que separa el autobús 7114 del clúster formado por el resto de los autobuses es solo ligeramente superior a la que separa al autobús 7109 del clúster formado por los autobuses 7107, 7121, 7111, 7119 y 7122.

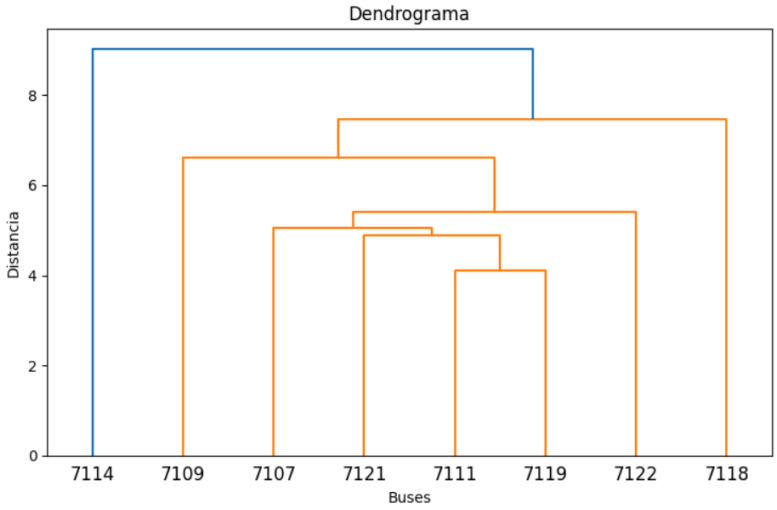

Figura 5.11 *Clustering* jerárquico para los autobuses del modelo 43 SCANIA tras excluir el 7117

Los autobuses 7111 y 7119 son los que mantienen la distancia más pequeña del grupo. Analizando las líneas recorridas por estos autobuses, el 7111 ha recorrido con mayor frecuencia la línea 11 y el 7119 ha recorrido más la línea 93. Ambas son líneas bastante largas y con varias paradas que van casi de extremo a extremo de la ciudad de Valencia, y son bastante similares entre sí. En el segundo nivel tenemos el autobús 7121 que también como recorrido predominante la línea 11.

Por otra parte, el autobús 7107 ha recorrido principalmente la línea 40 que es una línea corta de unos 5 km. Es normal que en distancias cortas la tasa de combustible aumente, lo que es equiparable a tener muchas paradas. El autobús 7122 no tiene una línea predominante ya que

tiene asignadas varias líneas, llegando a tener hasta 11 líneas distintas en un mes; ese es un indicador de que es un autobús de tipo apoyo que cubre los tiempos vacíos de otras líneas cuyos buses de tipo titulares no puede cubrir.

5.3.3 Discusión de resultados

Con los datos arrojados por los algoritmos se procede a describir las respuestas a las consultas hechas a los encargados de proporcionar los datos.

- El autobús 7117 es el que más días ha estado parado por lo que no se dispone de tantas lecturas como los otros autobuses.
- No se considera que el autobús 7117 tenga fallas, la separación detectada es debido a una falta de datos, por lo que se recomienda la exclusión de este en estudios futuros.
- Aunque esta familia tenga más autobuses que los estudiados en este apartado solo fueron incluidos nueve de ellos, porque son los únicos que disponen registros del combustible consumido. El hecho de que otros autobuses de la familia no tengan datos de consumo de combustible (es decir todos los que no entraron en este análisis) es porque en la configuración del sistema de recogida de datos en esos autobuses no está disponible las variables que se están estudiando.
- Al excluir el autobús 7117 del estudio, los algoritmos de *clustering* separan el autobús 7114. A pesar de que el valor del *silhouette* no es muy elevado, se procedió a revisar dicho autobús. Las verificaciones hechas en el autobús 7114 revelan un nivel de consumo de combustible normal. Por otro lado, en el registro de reparaciones reportados se rmencionan fallos en el sistema de filtrado de partículas, y tras este problema reportado que se repite varias veces a lo largo del registro se comenzaron a reportar fallos en el motor. Los fallos en el motor se reflejaron en problemas para cambiar de marcha, vibraciones en el volante y altas revoluciones en el motor, a esto se le conoce como avería APS.
- El algoritmo ha podido detectar un comportamiento inusual en el autobús 7114 pero, debido a la poca cantidad de autobuses disponibles para el estudio, la métrica de evaluación aplicada tiende a desestimar las agrupaciones conseguidas.

5.4 Análisis por modelos: 49 IVECO HEULIEZ GX 337

En este apartado se analiza los autobuses del modelo 49 IVECO, disponiendo únicamente de seis autobuses, por lo que al no tener gran cantidad de datos las agrupaciones pueden ser menos precisas.

5.4.1 Aplicación del algoritmo K-means

Aplicamos el algoritmo del K-means con agrupaciones de dos clústeres (C2) y tres clústeres (C3) clústeres, que se muestran en la Tabla 5.7.

BUS	C2	C3
9302	1	2
9303	0	2
9304	1	2
9305	1	1
9307	0	0
9308	1	0

Tabla 5.7 *Clustering* de los autobuses del modelo 49 IVECO

Los resultados del análisis del *silhouette* son los siguientes:

- Para dos clústeres, el valor medio del *silhouette* es 0.3
- Para tres clústeres, el valor medio del *silhouette* es 0.12

Los valores de *silhouette* obtenidos para esta familia de autobuses son bajos, lo que indica que la calidad del agrupamiento no es óptima, es decir, los autobuses de un clúster no son muy similares entre sí y disimilares de los autobuses de los otros clústeres. En Figura 5.12 se puede observar que la distancia que separa a algunos autobuses de este modelo llega hasta un valor máximo de 17.5. Esto contrasta, por ejemplo, con los autobuses del modelo 43 SCANIA de la sección 5.3, donde el valor máximo de DTW es 3.2, aproximadamente.

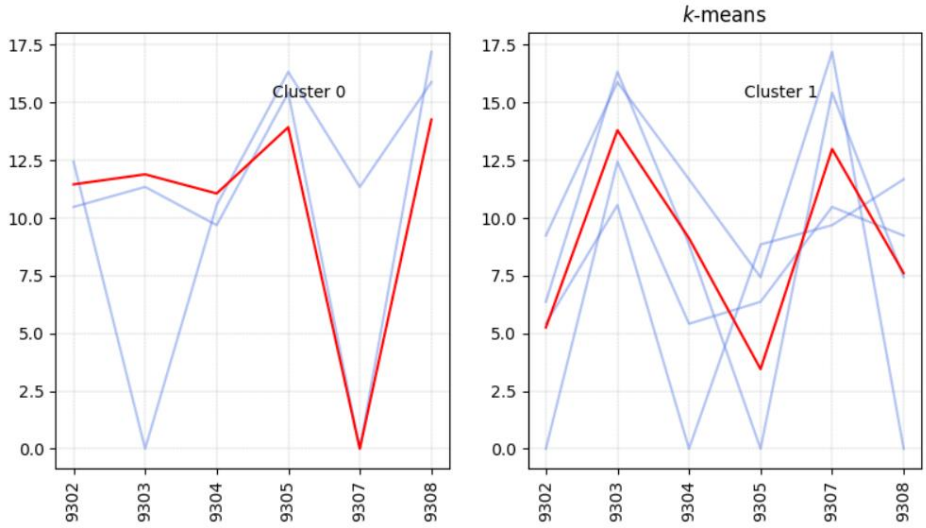

Figura 5.12 Configuración de los autobuses del modelo 49 IVECO en dos clústeres

La aplicación del algoritmo K-means no revela la formación de agrupamientos claramente definidos. Lo más llamativo es, precisamente, que la distancia entre los diferentes autobuses es alta. Sí se puede observar, como es lógico, que las distancias más grandes del centroide del clúster 1 se produce con respecto a los autobuses 9303 y 9307, que son los autobuses que están en el otro clúster. No se aprecia, por otro lado, diferencias significativas en las distancias que separan a los autobuses del clúster 1 de los autobuses 9303 y 9307.

El siguiente paso es pasar a un análisis cuantitativo para ver si algún autobús tiene una cantidad de datos que podría estar afectando el resultado de la evaluación. La cuantificación de datos puede apreciarse en la Tabla 5.8.

BUS	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	TOTAL
9302	0	0	270	220	40	85	615
9303	280	317	363	326	339	153	1778
9304	0	215	252	261	0	160	888
9305	3	88	203	255	209	7	765
9307	168	212	270	258	247	242	1397
9308	45	0	232	339	352	179	1147

Tabla 5.8 Distribución mensual de franjas horarias con lecturas de tasa de combustible 49 IVECO

Con una cuantificación de datos disponibles es notable que los autobuses 9302 y 9305 presenten casi la mitad de los datos que tienen sus pares. No se procedió a ejecutar nuevamente la evaluación sin estos autobuses ya que solo nos quedaríamos con cuatro autobuses a analizar. Por otro lado, se puede apreciar a partir de la Tabla 5.7 que las series temporales más largas son precisamente las correspondientes a los autobuses 9303 y 9307, particularmente destaca que son los autobuses que tienen un mayor número de datos en el mes de mayo.

5.4.2 Aplicación del algoritmo *clustering* jerárquico

Para poder reforzar o refutar los resultados obtenidos en el apartado anterior se procedió a ejecutar una evaluación de esta familia de autobuses con el algoritmo jerárquico con un valor del *threshold* 10. La representación de la configuración obtenida es mostrada en la figura 5.13

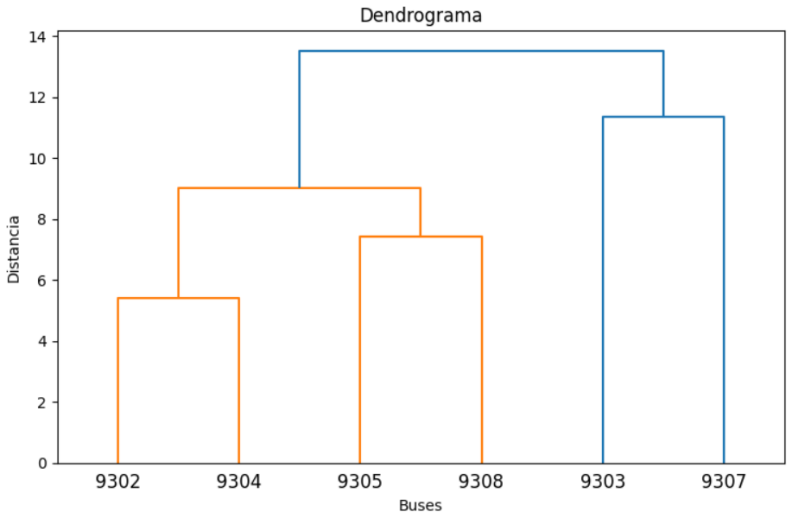

Figura 5.13 *Clustering* jerárquico para los autobuses del modelo 49 IVECO

Con una distancia de 10 la configuración jerárquica de esta familia tiene agrupaciones equilibradas, por ejemplo: los autobuses más cercanos son el 9302 y 9304 al tener la menor distancia, estos a su vez fueron agrupados con el 9308 y 9305 y marcados con un mismo color para luego ser agrupados con los autobuses restantes.

Estas agrupaciones no muestran un elemento anormal en la colección y esto puede deberse a lo pequeña que es esta familia de autobuses por lo que no es posible hacer separaciones óptimas de clústeres.

5.4.3 Discusión de resultados

La poca cantidad de datos puede ser un factor limitante a la hora de buscar particiones y por lo tanto anomalías en esta familia de autobuses. Por otra parte, es posible estudiar el comportamiento similar en estos autobuses. Para esta familia todos los autobuses recorren de forma predominante la línea 60 y a partir de esta similitud no es posible determinar otras características comunes entre otros grupos más pequeños de autobuses de esta familia; es decir, no hay un patrón de comportamiento reconocible en esta familia por lo que tampoco se puede encontrar anomalías.

Nótese que las distancias en el eje Y son considerablemente más grandes que en otras familias, indicador de que entre los autobuses de esta familia los comportamientos son menos similares entre ellos, menos de lo que se esperaría entre componentes de una misma familia

5.5 Análisis por modelos: CITARO

En este caso se decidió estudiar conjuntamente todos los autobuses del modelo CITARO, ya que los dos modelos de CITARO tienen pocos autobuses. Tras esta unión los datos siguen siendo escasos con una población de siete autobuses por lo que se espera que la calidad de los clústeres no sea de las mejores.

5.5.1 Aplicación del K-means

Como ya es habitual se presenta una evaluación de toda la familia y los resultados de la evaluación de la métrica del *silhouette* para dos y tres clústeres. Los datos se muestran en la Tabla 5.9 y en la Figura 5.14.

BUS	C2	C3
6224	0	1
6225	0	0
6237	0	0
6235	0	2
6234	0	0
6236	0	0
6238	1	1

Tabla 5.9 Configuración de los autobuses del modelo CITARO

Al aplicar la prueba del *silhouette* se obtuvieron los siguientes valores:

- Para dos clústeres, el valor medio del *silhouette* es 0.16
- Para tres clústeres, el valor medio del *silhouette* es 0.07

Las métricas para esta familia son bajas, lo que indica que no se ha encontrado un agrupamiento óptimo, o es consecuencia de los pocos individuos de estudio en esta familia. El bajo valor de la métrica del *silhouette* es coherente con la representación de la Figura 5.14 donde se muestra poca concordancia en la formación de los clústeres. En el clúster 0 se puede

ver que la forma del centroide no se ajusta de forma adecuada a la forma de las líneas azules de los autobuses, lo cual es un indicador de que este clúster no está bien definido. En cambio, el clúster 1 solo tienen un elemento, el autobús 6238 quedando totalmente aislado del resto.

Hay tres aspectos destacables:

1. Al igual que ocurre con el modelo SCANIA, los valores de la distancia DTW entre los autobuses son bajos, superando ligeramente el valor de 3.
2. En el clúster 0, y centrándonos en el bari-centro, se puede ver que está más alejado del autobús 6238, lo cual es obvio al estar este autobús en el otro clúster, pero también se observa que prácticamente a la misma distancia le sigue el autobús 6235, indicando que hay una diferencia de comportamiento en este autobús dentro de los elementos de su propio clúster 0.
3. En el clúster 1 se ve que el bari-centro está más alejado del autobús 6235, esto es un factor que puede llevar a suponer que el autobús 6235 es un posible candidato a revisión técnica.

Al hacer un análisis en cada clúster, los puntos 2 y 3 podrían apoyar la elección del autobús 6235 al estar este más distanciado en cuando al bari-centro se refiere. También debería analizarse si la separación del autobús 6238 en un clúster independiente responde a algún tipo de irregularidad de los datos, o anomalía del propio vehículo. Conviene recordar que al forzar la generación de dos clústeres, el algoritmo K-means ha separado el autobús 6238, pero dado que el *silhouette* es muy bajo y los autobuses del clúster 0 tampoco muestran un comportamiento muy cohesionado, esta separación podría no responder a ningún factor particular.

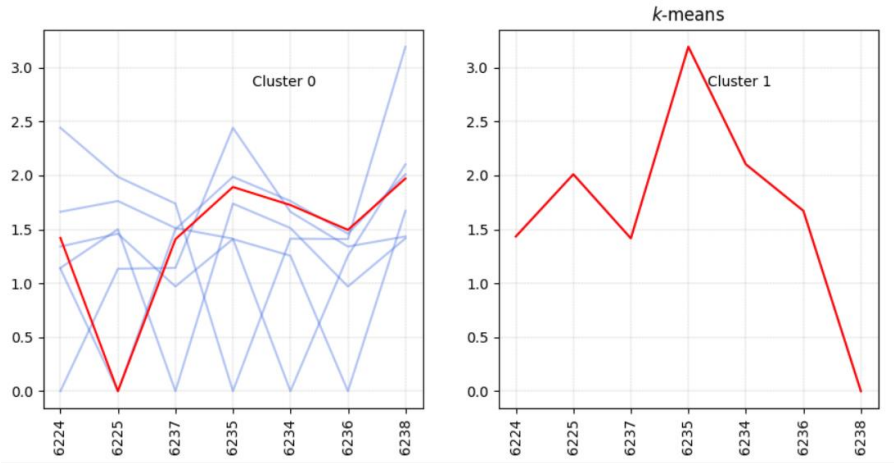

Figura 5.14 Configuración de los autobuses del modelo CITARO en dos clústeres

La Tabla 5.10 muestra la longitud de las series temporales para esta familia. A tenor del número de datos, se puede concluir que no hay series temporales con una ausencia pronunciada de datos. De hecho, en este caso, el número de autobuses es limitado, pero se dispone de un gran volumen de datos para cada autobús. En líneas generales, podríamos decir que las series están bastante equilibradas, aunque sí se puede observar que la serie más larga responde precisamente al autobús 6235. Respecto al autobús 6238, no hay ninguna indicación a partir de su serie temporal que pueda indicar un comportamiento distinto al resto de autobuses.

BUS	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	TOTAL
6224	88	149	140	55	255	175	862
6225	132	126	155	111	121	244	889
6237	158	225	203	39	215	136	976
6235	196	300	281	350	126	370	1623
6234	231	324	200	404	227	210	1596
6236	186	211	0	15	262	296	970
6238	169	169	216	36	161	146	897

Tabla 5.10 Distribución mensual de franjas horarias con lecturas de tasa de combustible CITARO

5.5.2 Aplicación del algoritmo *clustering* jerárquico

Para la evaluación del *clustering* jerárquico en esta familia de autobuses, se aplica la misma directiva que en familias anteriores, se evalúa a la familia en conjunto con un *thereshold*=1.5. La configuración de esta evaluación puede verse en la Figura 5.15.

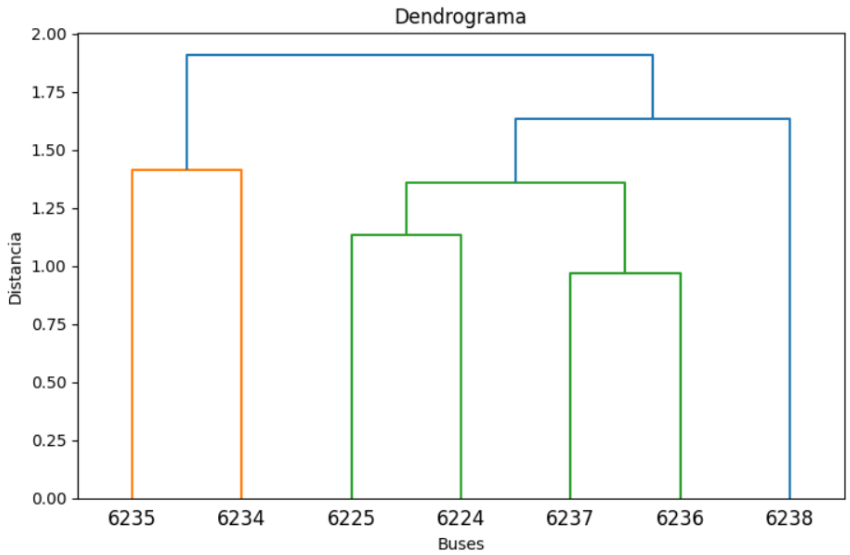

Figura 5.15 *Clustering* jerárquico para los autobuses del modelo CITARO

Con una distancia de 1.5 se observa que la línea naranja muestra que los autobuses 6236 y 6237 son los más cercanos entre sí, es decir, son los que tienen comportamiento más parecido entre sí. En el siguiente nivel aparecen juntos los autobuses 6225 y 6224 y a su vez esta dupla está agrupada al grupo anterior, estos cuatro autobuses están agrupados en un solo clúster de color verde. En el clúster de color azul se encuentra un solo elemento, el autobús 6238, coincidiendo con los resultados obtenidos con K-means expuestos en la figura 5.14. Finalmente, este clúster de color azul está conectado al clúster de color naranja, en el cual se encuentran los autobuses 6235 y 6234. El autobús 6234 no ha sido detectado en el apartado de K-means para esta familia, pero en base a lo obtenido en el *clustering* jerárquico se puede regresar a la figura 5.14 y ver que después del autobús 6235 los bari-centros se alejan del autobús 6234.

Cabe señalar que si bien los resultados de *clustering* jerárquico coinciden con el análisis hecho con el algoritmo K-means, el análisis jerárquico tiene resultados más claros y concisos que pueden leerse fácilmente en el dendrograma, a diferencia de la figura 5.14 obtenida con K-means que requirió más análisis.

Los autobuses 6235 y 6234 están agrupados porque ambos tienen asignadas líneas que en algún punto de su recorrido pasan por autovía, este escenario hace que recorran muchos kilómetros, pero gasten menos combustible al no haber paradas en que obliguen a tener ese consumo extra que exige un arranque en cada parada (ver más detalles en la siguiente sección). Sin embargo, el algoritmo K-means no ha detectado esta aparente peculiaridad como sí ha detectado la aplicación del algoritmo jerárquico. También es posible que esta particularidad se deba a que de entre todas las familias esta es la que menores distancias presentó en las gráficas resultantes de las pruebas, es decir, esta familia es la que tiene autobuses que más se parecen entre sí en cuanto a comportamiento se refiere.

5.5.3 Discusión de resultados

Como sucedió con la familia de los 49 IVECO, en esta familia se encuentra el mismo problema por los pocos individuos de estudio disponibles.

Con un inventario de las líneas más habituales recorridas por el conjunto de la familia se destacan que los autobuses 6234 y 6235 recorrieron predominante la línea 24 del Palmar que tiene 21 km. Los autobuses 6224 y 6225 recorrieron predominantemente la línea 60 que tiene entre 5 y 6 km. Los autobuses 6236 y 6237 recorren la línea 70 predominantemente, que tiene un recorrido entre 8 y 9 km. Estas distribuciones explican la razón de su agrupamiento. El autobús 6238 se distribuye entre varias líneas, pero se ve que recorre más las líneas 60 y 70 lo que explica su asociación con las líneas agrupadas en el clúster verde de la figura 5.15 ya que comparte comportamiento con los dos grupos formados por ese clúster. El hecho de estar aislado no implica en sí mismo un comportamiento anormal, esto se apoya en el valor muy bajo obtenido por el *silhouette* que indica la poca calidad de los *clusters* formados por el K-means, por lo que el análisis hecho en cada clúster es más fiable, es decir, hacer un análisis intra-cluster más que inter-cluster al tener un *silhouettte* tan bajo.

En el caso particular de los autobuses 6234 y 6235 se solicitó una revisión a las personas encargadas de proporcionar los datos y salió a relucir que estos dos autobuses son los que tienen niveles más bajos de consumo de combustible de la familia de los CITARO. Al recorrer la línea 24, de unos 20 km de longitud, estos pasan por un tramo de autovía, y las autovías se recorren a mayor velocidad por lo que reduce el consumo de combustible. Este detalle no indica un comportamiento particularmente anómalo indicador de falla en el motor, motivo por el cual el algoritmo K-means no pudo separarlos en un clúster, y el análisis tuvo que apoyarse con los resultados del algoritmo jerárquico.

5.6 Análisis global

Para finalizar con la aplicación de los algoritmos se decidió realizar un estudio global, este análisis permitirá tener una vista macro de la totalidad de buses que se están estudiando.

5.6.1 Aplicación del algoritmo K-means

Al tener bastantes datos en esta prueba no se consideró necesario exponer la tabla de asignación de clústeres, en cambio, sí se proporcionan los valores obtenidos en la prueba del *silhouette*.

- Para dos clústeres, el valor medio del *silhouette* es 0.24
- Para tres clústeres, el valor medio del *silhouette* es 0.15

En una vista global se puede ver la calidad de los clústeres no es buena comparada a la calidad obtenida con el estudio de familias por separado. La Figura 5.16 muestra los resultados de la aplicación del algoritmo K-means.

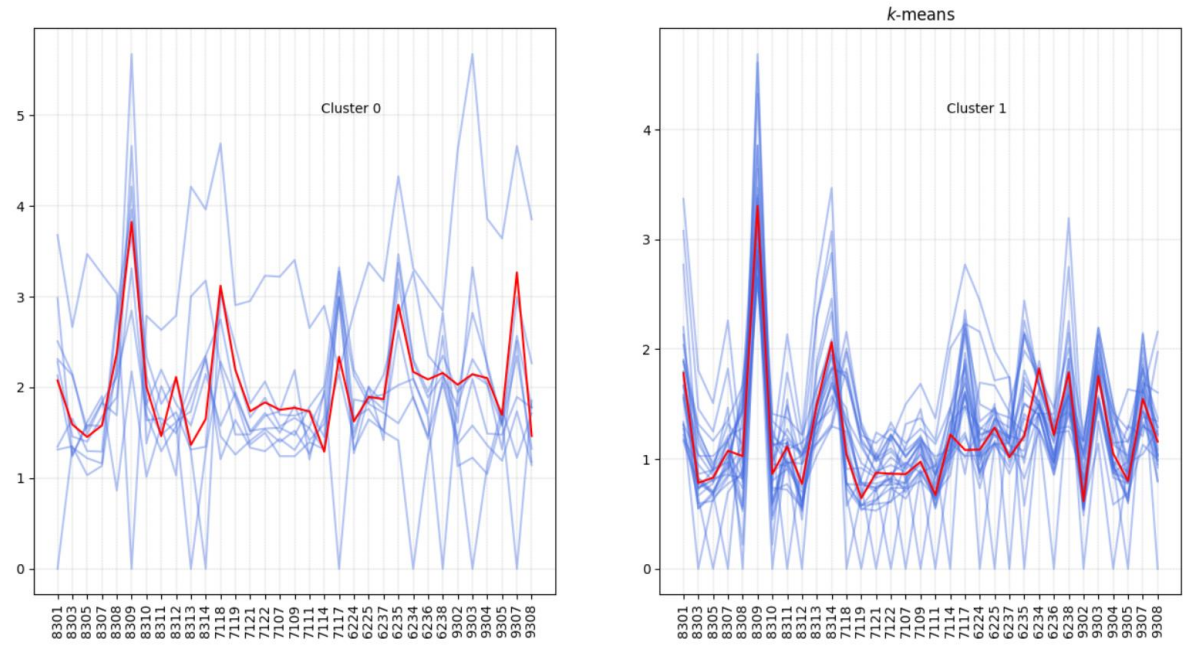

Figura 5.16 Configuración de dos clústeres de todos los autobuses

En este caso el clúster 0 tiene menos elementos que el clúster 1, pero se puede ver que las formas de las líneas azules se asemejan a la forma de los centroides rojos en ambos casos. En el clúster 0 están presentes varios de los autobuses identificados como anómalos o comportamiento discrepante en apartados anteriores. Se tiene al autobús 8309 identificado en la familia de los IVECO, el autobús 7117 de la familia de los 43 SCANIA, los autobuses 9303 y 9307 de los 49 IVECO y al autobús 6234 identificado en la familia de los CITARO. Si nos fijamos en los baricentros de ambos clústeres se puede ver que la mayor distancia se fija en el marcador del autobús 8309, indicando que aún en un análisis global este presenta un comportamiento muy diferenciado del resto de autobuses, poniéndolo como un candidato bastante robusto para una revisión técnica. De los autobuses CITARO solo el 6234 aparece en

el clúster 0, el 6235 aparece en el clúster 1, lo cual puede ser tema de discusión respecto a si el autobús 6235 tiene comportamiento anómalo o no. Es por eso que el análisis por familias es muy importante ya que detecta individuos que podrían pasar desapercibidos en un análisis global.

Estas apariciones en el clúster con menos elementos muestran de forma macro aquellos autobuses que podrían identificarse como autobuses de comportamiento anómalo y coinciden hasta cierto punto con los autobuses identificados en análisis más puntuales previamente realizados. A partir de eso se puede proponer que el análisis global sirva de apoyo al análisis por familias para poder validar los elementos discrepantes detectados.

5.6.2 Aplicación de *clustering* jerárquico

De forma semejante se somete a todos los autobuses de estudio a una evaluación por el algoritmo de *clustering* jerárquico, el resultado de muestra ilustrado en la figura 5.17, en este caso se utilizó un *threshold* de 2

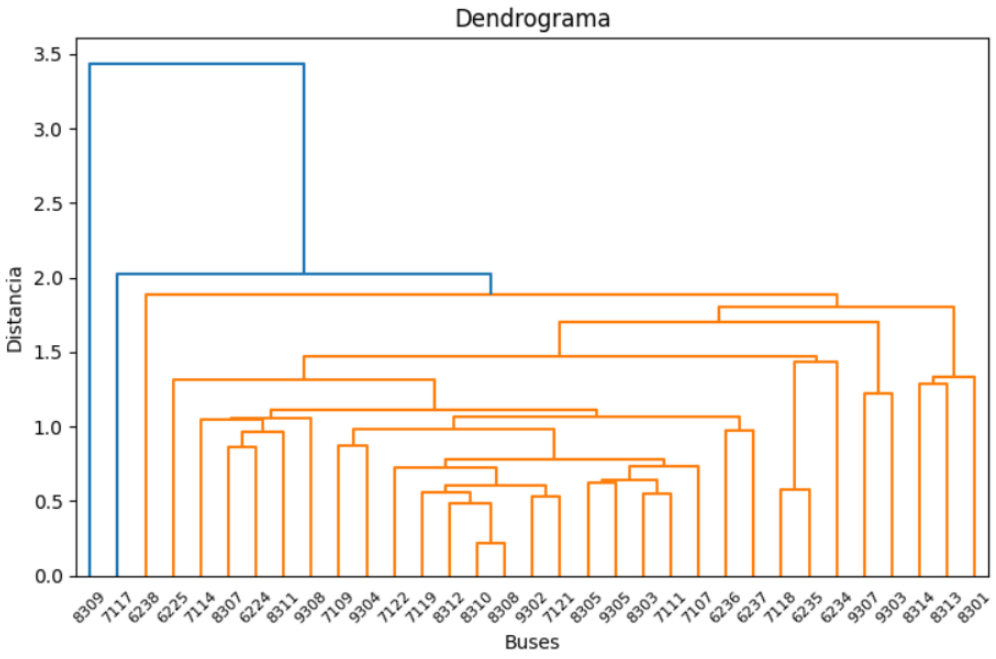

Figura 5.17 *Clustering* jerárquico para todos los autobuses

En la Figura 5.17 se puede ver que con un *threshold* de 2 los autobuses 8309 y 7117 fueron separados del grupo, en el caso del 8309 es más notoria la distancia a la que se encuentra, ya que la distancia es mayor que 3 cuando los otros niveles las distancias son mucho menores. En este análisis global los autobuses 8309 y 7117 de las familias 50 IVECO y 43 SCANIA respectivamente presentan una distancia considerable respecto a otras agrupaciones, esto podría señalar un comportamiento diferente en esos autobuses, mismos que fueron identificados en el análisis por familias.

5.6.3 Discusión de resultados

El análisis global ha permitido corroborar los resultados obtenidos en los análisis previos de las familias, en una primera instancia el algoritmo K-means ha separado varios autobuses en un clúster los cuales fueron identificados en mayor o menor medida en los análisis por familias; pero es con el algoritmo de *clustering* jerárquico que los autobuses 8309 y 7117 son marcados como diferentes de forma considerable. El autobús 8309 es el que está más distanciado del conjunto total, esto puede verse en la figura 5.15, como todas las líneas azules se alejan del marcador del autobús 8309, este detalle permite reforzar el supuesto de que el autobús 8309 requiere una revisión técnica ya que su comportamiento no es anómalo solo en su familia, sino también en la totalidad de los autobuses de este estudio.

5.7 Protocolo propuesto para la aplicación a demás casos

De acuerdo con el análisis realizado en los apartados anteriores podemos destacar varios aspectos. En primer lugar, los resultados coincidentes de la aplicación de los dos algoritmos de agrupamiento, K-means y *clustering* jerárquico, a cada una de las familias de autobuses. En segundo lugar, se puede concluir que el análisis revela la existencia de autobuses que tienen ciertas “peculiaridades” o que no se ajustan exactamente al patrón de comportamiento del resto de autobuses, siendo variadas las razones de este dispar comportamiento:

1. Por la irregularidad de la serie temporal del autobús (pocos datos o muchos datos) respecto al resto de autobuses
2. Por la existencia de una anomalía en el funcionamiento del autobús
3. Por la particularidad o especificidad de las rutas predominantes en los recorridos del autobús

Si bien no puede señalarse que la existencia de uno o varios autobuses que se desmarquen del resto sea debido a una anomalía del funcionamiento del autobús, sí podemos indicar que el señalamiento es debido a alguna causa de las indicadas arriba. Esta señal de aviso puede ser muy útil para el equipo encargado del mantenimiento de los autobuses con el fin de realizar un análisis más detallado que podría desvelar alguna anomalía.

En respuesta al análisis aplicado en el conjunto de datos proporcionados, este trabajo propone un protocolo, una serie de pasos a seguir por el personal técnico para una potencial detección de alguna anomalía. Este protocolo es presentado en forma de diagrama en la figura 5.18

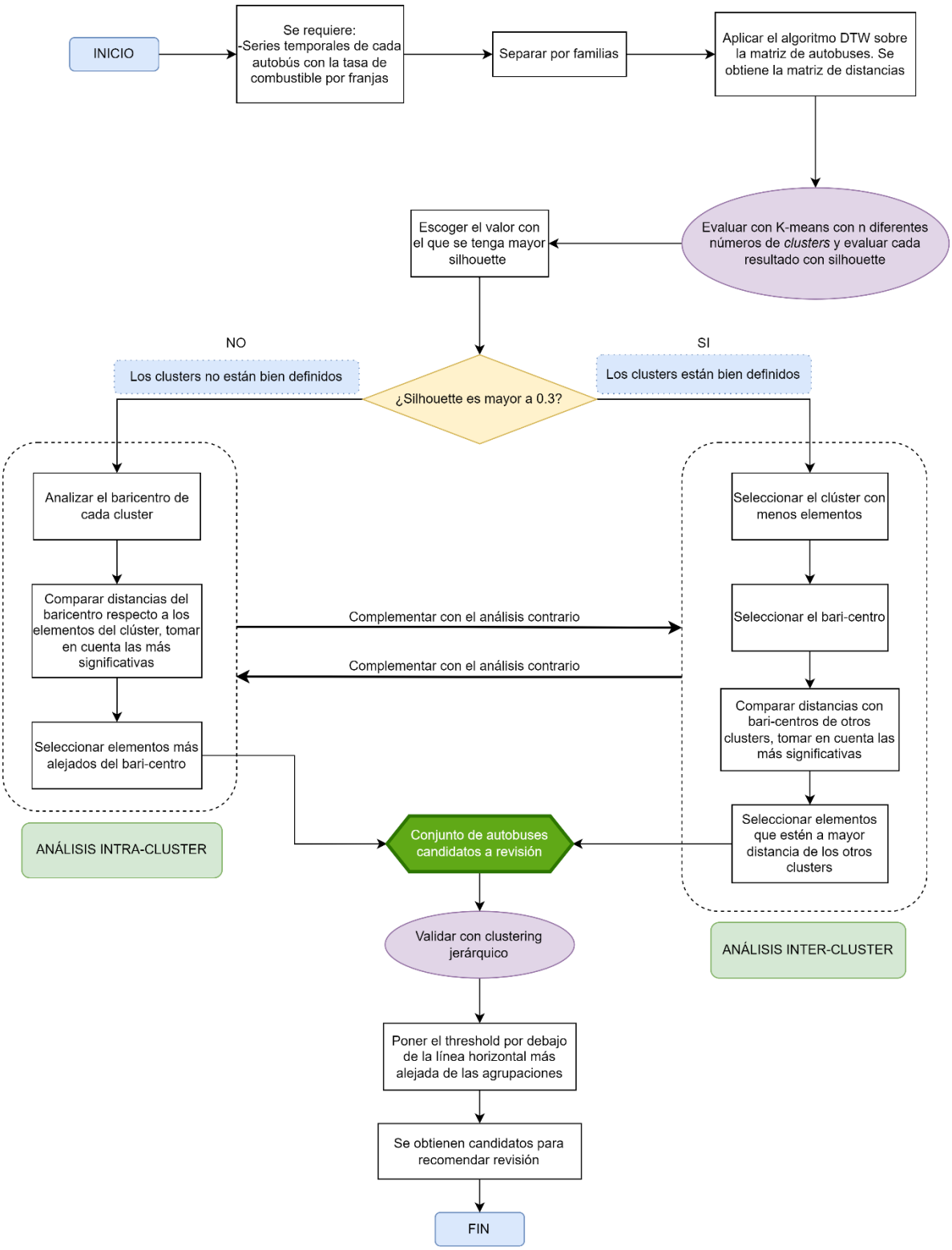

Figura 5.18 Protocolo propuesto para realizar el diagnostico en autobuses

La figura 5.18 describe los pasos y requisitos necesarios para poder identificar anomalías en una flota de autobuses, en resumen, y de forma general se abarcan las siguientes secciones

1. El tratamiento al que deben ser sometidas las series temporales para poder ser evaluadas con los algoritmos

2. La evaluación con el algoritmo K-mean y la toma de decisión sobre qué enfoque usar para seleccionar posibles candidatos.
3. El proceso de validación con el algoritmo de *clustering* jerárquico.

Identificar las distancias significativamente más grandes del bari-centro de un clúster a los elementos de los otros clústeres o a elementos del propio clúster puede ser un factor indicativo de un comportamiento inusual de dichos elementos. En otras palabras, el bari-centro de un clúster representa el comportamiento prototipo de las series del clúster y, por tanto, se plantean dos tipos de análisis:

- **INTRA-CLUSTER**. Si el valor del bari-centro se aleja comparativamente más de un elemento que de otros elementos del propio clúster puede ser un factor indicativo de alguna peculiaridad
    - En la Figura 5.2, el bari-centro del clúster 1 mantiene una distancia muy grande respecto al autobús 8308, perteneciente al mismo clúster 1 y, tal y como se ha visto, la serie del autobús 8308 es muy irregular por falta de datos.
    - En la Figura 5.11. el bari-centro del clúster 0 mantiene una distancia muy alta con el autobús 9303 perteneciente a dicho clúster; en el análisis ofrecido en la sección 5.4 se ha comprobado que la serie temporal del autobús 9303 es la más larga de toda la familia de autobuses 49 IVECO
    - Un caso similar se observa en la Figura 5.8, donde el autobús 7114 es el más alejado del bari-centro de su clúster, siendo el 7114 el autobús con mayor número de datos.
    - Lo mismo puede decirse en la Figura 5.4 respecto al autobús 8309
- **INTER-CLUSTER**. Si el valor del bari-centro se aleja comparativamente más de un elemento que de otros elementos pertenecientes a clústeres ajenos, esto puede ser un factor indicativo de alguna peculiaridad en dichos elementos
    - En la Figura 5.13 existe una mayor distancia del bari-centro del clúster 1 a los autobuses 6235 y 6234 que al resto de autobuses del clúster 0.
    - El mismo efecto se produce en la partición de dos clústeres de la familia IVECO, en este caso con respecto a los dos autobuses del clúster 0.

Lo que se debe identificar es a qué enfoque darle más prioridad en base al *silhouette* obtenido, pero no quiere decir que ambos análisis sean excluyentes entre sí. En el caso de que el *silhouette* sea alto, se hará un análisis inter-clúster como prioridad y de forma adicional podría hacerse el análisis intra-clúster para validar los resultados obtenidos, aunque este análisis no sea tan notorio ni significativo. De forma inversa se puede hacer un análisis inter-clúster como forma de apoyo al análisis intra-clúster si se obtiene un *silhouette* bajo.

6 Conclusiones y Trabajos Futuros

Para poder llegar a obtener el producto de este trabajo se determinó que la detección de comportamiento anormal en los autobuses es posible, siempre y cuando se inicie por un proceso de tratamiento de datos. Los datos que se manejan en la EMT deben ser revisados y procesados para determinar qué autobuses son válidos para ser analizados por el algoritmo planteado.

Las variables más importantes en el análisis fueron las señales 2, 3 y las coordenadas UTM que brindan información del combustible consumido y la distancia recorrida según sensor y según GPS respectivamente, estas señales son imprescindibles para el flujo de análisis ya que con ellas se calcula la tasa de combustible, esta variable es sobre la cual se comparan otros autobuses.

En la flota hay muchos autobuses que no disponen de la señal de consumo de combustible, esto se debe a que los sistemas de monitoreo de esos autobuses no tienen configurada la recogida de información de esa señal, esos deben ser descartados del análisis.

Con el algoritmo planteado al final del capítulo 5 se pudo evidenciar un funcionamiento anormal en el bus 8309, se recomiendo repetir el análisis tras el periodo en que se prevé solucionar los problemas reportados.

Un estudio por familias debe considerar dos puntos: la calidad de los clústeres formados y la cantidad de los autobuses estudiados. Una desventaja grande es la cantidad de individuos para el análisis ya que el modelo puede no ser preciso; pero esto permite el análisis más detallado individuo por individuo lo cual permite un análisis más detallado de los resultados.

Cuando se tiene un *silhouette* mayor a 0.3 es recomendable analizar el clúster que menos elementos tenga. Por otra parte, si el *silhouette* es menor a 0.3 se debe considerar la cantidad de autobuses estudiados, porque una cantidad pequeña puede afectar esta métrica, en ese caso se recomienda analizar cada clúster obtenido e intentar identificar cual es el elemento más distante a otros elementos, ese puede ser un indicador de que el autobús que cumpla con esa característica sea candidato a pasar por taller.

Se espera que bajo el supuesto de que en un futuro se tengan datos más regulares y se limite el problema de tener muchos valores nulos o con valores cero la detección de anomalías pueda ser hecha en periodos más cortos de tiempo. El objetivo de este trabajo no es predecir futuras fallas, sino detectar comportamiento anormal de forma temprana y recomendar revisiones a los autobuses resultantes de la aplicación del protocolo el cual se considera el producto final de este trabajo.

Al ser este un prototipo, se propone una serie de recomendaciones para expandir el estudio:

- **Evolucionar a un enfoque multivariable**. Se puede adicionar variables al estudio, como las variables relacionadas con la velocidad registrada, la variable de revoluciones del motor o incluso una variable dependiendo la época del año, por mencionar algunas.

- **Definir patrones de comportamiento.** Con ayuda de expertos se pueden definir ciertos tipos de comportamiento, por ejemplo: autobuses de rutas de más de *n* km, autobuses que recorren líneas con *n* número de paradas, autobuses que trabajan durante la mañana o tarde, etc. De esta forma se puede generar etiquetas que no requieren gran esfuerzo computacional para ser asignadas a cada bus, lo que permitiría aplicar algoritmos de clasificación supervisada e identificar mejor aquellos autobuses que no se ajusten a un patrón.
- **Aplicar otros algoritmos de clasificación**. Este trabajo se centró en analizar los datos con dos algoritmos principales, K-means y *clustering jerárquico*; pero aplicar otros algoritmos de aprendizaje no supervisado puede mejorar el protocolo propuesto. Un ejemplo de estos algoritmos puede ser los basados en densidad como el DBSCAN.

7 REFERENCIAS

[ALONSO20] Alonso del Saso, Javier. **Métodos de detección de anomalías y clustering en series temporales**. Universidad de Cantabria, Facultad de Ciencias. Trabajo de fin de máster. Septiembre de 2020.

[Chnadola09] Varun Chandola, Arindam Banerjee, Vipin Kumar. **Anomaly Detection: A Survey.** ACM Computing Surveys Volume 41, Issue 3, July 2009. Article No.: 15 pp 1–58. https://doi.org/10.1145/1541880.1541882

[CMT21] CMT-MOTORES TÉRMICOS, **Informe sobre la selección de las señales del BUS-CAN para su posible uso en mantenimiento predictivo y diagnóstico remoto.** Universitat Politècnica de Valencia. 8 de febrero de 2021.

[Everitt09] Everitt, B. S.; Landau, S.; Leese, M. (2001). **Cluster Analysis**. 4th Edition. Wiley, 2009.

[GVA21] Página del Institu cartografic Valencia. Disponible en: https://icv.gva.es/es/geocodificador.

[HDEI17] HDEI/BCEI Task Force. **FMS-Standard description.** Version 04, October 13, 2017

[ISO11898] ISO. **Road Vehicles – Controller Area Network (CAN), part 1: Data link layer and physical signalling.** Available in: https://www.iso.org/standard/63648.html. 2015.

[Kumar21] Ajitesh Kumar, **Elbow Method vs Silhouette Score – Which is Better?**. Available in: https://vitalflux.com/elbow-method-silhouette-score-which-better/#:~:text=The%20elbow%20method%20is%20used,cluster%20or%20across%20different%20clusters.

[Laptev15] Nikolay Laptev, Saeed Amizadeh, Ian Flint. Generic and Scalable Framework for Automated Time-series Anomaly Detection. KDD 2015: 1939-1947

[Lopez-Avila19] López-Avila, L., Acosta-Mendoza, N., gago-Alonso, A., López-Avila, L., Acosta-Mendoza, N. and gago-Alonso. **Detección de anomalías basada en aprendizaje profundo: Revisión**. Rev cuba cienc informat vol.13 no.3 La Habana jul.-set. 2019 Scielo.sld.cu. Available at: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992019000300107

[Massaro19] Alessandro Massaro, Sergio Selicato, Angelo Galiano. (2019). **Predictive Maintenance of Bus Fleet by Intelligent Smart Electronic Board Implementing Artificial Intelligence**. IoT. 2. 180-197. 10.3390/iot1020012.

[Parra19] Parra Francisco. **Estadística y machine learning con R.** Available in: https://bookdown.org/content/2274/portada.html. 25 de enero de 2019.

[Petitjean11] François Petitjean, Alain Ketterlin, Pierre Gançarski. A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition 44(3): 678-693, 2011.

[PREDICBUS21] **¿Una avería en el vehículo? La inteligencia artificial te avisa antes de que ocurra**. https://idescubre.fundaciondescubre.es/noticias/trabajan-en-un-programa-informatico-capaz-de-emitir-alertas/

[Smith21] G. M. Smith. **¿Qué es BUS-CAN y cómo se compara con otras redes de bus de vehículos?** Dewesoft.com. Disponible en: https://dewesoft.com/es/daq/que-es-el-bus-can. 6 de mayo de 2021

[StastsModel21] statsmodels v0.13.2. **Statistical models, hypothesis tests, and data exploration**. https://www.statsmodels.org/stable/index.html [accessed Enero 2021].

[Thudumu20] Thudumu, S., Branch, P., Jin, J. et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7, 42 (2020). https://doi.org/10.1186/s40537-020-00320-x.

[UV22] **Introducción al análisis de clúster**. Universitat de Valencia. https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm [accessed Febrero 2021].