Trabajo Fin de Máster

Machine Learning e IIoT aplicado al mantenimiento industrial

Alicia Rodríguez Oliveira

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Machine Learning e IIoT aplicado ao mantemento industrial
Título en español: Machine Learning e IIoT aplicado al mantenimiento industrial
English title: Machine Learning and IIoT applied to industrial maintenance.
Modalidad: Modalidad B
Autor/a: Alicia Rodríguez Oliveira, Universidade da Coruña
Director/a: Salvador Naya Fernández, Universidade da Coruña; Javier Tarrío Saavedra, Universidade da Coruña
Tutor/a: José González Pichel, Financiera Maderera S.A. (FINSA); ,
Breve resumen del trabajo: Este trabajo, desarrollado en conjunto con la empresa FINSA (Financiera Maderera S.A.), tiene como objetivo principal la detección de anomalías en equipos industriales, más concretamente en un ventilador industrial. Se entenderán las anomalías como la identificación de observaciones susceptibles de estar relacionadas con un mal funcionamiento de las instalaciones, alejándose del comportamiento esperado. En este trabajo se han utilizado técnicas univariantes para detectar anomalías tanto en el dominio del tiempo como en el de la frecuencia, además de técnicas multivariantes, únicamente en el dominio del tiempo, entre las que se encuentran diversos tipos de gráficos de control y de técnicas pertenecientes al *machine learning*.
Recomendaciones:
Otras observaciones:

Don/doña Salvador Naya Fernández, cargo 1 de la Universidade da Coruña, don/doña Javier Tarrío Saavedra, categoría 2 de la Universidade da Coruña, don/doña José González Pichel, Técnico de ingeniería de Financiera Maderera S.A. (FINSA), y don/doña , de , informan que el Trabajo Fin de Máster titulado

Machine Learning e IIoT aplicado al mantenimiento industrial

fue realizado bajo su dirección por don/doña Alicia Rodríguez Oliveira para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 05 de junio de 2023

Salvador Naya — Firmado digitalmente por Salvador Naya Fecha: 2023.06.05 08:45:20 +02'00'

El/la director/a:
Don/doña Salvador Naya Fernández

TARRIO SAAVEDRA JAVIER - DNI 53164914T — Digitally signed by TARRIO SAAVEDRA JAVIER - DNI 53164914T DN: c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=CERTIFICADO ELECTRONICO DE EMPLEADO PUBLICO, ou=MATEMATICAS, ou=18373, serialNumber=IDCES-53164914T, sn=TARRIO SAAVEDRA, givenName=JAVIER, cn=TARRIO SAAVEDRA JAVIER - DNI 53164914T Date: 2023.06.04 23:33:41 +02'00'

El/la director/a:
Don/doña Javier Tarrío Saavedra

GONZALEZ PICHEL, JOSE (FIRMA) — Firmado digitalmente por GONZALEZ PICHEL, JOSE (FIRMA) Fecha: 2023.06.05 18:52:21 +02'00'

El/la tutor/a:
Don/doña José González Pichel

RODRIGUEZ OLIVEIRA ALICIA - 53797154Q — Firmado digitalmente por RODRIGUEZ OLIVEIRA ALICIA - 53797154Q Fecha: 2023.06.03 19:41:04 +02'00'

El/la autor/a:
Don/doña Alicia Rodríguez Oliveira

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el/la autor/a declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Finsa

Santiago de Compostela a 29 de mayo del 2023

Desde el área técnica de Finsa, destacamos la contribución realizada por Alicia Rodríguez Oliveira a nuestro equipo durante la realización del Trabajo Fin de Máster "Machine Lerning e IoT aplicado al mantenimiento industrial". Resaltamos, en la perspectiva técnica, su capacidad de exploración y pensamiento crítico en la búsqueda de planteamientos para diseñar el enfoque del trabajo y el proceso de recogida de datos y análisis de los mismos, y en el aspecto más relacional, mencionamos especialmente su capacidad de adaptación y trabajo en equipo.

José González Pichel

Tutor empresarial del TFM

Agradecimientos

Este es un espacio ligeramente más informal que el resto del trabajo. No me gusta expresar demasiado afecto en público y menos sino conozco a las personas que lo leerán. Aún así me gustaría aprovechar este espacio para homenajear a las personas que están ahí día a día y a las que hicieron posible este trabajo.

En primer lugar a mis padres, ya que sin su apoyo no podría haber llegado nunca hasta aquí, ellos saben que son un pilar muy importante. A Manu, por estar en el día a día, que no siempre es fácil. A toda mi familia, la que viene predeterminada y la que se elige, ya que gracias a ambas soy como soy y estoy hoy escribiendo esto, quizá sin alguna de las personas que forma parte de ella mi vida hubiese sido muy diferente; por eso, gracias por ser parte de mi vida, por hacerme feliz en momentos cotidianos y por estar cuando hace falta, cada uno a su manera ha dejado un trocito en mi y, con lo cual, en este trabajo.

A Vero, por escuchar aunque no entienda de que le hablo y aportar su punto de vista siendo crítica y sincera. Gracias por estar, amiga.

También me gustaría agradecer a la empresa FINSA por su acogida, por acompañarme durante todo el proceso, por facilitarme siempre las cosas y hacerme sentir tan cómoda durante este periodo. En particular, me gustaría agradecer a Pichel y Gregorio por su ayuda, paciencia y apoyo en todo el trabajo; a Chus por los viajes, la compañía y todos los detalles; a Pilar por su punto de locura, los consejos y las galletitas; a Fran por amenizar los días con las conversaciones de café; al Fincar, a todo el equipo de OT, a Óscar, Lema, Julio, Lorena, y a todas las personas –me dejo muchas atrás– que hicieron el día a día de trabajo más agradable.

También a mis compañeros de máster: Álex, Sergio Antón, Andrea, Erik, Sergio, Fer y Jorge, por hacer más llevadera esta etapa de nuestras vidas. ¡Venga chicos que lo conseguimos!

Índice general

Resumen

Resumen en español

Este trabajo se ha desarrollado en conjunto con la empresa FINSA (Financiera Maderera S.A.), el objetivo principal es la detección de anomalías en equipos industriales. Entendiendo las anomalías como la identificación de observaciones susceptibles de estar relacionadas con un mal funcionamiento de las instalaciones, alejándose del comportamiento esperado. Esta es la primera investigación de un proyecto más ambicioso, enfocándose en el análisis de un ventilador con un solo flujo de trabajo –con un solo modo de trabajo–. La idea es poder extender el estudio a diversas máquinas con uno o varios flujos de trabajo para, por un lado, detectar las anomalías y poder realizar mantenimiento predictivo y, por otro lado, conocer el tiempo de vida que le queda a cada pieza para optimizar las paradas de mantenimiento preventivo. En esta investigación se ha trabajado con las mediciones de temperatura y vibraciones –ejes X, Y y Z– en 4 sensores del ventilador y con las mediciones del consumo del motor. Para el análisis se han utilizado técnicas univariantes para detectar anomalías tanto en el dominio del tiempo como de la frecuencia. Y, además, a partir de los datos definidos en el dominio del tiempo, se ha dado un paso más y se han analizado también en un contexto multivariante, a través de la aplicación gráficos de control y de técnicas pertenecientes al ámbito del *machine learning* –Support Vector Machines (SVM), Random Forest (RF), además de otros algoritmos para la detección de anomalías–.

Resumo en galego

Este traballo desenvolveuse en colaboración coa empresa FINSA (Financiera Maderera S.A.). O obxectivo principal é a detección de anomalías en equipos industriais. Entendendo as anomalías como a identificación de observacións susceptibles de estar relacionadas cun mal funcionamento das instalacións, desviándose do comportamento esperado. Esta é a primeira investigación dun proxecto máis ambicioso, enfocándose na análise dun ventilador cun único fluxo de traballo –cun único modo de traballo–. A idea é poder estender o estudo a diversas máquinas con un ou varios fluxos de traballo para, dunha banda, detectar as anomalías e poder realizar mantemento predictivo e, doutra banda, coñecer o tempo de vida que lle queda a cada peza para optimizar as paradas de mantemento preventivo. Nesta investigación traballouse coas medidas de temperatura e vibración –eixes X, Y e Z– en 4 sensores do ventilador e coas medidas de consumo do motor. Para a análise utilizáronse técnicas univariadas para detectar anomalías tanto no dominio do tempo como no da frecuencia. Ademais, a partir dos datos definidos no dominio do tempo, deuse un paso máis e tamén se analizaron nun contexto multivariante, mediante a aplicación de gráficos de control e técnicas pertencentes ao ámbito do *machine learning* –Support Vector Machines (SVM), Random Forest (RF), ademais doutros algoritmos para a detección de anomalías–.

English abstract

This work has been developed in collaboration with the company FINSA (Financiera Maderera S.A.). The main objective is the detection of anomalies in industrial equipment. Anomalies are understood as the identification of suspicious observations that deviate from the expected behavior. This is the initial investigation of a more ambitious project. In this work, the analysis machine is a fan and only has one workflow -it only works in a specific manner-. The aim is to extend this approach to different machines with one or multiple workflows to, on one hand, detect anomalies and enable predictive maintenance. And, on the other hand, stimate the remaining lifespan of each component in order to optimize preventive maintenance stops. This research focuses on the temperature and vibration measurements (X, Y, and Z axes) from four sensors of the fan, as well as motor consumption data. Univariate techniques have been applied for anomaly detection in both the time and frequency domains. Moreover, an additional step has been taken with the time domain data by analyzing it in a multivariate context through control charts and machine learning techniques such as Support Vector Machines (SVM), Random Forest (RF) and some anomaly detection algorithms.

Capítulo 1

Introducción al análisis de vibraciones

1.1. Tipos de mantenimiento

En el ámbito industrial, el mantenimiento de la maquinaria es un campo muy importante, ya que permite el correcto desarrollo del proceso de producción. Tradicionalmente, esta actividad se ha realizado de dos formas:

En la primera, la máquina opera de forma continua hasta que se produce algún defecto o avería que impide que siga trabajando; en ese momento se interviene, revisando y/o arreglando la máquina. Este tipo de mantenimiento recibe el nombre de mantenimiento correctivo, ya que se realiza para corregir un fallo.

En la segunda forma, se analiza la historia de cada máquina y se programan revisiones cada cierto tiempo para comprobar que todo funciona adecuadamente y renovar algunas piezas de la máquina –filtros, aceite, etc– para evitar fallos causados por su deterioro. Este tipo de mantenimiento se denomina mantenimiento preventivo, ya que se realizan cambios y/o revisiones para prevenir paradas inesperadas.

A partir de los años 80, un tercer tipo de mantenimiento se hizo muy popular, el mantenimiento predictivo. Este se basa en la idea de que la mayoría de las partes de la máquina producen alguna advertencia antes de fallar y que es posible detectarlas a través del análisis de mediciones de diferentes magnitudes como pueden ser vibraciones, temperatura, sonido, consumo, etc.

Este análisis consta de 3 fases fundamentales: (1) detección: reconocer el problema; (2) identificación: localizar la causa del problema; y (3) corrección: solucionar/eliminar el problema y su causa.

El mantenimiento predictivo permite: incrementar la productividad sin necesidad de aumentar el personal de mantenimiento, evitar paradas imprevistas que detengan y/o atrasen la producción, reducir los costes gracias a la detección de fallos prematuros y aumentar la vida útil de las máquinas (A. Royo et al., s.f.; Palomino, 2008; White, 1990)

1.2. ¿Qué es una vibración?

Durante mucho tiempo los operarios han utilizado técnicas auditivas y/o de contacto para comprobar que las máquinas funcionaban adecuadamente. Esta metodología continúa vigente y, aún en la actualidad, las vibraciones siguen siendo el indicador más representativo del estado de las máquinas, permitiendo detectar e identificar a través de ellas fallos, ya sean en periodo de desarrollo o ya evolucionados.

Podemos definir una vibración, según A. Royo et al. (s.f.) y Palomino (2008), como toda variación de una magnitud en el tiempo, que describe el movimiento o la posición de una máquina –o de algún

elemento de ella– en cualquier dirección del espacio desde su posición de equilibrio.

Las vibraciones pueden ser de 3 tipos:

- Vibraciones armónicas: es el movimiento más sencillo, se caracteriza por una onda senoidal, que puede estudiarse a través de un círculo trigonométrico –vector rotatorio–.

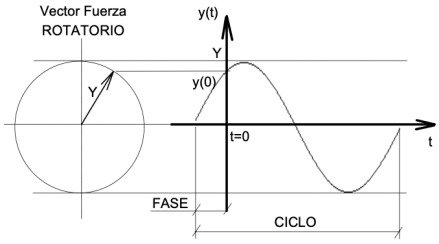


Figura 1.1: Representación de una vibración armónica a través de un círculo trigonométrico y del dominio del tiempo

Fuente: Palomino (2008, pp.11)

- Vibraciones periódicas: un movimiento que se repite cada cierto intervalo de tiempo.

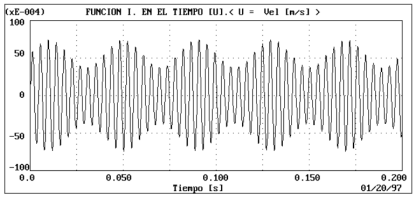


Figura 1.2: Representación de una vibración periódica

Fuente: Palomino (2008, pp.12)

- Vibraciones aleatorias: ocurren al azar y de forma continua.

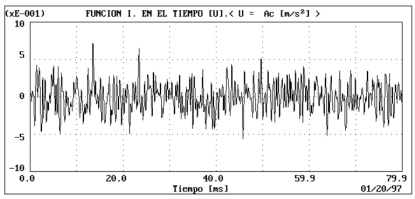


Figura 1.3: Representación de una vibración aleatoria

Fuente: Palomino (2008, pp.13)

En una situación ideal en la que las vibraciones de las máquinas no se viesen afectadas por factores externos, su representación en el dominio del tiempo –concepto en el que se ahondará en el apartado 2.2– debería seguir el movimiento armónico, es decir, serían vibraciones armónicas como las representadas en la Figura 1.1. En el mundo real suelen estar combinadas con vibraciones aleatorias (Figura 1.3).

Las vibraciones pueden observarse en el tiempo o en la frecuencia. Para poder medir los niveles de vibración es importante definir previamente las características de vibración, estas son: el desplazamiento, la velocidad, la aceleración y la frecuencia.

Según White (1990), el desplazamiento es un cambio de posición respecto a un sistema de referencia, la velocidad es la proporción de cambio en el desplazamiento –en otras palabras, lo rápido que cambia su posición– y la aceleración es la proporción de cambio de la velocidad. Estas unidades de vibración están relacionadas y se pueden obtener unas a partir de las otras: para convertir el desplazamiento en velocidad o la velocidad en aceleración, se debe hacer una diferenciación, es decir, se debe derivar; si, por el contrario, queremos convertir la velocidad en desplazamiento o la aceleración en velocidad, se debe integrar.

La frecuencia es el número de ciclos completos por unidad de tiempo. Su unidad natural es el ciclo por segundo –equivalente al hertzio (Hz)–, aunque las unidades de medida que se utilizan con más frecuencia son las revoluciones por minuto (RPM) –equivalente a ciclos por minuto (CPM)–. Estas unidades de medida están relacionadas, y para convertir los Hz en RPM se debe multiplicar por 60, para hacer la transformación inversa, de RPM a Hz, debe dividirse entre 60 (A. Royo et al., s.f.; White, 1990).

A veces, en vez de expresar las frecuencias en las unidades anteriores, es preferible hacerlo en órdenes o múltiplos de las RPM –el primer orden [1] es 1x, el segundo orden es 2x, etc.–. Esto se hace para poder realizar comparaciones entre mediciones tomadas en una máquina en diferentes momentos entre los que puede haber pequeñas variaciones de velocidad, ya que utilizando los órdenes, los armónicos de la velocidad se sitúan en las mismas posiciones en la gráfica pero no se tiene en cuenta la velocidad (White, 1990).

Las características descritas son importantes a la hora de analizar los problemas dado que la frecuencia nos indicará qué está mal en la máquina, mientras que las amplitudes y la velocidad informarán sobre la gravedad del problema (A. Royo et al., s.f.).

1.3. Transductores de vibración

Atendiendo a las técnicas tradicionales, ya mencionadas, utilizadas por los operarios para detectar fallos en las máquinas, existen dos tipos de mediciones: mediciones acústicas y mediciones de superficie. Se utiliza una u otra dependiendo de la magnitud que se quiera medir y/o la forma de recoger la energía de las máquinas.

La vibración acústica presenta una gran ventaja, y es que puede medir las vibraciones en todos los puntos de la maquinaria, aunque también tiene un inconveniente, y es que muchas veces en el ámbito industrial el sonido externo es igual o superior al sonido propio de la máquina que se quiere analizar. Cabe destacar, que no es lo mismo la señal sonora de la máquina escuchada a través de unos auriculares que la medición acústica que se produce por los cambios de presión del aire, provocados por las variaciones de forma y de posición de las piezas de la máquina.

Por otro lado, la medición de superficie debe ser realizada tanto de forma discreta –en algunos puntos concretos de la máquina– como de forma directa –entre la máquina y el dispositivo de medición– . Además, existen algunos instrumentos que permiten realizar una medición indirecta de superficies. Estos no tienen contacto directo entre la máquina y el dispositivo sino que cuantifican las vibraciones desde un punto de vista relativo y/o absoluto (Palomino, 2008).

Para medir las vibraciones suelen utilizarse transductores, estos son dispositivos que reciben las vibraciones o señales de las máquinas y producen una señal eléctrica que es una réplica del movimiento

[1] Estos órdenes son, por ejemplo, en una máquina que gira a 1500 RPM o CPM, cuando pase 1 minuto la máquina habrá completado un ciclo y realizado 1500 giros, 1500 será el primer orden (1x), cuando pasen 2 minutos, la máquina habrá realizado 3000 giros y ese será el segundo orden (2x) y así sucesivamente. Si nos interesa la medida en Hz o RPS, deben dividirse los RPM/60, siguiendo con el ejemplo anterior, 1500RPM/60=25RPS o Hz, 1x será 25Hz, 50 el 2x, etc

vibratorio de la máquina. Uno de los primeros transductores fue el dedo humano.

Siguiendo a Palomino (2008) y White (1990), podemos diferenciar 3 clases de transductores:

- Transductor de desplazamiento: Son de gran utilidad en la industria, mide el desplazamiento relativo entre dos partes de la máquina. Pueden ser de dos tipos:
  - Transductores de desplazamiento por contacto: Necesita del contacto físico con la superficie que vibra.

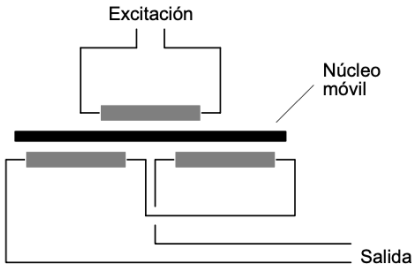


Figura 1.4: Representación de un transductor de desplazamiento por contacto, en concreto de un LVDT –transductor de desplazamiento lineal variable–

Fuente: Palomino (2008, pp.34)

  - Transductores de desplazamiento sin contacto: Utilizan la medición indirecta de superficies, su funcionamiento *"se basa en el encapsulamiento de un enrollado en su extremo libre, que al ser conectado a la unidad de alimentación del propio sensor, genera una señal de alta frecuencia que es transmitida (sin contacto) hacia la superficie del elemento cuyo desplazamiento se desea medir. Esto hace que se produzcan corrientes de Eddy cuya componente de directa (DC) es proporcional a la distancia entre el extremo del transductor y la superficie que vibra"* (Palomino, 2008, pg.34).

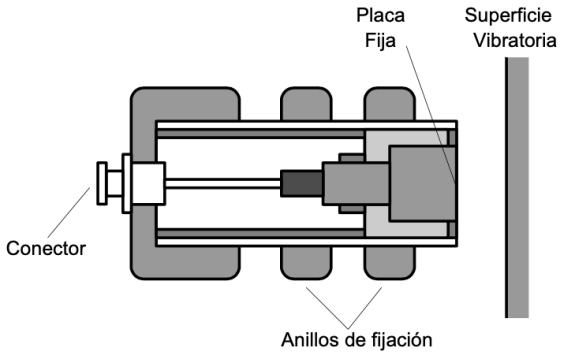


Figura 1.5: Representación de un transductor de desplazamiento sin contacto, en concreto de un transductor capacitativo

Fuente: Palomino (2008, pp.35)

- Transductor de velocidad: Fue uno de los primeros transductores de vibración, su idea básica consiste en una bobina y un imán permanente que se fija a la estructura del transductor de forma que crea un campo magnético. El movimiento relativo entre el campo magnético y la bobina produce una corriente proporcional a la velocidad del movimiento (puede verse en la Figura 1.6).

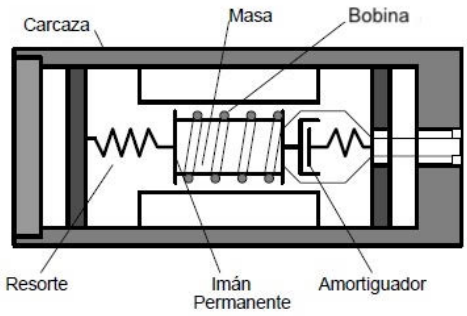


Figura 1.6: Representación de un transductor de velocidad
Fuente: Palomino (2008, pp.37)

- Acelerómetro: El acelerómetro piezoeléctrico es el transductor estándar para la medición de vibraciones en máquinas. Como se ve en la Figura 1.7, el elemento piezoeléctrico está sujeto entre la masa y la base, y cuando un elemento (una materia) está sujeta a una fuerza, se genera una carga eléctrica entre sus superficies que, en este diseño, será proporcional a la aceleración del transductor.

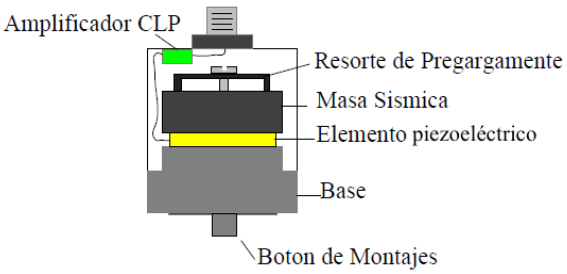


Figura 1.7: Representación de un acelerómetro
Fuente: White (1990, pp.55)

Capítulo 2

Análisis de vibraciones en ventiladores industriales

Los ventiladores tienen un rol relevante en la industrial actual, se podría decir incluso que son esenciales para ella. Su tarea habitual es el *"transporte o impulsión de aire, gases o vapores en sistemas de ventilación, sistemas de intercambio de calor o procesos que involucran combustión de gases. Pero también se puede aprovechar el flujo de aire para transportar otro tipo de productos, como polvos o pequeños sólidos"* (Trocel, 2021).

Los ventiladores industriales, en general, pueden clasificarse en 2 grandes grupos: los ventiladores axiales [1] y los ventiladores centrífugos. Cada uno tiene unas características específicas y unos fallos particulares. En este trabajo vamos a prestar especial atención a los ventiladores centrífugos, ya que es la máquina sobre la que se realiza la investigación.

2.1. Ventiladores centrífugos

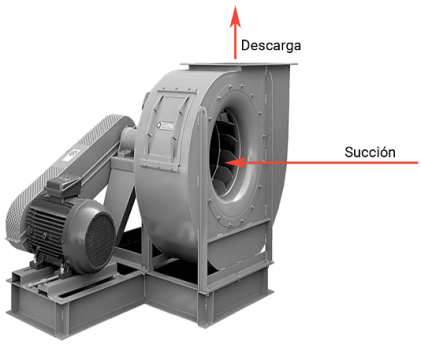


Figura 2.1: Flujo de aire/gas en ventiladores centrífugos

Fuente: Trocel (2021)

En la industria, un ventilador es *"una máquina que produce flujo de gas creando una diferencia de presión mediante el intercambio de momento de los álabes del ventilador con las partículas de gas. El impulsor del ventilador convierte la energía mecánica rotativa en energía cinética dentro del fluido gaseoso, energía que luego parcialmente se transforma en presión estática. En un ventilador del tipo centrífugo el fluido ingresa longitudinal al eje de rotación y se descarga en dirección perpendicular o radial, realizando el fluido un cambio de dirección de 90°"* (Trocel, 2021). Para tener una idea más visual puede consultarse la Figura 2.1.

Un ventilador centrífugo es una máquina relativamente sencilla, sus partes principales son (véase representación en la Figura 2.2):

[1] También se conocen como ventiladores de flujo axial.

1. Base o soporte estructural: puede ser de 2 tipos, base flexible o base rígida.
2. Impulsor y álabes: el impulsor, también conocido como rodete, está acoplado al rotor o eje del ventilador, está formado por una serie de álabes o paletas que son opcionales y su forma varía dependiendo del tipo de ventilador.
3. Carcasa: también conocida como voluta, es donde se direcciona el fluido desde la succión a la descarga.
4. Boca de visita: cuando se realiza el mantenimiento se utiliza para inspeccionar el equipo.
5. Rodamientos o cojinetes: dependiendo de las características del equipo pueden ser rodamientos o cojinetes planos.
6. Sistema de transmisión: puede ser un sistema de poleas/correas o un sistema de transmisión directa, depende de la velocidad a la que trabaje el ventilador.
7. Equipo conductor: puede ser un motor eléctrico de corriente alterna (AC), un motor eléctrico de corriente continua (DC) o turbinas de vapor.

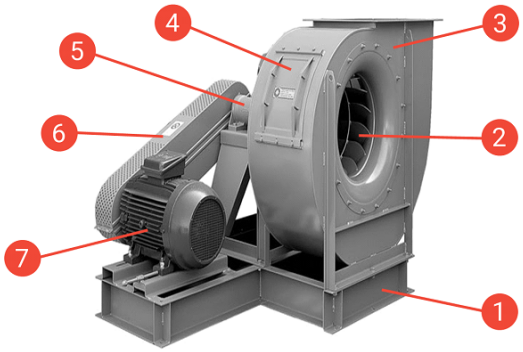


Figura 2.2: Partes principales de un ventilador centrífugo
Fuente: Trocel (2021)

En el caso del ventilador de estudio, utiliza un motor de inducción a corriente alterna (AC), puede verse en la Figura 2.3, su rotor es un electroimán, tiene *"barras de conducción en todo su largo, incrustadas en ranuras a distancias uniformes alrededor de la periferia. Las barras están conectadas con anillos (en cortocircuito como dicen los electricistas) a cada extremidad del rotor. Están soldadas a las extremidades de las barras. Este ensamblado se parece a las pequeñas jaulas rotativas para ejercer a mascotas como hamsters y por eso a veces se llama "jaula de ardillas", y los motores de inducción se llaman motores de jaula de ardilla."* (White, 1990, pp.119).

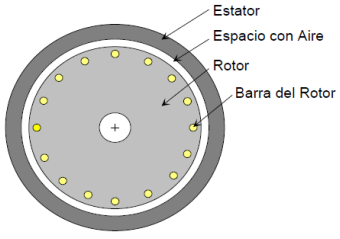


Figura 2.3: Representación motor de inducción AC
Fuente: White (1990, pp.73)

Las mediciones en este tipo de motores, siempre que sea seguro, deben hacerse en todos sus cojinetes o rodamientos y deben tomarse en 3 direcciones en cada apoyo del rotor, estas direcciones -véase Figura 2.4- son:

1. La dirección axial, la cual es paralela a la máquina o elemento en el que se mide -esta suele recoger golpes-.
2. La dirección radial o vertical, es la dirección desde el transductor hacia el centro de la máquina o parte de interés.
3. La dirección tangencial u horizontal, es tangente a la máquina o parte de interés.

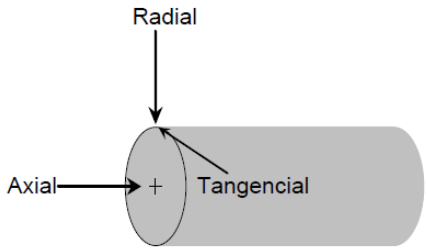


Figura 2.4: Alineación ejes de vibración
Fuente: White (1990, pp.120)

Las vibraciones pueden observarse en dos dominios: el dominio del tiempo y el dominio de la frecuencia. Cuando se está realizando un mantenimiento predictivo, el dominio del tiempo se utiliza generalmente para la fase de detección y el dominio de la frecuencia para la fase de identificación.

Esencialmente los datos son recogidos en el dominio del tiempo, y gracias a la transformada rápida de Fourier pueden ser "transportados.ªl llamado espectro o dominio de la frecuencia, este es el modo de relacionar los dos dominios (Palomino, 2008).

2.2. Dominio del tiempo

Se debe tener en cuenta que cada máquina tiene unas características diferentes y, por lo tanto, unos valores aceptables -rango a partir del que se consideran valores atípicos- diferentes. Esto no sucede sólo entre máquinas sino que ocurre en cada elemento dentro de una sola máquina, es decir, cada elemento tiene sus propios valores normales. Por ese motivo es importante realizar un análisis exploratorio previo, para el que se necesita disponer de un histórico de datos con el que realizarlo; en caso de que no se disponga de este, se puede evaluar la severidad de las vibraciones a través de la norma ISO 10816-1 [2].

Para poder localizar fallos en las máquinas o en elementos particulares de ellas, es muy importante determinar previamente los valores aceptables y las frecuencias de diagnóstico para cada elemento y/o máquina de interés. Una vez se conocen, se puede comenzar con el análisis.

El análisis en el dominio del tiempo es el uso de la representación gráfica de las vibraciones en función del tiempo para ayudar a diagnosticar problemas en las máquinas.

Generalmente se presta especial atención al valor pico (peak), al valor pico-pico (peak-peak), a la media bruta (average), al RMS, a la kurtosis y el factor de cresta -puede consultarse Figura 2.5-.

[2]En la práctica generalmente no se utiliza para las máquinas de estudio, por ese motivo no se va a profundizar en la norma ISO 10816-1, si fuese de vuestro interés puede consultarse https://es.scribd.com/document/377556606/Carta-de-Severidad-de-Vibracion-Norma-Iso-10816-1 .

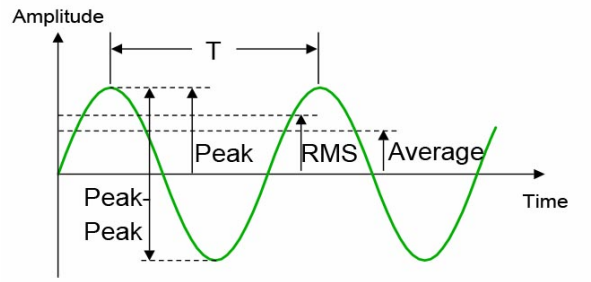


Figura 2.5: Gráfica explicativa medidas en el dominio del tiempo
Fuente: Universidad de Córdoba (2018)

En la Figura 2.5, T es el periodo de onda, definido por Palomino (2008, pg.17) como *"el tiempo necesario para un ciclo, o para un viaje ida y vuelta, o de un cruce del nivel cero hasta el siguiente cruce del nivel cero en la misma dirección (...) Se mide en segundos o milisegundos dependiendo de que tan rápido se cambie la onda"*.

Según la norma ISO 2041 (en Palomino, 2008, pp.30-31) *"el valor PICO es el valor máximo de una magnitud que varía durante cierto intervalo de tiempo"* y, *"el valor PICO-PICO (de un evento oscilatorio) es la diferencia algebraica entre los valores extremos de una magnitud que varía durante cierto intervalo de tiempo"*.

La media bruta, es un indicador estadístico que indica el valor promedio de cada variable:

$$Mean = \frac{\sum_{i=1}^{n} x_i}{N}.$$

El RMS -error cuadrático medio-, sirve para estudiar los errores de medida en los casos en los que interesa conocer la proporción del error y no si este es positivo o negativo. Además, el RMS otorga un mayor peso a los valores grandes que a los pequeños, ya que se calcula el cuadrado de todos los valores y, los de los números grandes son mucho mayores que los correspondientes a los valores pequeños:

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{N}}.$$

La curtosis -o kurtosis-, siguiendo lo expuesto por Metravib technologies (s.f.), es un indicador estadístico utilizado para caracterizar los pulsos [3] de una señal, es un parámetro adimensional que caracteriza el aplanamiento de una señal de densidad de probabilidad. Puede entenderse como un factor de forma, ya que su valor no se ve afectado por la amplitud de la señal; su expresión e interpretación son las siguientes::

$$K = \frac{1}{N} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)$$

- K= 1 implica señal cuadrática.
- K= 1.5 implica señal seno.
- K= 3 implica señal gaussiana.
- K >3 implica señal pulso.

El factor de cresta (FC) también es un parámetro adimensional. Sirve como indicador de la distorsión armónica, es decir, del impacto que tienen los fallos, el deterioro de piezas, etc. en la forma de

[3] Los pulsos son el resultado de la interacción entre dos ondas con frecuencias ligeramente diferentes.

onda. Cuanto mayor sea el valor del FC mayor será la distorsión, considerándose un valor alto a partir de 1.8. Además, también se puede utilizar para detectar posibles anomalías no cíclicas (Elvatron, 2020; Nuñez, 2013).

$$FC = \frac{valorpico}{RMS}.$$

2.2.1. Tipos de señales

Según White (1990), las señales del dominio del tiempo se pueden clasificar en los siguientes grupos:

- Señal estacionaria: Sus parámetros estadísticos son constantes en el tiempo.
  - Deterministas: *"Tienen un contenido de frecuencia y de nivel relativamente constante por un largo periodo de tiempo"* (White, 1990, pg.35). Son generadas por maquinaria rotativa, instrumentos musicales, y generadores de funciones eléctricas.
    - Periódicas: Tienen forma de onda con un patrón que se repite cada determinado intervalo tiempo.
    - Casi-periódicas: Tienen forma de onda que al observador puede parecerle periódica pero se repite en intervalos de tiempo variables.
  - Aleatorias: Son impredecibles en cuanto a su nivel (tendencia) y su amplitud (varianza/heterocedásticas), pero presentan características estadísticas relativamente uniformes en tiempo. Por ejemplo, la lluvia sobre el tejado.
- Señal no estacionaria: Sus parámetros estadísticos no son constantes.
  - Continuas: Son la vibración producida por una perforadora manual, y el sonido de fuegos artificiales.
  - Transitorias: *"Son señales que empiezan y terminan al nivel cero y duran una cantidad de tiempo finita. Pueden ser muy breves o bastante largos"* (White, 1990, pg.36). Por ejemplo: un golpe de un martillo, el ruido de un avión que pasa, o la firma de vibración de una máquina arrancando o terminando de funcionar.

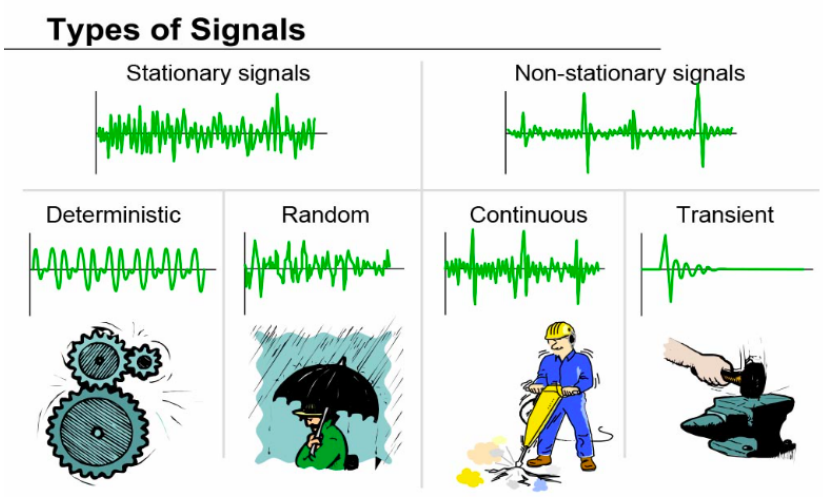


Figura 2.6: Tipos de señales
Fuente: Universidad de Córdoba (2018)

2.3. Transformada de Fourier

La transformada de Fourier tiene su origen en el siglo XIX. Es una transformación matemática empleada para convertir una señal representada en el dominio del tiempo en una señal representada en el dominio de la frecuencia (White, 1990). Tiene 4 formas:

- La serie de Fourier: *"Transforma una señal infinita periódica en un espectro de frecuencia infinito discrecional"* (White, 1990, pg.59). Funciona muy bien con señales periódicas (deterministas) pero no tan bien con señales aleatorias o continuas.
- La transformación integral de Fourier: Transforma cualquier señal continua de tiempo en un espectro continuo con extensión de frecuencias infinitas. Una característica interesante de esta es que *"un evento que abarca un periodo de tiempo corto se extenderá sobre un largo rango de frecuencias o viceversa"* (White, 1990, pg.59).
- La transformada discrecional de Fourier (TDF): Toma una señal de muestra discreta en el dominio de tiempo y a partir de ella genera un espectro de muestras discreto en el dominio de la frecuencia. Si la muestra tomada es representativa, el espectro producido será muy similar al verdadero espectro de "la población". Fue concebida por Gauss teóricamente en el siglo XIX, pero no se pudo llevar a la práctica hasta la llegada de los ordenadores.
- La transformada rápida de Fourier (TRF): Es un algoritmo que adapta la TDF para poder ser calculada por los ordenadores de forma rápida y eficaz.

En conclusión, la transformada de Fourier pasa señales del dominio del tiempo al dominio de la frecuencia representando en un gráfico de frecuencias la amplitud de las ondas simples. Gracias a esto nos permite saber cuál de las ondas tiene una mayor amplitud e interpretar los datos a través del espectro (Restrepo, 2020).

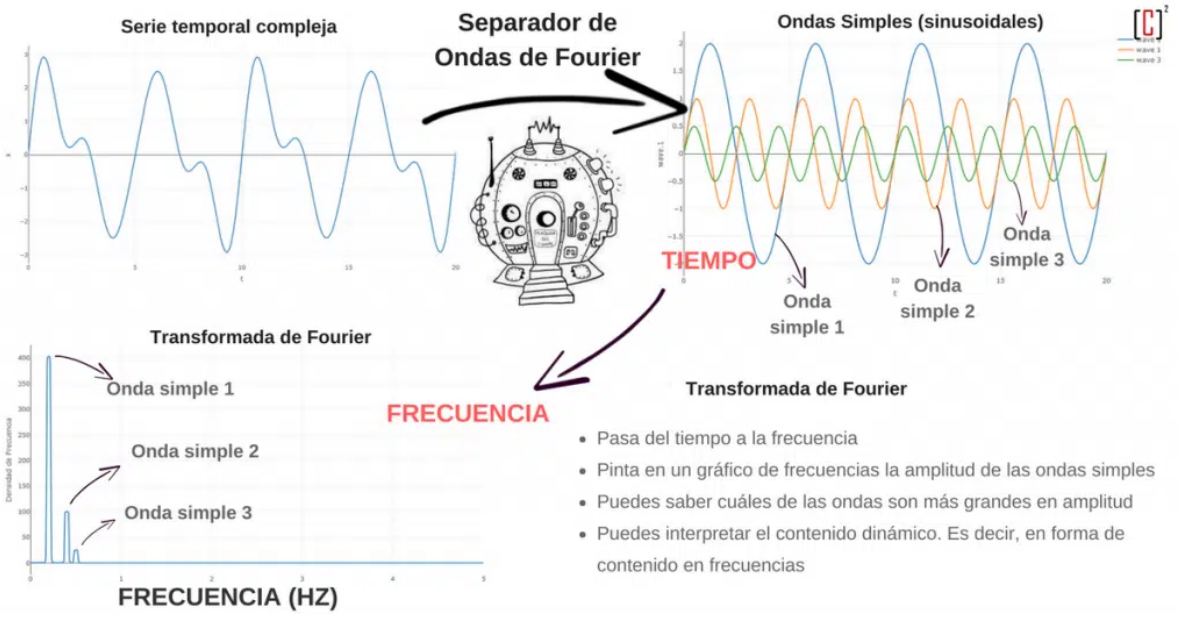


Figura 2.7: Representación explicación Transformada de Fourier
Fuente: Ollé (2017)

2.4. Dominio de la frecuencia

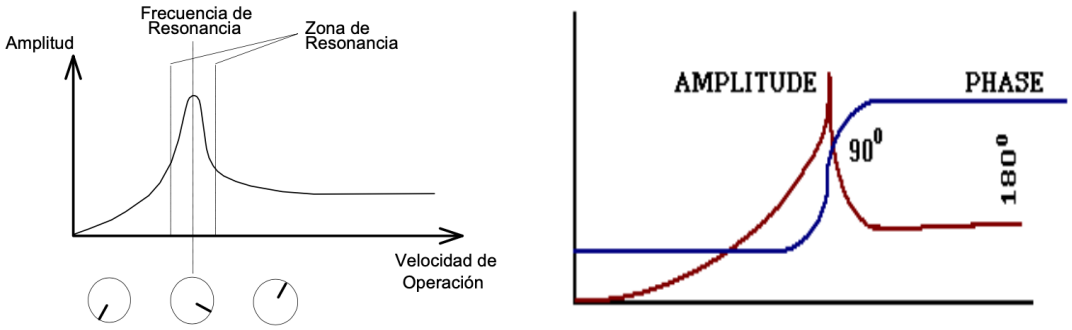
El análisis en el dominio de la frecuencia es el uso de la representación gráfica de los espectros para determinar si las anomalías detectadas en el dominio del tiempo son realmente fallos y, si es así, determinar sus causas y, si es posible, su gravedad.

Para la realización de este análisis se debe tener en cuenta la fuerte relación que existe entre la frecuencia y la velocidad angular de los elementos rotativos, ya que los problemas en las máquinas pueden ser detectados a través de las frecuencias iguales a la velocidad de giro o a sus órdenes (A. Royo et al., s.f.)

2.4.1. Tipos de frecuencias

Según Palomino (2008), dependiendo de las causas del origen de sus vibraciones existen 3 tipos de frecuencias:

- Frecuencias generadas: Son las que produce la máquina en su funcionamiento habitual, a veces se identifican como frecuencias forzadas o frecuencias de diagnóstico.
- Frecuencias excitadas: Son las frecuencias de resonancia de los elementos. La resonancia es el aumento sustancial inexplicable de la amplitud de las vibraciones en máquinas que presentan un estado favorable; este fenómeno se produce cuando coinciden alguna frecuencia de excitación con alguna frecuencia natural [4].



(a) Comportamiento de la fase en la zona de resonancia. Fuente: Palomino(2008, pp.89)

(b) Efecto de la resonancia en la fase. Fuente: Ssusera1e9de (2022, pp.18)

Figura 2.8: Representación frecuencias de resonancia

- Frecuencias producidas por fenómenos electrónicos: Hay que tener cuidado porque a veces se pueden observar frecuencias falsas o fuera de su ubicación correcta.

Además, dependiendo del tipo de señal de onda a partir de la cual se produzca el espectro, este será diferente. Si las señales son deterministas periódicas, producirán espectros con frecuencia discreta que serán una serie de armónicos [5], en cambio si son casi-periódicas, no habrá armónicos. Normalmente las vibraciones de las máquinas producen una combinación de estas (White, 1990).

[4] La frecuencia natural es la frecuencia a la que vibra un objeto/masa cuando se altera su posición de descanso/reposo.

[5] Esto quiere decir que todas las frecuencias serán múltiplos enteros de una frecuencia natural, es decir, se distinguirá claramente los picos en los órdenes (1x, 2x, etc.).

Para detectar las causas de las vibraciones, como anticipa Palomino (2008), se debe tener en cuenta que:

- En el ámbito industrial, la velocidad de operación suele medirse en RPM o FPM -frecuencia por minuto-.
- La frecuencia de las vibraciones se mide en CPM o Hz.
- Las frecuencias identifican el problema.
- Las amplitudes identifican la severidad relativa del problema. Estas pueden verse amplificadas por los efectos de solturas o resonancias, en cambio pueden atenuarse gracias a la influencia de la masa, la rigidez y/o el amortiguamiento presentes en el sistema de la máquina.

2.4.2. Causas más frecuentes de las vibraciones

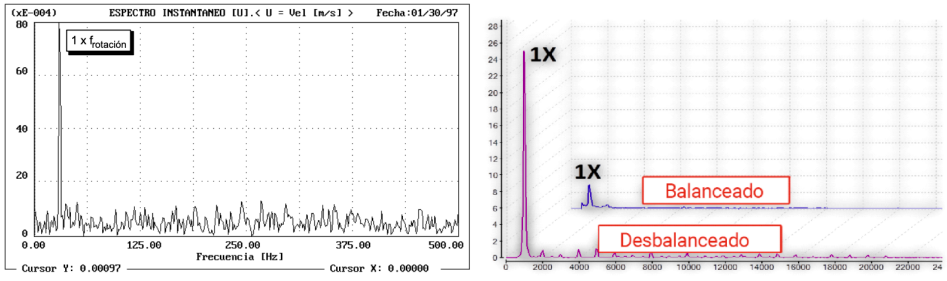
Según A. Royo et al. (s.f.), la mayor amplitud de vibración se encuentra en los puntos donde se localiza el problema, pero hay que tener cuidado, ya que muchas veces las vibraciones se transmiten a otros puntos de la máquina en los que no está presente. El análisis gráfico de los espectros permite identificar el tipo de defecto existente, aunque muy pocas veces aparece solo uno, esto complica la interpretación de la gráfica. La experiencia y el conocimiento de la máquina son dos factores clave para ser capaz de identificar la causa que produce una vibración inusual, además, conocer la apariencia de cada fallo puede facilitar esta tarea. Por ese motivo, a continuación se dan unas nociones básicas sobre las causas más frecuentes que producen alteraciones en las vibraciones y en su representación en el espectro:

Desbalance

Es la causa principal de aproximadamente el 40 % de las vibraciones excesivas. Puede definirse como la no concordancia entre eje radial y el eje tangencial principal de inercia del rotor (Palomino, 2008).

En ventiladores centrífugos suele producirse por deformaciones térmicas, pérdida de material por desgaste, erosión o corrosión, adherencia de partes extrañas o suciedad en el rotor, deformaciones por sobrecarga o torque, tratamientos de superficie o pintura o por la realización de procedimientos inadecuados de balanceo en el taller (Trocel, 2021).

Puede ser detectada porque se producirá, en la gráfica del espectro, un fuerte pico en 1x en la dirección radial, este pico tendrá una amplitud proporcional a la gravedad del desbalance y al cuadrado de la RPM (A.Royo et al., s.f.; Palomino, 2008; White, 1990).



(a) Fuente: Palomino(2008, pp. 65)

(b) Fuente: Trocel(2021)

Figura 2.9: Representación de un desbalanceo en el espectro

Desalineamiento

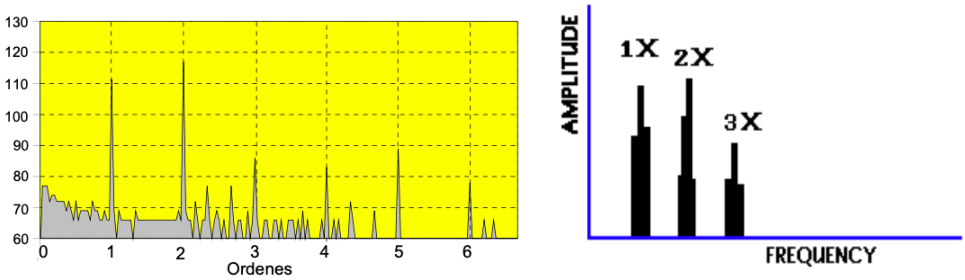
Es un problema muy común, ya que alinear dos ejes y sus rodamientos de forma que no se originen fuerzas que produzcan vibraciones tiene una gran dificultad, pero es la causa más fácilmente controlable y corregible -aunque para esto se necesita un trabajo mecánico preciso- (A. Royo et al., s.f.; Palomino, 2008).

En ventiladores centrífugos puede producirse debido a procedimientos inadecuados por parte del personal de mantenimiento, aplicaciones inadecuadas de los estándares o tolerancias de alineación, expansión térmica, deficiencia de la base soporte, distorsión de la base del motor, fallas en el acople, excesivo runout o deterioro o, por último, desgaste o deterioro de poleas y/o correas (Trocel, 2021).

Puede detectarse el desalineamiento porque se producen 3 armónicos en la frecuencia de rotación, es decir, picos en los órdenes 1x, 2x y 3x, es probable que la amplitud de 2x sea superior a la de 1x. Siempre que estén presentes los 3 primeros armónicos en el espectro, independientemente de que se esté dentro de los valores aceptables, se considerará un desalineamiento (A.Royo et al., s.f.; Palomino, 2008; White, 1990).

Hay 3 tipos de desalineamientos:

- Desalineación paralela: Las 3 primeras órdenes son significativas en la dirección tangencial –desalineamiento en el plano vertical– o en la dirección radial –desalineamiento en el plano horizontal–
- Desalineación angular: Las 3 primeras órdenes son significativas en la dirección axial.
- Desalineación general: Es una combinación de los tipos anteriores, las 3 primeras órdenes serán significativas en las 3 direcciones.



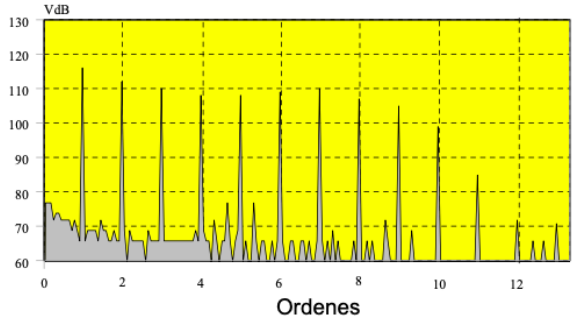
(a) Fuente: White (1990, pp.110) (b) Fuente: Ssusera1e9de (2022, pp.6)

Figura 2.10: Representación de un desalineamiento en el espectro

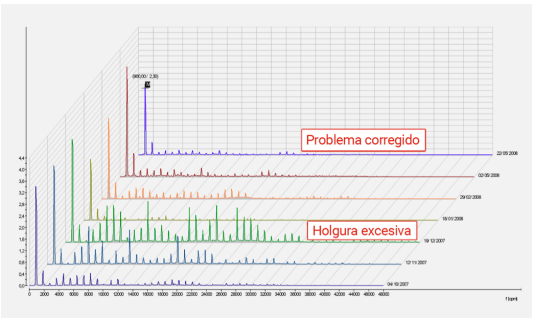
Hoguras rotativas

En ventiladores centrífugos suele originarse por desgaste de rodamientos y/o cojinetes, desgaste de cajeras o alojamiento de cojinetes y/o rodamientos, una mala calidad de la lubricación, un mal ensamblaje de partes como rodamientos, cojinetes, acoples, poleas o, por tener algunas partes defectuosas o inadecuadas (Trocel, 2021).

Puede identificarse por la aparición de armónicos excesivos en las 3 direcciones, que pueden extenderse desde 1x hasta por encima de 10x (White, 1990).



(a) Fuente: White(1990, pp. 117)



(b) Fuente: Trocel(2021)

Figura 2.11: Representación de holgura rotativa en el espectro

Problemas de rodamiento

Aquí pueden encontrarse muchos tipos de problemas, que en ventiladores centrífugos pueden deberse a fallos en la lubricación, un mal montaje de los componentes, existencia de algunas partes pueden deterioradas o con una mala calidad, un daño colateral de la desalineación excesiva o el desbalance o, simplemente, al desgaste por el funcionamiento habitual (Trocel, 2021).

Cada problema puede identificarse por unas características particulares, pero, para el análisis que aquí nos concierne, lo interesante es ver si existe un fallo en el rodamiento general –de este modo ya localizamos la causa y cambiaríamos el rodamiento– sin profundizar en el origen del fallo dentro del rodamiento. En general, lo que nos interesa es conocer que los problemas de rodamiento presentan frecuencias altas.

Resonancia

En el apartado 2.4.1 se ha definido lo que es la resonancia, aunque para explicar el problema que aquí nos concierne, tomaremos como sinónimos resonancia, frecuencia natural y velocidad crítica; aunque desde un punto de vista práctico no lo son, pueden utilizarse para describir la misma idea.

En ventiladores centrífugos, la resonancia se suele producir por: RPM de trabajo cercanas a la frecuencia natural de la base u otras partes del sistema, un debilitamiento estructural que modifica la rigidez de los soportes, influencia de máquinas externas sobre la frecuencia natural del sistema, cambios en las condiciones operacionales –RPM variables– (Trocel, 2021).

La resonancia puede identificarse a través de un aumento de la amplitud de las vibraciones de 10 a 30 veces respecto a los niveles de vibraciones habituales. En ventiladores, puede solucionarse la resonancia variando la sintonía de las partes resonantes, por ejemplo, la velocidad del motor, la masa o rigidez de los elementos, etc. (Palomino, 2008; Ssusera1e9de, 2022).

Fuerzas hidráulicas y aerodinámicas

Siempre están presentes en los gráficos de frecuencias de las vibraciones de ventiladores, bombas, turbinas, etc. Se produce por el paso de los álabes de los ventiladores como resultado de la fuerza aerodinámica sobre estos.

Las vibraciones que producen generan la conocida como frecuencia de paso (BPF [6]), la cual se define como el número de álabes multiplicado por la frecuencia de rotación del rotor portador (Palomino, 2008).

[6] La frecuencia de paso se conoce como BPF, por las siglas inglesas *Blade Pass Frequency*.

La frecuencia de paso puede tener una gran amplitud si el espacio entre la carcasa y el impulsor no es el adecuado, puede ser alta si el anillo del impulsor está desgastado y se agarrota en el eje y puede ser excesiva debido al rotor excéntrico[7] (Ssusera1e9de, 2022).

En ventiladores también pueden producirse turbulencias de flujo, estas son debidas a cambios de presión o de velocidad del aire en los conductos. En el gráfico de espectros se detecta porque aparecen vibraciones aleatorias de baja frecuencia posiblemente en el rango 50-2000 CPM.

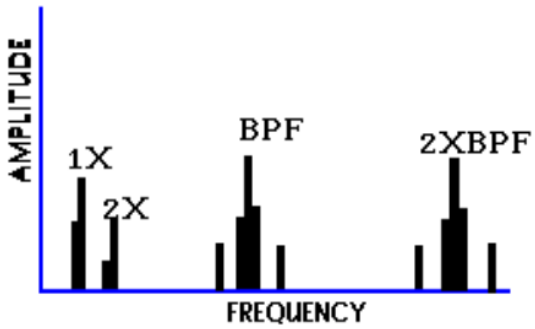


Figura 2.12: Representación de las fuerzas hidráulicas y aerodinámicas en el espectro
Fuente: Ssusera1e9de(2022, pp. 23)

[7]Según Ssusera1e9de (2022) un rotor excéntrico puede provocar largas vibraciones 1x en dirección axial. Para saber si es un rotor excéntrico se deben comparar las lecturas de los datos en la fase 0º y la de 180º

Capítulo 3

Introducción a las técnicas estadísticas aplicadas

Los métodos mencionados en el apartado anterior sobre el tratamiento de los datos en el dominio del tiempo y en el dominio de la frecuencia se han llevado a cabo en contextos univariantes. Pero, en el dominio del tiempo, se ha dado un paso más y se ha tenido en cuenta la información aportada por varias variables, además de la relación entre ellas, para poder detectar anomalías atendiendo. Este enfoque se ha llevado a cabo mediante la utilización de gráficos de control y la aplicación de algunas técnicas habitualmente englobadas dentro del *machine learning*.

3.1. Gráficos de control

Los gráficos de control –según Cabrera (2012)– *son herramientas utilizadas para controlar el comportamiento de una característica de calidad durante el proceso de fabricación.* Para que su utilización sea intuituva y fácilmente interpretable suelen realizarse gráficos, cuyos elementos principales son:

- Eje horizontal (X) y vertical (Y): Representan las observaciones en función del tiempo. Representando en el eje X el tiempo y en el eje Y el valor de las observaciones.
- Línea Central (CL): Indica el valor sobre el que suelen oscilar las observaciones.
- Límite de control superior o línea superior de control (UCL): Indica el valor máximo que deben tomar las observaciones en una situación normal, comúnmente definidos añadiendo tres veces la desviación típica a la línea central.
- Límite de control inferior o línea inferior de control (LCL): Indica el valor mínimo que deben tomar las observaciones en una situación normal, tradicionalmente definidos mediante la sustracción de un valor de tres veces la desviación típica respecto al valor central.

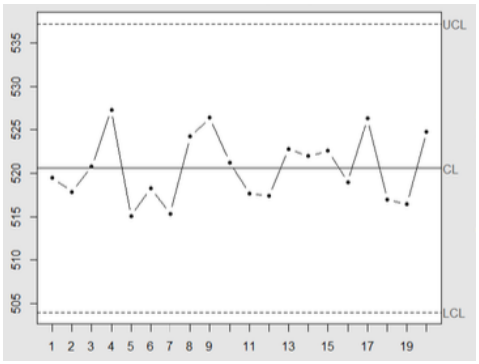


Figura 3.1: Representación de un gráfico de control bajo control

Fuente: Naya y Tarrío (2021)

Si se sobrepasan los límites de control podríamos estar ante una posible anomalía.

Para realizar correctamente el monitoreo estadístico de procesos, como se explica en Barbeito et al. (2017), se realiza el control estadísico de la calidad en 2 fases:

FASE 1 o fase retrospectiva: Es la etapa exploratoria sobre el histórico de datos. En ella se recolectan, estudian y analizan los datos para obtener información sobre la máquina o proceso estudiado. Se intenta establecer los valores normales y los valores aceptables utilizando gráficos de control y manteniendo una tasa de falsas alarmas aceptable para la situación de estudio. Esta fase suele realizarse mediante la técnica de prueba-error para seleccionar puntos fuera de control, establecer las causas de porqué están fuera de control -lo más común es que se trate de una posible anomalía, se deba a la falta de gaussianidad o sea debido al azar- y llevar a cabo medidas correctivas -generalmente la eliminación del punto fuera de control-. Se repetiría el proceso hasta que se consiga que todos los puntos estén dentro de los límites de control –proceso bajo control–, definiendo de este modo el conjunto de datos de referencia utilizados para ajustar la distribución de los datos –muestra de calibrado–, así como sus parámetros desconocidos –si se trata de una familia paramétrica– y comenzar la fase 2. En esta fase se estiman los límetes de control naturales del proceso, UCL y LCL.

FASE 2 o fase prospectiva: En ella se monitorean nuevos datos y se comprueba que el proceso continua bajo control. Las muestras tomadas en esta fase se representan en un gráfico con los límites de control fijados en la fase 1 y se estudia la variabilidad del proceso en función de los parámetros fijados en la fase 1.

Cabe resaltar que los límites de control son diferentes a los límites de especificación, estos últimos son los límites fijados por las personas que representan el target, es decir, a las que se dirige el estudio o, alternativamente, aquellos límites fijados por la empresa o por la normativa.

Los gráficos de control utilizados en este trabajo han sido multivariantes, en concreto, se ha utilizado el gráfico T^2 de Hotelling y el gráfico MEWMA.

3.1.1. T^2 de Hotelling

Este gráfico discrima observaciones según su probabilidad de ocurrencia, es decir, considera como un dato anómalo aquel que se aleja de los valores normales; parte de la idea de que si se conoce la distribución de una variable aleatoria se pueden conocer también los valores que tienen alta y baja probabilidad de ocurrencia, pudiendo, de este modo, definir un intervalo con los valores aceptables para cada variable. En el caso de estudio puede indicar que algo en el ventilador no funciona correctamente y ayudar a identificar así un posible fallo.

El gráfico T^2 de Hotelling, fue desarrollado por Hotelling en 1947, se utilizó por primera vez en la segunda guerra mundial y, en la actualidad, sigue siendo uno de los métodos de análisis de control de calidad más utilizados con datos multivariantes. Generalmente se denominan gráficos de Hotelling.

Consideremos un proceso de control con d variables aleatorias que se distribuyen en $\mathbb{R}^d$ según una normal d-variante con vector de medias μ y matriz de varianzas-covarianzas Σ. En la práctica se parte de la idea de que μ y Σ son desconocidas y se estiman mediante una muestra de calibrado -muestra de la fase 1- por $\bar{x}$ -vector de medias estimado- y S -matriz de varianzas-covarianzas estimada-.

La generalización multivariante, suponiendose normalidad, del estadístico t es

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}},$$

donde n es el tamaño muestral y $\bar{x}$ el vector de medias de la muestra de estudio; el numerador de la

ecuación representa la diferencia a probar y el denominador el error estándar [1]. Si se eleva al cuadrado la ecuación anterior obtenemos el siguiente estadístico,

$$t^2 = \frac{(\bar{x} - \mu)^2}{\frac{S^2}{n}} = n(\bar{x} - \mu)(S^2)^{-1}(\bar{x} - \mu),$$

que es posible generalizarlo como

$$T^2 = n(\bar{X} - \bar{\bar{X}})^t(S)^{-1}(\bar{X} - \bar{\bar{X}}).$$

Como se ha demostrado, el estadístico T^2 de Hotelling es la generalización multivariante de la T de Student. En cuanto a su significado, el estadístico T^2 de Hotelling mide la distancia de Mahalanobis entre el vector de medias esperado ($\bar{\bar{X}}$) y el vector de medias observado ($\bar{X}$) teniendo en cuenta la matriz de varianzas-covarianzas (S).

El estadístico sigue una distribución F de Snédecor con d y $(mn-m-p+1)$ grados de libertad, para un nivel de significación α; esta distribución es asimétrica, por ese motivo para obtener la probabilidad de que un dato sea considerado como una anomalía se define un cuantil superior denominado límite de control. Además, se supone que la estimación de S es no singular, que la matriz de datos X no contiene valores perdidos, que el número de variables o características a estudiar oscila entre 2 y 10 y que, como mínimo, se toman 20 muestras de cada variable.

El cálculo del límite de control depende de la fase del proceso de control en la que nos encontremos, si estamos en fase 1 (fase retrospectiva), el límite de control se calcula como:

$$UCL = \frac{d(m-1)(n-1)}{(mn-m-d+1)} F_{\alpha,d,mn-m-d+1},$$

en cambio, si estamos en fase 2 (fase prospectiva), se calcula como:

$$UCL = \frac{d(m+1)(n-1)}{(mn-m-d+1)} F_{\alpha,d,mn-m-d+1}.$$

En esta fase, el criterio para decidir si una observación se considera como anomalía o no se basa en comparar el valor de T^2 con el límite de control establecido; si el valor de T^2 es superior al límite de control se considerará como anomalía, en caso contrario se considerará como observación que sigue un comportamiento normal.

La monitorización de las máquinas suele realizarse en un largo periodo de tiempo, esto deja abierta la posibilidad a que los comportamientos habituales de las máquinas se vean modificados, ya sea por el desgaste de alguna pieza o incluso porque tiene diversos flujos de trabajo. Por ese motivo, es habitual trabajar con submuestras que permitan controlar que el proceso se desarrolla en condiciones aceptables; el tamaño de las submuestras puede ser siempre el mismo –tamaño unitario– o variar según la situación –tamaño variable–. En este caso, cada submuestra obtendrá sus propios valores para los parámetros estimados y se utilizarán para calcular el estadístico T^2 correspondiente así como el UCL que se aplicará al gráfico de control.

Como se ha enunciado al principio, el gráfico T^2 de Hotelling presupone que los datos de estudio siguen una distribución normal multivariante –o univariante– y si es así funcionará de forma adecuada aportando resultados satisfactorios. Pero, en la práctica, los datos obtenidos de mediciones de maquinaria industrial no acostumbran a seguir esta distribución. Esto no quiere decir que no se puedan utilizar estos gráficos, pero debe tenerse en cuenta ya que puede provocar que se cometan más errores; los analistas deben presuponer que algunas de las probabilidades calculadas pueden no ser correctas, existiendo algunas observaciones que se clasifiquen como anomalías sin serlo –por tener una probabilidad más baja que la real- y anomalías que no se clasifiquen como tal pero si lo sean -por asignarsele una probabilidad más alta que la real– (Naya y Tarrío, 2021; Sebas, 2020; Vaamonde, 2019).

[1]. *El error estándar de medida es la desviación estándar de los errores de medida asociados a las puntuaciones observadas de un test, para un grupo particular de examinados"*(Gempp, 2022, pp.117).

3.1.2. MEWMA

MEWMA son las siglas de *Multivariate Exponentially Weighted Moving Avarage Control Chart*, es decir, es el gráfico de control de medias móviles ponderadas exponencialmente para el caso multiariante; es la extensión de los gráficos de control univariantes EWMA. Este se dice que es un gráfico con memoria, ya que para evaluar si un proceso se encuentra bajo control se basa en la media ponderada de los vectores observados, es decir, de todas las submuestras estudiadas.

La extensión multivariante propuesta por Lowry(1992, en Naya y Tarrío (2021)) toma la siguiente expresión:

$$Z_i = \Lambda X_i + (I - \Lambda) Z_{i-1}, \qquad 1 \leq i \leq n,$$

donde X_i es el vector de medias muestrales, Λ es la matriz diagonal formada por los valores $\lambda \in (0, 1)$ para las distintas variables -en la práctica toma el mismo valor para todas las variables, por defecto acostumbra ser 0.1- e I es la matriz identidad de dimensión d -número de variables-.

La información de los Z_i se recogen en el siguiente estadístico:

$$T_i^2 = Z_i^t \Sigma_{Z_i}^{-1} Z_i,$$

siendo $\Sigma_{Z_i}^{-1}$ la matriz inversa de varianzas-covarianzas de los Z_i, que puede ontenerse a través de Σ_{Z_i}, esta puede calcularse como:

$$\Sigma_{Z_i} = \frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2_i})]\Sigma,$$

en la que Σ es la matriz de varianzas-covarianzas original.

En la práctica, Σ es estimada a través de la muestra de calibrado -en la fase 1- por S; permitiendo obtener el límite de control de una forma similar al caso de la T^2 de Hotelling. Tras esto, se realiza el gráfico de control representando los valores de T_i^2 junto al límite calculado.

Los gráficos MEWMA presentan propiedades análogas a los gráficos EWMA del contexto univariante. Es decir, los gráficos MEWMA son gráficos con memoria [2], por lo que detectarán de forma más eficiente pequeñas desviaciones o cambios con respecto a la media –vector de medias– a la que se supone que se supone que el proceso está bajo control. En el ámbito multivariante, los gráficos con memoria son más eficientes para detectar cambios con una magnitud menor a dos desviaciones típicas, mientras que los gráficos de Shewhart son más eficientes para detectar cambios mayores a dicha cantidad. Por otro lado, se ha estudiado el comportamiento de los gráficos tipo EWMA y Shewhart en escenarios de no normalidad, resultado más robustos los primeros. Teniendo en cuenta lo anterior, con el objetivo de detectar el mayor número de verdaderas alarmas y la menor cantidad de alarmas falsas, estos dos tipos de gráficos suelen aplicarse en conjunto (Montgomery, 2020; Naya y Tarrío, 2021; Sebas, 2020; Vaamonde, 2019).

3.1.3. Información complementaria

Los gráficos anteriores pueden aportar una gran contribución para la detección de anomalías, pero para su correcto uso e interpretación deben tenerse en cuenta varias cuestiones (Vaamonde, 2019).

En primer lugar, el correcto calibrado de los gráficos de control ya que:

- No se contruyen bajo el concepto de anomalía, sino sobre la teoría estadística de la probabilidad. Es decir, se encargan de detectar anomalías basándose en los cambios de estructuras de los datos. Para que esta acción se realice de forma correcta debe existir relación entre el comportamiento de los datos de la muestra de la fase 1 –con la que se fijan los parámetros y se establece el UCL– y

[2] La cualidad de memoria del gráfico MEWMA varía en función del valor del parámetro λ seleccionado. Cuanto más próximo a 1 se encuentre λ, su capacidad de memoria será menor e identificará bien cambios grandes, por el contrario, si el valor de λ está próximo a 0, su capacidad de memoria será mayor y trabajará bien detectando desajustes pequeños.

el comportamiento esperado por la maquinaria en la fase 2. Si la máquina tiene diferentes flujos de trabajo deben ajustarse diferentes modelos de comportamiento sino se estarán detectando falsas anomalías, ya que la máquina funcionaría correctamente pero serían los parámetros los que estarían mal ajustados.

- No se deben añadir anomalías en la muestra de calibrado. Para esto se deben aplicar los gráficos de control a la propia muestra de calibrado antes de utilizarlos con nuevas observaciones. Si se observa algún valor atípico en la fase 1, deben buscarse sus causas, corregirlo y calibrar de nuevo el gráfico.

Como se ha recalcado en apartados anteriores, lo más importante es conocer bien la máquina. Por ese motivo, para asegurarnos de que los gráficos de control representan el funcionamiento normal de la máquina se debe observar la evolución del equipo en un intervalo grande de tiempo, esta es la única forma de conseguir ajustar un modelo de detección de anomalías efectivo.

En segundo lugar, la correcta interpretación de resultados ya que no siempre que un punto aislado sobresale de los límites de control se trata de una anomalía, existen algunas excepciones:

- La falta de normalidad de los datos provoca que los gráficos de control funcionen como una aproximación. Esto puede provocar que en algunas ocasiones cuando algún punto sobrepase los límites de control se deba a esto. Solo será atribuible a esta causa si la distancia entre los puntos que sobresalen y el límite de control es pequeña.
- Alguna anomalía puede deberse a eventos concretos, por ejemplo a la realización de pruebas en el equipo. No todas las anomalías serán fallos; por ese motivo la labor de los técnicos es fundamental, ya que serán los encargados de revisar la máquina y confirmar si realmente los avisos de anomalías son fallos.
- La teoría de la probabilidad sobre la que se construyen los gráficos de control establece que algunos puntos se saldrán fuera de los límites de control debido al azar; el número de puntos que puede sobresalir depende del valor de α preestablecido.

Es importante resaltar en este apartado, que los equipos industriales no suelen deteriorarse de forma instantánea, suele ser un proceso progresivo, por ese motivo no deberían localizarse puntos aislados que sobrepasen los límites de control, sino que deberían localizarse tendencias en los puntos.

También es importante la distancia de los puntos sobresalientes a los límites de control, ya que cuanto más grande sea esta más preocupante es ese punto, porque su comportamiento se aleja en mayor medida del esperado.

3.2. Transformación de los datos

Cuando se trabaja con gráficos de control tanto univariantes como multivariantes se requiere que los datos sigan una distribución normal y sean independientes; además, en el caso multivariante se aconseja utilizar entre 2 y 10 variables explicativas –dado que su aplicación a un número mayor de variables podría favorecer el incremento del *Average Run Length* o número de observaciones hasta detectar la una anomalía–. En la práctica los datos no siempre tienen una estructura óptima para poder aplicar estas técnicas, esto produce que sea habitual realizar algunas transformaciones estadísticas.

En este caso se ha comprobado la normalidad de los datos mediante el test de Shapiro-Wilks, el test de Jarque-Bera, el test de asimetría, el test de kurtosis; se ha aplicado la transformación Box-Cox a los datos para intentar que su distribución se asimilara en mayor medida a una distribución gaussiana, se ha realizado un análisis de componentes principales (PCA [3]) para reducir la dimensionalidad de

[3]Siglas de Principal Component Analysis

los datos y también se han ajustado los datos con un modelo de series temporales para reducir la autocorrelación presente en los mismos. En este apartado se explicarán con mayor profundidad.

3.2.1. Estandarización

A menudo, cuando se trabaja con múltiples variables, estas no tienen por qué estar medidas en la misma escala, por ejemplo, si trabajamos con temperaturas y con vibraciones, como es el caso de esta investigación. Aunque, para utilizar algunos métodos estadísticos, es preferible que todas estén en una escala similar, ya que sino los resultados pueden verse alterados y las variables medidas en escalas con mayores magnitudes obtendrán mayor repercusión. Por este motivo, surge la estandarización de las variables que permite transformar la escala de las variables para adaptarlas a una escala común.

La estandarización univariante clásica -según Francisco (2021)- consiste en transformar cada valor en su puntuación típica. Dada una muestra $x = x_1, x_2, ..., x_n$, su estandarización o tipificación se define como

$$Z = \frac{x - \mu}{\sigma},$$

siendo μ y σ aproximados por sus estimadores muestrales $\bar{x}$ y $\hat{\sigma}$, respectivamente,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad y$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \frac{1}{\sqrt{n-1}}\sum_{i=1}^{n}(x_i - \bar{x})$$

La extensión multivariante de la estandarización –según Vaamonde (2019)– parte de una muestra de vectores en $\mathbb{R}^d, (x_1, x_2, ..., x_n)$ y se define como:

$$y_i = \Sigma^{-1/2}(x_i - \mu).$$

En la práctica, μ y Σ son aproximados por sus estimadores muestrales $\bar{x}$ y S, respectivamente

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t.$$

3.2.2. Normalidad

Muchos métodos univariantes y multivariantes asumen que la distribución de los datos es normal; esto puede deberse a que muchos dato, aunque no sigan exactamente esta distribución, gracias al Teorema Central del Límite, pueden ser una buena aproximación. Además, la distribución normal multivariante cuenta principalmente con dos ventajas: (1) puede describirse utilizando únicamente 2 parámetros: su vector de medias y su matriz de covarianzas; y (2) las combinaciones lineales de varias distribuciones normales siguen siendo distribuciones gaussianas (Pereira y Sánchez (2022)).

Para evaluar la normalidad, en el caso univariante, puede hacerse a través de métodos gráficos como histogramas o QQ-plots, o a través de test de normalidad como Shapiro-Wilks, χ^2, Kolmogrov-Smirnov, etc. Pero, aunque los datos sean normales en el análisis individual, esto no quiere decir que sigan una distribución gaussiana cuando los analizamos en conjunto, por eso es muy importante comprobar la normalidad multivariante de los datos. Para ello existen muchos test de normalidad, en este trabajo se han empleado 3: el test de Jarque-Bera, el test de asimetría y el test de kurtosis.

Vamos a explicar detalladamente estos tres test, pero previamente explicaremos el test de Shapiro-Wilks univariante, los histogramas y los qq-plots, ya que se han utilizado para la comprobación de normalidad de los datos univariantes.

- **Test de Shapiro Wilks univariante**
  Este test plantea como hipótesis nula que los datos de una muestra $x = x_1, x_2, ..., x_n$ provienen de una distribución gaussiana. Se trata de un procedimiento general que puede detectar también problemas de asimetría o kurtosis. Este test parte de la idea de que los datos están estandarizados $z_i = (x_i - \bar{x})/S, i \in 1, ..., n$ y el estadístico de Shapiro Wilks se construye como

$$W = \sum_{i=1}^{n/2} a_{i,n}(z_{(n-1+1):n} - z_{i:n}),$$

  "siendo $z_{1:n} < ... < z_{n:n}$ la muestra ordenada de datos estandarizados y $a_{i,n}$ ciertas constantes. Consiste en calcular las distancias entre los datos de la muestra ordenada, simétricos respecto a la mediana, esto es, la distancia entre el primero y el último, el segundo y el penúltimo, y así sucesivamente; en general el $z_{i:n}$ y el $z_{(n-1+1):n}$. El propósito es comparar estas distancias con las que habría en una muestra de observaciones normales" (Pateiro y Sánchez, 2022, pp.70).

- **Histogramas**
  El histograma es uno de los métodos gráficos más habituales para representar datos continuos, este representa mediante barras la frecuencia con las que aprecen ciertos valores, agrupados en intervalos. Es muy útil para observar la distribución de los datos aunque depende en gran medida de la venta o ancho de banda [4] y del origen. Existe mucha literatura sobre los histogramas ya que se trata de uno de los estimadores no parámetricos de la densidad de probabilidad más simples (Roca, 2017).

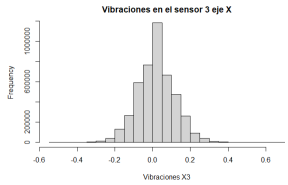


Figura 3.2: Representación de un histograma
Fuente: Elaboración propia

- **QQ-Plot**
  Este gráfico compara mediante un diagrama de dispersión, los cuantiles de una muestra con los cuantiles teóricos de una distribución específica, para comprobar la normalidad de los datos deben ser una distribución gaussiana estandarizada (Vilar, 2021).

  Si los datos siguen un patrón lineal en el QQ-Plot, esto quiere decir que los datos de la muestra pertenece a una familia paramétrica. En este caso la escala de las variables no afecta a la representación del gráfico.

- **Test de Jarque-Bera**
  El test de Jarque-Bera se basa en el comportamiento de los estimadores por el método de los momentos, es decir, este tiene en cuenta la asimetría y la kurtosis de las variables. Su hipótesis nula verifica que la muestra sigue una distribución normal (Vilar, 2021).

[4] La ventana o ancho de banda, generalmente se denomina como h, hace referencia al intervalo que se fija para dibujar las barras del histograma, por ejemplo h=5 (1-5, 6-10, 11-15, ...)

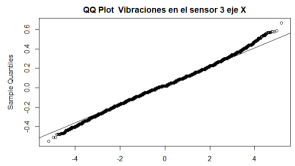


Figura 3.3: Representación de un QQ-Plot
Fuente: Elaboración propia

- Coeficiente muestral de asimetría:

$$b_1 = \frac{1}{nS^3}\sum_{i=1}^{n}(X_i - \bar{X})^3, \qquad b_1 \sim N(0, \sqrt{6/n}).$$

- Coeficiente muestral de Kurtosis:

$$b_2 = \frac{1}{nS^4}\sum_{i=1}^{n}(X_i - \bar{X})^4, \qquad b_2 \sim N(3, \sqrt{24/n}).$$

El estadístico de Jarque-Bera se construye como

$$JB = \frac{n}{6}b_1^2 + \frac{n}{24}(b_2 - 3)^2, \qquad JB_n \sim \chi^2_2.$$

JB se aproxima por una distribución Chi-cuadrado -χ^2- con 2 grados de libertad, toma siempre valores no negativos, rechaza la H_o para un α concreto si $\hat{JB} > \chi^2_{2,1-\alpha}$. Con muestras pequeñas este test puede presentar errores de Tipo I grandes, es decir, rechazará la hipótesis nula cuando esta es cierta (Vilar, 2021).

La generalización multivariante –siguiendo a Kim (2016)– de este test se realiza utilizando la ortogonalidad o la estandarización empírica de los datos. Dada $X_1, X_2, ..., X_n$ procedente de un vector aleatorio independiente e identicamente distribuido (iid), supongamos que cada X_i es una muestra de una distribución normal multidimensional $N_d(\mu, \Sigma)$, siendo el vector de medias μ y el vector de convarianzas Σ donde $Z_1, ..., Z_n$ con $Z_i = S^{*t}(X_i - \bar{X}), i = 1, ..., n$, obteniendo de este modo los residuos escalados.

Siguiendo una $N_d(0, I)$ asintótica, donde 0 es el vector nulo de orden d e I es la matriz identidad dxd, donde $\bar{X} = (1/n)\sum_{j=1}^{n} X_j$, $S = (1/n)\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})^t$ y $S^* = S^{*t}SS^* = I$. Los componentes de Z_i son denominados como $Z_{1i}, Z_{2i}, ..., Z_{di}$ son independientes y bajo hipótesis nula siguen una distribución $N(0, 1)$.

Ahora, se debe calcular la estadística de prueba univariante (JB) para cada componente de los vectores transformados para construir el test multivariante tal que

$$JB_M = \sum_{k=1}^{d} JB(k),$$

siendo $JB(k)$ el test estadístico para cada coordenada $Z_{k1}, Z_{k2}, ..., Z_{kn}, k = 1, ..., d$. JB_M sigue una distribución Chi-cuadrado χ^2 con $2 * d$ grados de libertad, χ^2_{2d}

- **Coeficiente de asimetría multivariante** Se define como

$$A_m = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^3,$$

siendo

$$r_{ij} = (x_i - \bar{x})^t S^{-1}(x_j - \bar{x}) = z_i^t z_j \quad y \quad z_i = S^{-1/2}(x_i - \bar{x}).$$

Si la muestra analizada es simétrica, el coeficiente de simetría será algo mayor que cero, si por el contrario la muestra tiene un comportamiento no simétrico, el coeficiente de asimetría sería más grande.

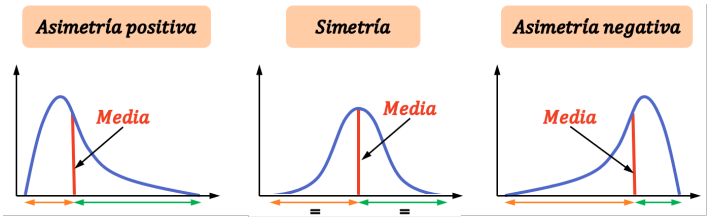


Figura 3.4: Tipos de asimetría
Fuente: Imagen de google: https://www.probabilidadyestadistica.net/tipos-de-asimetria/

La distribución normal es simétrica, por eso el contraste de normalidad rechazará esta si el coeficiente de A_m es demasiado grande en comparación con ciertos valores que están tabulados y se corresponden con distribuciones normales, pueden aproximarse estos valores por simulación. (Pateiro y Sánchez, 2022)

- **Coeficiente de kurtosis multivariante** Se define como

$$K_m = \frac{1}{n}\sum_{i=1}^{n} r_{ii}^2,$$

siendo r_{ii} la distancia de Mahalanobis de las observaciones al vector de medias, puede expresarse como

$$r_{ii} = (x_i - \bar{x})^t S^{-1}(x_i - \bar{x}) = z_i^t z_i.$$

Basándonos en la definición anterior, puede decirse que el coeficiente de kurtosisi se basa únicamente en los valores r_{ii}. El r_{ii} es la distancia del dato i-ésimo al vector de medias, los valores del coeficiente de kurtosis son r_{ii}^2, podría decirse que son las potencias cuartas de estas distancias (Pateiro y Sánchez, 2022).

La kurtosis trata de detectar si los datos se agrupan en torno a la $Media \pm DesviacinTpica$. Para la kurtosis es importante la magnitud de la desviación, no tiene en cuenta el sentido o la dirección de ésta.

La distribución univariante presenta un valor intermedio de kurtosis, esto se denomina mesocúrtica. Como enuncian Pereira y Sánchez (2022), un vector normal $d-$dimensional tiene kurtosis igual a $d(d+2)$. Los datos procedentes de uniformes o de mixturas de normales obtendrán valores de K_m menores a $d(d+2)$, esto se denomina platicúrticas. Los datos originados por exponenciales son leptocúrticas, esto quiere decir que obtienen valores de K_m mayores a $d(d+2)$.

Para el contraste de normalidad, esta se rechazará si los valores de K_m son demasiado grandes o demasiado pequeños. Los valores para las muestras normales están tabulados -al igual que en el caso de la asimetría- pero pueden aproximarse por simulación.

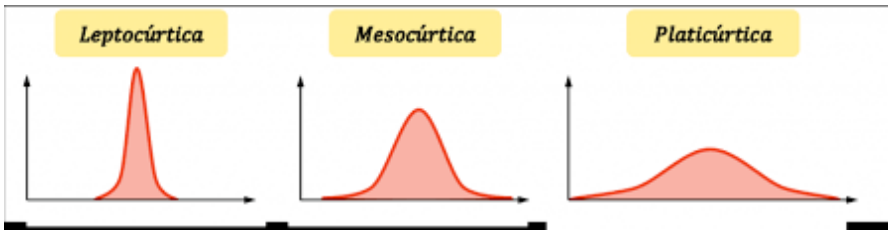


Figura 3.5: Tipos de kurtosis
Fuente: Imagen de google: https://www.probabilidadyestadistica.net/curtosis/

3.2.3. Transformación Box-Cox

Cuando los datos presentan una varianza poco constante o no siguen una distribución normal, en estadística es común aplicarles transformaciones potenciales, ya que es una forma de modificar la estructura de los datos manteniendo la información que nos proporcionan sobre su variabilidad.

Una de las transformaciones más habituales es la propuesta por George E.P. Box y David Cox en 1964 –véase Sebas (2020) y Vaamonde (2019)–, se denomina transformación Box-Cox. Su idea principal consiste en aplicar a una muestra unidimensional $(x_1, x_2, .., x_n)$ la siguiente función:

$$y_i \begin{cases} \frac{x_i^\lambda - 1}{\lambda} \text{ cuando } \lambda \neq 0 \\ log(x_i) \text{ cuando } \lambda = 0, \end{cases}$$

en la que x_i es el dato original, y_i el dato transformado y λ es un parámetro muy importante; dependiendo del valor elegido para este parámetro la transformación puede variar considerablemente. El método más habitual para la elección de λ es utilizar el método de máxima verosimilitud.

Esta es la transformación que se ha utilizado en el trabajo para intentar aproximar la distribución de los datos en mayor medida a una distribución gaussiana; el procedimiento se le ha realizado a las variables individualmente. Una de las principales ventajas de esta transformación es la sencillez de su implementación y una de sus desventajas es que no permite valores negativos, pudiendo solucionarse mediante la adición de una constante a los datos originales(Sebas, 2020 ; Vaamonde, 2019).

3.2.4. Análisis de Componenetes Principales

El análisis de componentes principales (PCA), es una técnica de reducción de la dimensión; es capaz de convertir un gran número de variables incorreladas en unas pocas componentes principales. Consiste en buscar combinaciones lineales de las variables originales que reproduzcan lo mejor posible la variabilidad de los datos. Este método es útil porque permite entender la información proporcionada por los datos a través de unas pocas componentes principales y las estructuras de correlación presentes en los datos. Además, las componentes producto de este análisis, pueden utilizarse en análisis estadísticos posteriores. De hecho, es un paso previo habitual a la aplicación de modelos de regresión y clasificación, además de gráficos de control. (Pateiro y Sánchez, 2022).

Para realizar el análisis de componentes principales se parte de un vector aleatorio $d-$dimensional $x= (x_1, ..., x_d)^t$ con media $\mu = E(x)$ y matriz de covarianzas $\Sigma = E((x-\mu)(x-\mu)^t)$. La primera componente principal de x se define como una variable aleatoria z_1 tal que

$$z_1 = v_1^t x = v_{11}x_1 + ... + v_{d1}x_d$$

con

$$v_1 = (v_{11}, ..., v_{d1})^t \in \mathbb{R}^d,$$

siendo

$$Var(z_1) = \text{máx}\{Var(v^t x) : v \in \mathbb{R}^d, v^t v = 1\} = \lambda_1$$

La primera componente principal es la combinación lineal normalizada de todas las variables de x que tiene mayor varianza. λ_1 es el mayor autovalor de Σ y v_1 es el autovector asociado a λ_1 de norma uno $v_1^t v_1 = 1$, es decir, los coeficientes asociados a la primera componente.

La segunda componente principal es la combinación lineal de las variables de x formada por vectores unitarios ortogonales a v_1 que tiene mayor varianza. Pudiendo definirse z_2 tal que

$$z_2 = v_2^t x = v_{12} x_1 + ... + v_{d2} x_d$$

con

$$v_2 = (v_{12}, ..., v_{d2})^t \in \mathbb{R}^d,$$

siendo

$$Var(z_2) = \text{máx}\{Var(v^t x) : v \in \mathbb{R}^d, v^t v = 1, v^t v_1 = 0\} = \lambda_2$$

λ_2 el segundo autovalor de Σ mientras que v_2 es el vector asociado, de norma uno ($v_2^t v_2 = 1$) y ortogonal -perpendicular- a $v_1 (v_1^t v_2 = 0)$

Generalizando, se podrían definir las d componentes principales de x como las variables aleatorias $(z_1, ..., z_d)$ tales que

$$z_1 = v_1^t x, ..., z_d = v_d^t x,$$

con $v_1, ..., v_d \in \mathbb{R}^d$. Estas d componentes adoptan la forma

$$z_j = v_j^t x, \qquad j \in \{1, .., d\},$$

siendo $\lambda_1 \geq ... \geq \lambda_d \geq 0$ los d autovalores ordenados de Σ y $v_1, ..., v_d$ los autovectores asociados normalizados. Además $Cov(z_j, z_k) = 0$ si $j \neq k$ y

$$Var(z_j) = \lambda_j, \qquad j \in \{1, ...d\}.$$

Definiendo $z = (z_1, ..., z_d)^t$ y $V = (v_1, ..., v_d)$. Además, las columnas de la matriz V son los autovectores de Σ o dicho de otra forma, los coeficientes de las d componentes. Entonces,

$$Cov(z, z) = Cov(V^t x, V^t x) = V^t Cov(x, x) V = V^t \Sigma V.$$

La proporción de variabilidad explicada por las r primeras componentes principales, es

$$\frac{\lambda_1 + ... + \lambda_r}{\lambda_1 + ... + \lambda_r + \lambda r + 1 + ... + \lambda_d}.$$

Es habitual utilizar alguno de los siguientes criterios para fijar el número de componentes principales:

- Criterio de la varianza explicada: Se seleccionan el número de componentes principales que conjuntamente expliquen un 90-95 % de la proporción de varianza.
- Criterio del valor propio: Consiste en seleccionar los componentes cuyos autovalores sean mayores a la media.
- Gráfico de sedimentación o *screeplot*: Se representan gráficamente los λ en orden decreciente y se busca el punto a partir del cual los valores de λ se estabilizan, es decir, son muy similares entre sí y más pequeños que los anteriores. Este punto se denomina como codo. Si se observa la Figura 3.6, el codo se puede ver en la segunda componente, a partir de la cual los valores se estabilizan y se diferencian de la componente 1 que tiene un valor mucho más elevado que el resto.

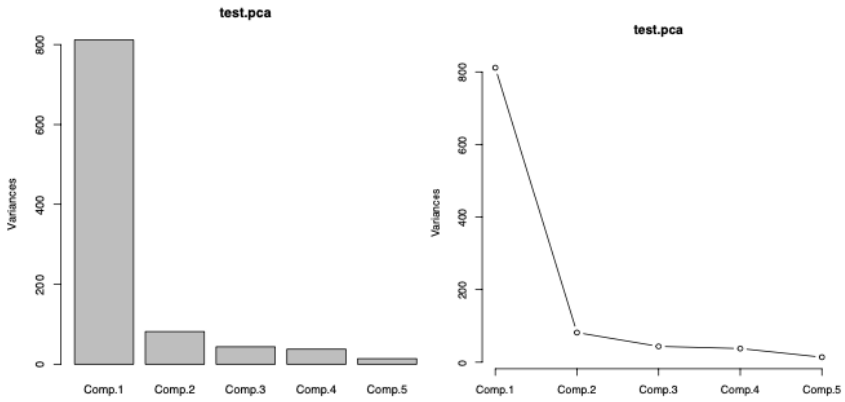


Figura 3.6: Gráfico de sedimentación
Fuente: Pateiro y Sánchez, 2022, pp.96

- Número preestablecido: Es habitual seleccionar dos componentes principales, ya que se pueden representar gráficamente.

El PCA establece dos premisas: (1) Las variables deben ser dependientes y (2) seguir una combinación lineal. Además, cabe señalar que sus resultados dependen de la escala de medida de las variables originales. Si se modifica la escala de las variables se modificarán las componentes principales e incluso su interpretación, ya que si una variable tiene una escala mayor al resto, esta aumentará su varianza y su aportación a la varianza total, captando la primera componente principal. Esto puede solucionarse de dos formas:

1. Si las variables son de la misma naturaleza, pueden medirse en la misma escala.
2. Aplicando el PCA sobre las variables estandarizadas, trabajando sobre la matriz de correlaciones en vez de sobre la de covarianzas.

Si se utiliza el PCA para realizar análisis estadísticos posteriores, es interesante y necesario relacionar las componentes principales con las variables originales, esto se hace calculando la covarianza o la correlación –depende de si el PCA se realiza sobre las variables originales o estandarizadas– entre ambas (Pereira y Sánchez, 2022).

El PCA se ha utilizado en este trabajo para reducir el número de variables, ya que, para la realización de gráficos de control no es recomendable trabajar con más de 10 variables explicaivas y, en nuestro caso, se cuenta con 17.

3.2.5. Series de tiempo

Al realizar mediciones en maquinaria industrial sobre la misma máquina o pieza, es muy probable que exista autocorelación entre las mediciones tomadas. Esto, si no se tiene en cuenta, puede ser perjudicial para la aplicación de ciertas técnicas estadísticas –como el PCA y los gráficos de control– ya que los resultados finales pueden verse alterados.

En este trabajo se ha realizado ha aplicado este tipo de modelos de forma previa al uso de gráficos de control. De hecho, si no se tuviese en cuenta la autocorrelación entre las variables y/o las observaciones, se podría incurrir en la detección de un gran número de falsas alarmas, dado que estaríamos subestimando la variabilidad real del proceso (Barbeito et al., 2017).

La literatura existente sobre el tema se centra generalmente en la fase de control prospectiva -fase 2-, pero en este trabajo se trabaja con datos de fase retrospectiva -fase 1-. Por eso, para entender la

metodología de trabajo a llevar a cabo con datos autocorrelados en fase 1 se ha consultado Barbeito et al.(2017), la cual enuncia dos enfoques principales para trabajar con los datos.

El primer enfoque se basa en ajustar los datos mediante un modelo de serie de tiempo y monitorear los resiudos resultantes en un gráfico de control, estos gráficos se conocen como gráficos de causas especiales o gráficos residuales.

El segundo enfoque consiste en monitorear directamente las observaciones asumiendo modelos específicos de series de tiempo para los datos. En este caso, para la construcción de los límites de control es de gran importancia atender a la estructura de la varianza-covarianza. Este tipo de gráficos se conocen como gráficos de causa común.

Para este trabajo se ha optado por la primera opción, por ese motivo es perteninte realizar una breve introducción a las series temporales y a algunos modelos de las mismas.

Un proceso estocástico, como define Aneiros (2022), es un conjunto de variables aleatorias, $\{T_t\}_{t\in\mathbb{Z}}$, definidas sobre un mismo espacio de probabilidad,

$$..., X_{-2}, X_{-1}, X_0, X_1, X_2, ...$$

siendo el subíndice t de cada variable el instante de tiempo en que se observan.

Una observación del proceso estocástico, se conoce como una realización o trayectoria del proceso estocástico, y se denota como

$$..., x_{-2}, x_{-1}, x_0, x_1, x_2, ...$$

Una serie de tiempo

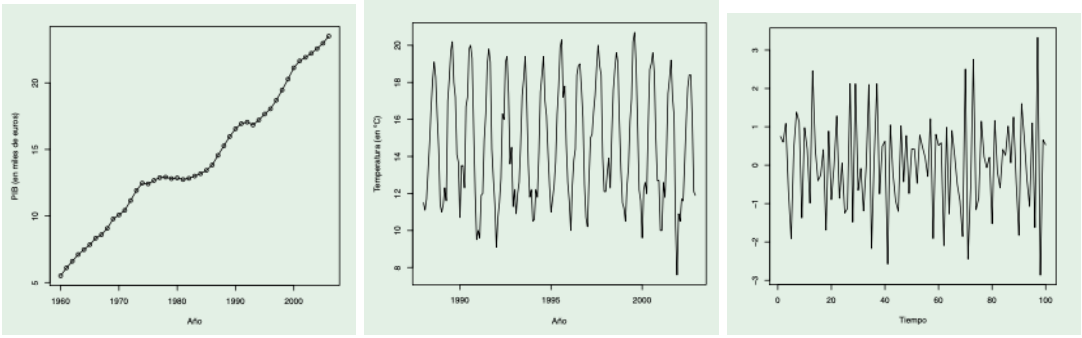
$$x_1, x_2, ..., x_T,$$

es una realización o trayectoria parcial de un proceso estocástico, esto puede definirse como un conjunto de observaciones secuenciales de una variable, X, a lo largo del tiempo.

Para el análisis de las series de tiempo, lo primero es realizar una representación de las series mediante un gráficos secuencial. El gráfico secuencial permite observar como evolucionan las series en el tiempo y si presentan alguna de sus características principales, que se describen brevemente en las siguientes líneas.

- **TENDENCIA**: Comportamiento de la serie a largo plazo. La existencia de tendencia provoca que la serie no sea constante. Si una serie presenta tendencia debe diferenciarse [5] para eleminársela.
- **ESTACIONALIDAD**: O componente estacional, es el comportamiento periódico de la serie. Si existe estacionariedad la serie tendrá un comportamiento repetitivo. Esto puede deberse a que la serie está afectada por factores estacionarios, anuales, mensuales, etc.; si se identifica este patrón debe indicarse al modelar la serie de tiempo.
- **HETEROCEDASTICIDAD**: Comportamiento de la variabilidad de la serie. Si la serie es heterocedástica quiere decir que su variabilidad no es constante. Generalmente se estabiliza la variabilidad a través de la aplicación del logaritmo neperiano a la serie pero existen otros métodos.

[5] Diferenciar una serie de tiempo es restar a cada valor de la serie el valor anterior. Consiste en pasar de la serie x_t a la serie $y_t = x_t - x_{t-1}$.



(a) Tendencia (b) Estacionalidad (c) Heterocedasticidad

Figura 3.7: Representación de las características principales de las series de tiempo. Fuente: Aneiros, 2022, pp.10-11, 22

La clase de modelos de series temporales que se utiliza generalmente para eliminar la autocorrelación, según Vaamonde (2019), son los modelos ARMA. Estos son modelos que combinan en un mismo proceso la estructura autorregresiva (AR) y de medias móviles (MA), pueden representarse como

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_q a_{t-q},$$

donde c, $\phi_1, ..., \phi_p (\phi \neq 0)$ y $\theta_1, ..., \theta_q$ son constantes. La representación puede compactarse como

$$\phi(B) X_t = c + \theta(B) a_t,$$

donde

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \quad y$$
$$\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)$$

siendo B el operador retardo, tal que $BX_t = X_{t-1}$ ϕ es el parámetro del proceso autorregresivo de orden p (AR(p)) y θ es el parámetro del proceso de medias móviles de orden q. Al combinar ambos procesos, el modelo ARMA, definido por dos parámetros p y q denominándose normalmente como ARMA(p,q). Si se precisan modelos más flexibles para estimar la variabilidad y tendencia real, dentro de los modelos Box-Jenkings, la alternativa son los modelos denominados ARIMA, definidos por tres parámetros, ARIMA (p,d,q), tal y como se indica a continuación:

$$\phi(B)(1 - B)^d X_t = c + \theta(B) a_t$$

A las series de tiempo, una vez se les aplican d diferencias, se convierten en modelos ARMA(p,q). En la representación, la parte $(1 - B)^d$ es la que hace referencia a la diferenciación (Aneiros, 2022).

En la práctica suele ser suficiente con aplicar un modelo AR(p)

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + a_t$$

Los modelos AR, MA, ARMA y ARIMA se utilizan para modelar datos univariantes, aunque para no perder la estrutura de covarianzas y correlaciones de los datos, en este trabajo, sería interesante ajustar algún modelo multivariante. Por ese motivo, se ha utilizado un modelo VAR(p) -Vectores autorregresivos-. Este podría decirse que es una extensión del modelo AR:

$$Y_t = A_o + \sum_{s=1}^{p} A_s Y_{t-s} + GW_t + u_t.$$

En cuya expresión Y_t es un vector columna $kx1$, p es el orden del modelo VAR, W_t es el vector de variables independientes y u_t es un vector $kx1$ de innovaciones, es decir, de procesos sin autocorrelación con $Var(u_t) = \Sigma$, constante. Se espera que al ajustarse la serie de tiempo a través de un modelo VAR(p), este cumpla un conjunto de especificaciones: estacionariedad, independencia de los residuos, homocedasticidad de la varianza de los residuos y, también es conveniente comprobar si los residuos siguen una distribución normal, aunque no es obligatorio (Novales, 2017 ; Mauricio, 2007).

Por otro lado, existen modelos que son adecuados cuando la varianza de los residuos no es constante. Estos modelos se denominan ARCH -modelo autorregresivo con heterocedasticidad condicional-. En el contexto univariante, el modelo ARCH más sencillo parte del esquema de un AR(1), tal que

$$y_t = \phi_1 y_{t-1} + \epsilon_t,$$

en el que ϵ_t es ruido blanco, esto es, un proceso idénticamente distribuido con media cero y varianza constante. Los errores o residuos también pueden obtenerse por un modelo de regresión con perturbaciones heterocedásticas,

$$y_t = x_t\beta + \epsilon_t$$

donde y_t es la variable dependiente, x_t son las variables dependientes e independientes retardadas y β es un vector de parámetros desconocidos. Siendo su extensión multivariante

$$y_t = \epsilon_t \Sigma_t$$

donde

$$\Sigma_t = \omega + \sum_{i=1}^{q} A_i vech(y_{t-1} y_{t-1}^t)$$

donde $vech(y_{t-1}y_{t-1}^t)$ es un operador que vectoriza la parte inferior de una matriz NxN como un vector de orden $N(N+1)/2$, ω es un vector de orden $N(N+1)/2$ y A_i es una matriz de orden $N(N+1)/2 + N(N+1)/2$ (Ruiz, 1994; Sáez y Pérez, 1994).

3.3. Machine Learning

Machine Learning (ML) es un término muy amplio, que engloba tanto la aplicación de técnicas estadísticas –denominados algoritmos– como la aplicación de procedimientos informáticos y computacionales, y cuya popoluraridad es creciente debido a su incorporación en el marco amplio de la intelligencia artificial. Por este motivo es conveniente comenzar este apartado con algunas definiciones que ayuden a entender mejor la información y utilidades que este grupo de procedimientos aporta.

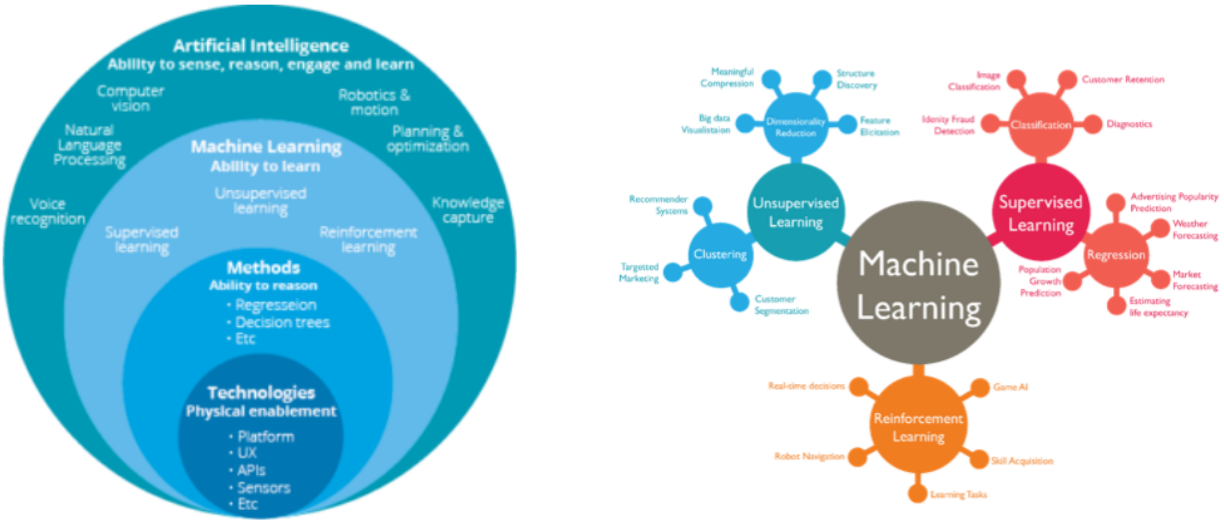
Machine Learning es una técnica utilizada desde 1959 en el campo de la inteligencia artificial. Es la parte que tiene la capacidad de aprender de los datos. Podría definirse como un conjunto de algoritmos automáticos de predicción (Fernández et al., 2022; López y Fernández, 2021).

Como se muestra en la Figura 3.8, dentro del ML pueden diferenciarse 3 grandes bloques:

- **Aprendizaje no supervisado**: Abarca los métodos exploratorios, aquellos en los que no hay una variable respuesta explícita. Estos métodos tratan de entender la relación entre los datos y sus estructuras. Algunos ejemplos son: análisis descriptivo, PCA, clúster, detección de valores atípicos o anomalías, entre otros.
- **Aprendizaje supervisado**: Contiene los métodos predictivos, en estos la variable respuesta está definida. Su finalidad es construir modelos que se utilicen posteriormente para hacer predicciones. Dependiendo del tipo de la variable respuesta pueden ser:
    - Clasificación: Variable dependiente categórica.

- Regresión: Variable dependiente numérica

- **Aprendizaje por refuerzo**: Es un aprendizaje sin supervisión humana, se basa en modelos que están en continuo aprendizaje, se van automodifacando a partir de los datos que recogen. Es habitual su utilización en juegos o simulaciones.



(a) Fuente: López y Fernández (2021, pp.74)

(b) Fuente: López y Fernández (2021, pp.75)

Figura 3.8: Visión global del Machine Learning

IIoT o Internet Industrial de las Cosas [6] es *ünа red de interconexión digital entre dispositivos, personas y la propia Internet que permite el intercambio de datos entre ellos, posibilitando que se pueda capturar información clave sobre el uso y el rendimiento de los dispositivos y los objetos para detectar patrones, hacer recomendaciones, mejorar la eficiencia y crear mejores experiencias para los usuarios"* en el contexto industrial (Zubia, 2021, pp.1).

Los dispositivos IIoT son conectados mediante un proceso M2M –Machine to Machine– utilizando todo tipo de conectividad –cable, wifi, bluetooth, etc.–. Desarrollan su trabajo sin intervención humana, únicamente necesitan dispositivos –emisores, receptores y chips– específicos integrados en la maquinaria industrial que generan una gran cantidad de datos, esta es recogida, procesada y analizada en una plataforma IIoT.

El ML en IIoT está contribuyendo a la evolución de la industria gracias a la inclusión de inteligencia artificial en los procesos continuos de monitoreo y control de las máquinas. Generalmente las máquinas que son monitorizadas producen patrones de datos bastante estables. Algunas veces surgen anomalías que en físico no son perceptibles, pero gracias a la sensorización mediante los dispositivos IIoT y al análisis de los datos recolectados, se pueden observar los cambios de los patrones habituales y detectar las anomalías antes de que se produzca daños en el equipo, paradas innecesarias, etc. (Wang et al., 2021; Zubia, 2021).

En este trabajo se utilizan técnicas de aprendizaje supervisado y de detección de anomalías. Aunque las primeras pueden ayudarnos a clasificar los datos en anomalías o no anomalías, no son capaces de hacerlo si no hay observaciones previamente identificadas con valores similares, es decir, las técnicas de aprendizaje supervisado son útiles si se cuenta con una base de datos en la que se hayan podido

[6] Siglas de: Industrial Internet of Things

identificar una amplia variedad de anomalías. Si, por el contrario, no se han podido identificar anomalías en esa base de datos o únicamente unos tipos muy concretos –pero se quieren identificar más clases–, es preferible utilizar métodos de detección de anomalías (Ng et al., 2012). En este trabajo se han probado ambas técnicas y, posteriormente, se explicarán en profundidad cada uno de los métodos utilizados. Pero antes, se van a exponer algunas ideas básicas comunes a ambos métodos.

Para elaborar los modelos de clasificación es habitual emplear toda la información disponible –es decir, el conjunto de datos al completo– para construir un modelo que refleje lo que ocurre en la máquina de la forma más real posible –modelo válido–. Para asegurarnos que el modelo es válido, primero debe entrenarse y luego comprobarse su precisión, bondad de ajuste y bondad de predicción (Fernández et al. 2022). Esto puede realizarse de dos formas:

1. Si la muestra es pequeña suelen entrenarse los modelos con todos los datos y se utilizan técnicas de remuestreo para evaluar su precisión.
2. Si la muestra es grande puede dividirse la base de datos en dos o incluso 3 conjuntos de datos:
   - Muestra de entrenamiento: Es la que se utiliza para elaborar el modelo
   - Muestra de validación (opcional): Se utiliza para la evaluación de la muestra o de los hiperparámetros seleccionados.
   - Muestra de test: Se emplea para estimar el rendimiento del modelo

   Estas muestras son de diferentes tamaños, lo más común es seleccionar al azar el 80 % de los datos y utilizarlos para la muestra de entrenamiento y el 20 % restante utilizarlos con la muestra de validación. Si se utilizan los 3 conjunto lo usual es 15 %-15 %-70 %, respectivamente.

Para evaluar la eficiencia de los modelos generalmente se hace a través de una matriz de confusión, esta es una tabla de contingencia elaborada con las predicciones para el conjunto de test frente los valores reales.

Observado/Predicción	Positivo	Negativo
Positivo	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Negativo	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Cuadro 3.1: Matriz de confusión

A partir de la matriz de confusión se pueden obtener las siguientes medidas de precisión:

- Sensibilidad (TPR): es la tasa de verdaderos positivos,

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}.$$

- Especificidad (TNR): es la tasa de verdaderos negativos,

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}.$$

- Precisión globar (ACC):

$$ACC = \frac{TP + TN}{P + N}.$$

- Precisión global balanceada (BA): Es una tasa que hace referencia a la precisión global del modelo cuando el número de observaciones en las clases de estudio están desequilibradas,

$$BA = \frac{TPR + TNR}{2}.$$

- Puntuación F_1: Es la media armónica de TPR y PPV

$$F_1 = \frac{2TP}{2TP + FP + FN}.$$

- Valor predictivo positivo (PPV): Es la tasa de positivos en la muestra test,

$$PPV = \frac{TP}{TP + FP}.$$

- Valor predictivo negativo (NPV): Es la tasa de negativos en la muestra de test,

$$NPV = \frac{TN}{TN + FN}.$$

3.3.1. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (SVM[7]), son métodos estadísticos que se comenzaron a desarrollar a mediados de 1960 por Vladimir Vapnik (en Fernández et al. 2022) para problemas de clasificación con dos categorías; se basan en la idea de separación de datos mediante hiperplanos. Estos métodos fueron evolucionando y, en la actualidad, existen extensiones para clasificación con más de dos categorías, para regresión y para detección de datos atípicos.

Hay varios tipos de clasificadores SVM que son adecuados según la situación de análisis, los cuales se describen brevemente a continuación.

CLASIFICADORES DE MÁXIMO MÁRGEN

Se trata de un método de clasificación binaria que se utiliza cuando hay una frontera lineal claramente definida, separando perfectamente los datos de entrenamiento de ambas categorías (Fernández et al. 2022). Para la utilización de este método se etiquetan las dos categorías como $+1/-1$, es decir, los valores de la variable dependiente $Y \in -1, 1$, suponiendo que existe un hiperplano

$$\beta_0 + \beta_1 X_1 + \beta_2 x_2 + ... + ...\beta_d x_d = 0,$$

siendo d el número de variables explicativas que separan los datos de entrenamiento según la categoría a la que pertenecen,

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) > 0,$$

para todo $i = 1, 2, ...n$ siendo n el tamaño de la muestra de entrenamiento.

Una vez se ha seleccionado el hiperplano, se deben clasificar las nuevas observaciones calculando el signo de

$$m(x) = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + ... + ...\beta_d x_d,$$

para cada nueva observación x, catalogándolo dentro de la clase $+1$ si el signo es positivo, y de la clase -1 si es negativo. También es importante el valor absoluto de $m(x)$ ya que nos ofrece una imagen de la distancia entre la observación y la frontera que define el hiperplano. Más concretamente

$$\frac{y_i}{\sqrt{\sum_{j=1}^{d} \beta_j^2}}(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}),$$

[7]Siglas de *Support Vector Machines*

siendo esta la distancia de la observación $i-$ésima al hiperplano. Se debe tener en cuenta que aunque los datos de entrenamiento se clasifiquen sin error, esto no garantiza que las nuevas observaciones se cataloguen correctamente. Cuando d es grande, es sencillo que se produca sobreajuste.

Si existe al menos un hiperplano que separa perfectamente los datos de entrenamiento de las dos categorías, habrá infinitos hiperplanos; el objetivo es seleccionar uno. Para ello, se deben calcular las distancias de todas las observaciones al hiperplano y se define el margen como la menor distancia. La técnica de clasificadores de máximo margen selecciona, de los infinitos hiperplano, el que tiene un mayor margen. Siempre van a exitir observaciones que equidistan del hiperplano el máximo margen y que su distancia será precisamente el margen, estas reciben el nombre de *vectores soporte*.

Matemáticamente, dadas n observaciones de entrenamiento $x_1, x_2, ...x_n$, el clasificador de máximo margen puede expresarse mediante el siguiente problema de optimización

$$max_{\beta_0,\beta_1,...,\beta_d} M$$

$$sujeto \;\; a$$

$$\sum_{j=1}^{d} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) \geq M \;\; \forall i.$$

Si los datos se pueden separar perfectamente mediante un hiperplano, la solución al problema anterior será $M > 0$, y M será el margen. Equivalentemente, utilizando $M = \frac{1}{\|\beta\|}$ con $\beta = (\beta_1, \beta_2, ..., \beta_d)$

$$min_{\beta_0,\beta} \parallel \beta \parallel$$

$$sujeto \;\; a$$

$$\sum_{j=1}^{d} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) \geq 1 \;\; \forall i$$

Cabe resaltar que, en este caso, el modelo únicamente depende de los vectores soporte y, si algún otro dato varía, no influirá en la construcción del modelo a no ser que la modificación de ese dato cambie el margen.

CLASIFICADORES DE SOPORTE VECTORIAL

Se trata de una extensión del problema anterior pero, en este caso, no existe ningún hiperplano que separe a la perfección ambas categorías. Este enfoque se basa en aceptar que algunos datos de entrenamiento estarán mal clasificados. Este tipo de clasificadores son más robustos que los clasificadores de máximo margen, por ese motivo será preferible su utilización aunque exista un clasificador de máximo margen(Fernández et al. 2022). Matemáticamente puede formularse como

$$max_{\beta_1,\beta_2,...,\beta_d,\epsilon_1,...,\epsilon_n} M$$

$$sujeto \;\; a$$

$$\sum_{j=1}^{d} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) \geq M(1 - \epsilon) \;\; \forall i,$$

$$\sum_{i=1}^{n} \epsilon_i \leq K,$$

$$\epsilon_i \geq 0 \forall i$$

donde las variables ϵ_i son variables de holgura, K se puede interpretar como la tolerancia al error –si K=0 estaríamos ante el caso de clasificadores de máximo margen– o como una penalización por la complejidad del modelo; en este último caso, en términos de varianza y sesgo, si se utilizan valores muy pequeños de K darán lugar a modelos muy complejos con riesgo de sobreajuste –mucha varianza y poco sesgo–, por la contra, valores muy grandes de K darán lugar a modelos con mucho sesgo y poca varianza. El valor óptimo para K puede seleccionarse mediante validación cruzada, bootstrap, etc.

Una forma equivalente de formular el problema es

$$min_{\beta_0,\beta} \parallel \beta \parallel$$

$$sujeto \;\; a$$

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) \geq 1 - \epsilon_i \;\; \forall i,$$

$$\sum_{i=1}^{n} \epsilon_i \leq K,$$

$$\epsilon_i \geq 0 \;\; \forall i$$

En la práctica, se utiliza equivalentemente

$$min_{\beta_0,\beta} 0{,}5 \parallel \beta \parallel^2 + C \sum_{i=1}^{n} \epsilon_i$$

$$sujeto \;\; a$$

$$y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 x_{2i} + ... + ...\beta_d x_{di}) \geq 1 - \epsilon_i \;\; \forall i,$$

$$\epsilon_i \geq 0 \;\; \forall i.$$

Aunque el problema a resolver es el mismo -cuantas observaciones permito tener dentro del margen-, el parámetro K se ha sustituido por el parámetro C, que tiene una interpretación inversa. El parámetro C es la penalización por mala clasificación, es decir, coste de que un dato de entrenamiento se encuentre mal clasificado.

El clasificador de soporte vectorial

$$m(x) = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + ... + ...\beta_d x_d,$$

puede representarse como

$$m(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i x^t x_i,$$

en el que $x^t x_i$ es el producto escalar entre el vector x del dato a clasificar y el vector x_i del datos de entrenamiento $i-$ésimo. Se asume que los coeficientes $\beta_0, \alpha_1, ..., \alpha_n$ se obtienen a partir de los productos escalares $x^t x_j$ de los distintos pares de entrenamiento y de las respuestas y_i.

CLASIFICADORES CON MÁS DE DOS CATEGORÍAS

Las formas más populares y más sencillas para realizar la clasificación cuando hay más de dos categorías son:

- Uno contra uno: Se basa en construir tantos modelos como pares de categorías existan. Cuando se añade una nueva observación se mira en qué categoría la ha clasificado cada uno de los modelos y se hace un recuento, ganará la categoría con más votos.

- Uno contra todos: Se contruye el modelo que considera esa categoría frente a todas las demás –que se agrupan en una única–, y para clasificar las nuevas observaciones se examina su distancia con la frontera. La observación se catalogará dentro de la categoría a la que exista mayor distancia.

Cuando los datos no se pueden separar utilizando particiones basadas en rectángulos, como en los métodos anteriores, los métodos SVM no serán adecuados. Una solución es sustituir el hiperplano –en esencial lineal– por otra función que dependa de las variables independientes $X_1, X_2, ..., X_n$, aunque esto puede ser computacionalmente muy costoso. Por eso Boser et al. en 1992 (en Fernández et all. 2022), propuesieron reemplazar, en todos los cálculos que llevan a la expresión

$$m(x) = \beta_0 + \sum_{i \in S} \alpha_i x^t x_i,$$

los productos $x^t x_i$, $x_i^t x_j$ por funciones kernel, obteniendo

$$m(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i).$$

Alguna de las funciones kernel que mas se utilizan son:

- Kernel lineal,
$$K(x, y) = x^t y.$$
- Kernel polinómino,
$$K(x, y) = (1 + \gamma x^t y)^p.$$
- Kernel radial,
$$K(x, y) = exp(-\gamma \parallel x - y \parallel^2).$$

Antes de construir el modelo, es recomendable estandarizar los datos para que todas las variables tengan la misma importancia y las escalas no alteren los resultados. En las expresiones anteriores, γ es el inverso al parámetro ventana. Para evitar sobreajustes o infrajustes es de gran importancia asignarle los valores óptimos tanto a γ como a C.

Siempre es conveniente conocer las limitaciones de los métodos utilizados, por eso, a continuación se enunciarán las ventajas y desventajas de SVM (Fernández et al. 2022).

Ventajas:

- Pueden adaptarse a fronteras no lineales, es decir, son muy flexibles, esto facilita la obtención de buenas predicciones en muchos casos.
- Si se suaviza el margen -a través del parámetro C-, son bastante robusto frente a valores atípicos.

Desventajas:

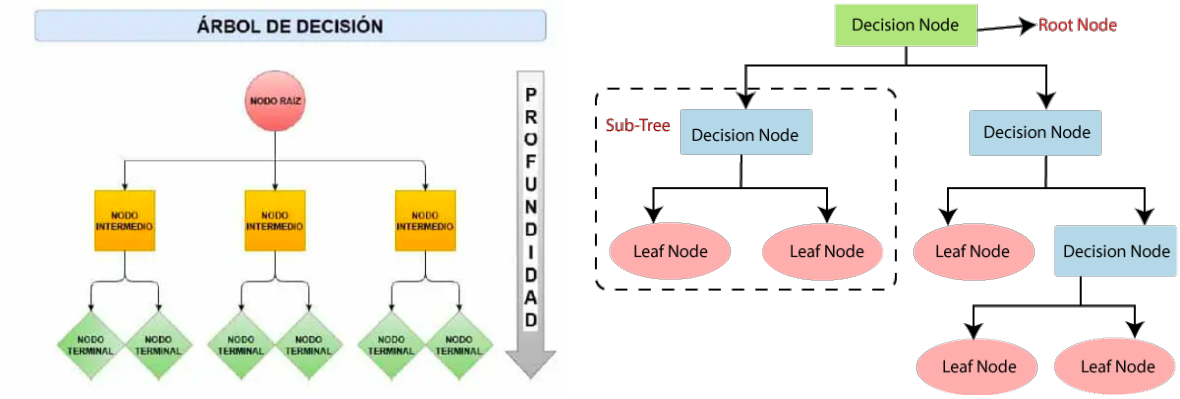
- Una vez se ajustan los modelos su interpretación es compleja.
- Cuando $n \gg d$ el tiempo de computación puede ser muy elevado.
- Como están pensados para variables explicativas numéricas, si se utilizan variables categóricas deben preprocesarse, es decir, transformarlas en variables dummy [8]

[8] Las variables dummy o indicadoras son variables ficticias que se utilizan para transformar las variables cualitativas con dos o más categorías de respuesta diferentes, generalmente se establece una etiqueta numérica (0,1,2,3,...) para cada categoría de respuesta cualitativa. Por ejemplo: Mujer(0)/Hombre(1), Sin estudios(0)/Estudios primarios(1)/Estudios secundarios(2)/Estudios superiores(3), Mañana(1)/Tarde(2)/Noche(3), etc.

3.3.2. Random Forest

Random forest (RF) o árboles aleatorios, son una variante de *bagging* diseñada para trabajar con árboles de decisión. Su idea general consiste en mezclar métodos de predicción débiles, es decir, con baja capacidad predictiva –en este caso árboles de decisión– para lograr un método robusto con gran capacidad predictiva (Fernández et al. 2022).

Los árboles de decisión son "*uno de los métodos más simples y fáciles de interpretar para realizar predicciones en problemas de clasificación y de regresión*"(Fernández et al. 2022, pp. 41). Su idea "*consiste en la segmentación (partición) del espacio predictor (es decir, del conjunto de posibles valores de las variables predictoras) en regiones tan simples que el proceso se pueda representar mediante un árbol binario. Se parte de un nodo inicial que representa a toda la muestra (se utiliza la muestra de entrenamiento), del que salen dos ramas que dividen la muestra en dos subconjuntos, cada uno representado por un nuevo nodo. (...) Este proceso se repite un número finito de veces hasta obtener las hojas del árbol, es decir, los nodos terminales, que son los que se utilizan para realizar la predicción. Una vez construido el árbol, la predicción se realizará en cada nodo terminal utilizando, típicamente, la media en un problema de regresión y la moda en un problema de clasificación*"(Fernández et al. 2022, pp. 41).



(a) Fuente: Ferrero (2020)

(b) Fuente: Fernández (s.f.)

Figura 3.9: Ejemplo de las partes de un árbol de decisión

El *bagging* "*es un método general de reducción de la varianza que se basa en la utilización del bootstrap junto con un modelo de regresión o de clasificación*"(Fernández et al. 2022, pp. 73). A grandes rasgos consiste en, a partir de muchas muestras de entrenamiento, "*utilizar cada una de ellas para entrenar un modelo que después nos servirá para hacer una predicción. De este modo tendremos tantas predicciones como modelos y por tanto tantas predicciones como muestras de entrenamiento. El procedimiento consistente en promediar todas las predicciones anteriores tiene dos ventajas importantes: simplifica la solución y reduce mucho la varianza*"(Fernández et al. 2022, pp. 73).

Habitualmente los árboles tendrán estructuras similares, sobre todo en la parte alta, esto se conocemo como correlación entre árboles y surge sobre todo cuando el árbol es un modelo adecuado para describir la relación entre variables independientes y dependiente o cuando una de las variables independientes o predictores es especialmente relevante. Esta correlación se traducirá en una correlación entre los predictores. Cuanto mayor sea esta correlación, se producirá una menor reducción de la varianza que se puede arreglar añadiendo aleatoriedad al proceso, una de las formas de hacerlo es mediante RF.

RF consiste en realizar cortes binarios en cada uno de los árboles aleatorios que forman el bosque, para cada corte se seleccionará una variable predictora. Pero, previamente a realizar cada uno de los cortes de las d variables predictoras, se deben seleccionar $m < d$ variables independientes -o predictoras- que serán las candidatas para el corte.

Para seleccionar m se pueden utilizar métodos de validación cruzada, bootstrap, etc. Como punto de partida es habitual considerar $m = \sqrt{d}$, en problemas de clasificación, y $m = p/3$, en problemas de regresión.

Los bosques aleatorios son computacionalmente más eficientes que el bagging, porque aunque utilice más árboles, la construcción de cada uno de ellos se realiza con unas pocas variables predictoras.

3.3.3. ALSO

Attribute wise learning for scoring outliers (ALSO) es un algoritmo no supervisado de detección de anomalías cuando no se tienen las anomalías clasificadas, es decir, cuando los datos no están etiquetados (Amat, 2020).

Este algoritmo entrena un modelo de regresión para cada variable en función del resto de variables. El grado de anomalía de cada observación *"se obtiene a partir el agregado del error cuadrático de los modelos al tratar de predecirla"* (Amat, 2020). ALSO reformula un problema no supervisado como un conjunto de problemas supervisados, permitiendo utilizar los algoritmos de aprendizaje supervisado como random forest, SVM, etc.

ALSO se basa en la idea de que una anomalía puede caracterizarse por separarse del comportamiento esperado, es decir, del patrón que siguen la gran parte de los datos. Si un modelo aprende este patrón, tendrá problemas para predecir las anomalías y su error cuadrático medio (RMS) será más elevado que el resto. Como en la práctica es habitual trabajar con más de una variable, ALSO realiza este proceso con cada una de ellas y añade el RMS de todos los modelos para crear una puntuación de anomalía.

Al utilizar este algoritmo, es importante tener en cuenta que:

- La partición muestra de entrenamiento-muestra test de cada modelo debe realizarse a través del método de validación cruzada, para evitar problemas de sobreajuste.
- Los datos deben estar estandarizados ya que la magnitud del error depende de las unidades de las variables.
- Como cada modelo es independiente de los demás, es fácilmente paralelizable.

Recapitulando, según Amat (2020), la idea base de ALSO es que las anomalías pueden detectarse porque se alejan del patrón general de los datos, pero para esto es necesario que exista dicho patrón. Siguiendo con este razonamiento, las variables que no identifican este patrón no deberían añadirse a la métrica final. Para evitarlo, ALSO pondera el error de cada modelo según su capacidad para identificar dicho patrón. El peso que se le da a cada modelo (w_k) se calcula como

$$w_k = 1 - min\{1, RMSE(k)\},$$

siendo k el modelo a estudiar y RMSE el error cuadrático medio estandarizado -ya que se parte de los datos estandarizados-. El RMSE del modelo que únicamente predice la media y no identifica ningún patrón, será 1; todo modelo que prediga igual o peor que el anterior tendrá $w_k = 0$.

La puntuación de anomalía s para una observación x_i que pertence a una base de datos con d variables será

$$s(x_i) = \sum_{k=1}^{d} w_k error_k^2(x_i).$$

Al tratarse de un método no supervisado, no se puede fijar el valor óptimo a partir del cual se debe considerar una observación como anomalía, ya que la puntuación que recibe cada observación es una medida relativa en relación a las demás observaciones. En la práctica, es habitual considerar como anomalía las observaciones que se sitúan por encima de un determinado cuantil.

Algunas ventajas del algoritmo ALSO son:

- Cuando se está trabajando con un alto número de variables, es decir, con modelos de alta dimensionalidad, en los cuales gran parte son no informativas, es un método robusto. Esto se debe, por un lado, a que ALSO no se emplea el concepto de distancia ni de densidad para medir el grado de anomalía y, por otro lado, a que ALSO tiene la capacidad de asignar peso 0 a las variables que no identifican ningún patrón.
- Cuando se utilizan modelos interpretables –como pueden ser los árboles de decisión, modelos lineales, etc.–, ALSO posibilita explicar el porqué de las puntuaciones obtenidas.

Una de sus fortalezas es también puede ser una desventaja, ya que, ALSO considera las anomalías en relación al error de predicción y no a lo aisladas que se encuentran las observaciones respecto al resto de la muestra, esto puede hacer que incurra en errores. Por ejemplo, si una observación está muy alejada del resto pero el modelo la predice correctamente, no se considerará como anomalía porque su puntuación será muy baja (Amat, 2020).

Capítulo 4

Contextualización de los datos

Los datos con los que se ha trabajado para la realización de esta investigación, han sido los correspondientes al Ventilador Aspiración Microondas, ventiladorMW [1]. Este se encuentra en Orember, la fábrica de Finsa localizada en Ourense (Galicia).

Esta fábrica se dedica a la realización de tableros MDF –tableros de fibra de densidad media–. Para su fabricación se astilla la madera previamente descortezada y se desfibra. Sobre la fibra todavía húmeda se añaden los adhesivos –cola– y aditivos, luego se transporta mediante ciclones industriales que rebajan su humedad hasta un 8 %. Tras esto, se va depositando sobre una cinta la cantidad de fibra necesaria para obtener el espesor de tablero deseado; al ir avanzando en la cinta se forma una manta –con la fibra y los adhesivos y aditivos– que pasa por el microondas donde la cola empieza a fraguarse, y posteriormente la manta se dirige a la prensa calefactada donde la cola se endurece y la manta se compacta; al salir de la prensa la manta es cortada por una sierra viajera para formar los tableros del tamaño deseado. Los tableros avanzan hacia unos volteadores en los que se enfrían para que no pierdan la forma. Por último, los tableros se lijan para eliminar las irregularidades de las superficies y, si es necesario, se escuadran para conseguir las dimensiones deseadas. Si se quiere profundizar más sobre el proceso de fabricación se puede consultar El Bosque Protector (2014).

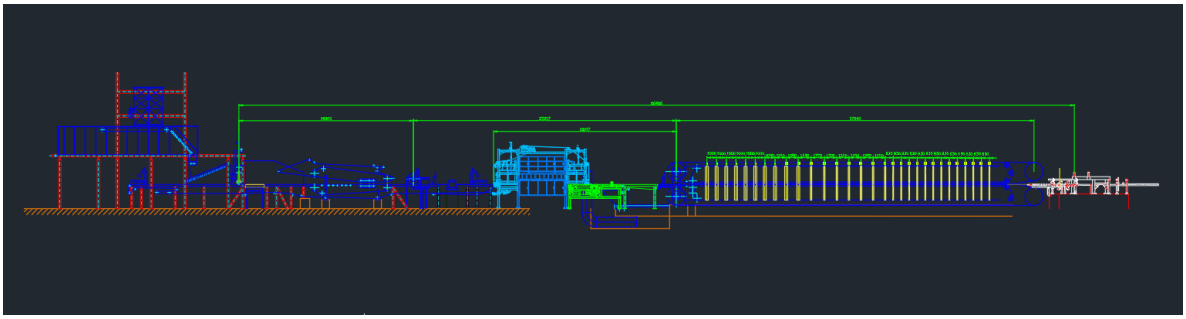


Figura 4.1: Imagen del esquema de la fábrica de Orember, proceso de fabricación de tableros MDF

Además de los tableros MDF, la Finsa fabrica tableros de aglomerado y tableros Superpan [2]. Los 3 pueden ser tableros normales, ignífugos o hidrófugos, y se puede elegir su color –al natural, rojo, amarillo, azul, negro o gris–, si se quiere que estén cubiertos con chapas de madera [3] –esto se hace en

[1] Ventilador Microwave.

[2] Más información en https://superpan.finsa.com/.

[3] Son láminas de madera de 3-4 cm que se pone en las caras de los tableros, es madera de verdad, con lo cual puede pintarse, barnizarse, etc.; pero, al igual que la madera sino está tratada se estropea con el agua, etc.

la fábrica de Pontecesures– o impregnados en papel de melamina [4], incluso pueden hacerse con relieve aplicando presión –únicamente se hace en Orember–. Pueden verse los diferentes tipos de tableros y sus acabados en el Apéndice A.

El VentiladosMW, ventilador en el que se han analizado los datos, se encarga de succionar las partículas de fibra que se sueltan en el microondas, representado en la Figura 4.1 en color azul claro.

Figura 4.2: VentiladorMW

Los datos del VentiladorMV fueron medidos a través de 4 sensores (véase Figura 4.3a). Dos de ellos se colocan en los rodamientos del motor (Figura 4.3b) y los otros dos en los rodamientos del eje de la turbina del ventilador (Figura 4.3c). Todos los sensores miden las vibraciones en 3 ejes (X,Y,Z) y la temperatura, pero además, se mide también el consumo del motor.

(a) Sensor (b) Sensores en el motor (c) Sensores en el eje de la turbina

Figura 4.3: Sensores y su colocación

[4]Si el papel de melamina imita a la madera puede llevar relieve que coincida con la forma del dibujo, pero el tacto sigue siendo de papel.

Los datos de los sensores se recogen en una caja denominada FCM –Finsa Condition Monitoring, véase Figura 4.4– que los envía a una plataforma IIoT, denominada ThingWorx, de dos formas:

1. En streaming: Los cuales son datos que se envían cada 2 segundos de forma constante.
2. En batch: Son datos recogidos a alta frecuencia en períodos de tiempo concreto -c-ada hora, día, semana, quincena, mes, etc.–; estos datos se envían en intervalos/paquetes de 5 segundos, cada paquete son 1000 datos de cada eje de cada sensor, es decir, cada 5 segundos envía 12000 datos; al ser tantos datos la latencia de red no permite enviarlos en streaming –en continuo– y por eso se envían en batch –en lotes–.

(a) FCM con sus componentes

(b) Colocación FCM

Figura 4.4: FCM

Los datos de ThingWorx se almacenan en una base de datos en InfluxBD y se visualizan a través de Grafana.

El ventilador de estudio fue el diseño piloto en el que se colocaron los sensores y la FCM. En él, el lector de consumo conecta la FCM y el motor mediante un cable exterior (véanse Figura 4.3b y Figura 4.4b). En otros diseños, dependiendo del tamaño de la máquina en la que se coloque, el lector de consumo puede ir dentro del propio motor (Figura 4.5) o no ser necesario porque mide el consumo el propio motor.

(a) Ampliación visión motor

(b) Visión general del ventilador con cable interno

Figura 4.5: Lector de consumo sin cable exterior

Los sensores, como ya se ha comentado, están colocados en los rodamientos del ventilador (ver Figura 4.6).

Figura 4.6: Enumeración de los sensores

Los ejes de los sensores 1 y 2 se definen del siguiente modo: el eje X se corresponde con el eje radial, el eje Y con el eje tangencial y el eje Z con el eje axial; esto se debe a la colocación de los sensores en el motor, véase Figura 4.7.

(a) Ampliación visión motor (b) Colocación sensores en el motor

Figura 4.7: Colocación sensores en el motor

Los ejes de los sensores 3 y 4 se corresponden el eje X con el eje tangencial, el eje Y con el eje radial y el eje Z con el eje axial, esto, al igual que en el caso anterior es por la colocación de los sensores en el eje de la turbina, véase Figura 4.8.

(a) Ampliación visión turbina (b) Colocación sensores en el motor

Figura 4.8: Colocación sensores en el eje de la turbina

Según Royo et al. (s.f.), para el correcto desarrollo del mantenimiento predictivo son importantes 9 etapas:

1. Decidir la necesidad y el objetivo de este mantenimiento en la empresa.
2. Seleccionar las máquinas que se van a analizar.
3. Elegir las técnicas óptimas para la verificación, es decir, para decidir qué, cómo, cuándo, dónde se van a realizar las mediciones.
4. Implantar el mantenimiento predictivo .
5. Establecer los valores normales de las variables y los límites de control o valores aceptables de cada máquina/variable.
6. Obtener las mediciones de referencia.
7. Recopilar, registrar y analizar las tendencias.
8. Analizar las condiciones de las máquinas.
9. Corregir los fallos.

En este trabajo se ha abordado desde la etapa 5 a la etapa 8. La etapa 5 se ha realizado tras la limpieza de la base de datos, las etapas 6 y 7 se realizan en el dominio del tiempo y la etapa 8 en el dominio de la frecuencia. En los siguientes apartados se estudia con más detalle.

Capítulo 5

Análisis en el dominio del tiempo

Los datos para el análisis en el dominio del tiempo son los datos recogidos por la FCM en streaming –en continuo cada 2 segundos–. Se me han facilitado en ficheros csv de datos semanales, incluyendo datos desde el 4 de septiembre de 2022 hasta el 2 de febrero de 2023.

Esos datos se han limpiado medianta la aplicación de código programado en Python., eliminando las filas que tenían todos los valores nulos, las observaciones de cuando la fábrica estuvo de vacaciones –consumo=0 por un periodo superior a un día–, los datos atípicos –picos puntuales superiores e inferiores que no afectaron al funcionamiento de la máquina–. Además, tras realizar un primer análisis se ha observado que cuando la máquina está apagada y se enciende, el consumo produce un pico que dispara su valor y afecta también a los valores de las temperaturas de algunos sensores; estos picos no son un consumo real sino un cálculo matemático, por ese motivo se han borrado los datos de consumo puntual superior a 40 amperios –consejo de mi tutor en la empresa– ya que alterarían los resultados.

Tras esta limpieza, se han unificado los ficheros de forma que obtenemos bases de datos mensuales y una base de datos del conjunto total de los datos proporcionados. Las primeras se han utilizado para analizar la evolución de los valores normales de los datos más detalladamente y para realizar alguna representación gráfica en el análisis exploratorio que debido al volumen de datos de la BBDD general era demasiado costosa computacionalmente. La base de datos general ha sido la más utilizada, tanto en el análisis exploratorio como en el análisis de los datos en el dominio del tiempo.

Todas las bases de datos cuenta con 18 variables, el tiempo, la temperatura medida por los 4 sensores, el consumo del motor, las vibraciones medidas en el eje X por los 4 sensores, las vibraciones medidas en el eje Y por los 4 sensores y las vibraciones medidas en el eje Z por los 4 sensores. La base de datos de septiembre cuenta con 1.186.341 observaciones, la de octubre con 991.653, la de noviembre con 804.601, la de diciembre 501.099, la de enero con 1.035.938 y la general con 4.519.632 observaciones.

5.1. Análisis descriptivo de las variables

La base de datos cuenta con 18 variables, la primera es el tiempo, que es una variable continua con formato “año-mes-día hora:minutos:segundos” desde el 4 de septiembre de 2022 a las 11:31:02 hasta el 2 de febrero de 2023 a las 23:59:58, medida cada dos segundos.

Las variables 2, 3, 4 y 5, son las temperaturas medidas desde el sensor 1 al sensor 4 ($T1$, $T2$, $T3$, $T4$). Estas se ven afectadas por las circunstancias climatológicas externas, es decir, en verano la temperatura medida por los sensores es más alta que la medida en otoño o invierno, esto se ve claramente en la Figura 5.1, en la cual se recogen las temperaturas medias, las máximas y las mínimas en cada mes y la temperatura media de todo el período estudiado.

	SEPT 22	OCT 22	NOV 22	DIC 22	ENERO 23	Mean pob
Temperatura 1	mean: 22.09 min: 9.0 max: 36.1	mean: 18.77 min: 6.62 max: 31.4	mean: 13.3 min: 3.62 max: 23.1	mean: 15.77 min:2.5 max: 24.3	mean: 10.6 min:0.25 máx: 21	16.46
Temperatura 2	mean: 26.34 min: 13.6 max: 40.8	mean: 23.33 min: 10.3 max: 37.4	mean: 18.09 min: 7.5 max: 27.4	mean: 20.91 min: 5.62 max: 27.4	mean: 16.52 min: 4.0 máx: 29.6	21.36
Temperatura 3	mean: 59.92 min: 19 max: 75.1	mean: 59.52 min: 10.5 max: 75.9	mean: 50.45 min: 7.62 max: 65.6	mean: 55.78 min: 5.87 max: 65.4	mean: 50.9 min: 4.25 máx: 64	55.62
Temperatura 4	mean: 65.23 min: 18.6 max: 81.4	mean: 62.49 min: 10.9 max: 77.8	mean: 55.06 min: 7.87 max: 69.8	mean: 59.04 min: 6.25 max: 67.3	mean: 54.38 min: 6.25 máx: 71.9	59.65

Figura 5.1: Temperaturas medias, máximas y mínimas

Como se muestra en los histogramas y los gráficos QQ-plot de la Figura 5.2, ninguna de las temperaturas sigue una distribución normal aunque la Temperatura 1 y la Temperatura 2 se aproximan.

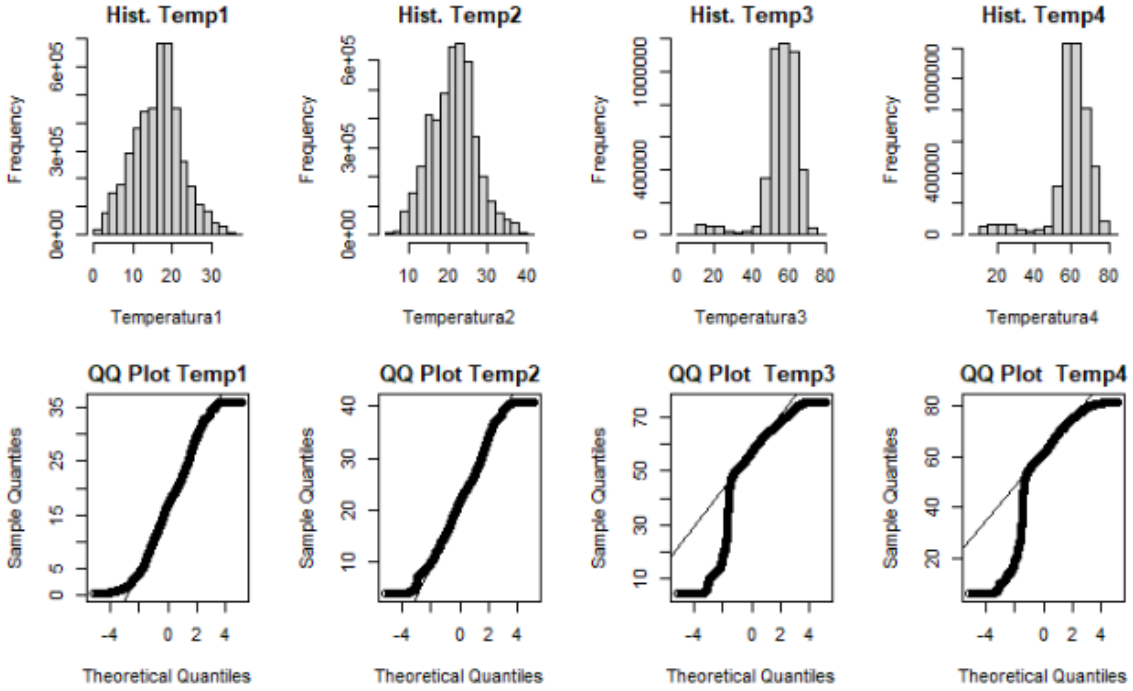


Figura 5.2: Histogramas y QQ-Plots de las temperaturas medidas en los 4 sensores

El rango de valores habituales y, por lo tanto, aceptables de las temperaturas son:
Para la temperatura 1 (T1), entre 2 y 34 °C.
Para la temperatura 2 (T2), entre 5 y 38 °C.
Para la temperatura 3 (T3), entre 45 y 70 °C.
Para la temperatura 4 (T4), entre 56 y 80 °C.

Como se muestra en el Cuadro 5.1, entre $T1$ y $T2$ y $T3$ y $T4$ hay correlaciones lineales positivas muy fuertes. Entre $T1$ y $T3$, $T1$ y $T4$, $T2$ y $T3$ y $T2$ y $T4$ también, pero menos fuertes. Esto se debe a que las mediciones de temperatura se realizan en el mismo momento por los 4 sensores, por ese motivo existe una relación y más fuerte cuanto más cerca se encuentran los sensores.

	T1	T2	T3	T4
T1	1	0.97	0.43	0.36
T2		1	0.56	0.49
T3			1	0.92
T4				1

Cuadro 5.1: Correlación entre las temperaturas medidas en los 4 sensores

La variable 6 es el consumo del motor (*Consumo*), el QQ-plot nos permite predecir que pertenece a una familia no gaussiana. Sus valores normales en activo oscilan entre 28 y 36 amperios (A), ya que cuando está parado su consumo es 0 y cuando se enciende sufre picos de hasta 60 A que fueron eliminados.

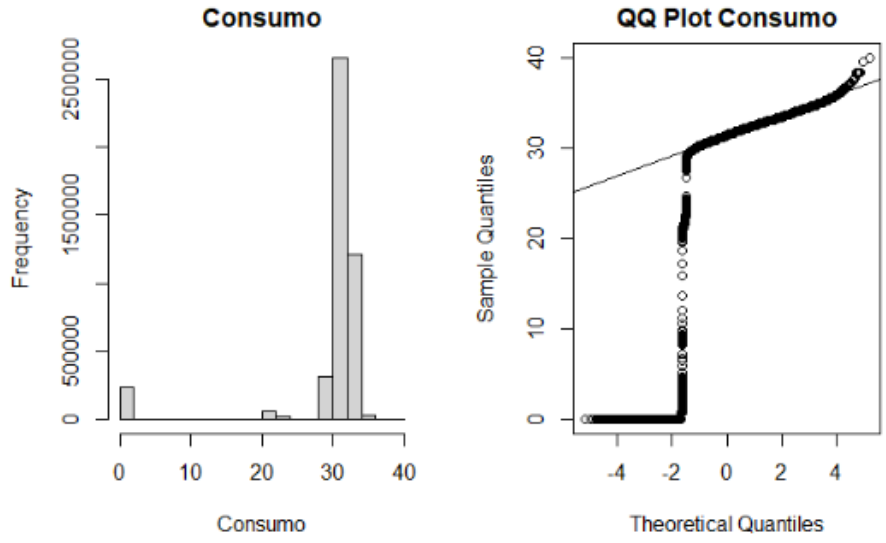


Figura 5.3: Histograma y QQ-Plot del consumo

Las variables 7, 8, 9 y 10 son las mediciones de las vibraciones en el eje X de los sensores 1, 2, 3 y 4 respectivamente ($X1$, $X2$, $X3$, $X4$). Se debe tener en cuenta que el eje X en los sensores 1 y 2 se corresponde con el eje radial, mientras que en los sensores 3 y 4 con el eje tangencial.

Si se observan los histogramas y los QQ-plot de la Figura 5.4, se puede ver que las vibraciones en el eje X medidas en el sensor 3 siguen una distribución gaussiana, en cambio, las medidas en el sensor 1, el sensor 2 y el sensor 4 no.

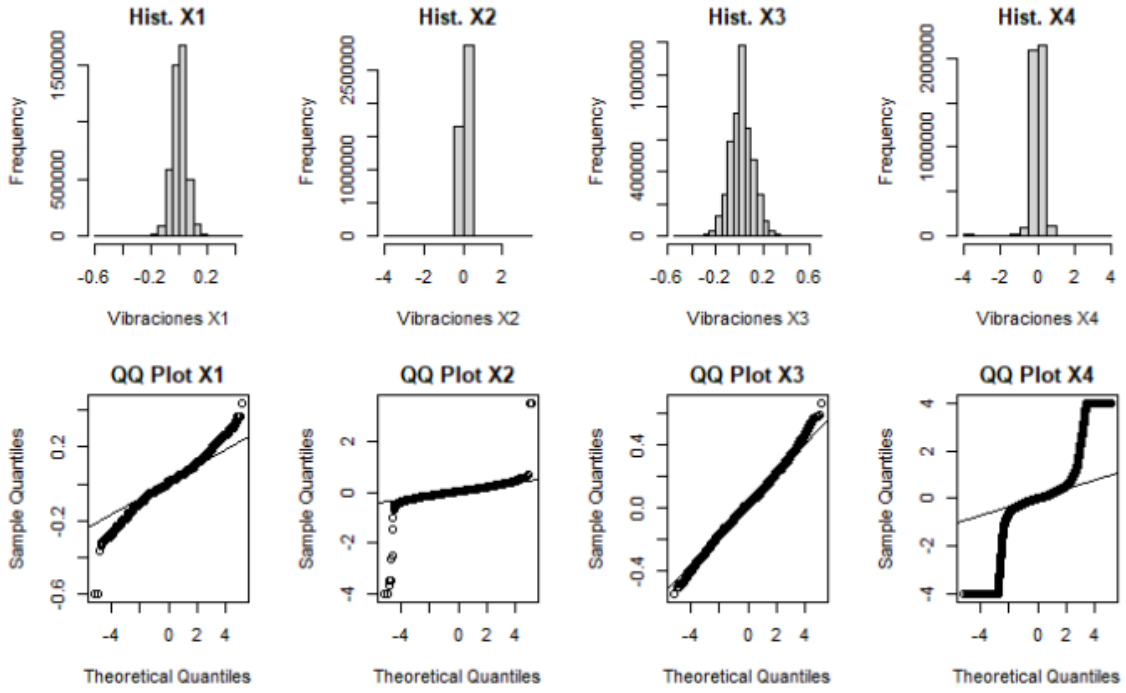


Figura 5.4: Histogramas y QQ-Plots de las vibraciones del eje X medidas en los 4 sensores

Los valores de las vibraciones en el eje X no dependen de factores temporales o estacionales, se mantienen estables en los meses estudiados. Sus valores normales y aceptables son:
Vibraciones en el eje X sensor 1: -0.25 y 0.25.
Vibraciones en el eje X sensor 2: -0.5 y 0.5.
Vibraciones en el eje X sensor 3: -0.45 y 0.5.
Vibraciones en el eje X sensor 4: -2 y 2.

Las variables 11, 12, 13 y 14 son las mediciones de las vibraciones del eje Y en los sensores 1, 2, 3 y 4 ($Y1$,$Y2$,$Y3$,$Y4$) respectivamente. Cabe recordar que el eje Y en los sensores 1 y 2 se corresponde con el eje tangencial y en los sensores 3 y 4 se corresponde con el eje radial.

Como puede verse en los histogramas y QQ-plots representados en la Figura 5.5, las vibraciones en el eje Y no siguen una distribución gaussiana. Sus valores aceptables son:
Vibraciones en el eje Y sensor 1: -0.5 y 0.5.
Vibraciones en el eje Y sensor 2: -0.5 y 0.6.
Vibraciones en el eje Y sensor 3: -0.3 y 0.3.
Vibraciones en el eje Y sensor 4: -0.75 y 0.75.

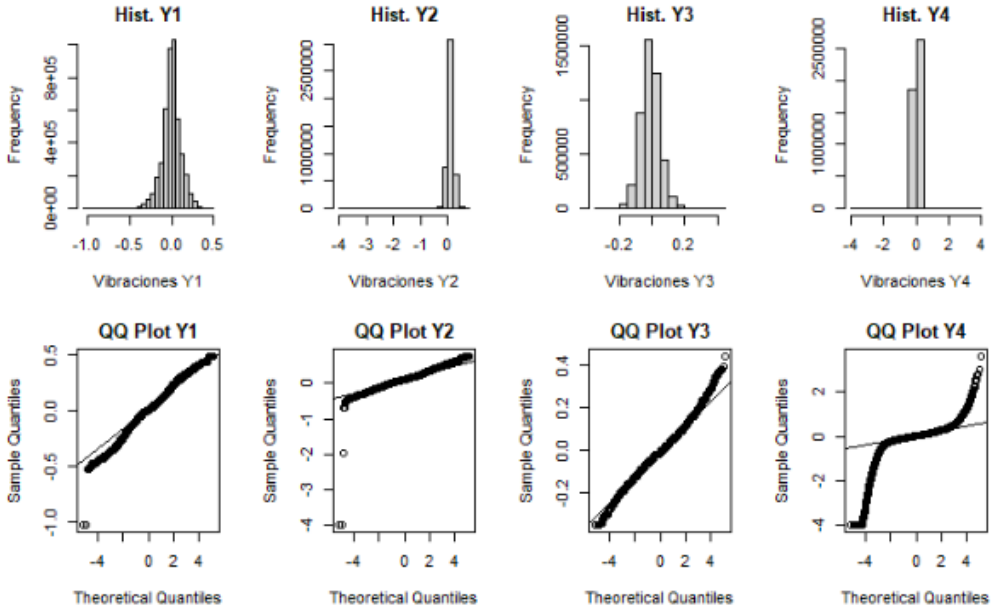


Figura 5.5: Histogramas y QQ-Plots de las vibraciones del eje Y medidas en los 4 sensores

Las variables 15, 16, 17 y 18 son las mediciones del eje Z provenientes de los sensores 1, 2, 3 y 4 ($Z1$, $Z2$, $Z3$, $Z4$) respectivamente, en todos los casos se corresponde con el eje axial.

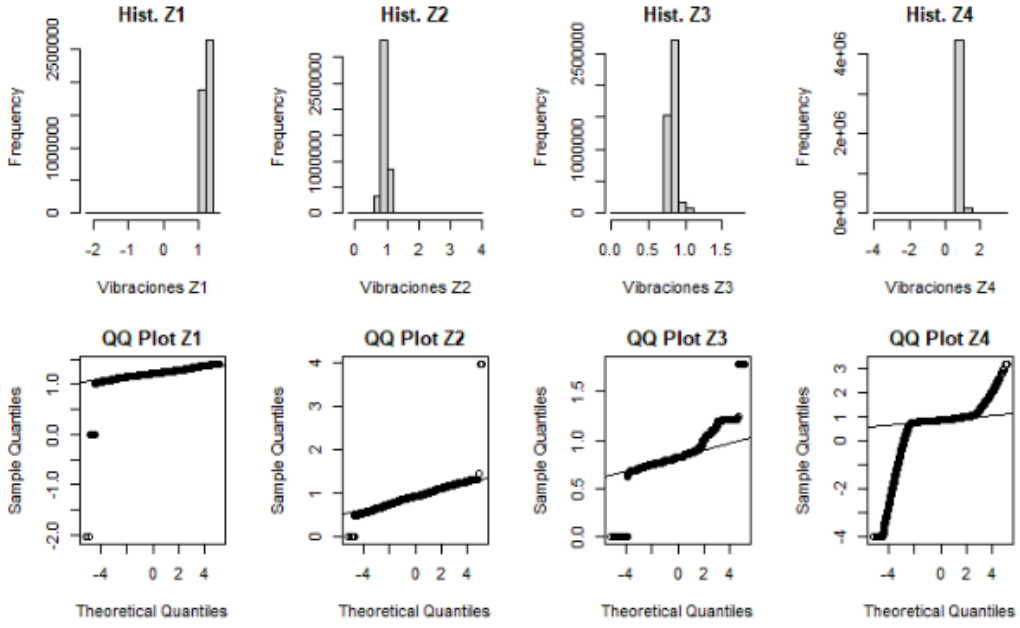


Figura 5.6: Histogramas y QQ-Plots de las vibraciones del eje Z medidas en los 4 sensores

Como puede verse en los histogramas y QQ-plots representados en la Figura 5.6, ninguna de las vibraciones del eje Z sigue una distribución normal. Sus valores aceptables o límites de especificación, definidos por los clientes, la empresa o la normativa son:
Vibraciones en el eje Z sensor 1: 1 y 1.4.
Vibraciones en el eje Z sensor 2: 0.5 y 1.3.
Vibraciones en el eje Z sensor 3: 0.6 y 1.
Vibraciones en el eje Z sensor 4: 0.5 y 1.2.

En cuanto a las correlaciones entre los ejes y los sensores, que se muestran en la Figura 5.7 y la Figura 5.8, en general son bajas. Aunque existen algunas correlaciones superiores a 0.5 que sería interesante destacar.

Por un lado, existen correlaciones dentro de los sensores: en el sensor 1, el eje Y y el eje Z, presentan una correlación negativa de -0.56; en el sensor 2, el eje Y y el eje Z, presentan también una correlación negativa de -0.75 (bastante fuerte); por último, en el sensor 4, el eje X y el eje Z, muestran una correlación positiva de 0.58.

	THF021305911VentiladorMW.X01	THF021305911VentiladorMW.Z01	THF021305911VentiladorMW.Y01
THF021305911VentiladorMW.X01	1.000000000	-0.009997138	0.02842798
THF021305911VentiladorMW.Z01	-0.009997138	1.000000000	-0.55623469
THF021305911VentiladorMW.Y01	0.028427979	-0.556234691	1.00000000
	THF021305911VentiladorMW.X02	THF021305911VentiladorMW.Z02	THF021305911VentiladorMW.Y02
THF021305911VentiladorMW.X02	1.0000000	0.1269607	-0.1535362
THF021305911VentiladorMW.Z02	0.1269607	1.0000000	-0.7472646
THF021305911VentiladorMW.Y02	-0.1535362	-0.7472646	1.0000000
	THF021305911VentiladorMW.X03	THF021305911VentiladorMW.Z03	THF021305911VentiladorMW.Y03
THF021305911VentiladorMW.X03	1.0000000	-0.2345670	0.4397644
THF021305911VentiladorMW.Z03	-0.2345670	1.0000000	-0.1617737
THF021305911VentiladorMW.Y03	0.4397644	-0.1617737	1.0000000
	THF021305911VentiladorMW.X04	THF021305911VentiladorMW.Z04	THF021305911VentiladorMW.Y04
THF021305911VentiladorMW.X04	1.0000000	0.58055317	-0.22123403
THF021305911VentiladorMW.Z04	0.5805532	1.00000000	0.07262169
THF021305911VentiladorMW.Y04	-0.2212340	0.07262169	1.00000000

Figura 5.7: Correlación entre los ejes de los mismos sensores

Por otro lado, existe relación entre diferentes sensores, en concreto, las vibraciones del eje Y del sensor 1 y del sensor 2 presentan una correlación negativa de -0.63.

	THF021305911VentiladorMW.X01	THF021305911VentiladorMW.X02	THF021305911VentiladorMW.X03	THF021305911VentiladorMW.X04
THF021305911VentiladorMW.X01	1.000000000	-0.029073930	-0.007061262	-0.002736289
THF021305911VentiladorMW.X02	-0.029073930	1.000000000	0.008472659	0.001926859
THF021305911VentiladorMW.X03	-0.007061262	0.008472659	1.000000000	0.010611120
THF021305911VentiladorMW.X04	-0.002736289	0.001926859	0.010611120	1.000000000
	THF021305911VentiladorMW.Y01	THF021305911VentiladorMW.Y02	THF021305911VentiladorMW.Y03	THF021305911VentiladorMW.Y04
THF021305911VentiladorMW.Y01	1.00000000	-0.63939526	0.03234948	0.02625877
THF021305911VentiladorMW.Y02	-0.63939526	1.00000000	-0.03202858	-0.02259363
THF021305911VentiladorMW.Y03	0.03234948	-0.03202858	1.000000000	0.006536072
THF021305911VentiladorMW.Y04	0.02625877	-0.02259363	0.006536072	1.000000000
	THF021305911VentiladorMW.Z01	THF021305911VentiladorMW.Z02	THF021305911VentiladorMW.Z03	THF021305911VentiladorMW.Z04
THF021305911VentiladorMW.Z01	1.00000000	-0.30969663	0.05669450	0.02612539
THF021305911VentiladorMW.Z02	-0.30969663	1.00000000	-0.09697182	-0.03518770
THF021305911VentiladorMW.Z03	0.05669450	-0.09697182	1.00000000	0.03593181
THF021305911VentiladorMW.Z04	0.02612539	-0.03518770	0.03593181	1.00000000

Figura 5.8: Correlación entre los ejes de los diferentes sensores

Por último se añadió una variable 19, indicadora de la existencia o no de anomalías, por eso se denomina .ªnomalía". Se trata de una variable dicotómica con categorías de respuesta 0 -no anomalía- y 1 -anomalía-. Como es habitual en estos casos, en los cuales las anomalías son menos frecuentes que los estados de funcionamiento normal, las clases están desbalanceadas.

Vamos a analizar ahora la relación entre las diferentes variables, comenzando con la temperatura.

La Figura 5.9 muestra que la temperatura puede estar relacionada con las vibraciones en los dife-

rentes ejes y sensores. Más concretamente, se observa que en el sensor 1 el eje Y tiende a vibrar más –mayor amplitud– a partir de 7°C; en el sensor 3 el eje Z al aumentar la temperatura disminuye el nivel de las vibraciones hasta los 35°C aproximadamente que se estabilizan, en cambio los ejes X e Y a partir de 40°C aumentan sus vibraciones –mayor rango–. Por último, en el sensor 4, entre los 20 y los 40 grados, se encuentran las vibraciones más fuertes –mayores amplitudes– en los 3 ejes, aunque, en el eje X, en torno a los 60°C vuelve a aumentar.

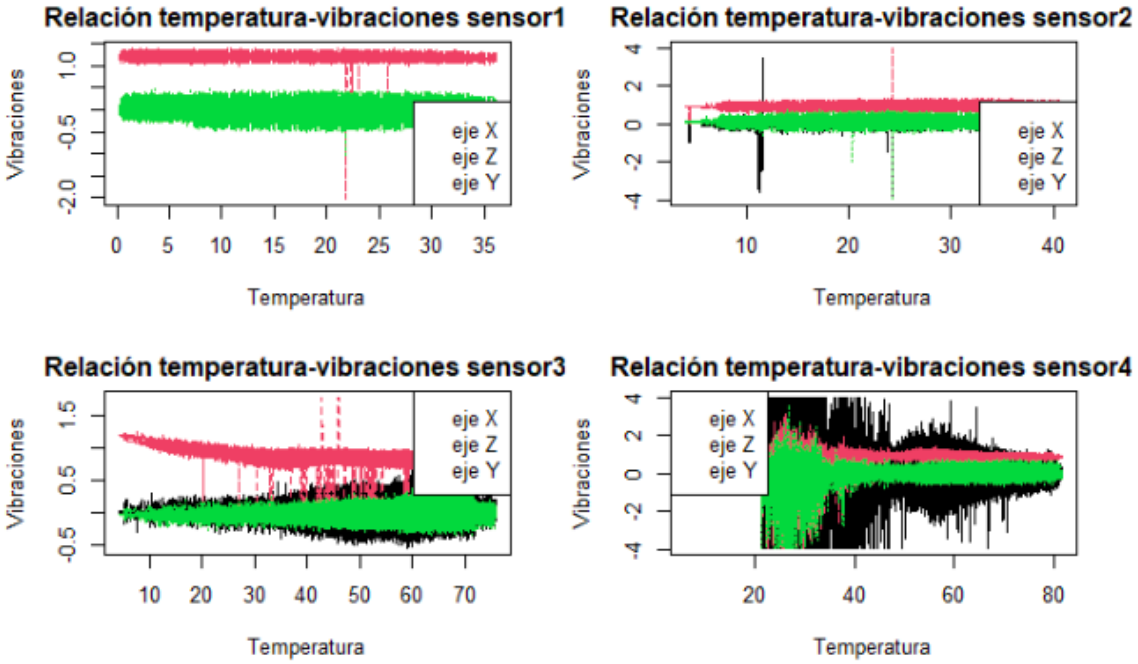


Figura 5.9: Relación entre la temperatura y las vibraciones en los diferentes ejes

En cambio si analizamos las correlaciones –Cuadro 5.2–, podemos observar que en general son bajas, aunque destaca una correlación negativa de valor -0.81, esta se produce entre la temperatura medida en el sensor 3 y la vibración del eje Z en el mismo sensor.

	T1		T2		T3		T4
X1	-0.11	X2	0.03	X3	0.03	X4	0.11
Z1	-0.16	Z2	0.24	Z3	-0.81	Z4	0.07
Y1	-0.01	Y2	0.03	Y3	0.05	Y4	0.07

Cuadro 5.2: Correlaciones de Pearson entre las temperaturas y las vibraciones medidas en los mismos sensores

Al contrario de lo que pudiésemos pensar de las correlaciones entre temperatura y consumo [1]

[1] Debido a la colocación de los sensores, se podría pensar que $T1$ y $T2$ tendrían mayores correlaciones con el consumo.

(Cuadro 5.3) las correlaciones más altas –aunque siguen siendo bajas– son entre el consumo y la temperatura 3 y la temperatura 4.

	T1	T2	T3	T4
Consumo	-0.19	-0.05	0.55	0.53

Cuadro 5.3: Correlaciones de Pearson entre las temperaturas y las vibraciones medidas en los mismos sensores

Entre el consumo y las vibraciones en los diferentes ejes y sensores (Cuadro 5.4), no hay ninguna correlación notable, la más alta es de -0.46, surge entre el consumo y las vibraciones del eje Z del sensor 3.

	Consumo
X1	0.015
Z1	0.035
Y1	0.003
X2	0.014
Z2	-0.005
Y2	-0.032
X3	0.018
Z3	-0.456
Y3	0.03
X4	-0.012
Z4	0.024
Y4	0.002

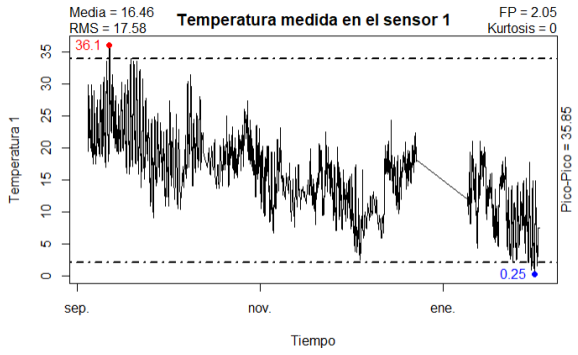
Cuadro 5.4: Correlaciones de Pearson entre el consumo y las vibraciones medidas en los mismos diferentes ejes y sensores

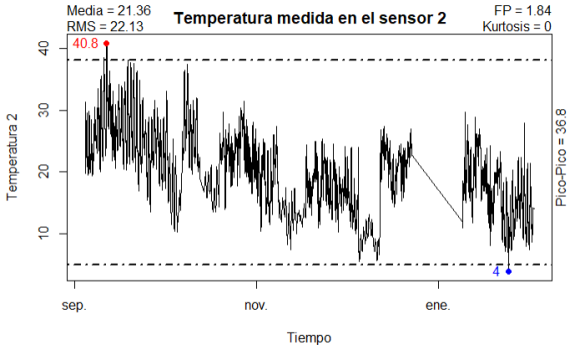
5.2. Análisis univariante

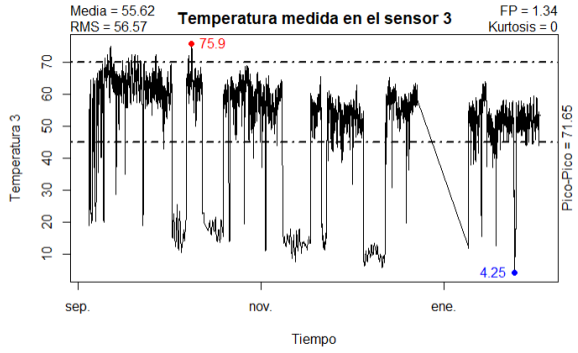
Para la detección de problemas en las máquinas de forma univariante, se ha utilizado el entorno de RStudio, con el fin de realizar la representación gráfica de las vibraciones en función del tiempo y el análisis de los indicadores estadísticos explicados en el marco teórico. Esto se lleva a cabo a través de gráficos secuenciales, en los cuales se han establecido los límites de especificación a través de la experiencia –del análisis exploratorio sobre la muestra de la que se disponía–.

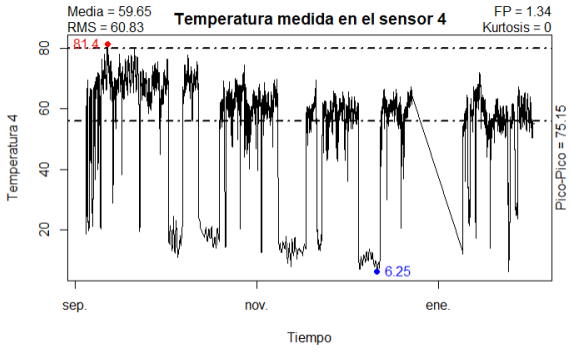
Es interesante resaltar que la empresa todavía no ha implementado métodos del control estadístico de procesos por los cuales se pueden estimar las distribuciones de las variables críticas para la calidad de los procesos y, por tanto, sus valores límite, partiendo de la realidad de la empresa, más que de unos límites fijados externamente –límites de especificación–.

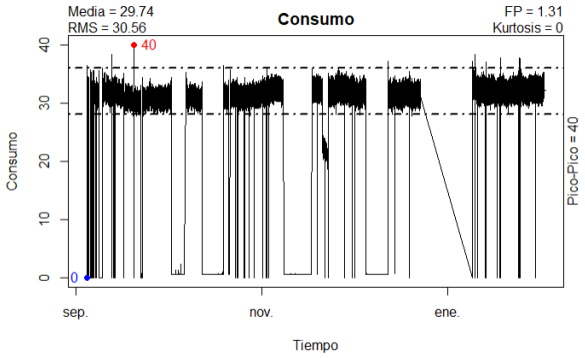
A continuación se muestran los gráficos secuenciales obtenidos para todo el periodo de estudio de cada una de las variables.

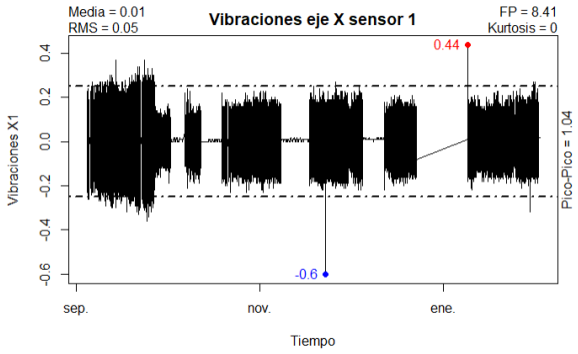


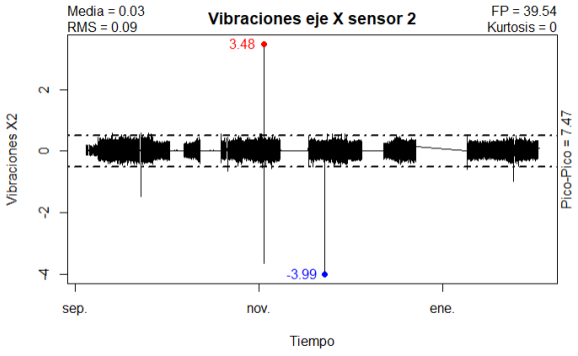


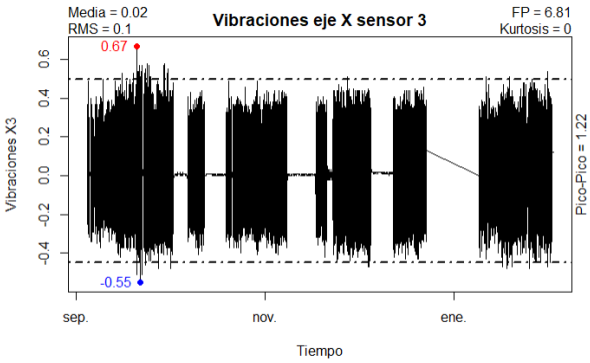


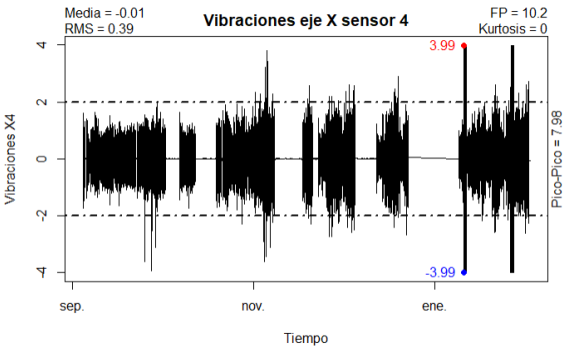


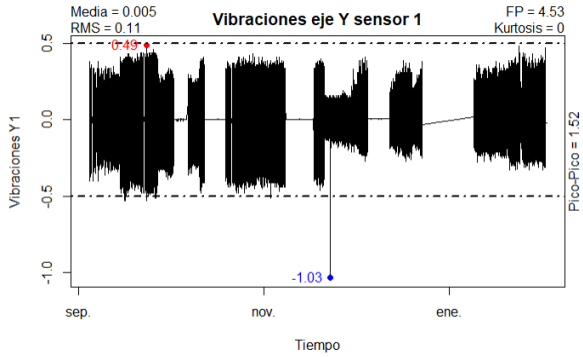

Media = 29.74
RMS = 30.56
Consumo
FP = 1.31
Kurtosis = 0
40
0
Consumo
Pico-Pico = 40
sep.
nov.
ene.
Tiempo

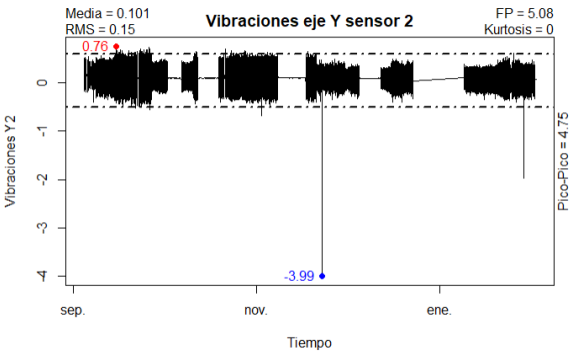

Media = 0.01
RMS = 0.05
Vibraciones eje X sensor 1
FP = 8.41
Kurtosis = 0
0.44
-0.6
Vibraciones X1
Pico-Pico = 1.04
sep.
nov.
ene.
Tiempo

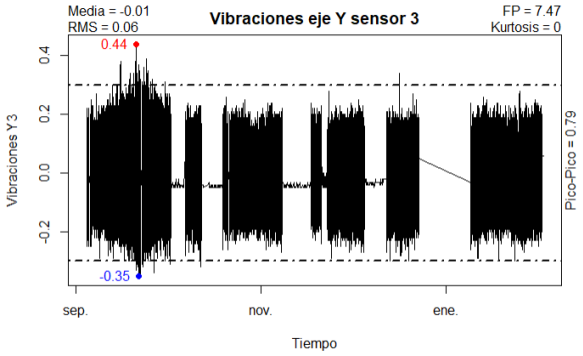

Media = 0.03
RMS = 0.09
Vibraciones eje X sensor 2
FP = 39.54
Kurtosis = 0
3.48
-3.99
Vibraciones X2
Pico-Pico = 7.47
sep.
nov.
ene.
Tiempo

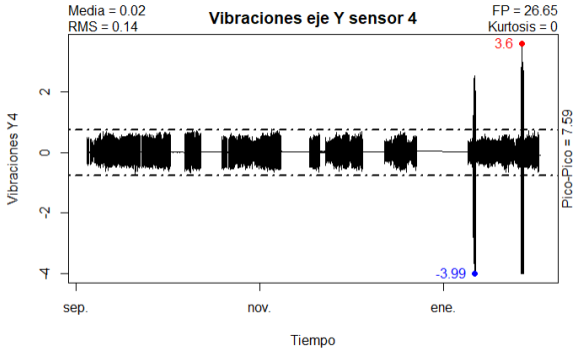

Media = 0.02
RMS = 0.1
Vibraciones eje X sensor 3
FP = 6.81
Kurtosis = 0
0.67
-0.55
Vibraciones X3
Pico-Pico = 1.22
sep.
nov.
ene.
Tiempo


Media = -0.01
RMS = 0.39
Vibraciones eje X sensor 4
FP = 10.2
Kurtosis = 0
3.99
-3.99
Vibraciones X4
Pico-Pico = 7.98
sep.
nov.
ene.
Tiempo


Media = 0.005
RMS = 0.11
Vibraciones eje Y sensor 1
FP = 4.53
Kurtosis = 0
0.49
-1.03
Vibraciones Y1
Pico-Pico = 1.52
sep.
nov.
ene.
Tiempo


Media = 0.101
RMS = 0.15
Vibraciones eje Y sensor 2
FP = 5.08
Kurtosis = 0
0.76
-3.99
Vibraciones Y2
Pico-Pico = 4.75
sep.
nov.
ene.
Tiempo


Media = -0.01
RMS = 0.06
Vibraciones eje Y sensor 3
FP = 7.47
Kurtosis = 0
0.44
-0.35
Vibraciones Y3
Pico-Pico = 0.79
sep.
nov.
ene.
Tiempo


Media = 0.02
RMS = 0.14
Vibraciones eje Y sensor 4
FP = 26.65
Kurtosis = 0
3.6
-3.99
Vibraciones Y4
Pico-Pico = 7.59
sep.
nov.
ene.
Tiempo

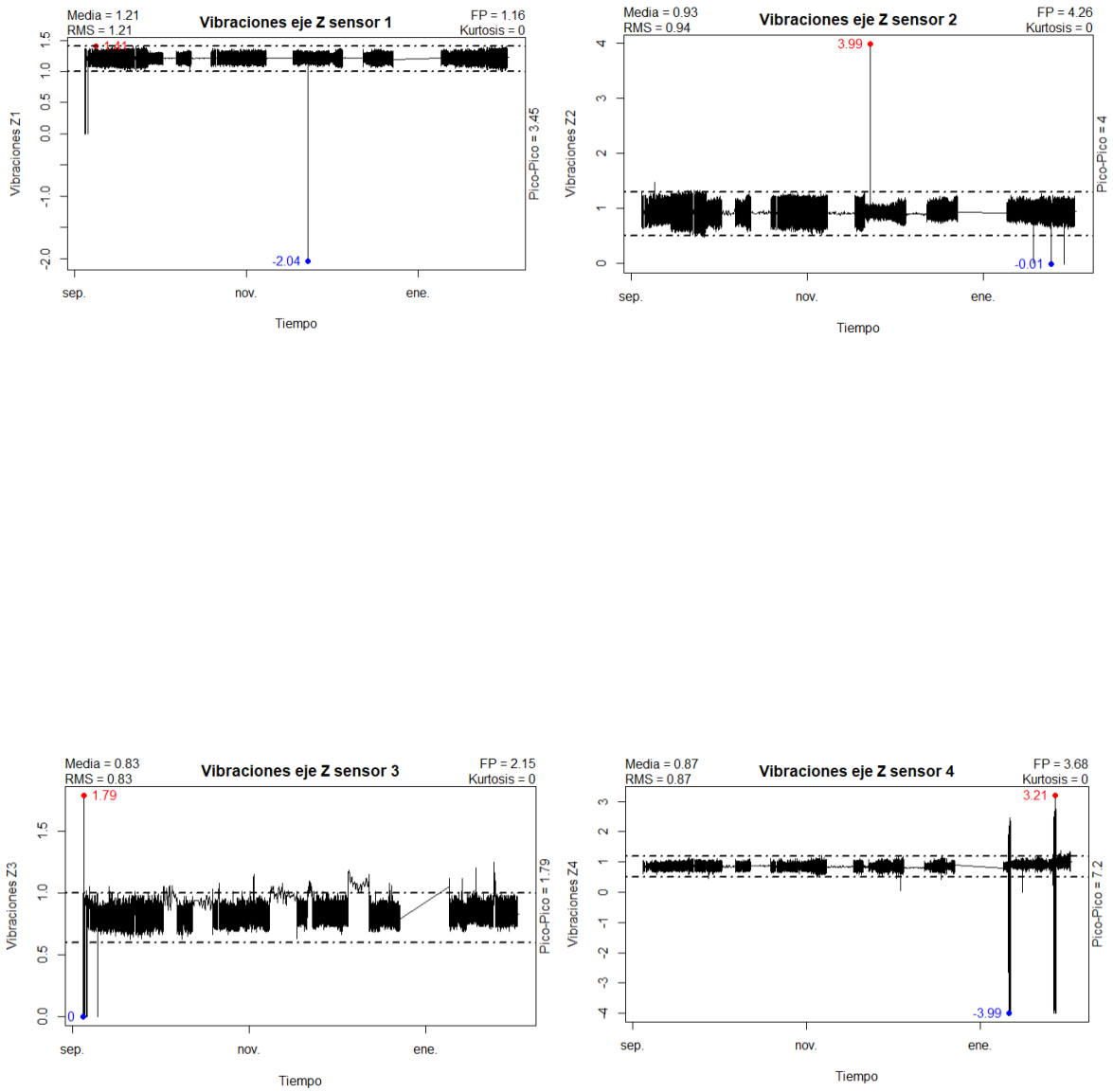


Figura 5.10: Gráficos secuenciales de las variables $T1$, $T2$, $T3$, $T4$, $Consumo$, $X1$, $X2$, $X3$, $X4$, $Y1$, $Y2$, $Y3$, $Y4$, $Z1$, $Z2$, $Z3$, $Z4$

En los datos analizados se han detectado 3 momentos en los que el ventilador ha fallado y se ha podido identificar el motivo:

- El 21 de noviembre de 2022 de 12:23 a 15:49 se han detectado valores anómalos debido a la ruptura de la correa. En este caso, el valor del consumo se sitúa entre los 15 y los 28 amperios, la temperatura del sensor 3 y del sensor 4 se sitúa por debajo de sus valores normales, las vibraciones de $X1$, $X2$, $X4$, $Y1$, $Y2$, $Y3$, $Y4$ y $Z1$ disminuyen su amplitud.

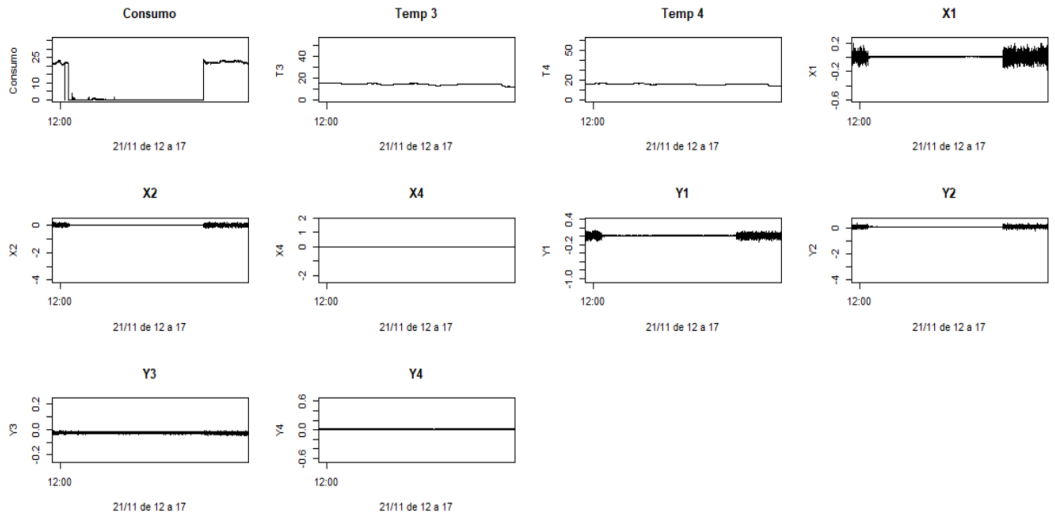


Figura 5.11: Valores anómalos ruptura de correa

- Del 10 de enero de 2023 a las 17:25 al 11 de enero de 2023 a las 12:55 se ha detectado necesidad de engrasar el rodamiento. En este caso la temperatura en el sensor 3 y en el sensor 4 descienden superando sus valores aceptables inferiores y las vibraciones en el sensor 4 en los 3 ejes ($X4$, $Y4$ y $Z4$) aumentan su amplitud.

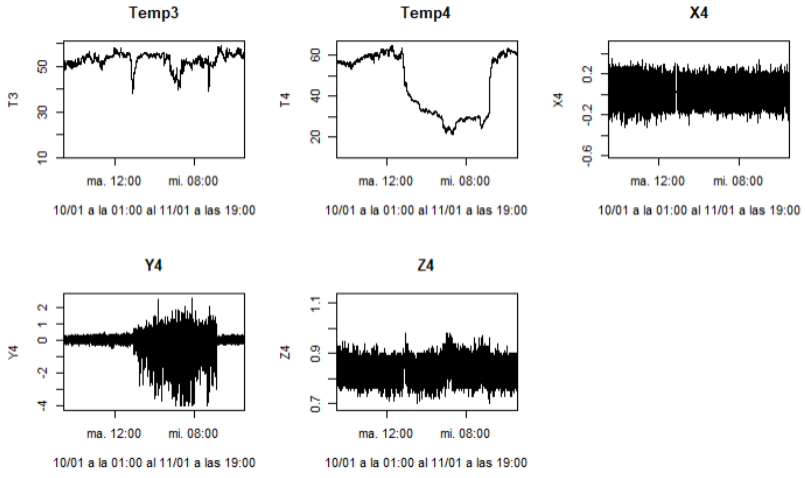


Figura 5.12: Valores anómalos falta de grasa

- Del 26 de enero de 2023 a las 13:01 al 27 de enero de 2023 a las 11:38 se detecta de nuevo falta de engrasar el rodamiento. Se repite el patrón anterior pero la temperatura 3 desciende en menor medida.

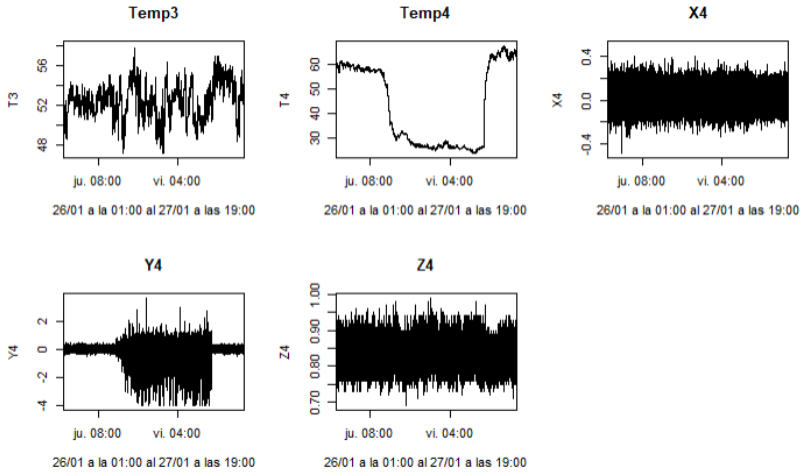


Figura 5.13: Valores anómalos falta de grasa

Tras analizar detenidamente las gráficas se han localizado algunas anomalías no clasificadas, estas podrían ser fallos no clasificados, pruebas en la máquina o no ser nada. Vamos a ir estudiando los posibles fallos por orden cronológico, en el mes de septiembre se detectaron 3 anomalías.

El primer funcionamiento anómalo se ha detectado entre el 10 y el 15 de septiembre. En la Figura 5.14, se puede observar que la temperatura medida en los 4 sensores aumenta –superando los límites superiores–, es más, el sensor 1, el sensor 2 y el sensor 4 obtienen en este intervalo sus temperaturas máximas –pico máximo–. También puede verse que las vibraciones en $X1$ –eje X sensor 1– y en $Y4$ –eje Y sensor 4– supera sus valores aceptables tanto superiores como inferiores, y las vibraciones en X2 rozan el límite aceptable superior.

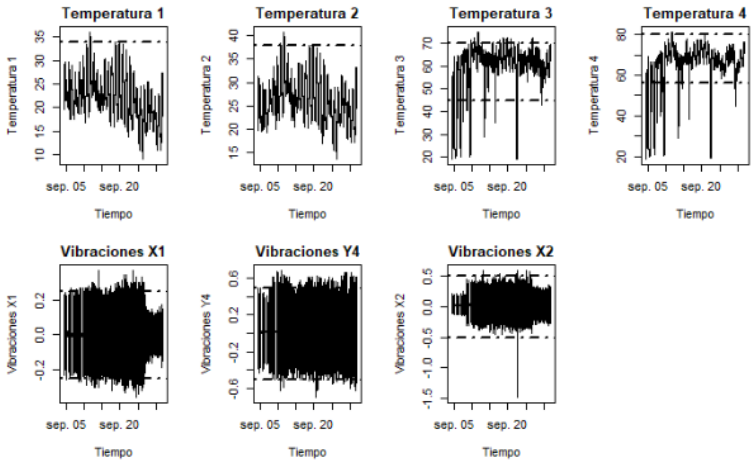


Figura 5.14: Primera anomalía localizada en el mes de septiembre

El segundo posible fallo localizado en septiembre, se produce entre el 20 y el 25 –cuarta y quinta línea del eje X de las gráficas de la Figura 5.15–. En este caso las vibraciones en $X1, X3, Y3$ e $Y4$ superaron sus valores normales tanto superiores como inferiores, de hecho tanto $X3$ como $Y3$ obtienen su pico máximo y su pico mínimo; las vibraciones en $X2$ e $Y2$ superan su límite superior, $Y2$ obtiene su valor máximo –pico máximo– y las vibraciones en $Z2$ aumentan su varianza y rozan los valores aceptables superiores e inferiores.

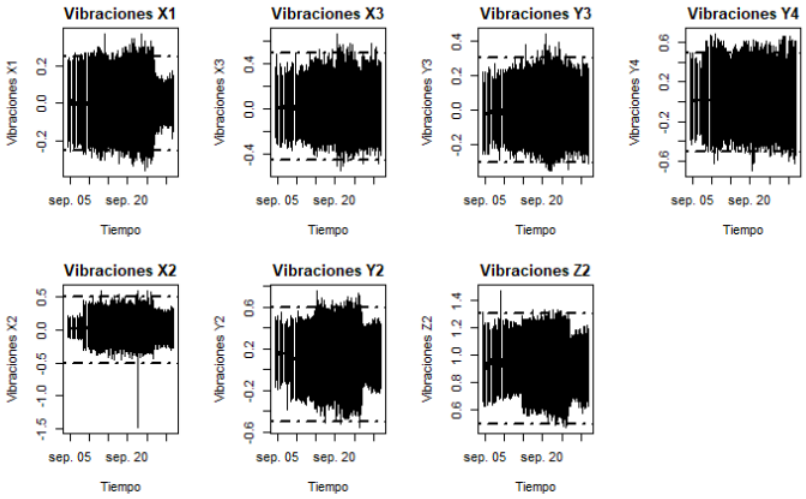


Figura 5.15: Segunda anomalía localizada en el mes de septiembre

La última posible anomalía detectada en septiembre se produjo en torno al día 30 –última línea del eje X de las gráficas de la Figura 5.16–, en ella, la temperatura medida en el sensor 3 y en el sensor 4 superan su límite inferior, las vibraciones medidas en $X3$ superan ligeramente su límite superior, las vibraciones de $X4$ superan ligeramente sus valores aceptables inferiores y las vibraciones de $Y4$ superan los valores aceptables tanto superiores como inferiores.

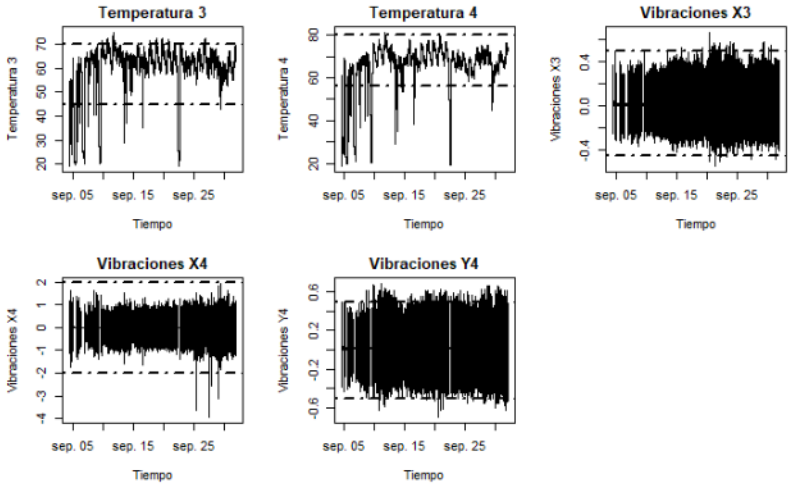


Figura 5.16: Tercera anomalía localizada en el mes de septiembre

En octubre se ha detectado únicamente una posible anomalía entre el día 5 y el día 15 –primera y segunda línea del eje X de las gráficas de la Figura 5.17–. Esta ha afectado únicamente al eje Y, el sensor 2 ha superado su límite superior, el sensor 3 supera ligeramente su límite inferior y el sensor 4 supera ambos límites –superior e inferior–.

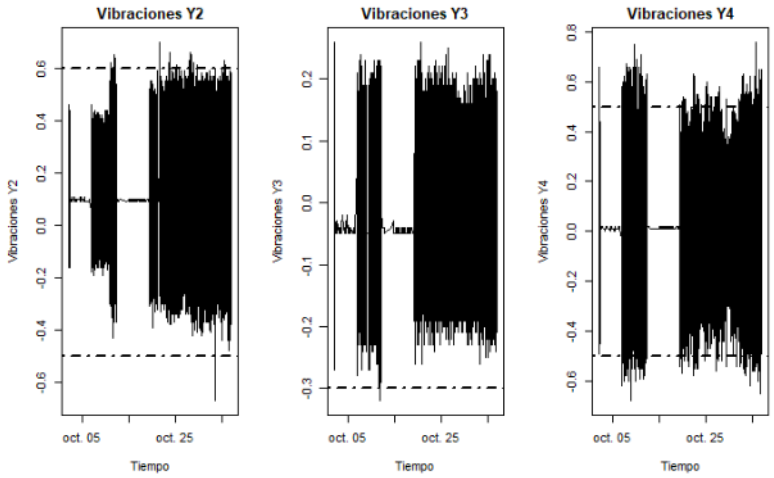


Figura 5.17: Anomalía localizada en el mes de octubre

En noviembre se han localizado 3 posibles fallos. El primero fue entre el 4 y el 5 de noviembre, puede verse en el Figura 5.18, aunque parece que representa el mes de octubre estos gráficos incluyen hasta el día 5 de noviembre –la última línea del eje X señala el día 4 de noviembre–. Este fallo afectó al sensor 4, tanto en el eje X como en el eje Y las vibraciones superaron los límites superiores e inferiores y en el eje Z las vibraciones rozan el límite inferior.

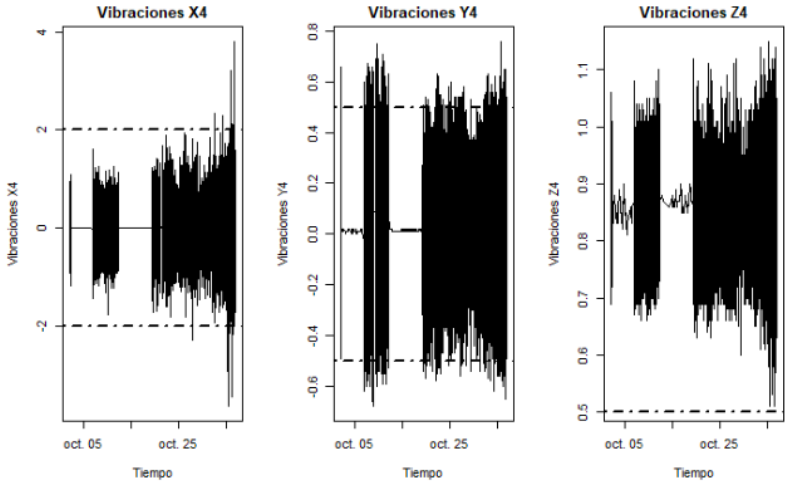


Figura 5.18: Primera anomalía localizada en el mes de noviembre

Las segundas incidencias localizadas en noviembre están muy próximas a las anteriores, se localizan entre los días 6-7 de noviembre –inicio de las gráficas mostradas en la Figura 5.19–. En este momento,

las temperaturas medidas por los sensores 3 y 4 junto con las vibraciones medidas en $X4$ sobrepasan sus límites inferiores, las vibraciones medidas en $Y4$ superan sus valores aceptables tanto inferiores como superiores y, las vibraciones en $Z4$ superan ligeramente su límite superior.

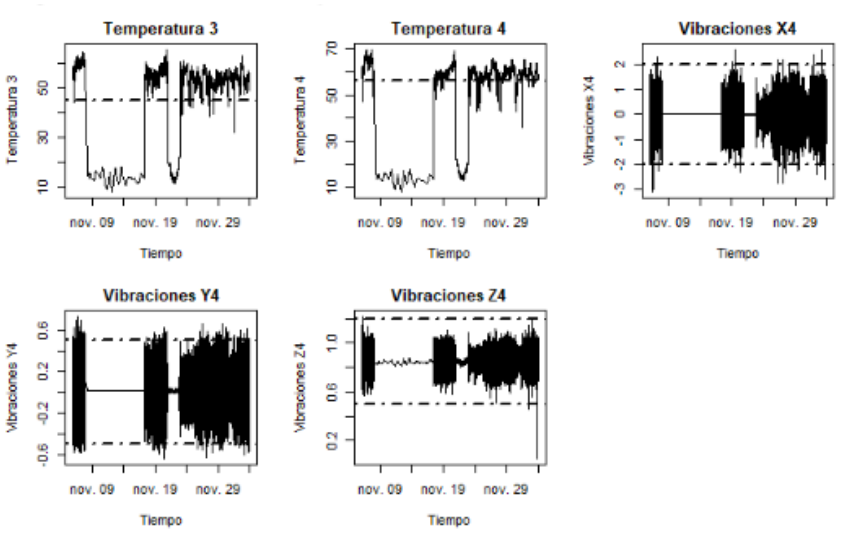


Figura 5.19: Segunda anomalía localizada en el mes de noviembre

El tercer momento es la primera anomalía clasificada, se trata de la ruptura de correa, esto sucedió el día 21 de noviembre.

En diciembre se han localizado 2 posibles fallos, el primero se ha localizado en el día 4 de diciembre en el sensor 4, la temperatura supera ligeramente el límite inferior, las vibraciones en los ejes X e Y superan los valores aceptables tanto superiores como inferiores y las vibraciones en el eje Z supera el límite inferior.

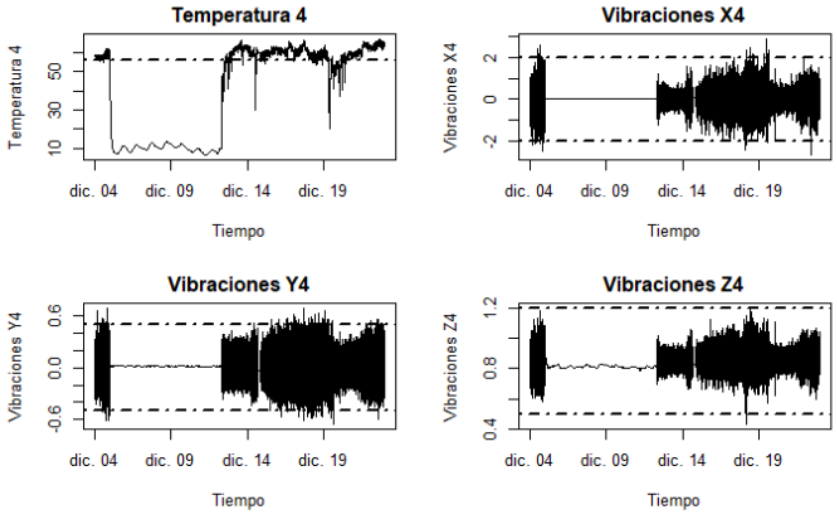


Figura 5.20: Primera anomalía localizada en el mes de diciembre

El segundo fallo de diciembre se ha localizado cerca del día 19 –cuarta línea del eje X en los gráficos de la Figura 5.21–, aquí tanto la temperatura 3 como la temperatura 4 superan su límite inferior, las vibraciones en $X4$, $Y4$ y $Z3$ superan su límite superior y luego disminuyen su varianza. Las vibraciones en $Y1$, $Z2$ y $Z4$ no superan los valores aceptables pero disminuyen su varianza.

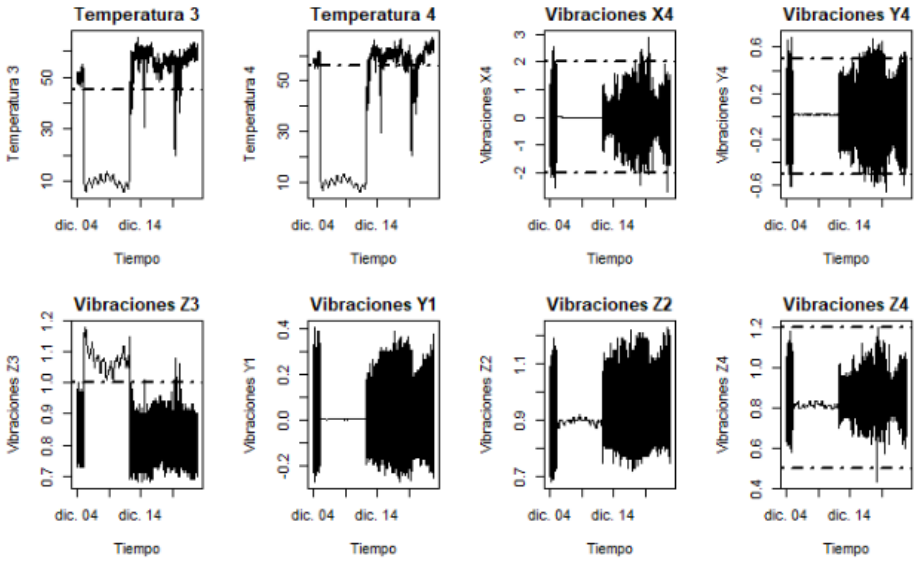


Figura 5.21: Segunda anomalía localizada en el mes de diciembre

El último mes estudiado es enero, aquí se han detectado dos anomalías que coinciden con los dos fallos etiquetados en este mes por falta de engrasar el rodamiento.

Una vez detectadas las anomalías del periodo de estudio, se ha programado un código en Python que permite localizar univariantemente las anomalías y, es capaz de clasificar si se trata de uno de los fallos ya definidos o de una anomalía nueva -en este caso, manda revisar la máquina-.

```
import pandas as pd
import math as mt
import numpy as np

filename= "C:\\Users\\aro\\Desktop\\Datos\\VentiladorMW_sept.csv"
oct= pd.read_csv(filename, header=0)
oct= oct.dropna()

datos= oct[107194:107780] #Una prueba, se ha probado con varias

#Limpiamos BBDD:
    #Consumo>40
datos= datos.drop(datos[datos["THF021305911VentiladorMW.Amps"]>40].index)
    #Picos en cualquier variable Temp, X, Y o Z superiores a 100
datos= datos.loc[(datos.iloc[:,1:18] < 100).all(axis=1), :]
    #Picos en cualquier variable Temp, X, Y o Z inferiores a -10
datos= datos.loc[(datos.iloc[:,1:18] > -10).all(axis=1), :]

col_names= ["FechaHora","Temp1","Temp2","Temp3","Temp4","Consumo","x1","x2","x3",
            "x4","Y1","Y2","Y3","Y4","z1","z2","z3","z4"]

    #Calculamos RMS, FC y avisamos si FC es mayor a 1.8
for col in datos.iloc[:,1:]:
    signal = datos[col]
```

```
    RMS = max(signal) * mt.sqrt(2)
    FC = max(signal) / RMS
    if FC > 1.8:
        col_num = col + 1
        print(f"Vigilar variable {col_names[col_num]}, posibles anomalias")
    else:
        continue

    #Calculamos kurtosis y avisamos si es mayor que 3
kurt= datos.kurt(axis = 0, skipna = True, numeric_only=bool)
    #skipna=T por si hay valores perdidos
    #numeric_only para que no haga la kurtosisi en la fecha
for i in range(len(kurt)):
    if kurt[i] > 3.0:
        index = i + 1
        print(f"Ojo, posibles fallos en {col_names[index]}")

    ######################
    #Deteccion de fallos
    ######################

    #Valores analizados para cada variable
T1= datos.iloc[:,1] ; T2= datos.iloc[:,2] ; T3= datos.iloc[:,3] ; T4= datos.iloc[:,4]
consumo= datos.iloc[:,5] ; x1 = datos.iloc[:,6]; x2= datos.iloc[:,7]
x3= datos.iloc[:,8] ; x4= datos.iloc[:,9] ; y1= datos.iloc[:,10]
y2= datos.iloc[:,11] ; y3= datos.iloc[:,12] ; y4= datos.iloc[:,13]
z1= datos.iloc[:,14];z2= datos.iloc[:,15];z3= datos.iloc[:,16];z4= datos.iloc[:,17]

#Aviso cuando falla la correa (valores conocidos)
correa = np.array([(consumo < 28) & (consumo > 15)] and [(T3 < 45)] and [(T4 < 56)]
    and [(-0.01 < x1) & (x1 < 0.03)] and  [(0 < x2) & (x2< 0.04)] and [(-0.01 < y1) & (
    y1 < 0.03)] and [(-0.04 < y3) & (y3 < -0.01)] and [(-0.01 < y4) & (y4 < 0.03)] and
    [(0.09 < y2) & (y2 < 0.12)] and [(1.19 < z1) & (z1 < 1.23)] and [(-0.02 < x4) & (x4
     < 0.03)])

#Aviso cuando necesita engrasar (valores conocidos)
engrasar = np.array([(T4 < 56)] and [(-2.1 > x4) | (x4 > 2.1)] and [(-0.7 > y4) | (y4
    > 0.7)] and [(0.4 > z4) | (z4 > 1.4)])

#Establecemos valores aceptables para cada variable
indi_x1= np.array((x1 < -0.25) | (x1 > 0.25))
indi_x2 = np.array((-0.5 > x2) | (x2 > 0.5))
indi_x3= np.array((-0.45 > x3) | (x3 > 0.5 ))
indi_x4= np.array((-2 > x4) | (x4 > 2))
indi_y1= np.array((-0.5 > y1) | (y1 > 0.5))
indi_y2= np.array((-0.5 > y2) | (y2 > 0.6))
indi_y3= np.array((-0.3 > y3) | (y3 > 0.3))
indi_y4= np.array((-0.5 > y4) | (y4 > 9.5))
indi_z1= np.array((1 > z1) | (z1 > 1.4))
indi_z2= np.array((0.5 > z2) | (z2 > 1.3))
indi_z3= np.array((0.6 > z3) | (z3 > 1))
indi_z4 = np.array((0.5 > z4) | (z4 > 1.2))

#Agrupamos los valores de las variables en una matriz por filas (cada fila es una
    variable)
indi = np.vstack((indi_x1, indi_x2, indi_x3, indi_x4, indi_y1, indi_y2, indi_y3,
    indi_y4, indi_z1, indi_z2, indi_z3, indi_z4))
    #Contar el numero de valores fuera de los limites (True) en cada fila
Num_noacept = np.sum(indi, axis=1)
    #Numero de filas con mas de 30 valores fuera de los limites
VariablesAlarma = np.sum(Num_noacept >= 30)

if (correa == True).sum() > 30:
    print("REVISAR CORREA!")

```

```
83 elif (engrasar == True).sum() > 30:
84     print("NECESITA ENGRASAR!")
85 
86 elif np.sum(VariablesAlarma) >= 3:
87     print("Revisar maquina")
88     #Si mas de tres variables tienen valores fuera de los limites
89 
90 #Nota:si se recogen observaciones cada 2 segundos, en caso de que empiece a fallar, si
       se detecta con 30 fallos, nos daremos cuenta al minuto
```

Listing 5.1: Código de Python para detectar anomalías según los patrones univariantes detectados

5.3. Análisis multivariante

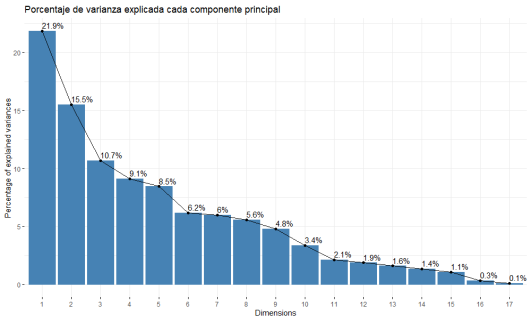
Como se comentó en anteriores secciones, se proponen diversas formas para llevar a cabo el análisis. Concretamente, en este trabajo se aplicarán gráficos de control multivariantes, además de técnicas enmarcadas en el ámbito del *machine learning*.

5.3.1. Gráficos de control

Para realizar el análisis multivariante a través de gráficos de control no es conveniente utilizar más de 10 variables explicativas (algunos estudios indican que aumentaría el ARL o número de observaciones necesarias hasta darnos cuenta que el proceso está fuera de control). Dado que en nuestro caso tenemos 17 –$T1, T2, T3, T4, Consumo, X1, X2, X3, X4, Y1, Y2, Y3, Y4, Z1, Z2, Z3, Z4$–, se ha optado por reducir la dimensionalidad a través de la aplicación del PCA. Además, como las variables están medidas en diferentes escalas, es necesario estandarizarlas previamente, previniendo así que las variables con mayor varianza –mayor escala– prevalezcan sobre el resto.

Para poder realizar un PCA, las variables deben ser linealmente dependientes. La dependencia de las variables se contrasta mediante el test de Ljung-Box, utilizando la función *Box.test* de la librería *stats* –debe especificarse *type= "Ljung-Box"*–. La hipótesis nula del test indica que las variables son independientes. La linealidad se puede comprobar de 2 formas: (1) imprimiendo el PCA y examinando si se puede explicar una apreciable cantidad de información con menos componentes principales que varibles originales existen o (2) realizando un *summary* del PCA para verificar si las desviaciones típicas de las componentes tienen un valor igual a 0 o muy próximo. En este caso, se aceptan ambas premisas –consultar Apéndice B para ver las demostraciones–.

Se realiza el PCA a través de la función *PcaCov* de la librería *rrcov* –especificaremos *scale=TRUE* para estandarizar las variables–, y se analiza el número de componentes principales que se necesitan para explicar un 90 % de la variabilidad. En la Figura 5.22, se observa que son necesarias las 10 primeras componentes principales para conseguir explicar el 91.52 %.



(a) Representación gráfica de la importancia de cada componente

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
Standard deviation     1.9275213 1.6234360 1.3470404 1.24528434 1.19889501 1.02370003 1.01030373
Proportion of Variance 0.2185493 0.1550320 0.1067363 0.09121959 0.08454996 0.06164481 0.06004198
Cumulative Proportion  0.2185493 0.3735813 0.4803177 0.57153726 0.65608722 0.71773203 0.77777400
                          Comp.8    Comp.9   Comp.10   Comp.11   Comp.12   Comp.13
Standard deviation     0.97404155 0.90319812 0.75592972 0.60111883 0.56502570 0.52081301
Proportion of Variance 0.05580923 0.04798628 0.03361351 0.02125552 0.01877965 0.01595566
Cumulative Proportion  0.83358324 0.88156952 0.91518304 0.93643856 0.95521821 0.97117386
                          Comp.14   Comp.15   Comp.16   Comp.17
Standard deviation     0.48155366 0.42503525 0.2433087 0.135263870
Proportion of Variance 0.01364082 0.01062676 0.0034823 0.001076254
Cumulative Proportion  0.98481468 0.99544145 0.9989237 1.000000000
```

(b) Summary del modelo

Figura 5.22: Elección número de componentes principales

Una vez seleccionamos las 10 primeras componentes principales, es necesario conocer la importancia de cada variable original en cada componenteo nueva variable. Esto puede hacerse a través de un estudio de correlación, mostrado en la Figura 5.23, obtenido a partir de la función *corPlot* de la librería *psych*.

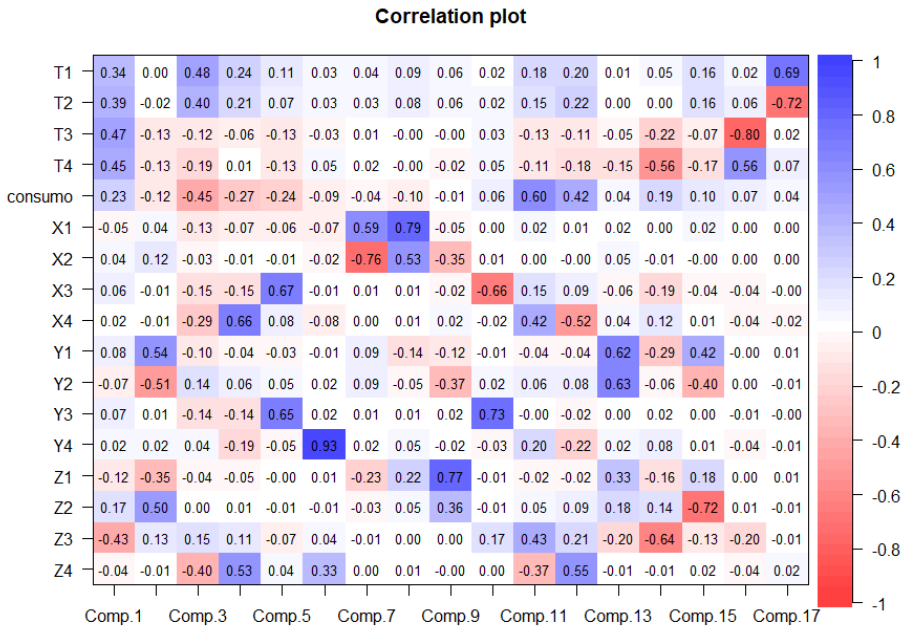


Figura 5.23: Gráfico de correlaciones. Importancia de cada variable en cada componente

Si se presta atención a las correlaciones mayores a 0.3 (en valor absoluto), se puede observar que las variables más importantes en las 10 componentes principales son:

- En la primera: $T3,\ T4,\ Z3,\ T2,\ T1$.
- En la segunda: $Y1,\ Y2,\ Z2,\ Z1$.

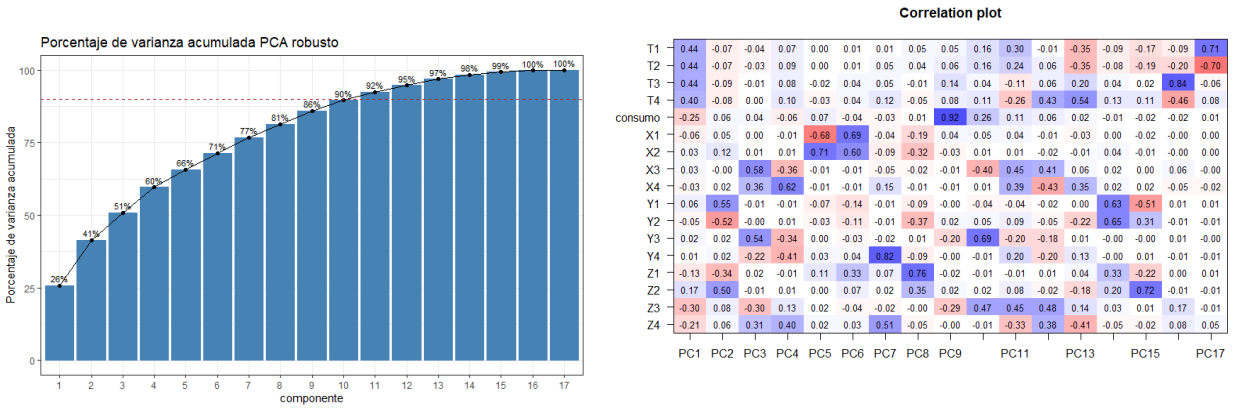
- En la tercera: $T1,\ Consumo,\ Z4,\ T2$.
- En la cuarta: $X4,\ Z4$
- En la quinta: $X3,\ Y3$
- En la sexta: $Y4,\ Z4$ (prácticamente solo $Y4$)
- En la séptima: $X2,\ X1$
- En la octava: $X1,\ X2$
- En la novena: $Z1,\ Y2,\ Z2,\ X2$
- En la décima: $Y3,\ X3$

Una vez seleccionadas las componentes principales, se procede a comprobar la eficacia del modelo. Para ello se elabora la matriz de confusión reconstruyendo las observaciones con las que se creó el PCA a través de las componentes principales seleccionadas, calculando el error cuadrático medio de la reconstrucción e identificando las observaciones anómalas a partir de los datos derivados de la reconstrucción de las observaciones -en anomalías y no anomalías- a partir de dicha reconstrucción.

En este trabajo se han considerado las anomalías como "Negativosz las NO anomalías como "Positivos". Siguiendo este criterio, la sensibilidad del modelo está próxima al 1 (0.99996), la especificidad y el NPV son 0 y el balance accuracy(BA) es 0.5. El modelo no es capaz de detectar las anomalías, esto puede deberse a que el PCA es sensible a los valores atípicos. Para intentar solventar este problema se ha utilizado el PCA Robusto. Este es una extensión del PCA que utiliza medidas de dispersión robustas –obteniendo los autovalores y autovectores a través de una matriz de covarianzas robustas o calculando cada componente a través de un método robusto– y se aplica añadiendo la opción *cov.control = CovControlMcd()*, a la función *PcaCov*.

Para realizar el PCA robusto también deberían verificarse las premisas de linealidad y dependencia. Como sigue siendo la misma BBDD, se cumplen ambas premisas.

Con las variables estandarizadas, se realiza el PCA Robusto y, como se muestra en la figura 5.24, son necesarias 10 componentes principales para obtener el 90 % de la varianza explicada.



(a) Representación gráfica de la importancia de cada componente

(b) Gráfico de correlaciones. Importancia de cada variable en cada componente

Figura 5.24: Elección número de componentes principales e importancia de cada variable en cada componente

Las variables más importantes para explicar cada una de las 10 componentes principales son:

- Componente 1: $T1,\ T2,\ T3,\ T4,\ Z3$.
- Componente 2: $Y1,\ Y2,\ Z2,\ Z1$.
- Componente 3: $X3,\ Y3,\ X4,\ Z3$.
- Componente 4: $X4,\ Y4,\ Z4,\ X3,\ Y3$.
- Componente 5: $X2,\ X1$.
- Componente 6: $X1,\ X2,\ Z1$.
- Componente 7: $Y4,\ Z4$.
- Componente 8: $Z1,\ Y2,\ Z2,\ X2$.
- Componente 9: $Consumo$
- Componente 10: $Y3,\ Z3,\ X3$.

Para evaluar la eficacia del modelo se sigue el método utilizado en el PCA. En este caso los resultados son ligeramente mejores para el objetivo del estudio –ser capaz de detectar las anomalías–, dado que la sensibilidad está muy próxima a 1, la especificidad próxima a 0 –pero no es cero (0.00049)–, la precisión balanceada del modelo es 0.5 y el NPV -dato muy importante- es 0.2.

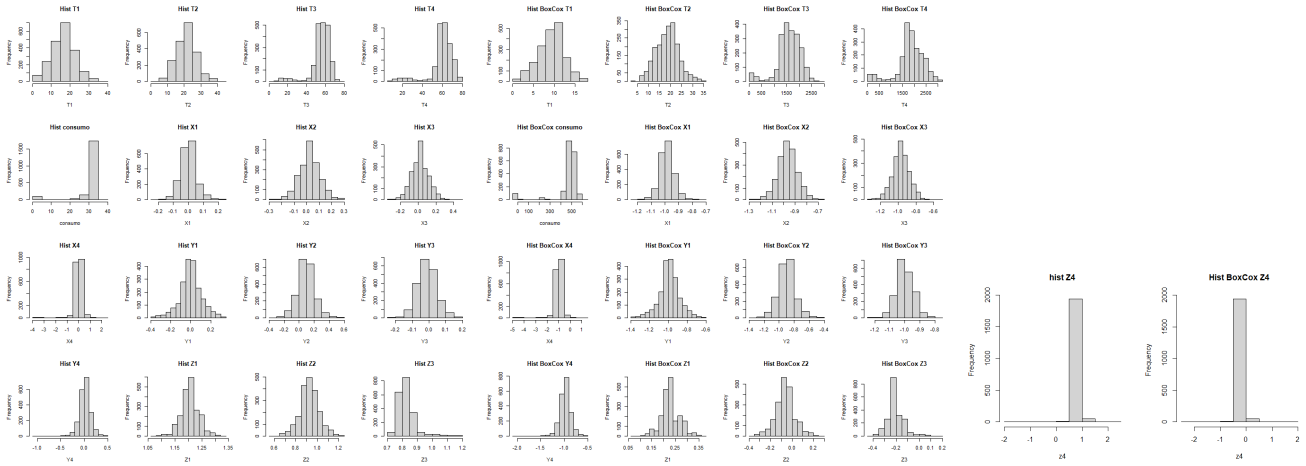
Los resultados obtenidos con los modelos anteriores no son los deseados, además, trabajar con estas bases de datos computacionalmente es muy costoso [2]. Por ese motivo, no se continuó con su análisis sino que se intentó suavizar los datos tomando medias de los mismos en intervalos de tiempo concretos, cada 30, 60 y 75 minutos.

- **Medias de datos cada 75 minutos**

Vamos a comenzar con el análisis de la media de los datos agrupados en intervalos de 75 minutos. Esta base de datos cuenta con 2008 observaciones de las cuales 40 están clasificadas como anomalías.

Al realizar esta transformación a los datos las variables se aproximan en mayor medida a una distribución gaussiana pero todavía no siguen esta distribución. Para intentar que cumplan las hipótesis de partida para la aplicación de gráficos de control, se les ha realizado una transformación Box-Cox -véase Figura 5.25-.

[2] Debido a que se trabaja con 4519632 de observaciones.

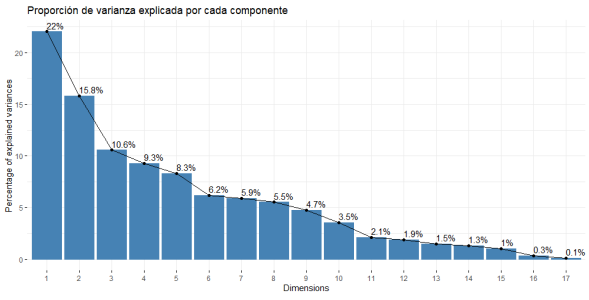


(a) Histogramas de las variables originales (b) Histogramas de las variables con la transformación BoxCox (c) Histogramas de $Z4$

Figura 5.25: Histogramas de las variables a las que se les ha realizado la media cada 75 minutos originales y con la transformación Box-Cox

Los cambios en los datos no provocan una mejora significativa, por ese motivo se seguirá trabajando con las variables sin la transformación Box-Cox. Los datos tienen combinaciones lineales –se puede explicar el 90 % de la variabilidad con 10 componentes– y $X1$, $X2$, $X3$, $Y1$, $Y2$, $Y3$, $Y4$ y $Z1$ son independientes.

Se realiza el PCA y se seleccionan las 10 primeras componentes principales para explicar un 90 % de la variabilidad (véase Figura 5.26).



(a) Representación gráfica de la importancia de cada componente

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
Standard deviation     1.9347909 1.6383542 1.3400979 1.25431542 1.18816813 1.02477544
Proportion of Variance 0.2202009 0.1578944 0.1056390 0.09254748 0.08304374 0.06177439
Cumulative Proportion  0.2202009 0.3780953 0.4837343 0.57628174 0.65932547 0.72109987
                          Comp.7     Comp.8     Comp.9    Comp.10    Comp.11
Standard deviation     0.99955437 0.96821296 0.89643555 0.77597342 0.59921672
Proportion of Variance 0.05877111 0.05514331 0.04727039 0.03541969 0.02112122
Cumulative Proportion  0.77987098 0.83501430 0.88228469 0.91770438 0.93882560
                          Comp.12    Comp.13    Comp.14    Comp.15    Comp.16
Standard deviation     0.56827687 0.50064910 0.47010382 0.41572859 0.23277699
Proportion of Variance 0.01899639 0.01474409 0.01299986 0.01016649 0.00318736
Cumulative Proportion  0.95782199 0.97256607 0.98556593 0.99573242 0.99891978
                          Comp.17
Standard deviation     0.13551288
Proportion of Variance 0.00108022
Cumulative Proportion  1.00000000
```

(b) Summary del modelo

Figura 5.26: Elección número de componentes principales

Si se observa la Figura 5.27, puede verse que las variables más importantes en las 10 componentes principales son:

- En la primera: $T3$, $T4$, $Z3$, $T2$, $T1$.
- En la segunda: $Y1$, $Y2$, $Z2$, $Z1$.

- En la tercera: $T1,\ Z4,\ Consumo,\ T2,\ X4$.
- En la cuarta: $X4,\ Z4,\ Consumo$.
- En la quinta: $X3,\ Y3$.
- En la sexta: $Y4,\ X1$ (prácticamente solo $Y4$).
- En la séptima: $X1,\ X2$.
- En la octava: $X2,\ X1,\ Y4,\ Z1$.
- En la novena: $Z1,\ Z2,\ X2,\ Y2$.
- En la décima: $Y3,\ X3$.

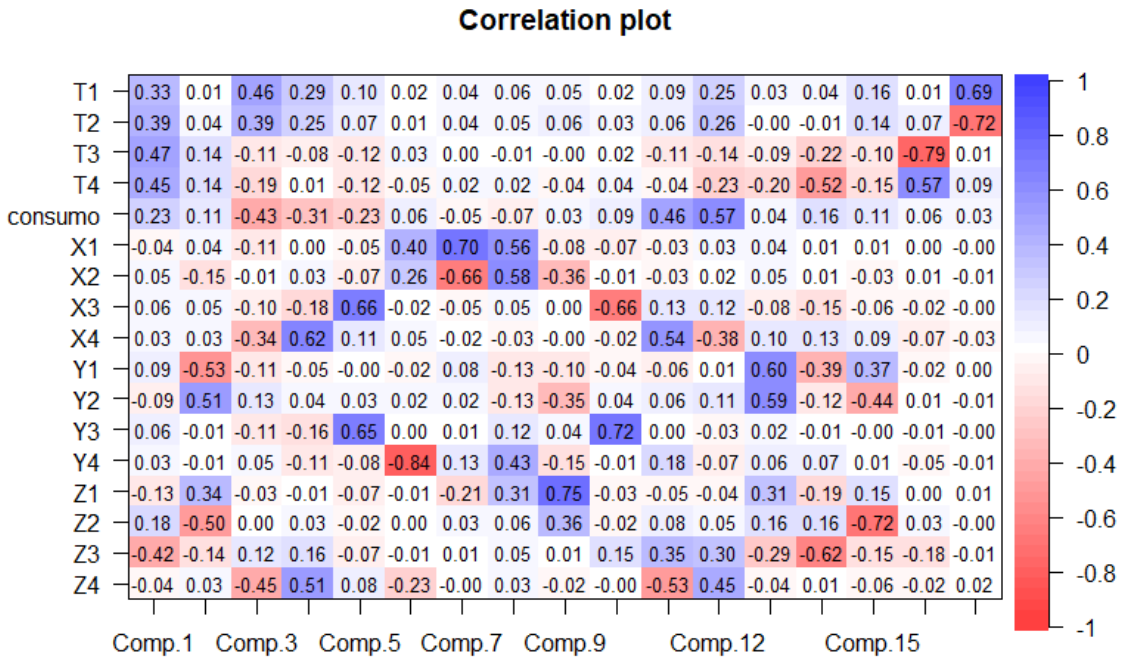


Figura 5.27: Gráfico de correlaciones. Importancia de cada variable en cada componente

Este modelo obtiene una sensibilidad de 0.93, una especificidad de 0.9, un BA de 0.91 pero un NPV de 0.2; es mejor que los modelos anteriores pero mantiene un bajo porcentaje (20 %) de anomalías reales clasificadas correctamente. El presente modelo tiene un problema de detección de falsas anomalías, detecta mal el 80 % de las anomalías clasificadas. Esto probablemente sea debido, en parte, al no cumplimiento de las hipótesis de independencia de observaciones y de normalidad de las variables.

A pesar de transformar los datos a través de las medias, algunas variables siguen manteniendo la autocorrelación, esto dificulta la aplicación de gráficos de control. Por ese motivo se llevará a cabo la metodología más habitual ante esta situación: ajustar un modelo de series de tiempo a las componentes principales seleccionadas para trabajar con los residuos.

La representación de las series de tiempo para las 10 primeras componentes principales pueden verse en la Figura 5.28.

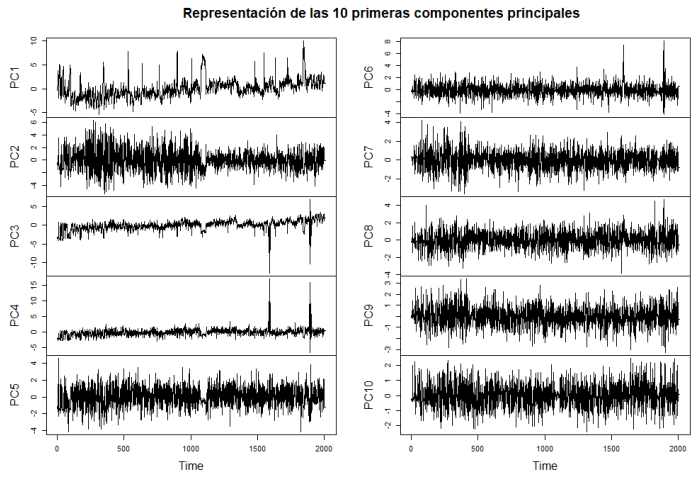


Figura 5.28: Representacion de las series de tiempo para las 10 primeras componentes principales

La componente 1 presenta tendencia y permite identificar lo que podría ser la ruptura de correa con un aumento en el nivel de las vibraciones, a partir de este momento la componente 2 disminuye la amplitud de las vibraciones y la componente 3 identifica este momento con una ligera bajada en el nivel de las vibraciones, además de lo que podría ser la falta de grasa con un aumento de la amplitud de las vibraciones. Esto también se detecta en la componente 4 y en la componente 6. Además, la componente 4, la componente 5 y la componente 10 identifican la posible ruptura de correa debido a una disminución en la amplitud de las vibraciones. En las componentes de las 7 a la 9 no se identifican patrones.

En este momento es necesario dividir la serie de tiempo en dos –la serie de entrenamiento o calibrado y la serie test o monitorizado– para poder posteriormente realizar los gráficos de control. A partir de este momento se trabajará con la serie de entrenamiento.

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     1      1      1      1

$criteria
                1           2           3           4           5           6           7
AIC(n) -1.1345110 -1.0631645 -0.9902239 -0.9074483 -0.8130383 -0.7019456 -0.6321846
HQ(n)  -0.9196845 -0.6530413 -0.3848039 -0.1067316  0.1829752  0.4893647  0.7544224
SC(n)  -0.5707698  0.0130687  0.5985013  1.1937688  1.8006708  2.4242555  3.0065085
FPE(n)  0.3215827  0.3453863  0.3715805  0.4037713  0.4439714  0.4965062  0.5329298
                8           9          10          11          12          13          14
AIC(n) -0.5303192 -0.4719530 -0.3963388 -0.3446305 -0.2649246 -0.1779478 -0.1388659
HQ(n)   1.0515846  1.3052475  1.5761585  1.8231636  2.0981662  2.3804397  2.6148184
SC(n)   3.6208659  4.1917241  4.7798303  5.3440306  5.9362284  6.5356972  7.0872711
FPE(n)  0.5908878  0.6275044  0.6782798  0.7161931  0.7781155  0.8520626  0.8899734
               15          16          17          18          19          20          21
AIC(n) -0.0771346 -0.02671397 0.02664752 0.07877553 0.170751  0.2355202  0.3094759
HQ(n)   2.8718465  3.11756386 3.36622211 3.61364688 3.900919  4.1609850  4.4302375
SC(n)   7.6614944  8.22440698 8.79026046 9.35488045 9.959348 10.5366091 11.1230568
FPE(n)  0.9515261  1.00664758 1.06895808 1.13470247 1.254607  1.3512712  1.4703404
               22          23          24
AIC(n)  0.3412132  0.403102  0.4625984
HQ(n)   4.6572715  4.914457  5.1692503
SC(n)  11.6672860 12.241667 12.8136552
FPE(n)  1.5354508  1.654452  1.7806063
```

Figura 5.29: Criterio para seleccionar el orden del modelo

Se comprueba la estacionariedad de las series de tiempo de cada componente a través del test de Phillips-Perron, para este todas las series son estacionarias y, por tanto, no es necesario diferenciarlas;

se realiza una segunda comprobación utilizando el comando ndiffs del paquete forecast que calcula el número de diferenciaciones necesarias para convertir una serie en estacionaria. En este caso, para el conjunto de series, confirma que la diferenciación no es necesaria. Tras esto se comprueba el orden del modelo VAR, a través del comando *VARselect* del paquete *vars*, obteniendo que el rezago óptimo es 1 (véase Figura 5.29), es decir, estimamos el modelo como un $VAR_{10}(1)$ con constante.

Ahora se debe comprobar las especificaciones del modelo: estacionariedad, independencia de los residuos, homocedasticidad de la varianza de los residuos y normalidad de los residuos.

La estacionariedad se comprueba a través de los roots del modelo, es decir, de los autovalores. Si todos son menores a 1, como en este caso, se puede aceptar el supuesto de estacionariedad y se confirma que el orden del modelo está seleccionado correctamente.

```
> roots(modVAR75)
 [1] 0.91591024 0.65029626 0.08198319 0.08198319 0.07654369 0.07654369 0.06780508
 [8] 0.05649477 0.05649477 0.03622080
```

Figura 5.30: Comprobación estacionariedad

La independencia de los residuos se comprueba con el test de Portmanteau, que se realiza a través de la función *serial.test* de la librería *vars*. La hipótesis nula de esta prueba afirma que los residuos se distribuyen de forma independiente, en este caso se rechaza.

```
> serial.test(modVAR75, lags.pt=1, type="PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object modVAR75
Chi-squared = 28.488, df = 0, p-value < 2.2e-16
```

Figura 5.31: Comprobación independencia de los residuos

La homocedasticidad de la varianza de los residuos se verifica a través del test ARCH-LM, esta se realiza a través de la función *arch.test* de la librería *vars*. Su hipótesis nula afirma que la varianza de los residuos es constante, es decir, homocedástica; en este caso se rechaza (véase Figura 5.32).

```
> #Comprobamos homocedasticidad de la varianza de los residuos
> #Ho: la varianza de los residuos es homocedástica (cte)
> arch.test(modVAR75, lags.multi = 1)

        ARCH (multivariate)

data:  Residuals of VAR object modVAR75
Chi-squared = 3801.5, df = 3025, p-value < 2.2e-16
```

Figura 5.32: Comprobación homocedasticidad de la varianza de los residuos

Por último, se examina si los residuos siguen una distribución gaussiana multivariante a través del test de Jarque-Bera, test de asimetría y test de kurtosis implementados en la función *normality.test* de la librería *vars*. Las tres pruebas plantean en su hipótesis nula que los residuos siguen una distribución gaussiana; en el caso de estudio es rechazada por los tres test (véase Figura 5.33).

```
> norm$jb.mul
$JB

        JB-Test (multivariate)

data:  Residuals of VAR object modVAR75
Chi-squared = 2990.1, df = 20, p-value < 2.2e-16


$Skewness

        Skewness only (multivariate)

data:  Residuals of VAR object modVAR75
Chi-squared = 216.99, df = 10, p-value < 2.2e-16


$Kurtosis

        Kurtosis only (multivariate)

data:  Residuals of VAR object modVAR75
Chi-squared = 2773.1, df = 10, p-value < 2.2e-16
```

Figura 5.33: Comprobación normalidad de los residuos

Este modelo no cumple las especificaciones de independencia de residuos ni de homogeneidad de la varianza de los residuos. Cuando la varianza de los residuos es heterogénea puede ser conveniente utilizar modelos ARCH.

Para ello se trabaja con el modelo ajustado previamente, $VAR_{10}(1)$ con constante. El primer paso es calcular los residuos del modelo al cuadrado y representarlos (véase Figura 5.34) para comprobar si la varianza es constante, en este caso no es así.

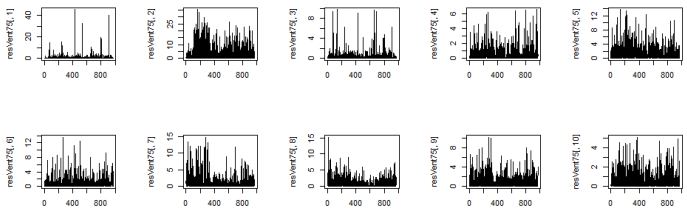


Figura 5.34: Representación de los residuos de las 10 primeras componentes principales

Para comprobar si existe efecto ARCH se puede hacer de dos formas. La primera consiste en realizar un modelo de regresión con los residuos de cada serie y realizar un *summary* del modelo para comprobar si los coeficientes son significativos, esto se lleva a cabo a través de la función *dynlm* de la librería *dynlm* que ajusta modelos de regresión sobre series de tiempo; en este caso la hipótesis nula indica que los coeficientes no son significativos y que no hay efectos ARCH. Con el modelo de estudio, en todos los casos se rechaza la hipótesis nula afirmando de este modo la existencia de efectos ARCH véase Apéndice B.

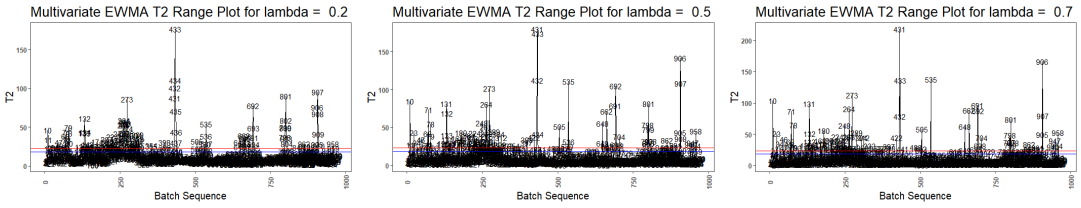
La segunda forma se basa en utilizar la prueba de Lagrange multivariante para ARCH a través de la función *ArchTest* de la librería *FinTS*, la hipótesis nula indica que no hay efecto ARCH, de nuevo se rechaza.

```
> #Otra forma de comprobarlo es con ArchTest
> #H0: No hay efectos ARCH p-value>0.05
> ArchTest(resVent75) #Rechazamos Ho, hay efecto ARCH

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  resVent75
Chi-squared = 678.55, df = 12, p-value < 2.2e-16
```

Figura 5.35: ArchTest

El modelo ajustado es un $VAR_{10}(1)$ con constante y efectos ARCH(1). Se van a estimar los parámetros de los gráficos de control con los residuos al cuadrado para reflejar los efectos ARCH(1). Se ha optado por utilizar el gráfico MEWMA, por ser de los más robustos cuando los datos no siguen una distribución normal, con diferencia el más utilizado cuando las observaciones son multivariantes (Montgomery, 2020).

El parámetro fundamental de los gráficos EWMA y MEWMA es el denominado factor de olvido o lambda. Se ha probado con diferentes valores de lambda –define el peso que las observaciones pasadas tienen en el cálculo del estadístico en el momento actual, el "nivel de memoria- para el gráfico MEWMA, pero en todos ellos el proceso está fuera de control. En este caso se ha considerado $\lambda = 0{,}2$ como el más adecuado –siguiendo las recomendaciones de Montgomery, 2020–.



(a) Gráfico MEWMA con $\lambda = 0{,}2$ (b) Gráfico MEWMA con $\lambda = 0{,}5$ (c) Gráfico MEWMA con $\lambda = 0{,}7$

El procedimiento habitual cuando un punto está fuera de control es comprobar porqué está fuera de control –es una anomalía, por falta de gaussianidad o es por aleatoriedad– y, si es posible, se elimina y se vuelve a realizar el gráfico para comprobar si todos los puntos están dentro de los límites. En este caso, hay demasiadas observaciones que superan los intervalos de confianza al 95 % –línea azul– y al 99 % –línea roja–. Este proceso no puede controlarse siguiendo por este camino. ¿Si utilizamos el PCA Robusto sucederá lo mismo?

Si utilizamos el PCA Robusto, al igual que en el PCA, los datos siguen una combinación lineal pero $X1,\ X2,\ X3,\ Y1,\ Y2,\ Y3,\ Y4$ y $Z1$ son independientes. Si observamos la Figura 5.36, puede verse que las variables importantes en cada componente son:

- Componente 1: $T1,\ T2,\ T3,\ T4,\ Z3$.
- Componente 2: $Y1,\ Y2,\ Z2,\ Z1$.
- Componente 3: $X3,\ Y3,\ Z3$.
- Componente 4: $X4,\ Z4,\ Y4$.
- Componente 5: $X1$.

- Componente 6: $X2$.
- Componente 7: $Y4,\ Z4$.
- Componente 8: $Consumo,\ Z1$.
- Componente 9: $Z1,\ Consumo,\ X2$.
- Componente 10: $Y3,\ Consumo,\ Z3,\ X3$.
- Componente 11: $X3,\ Z3,\ X4,\ Z4$.
- Componente 12: $Z3,\ T4,\ X4,\ X3,\ Z4$.
- Componente 13: $Z4,\ T4,\ X4,\ T2,\ T1$.
- Componente 14: $Y1,\ Y2,\ Z1$.
- Componente 15: $Z2,\ Y1,\ Y2$.
- Componente 16: $T3,\ T4$.
- Componente 17: $T1,\ T2$.

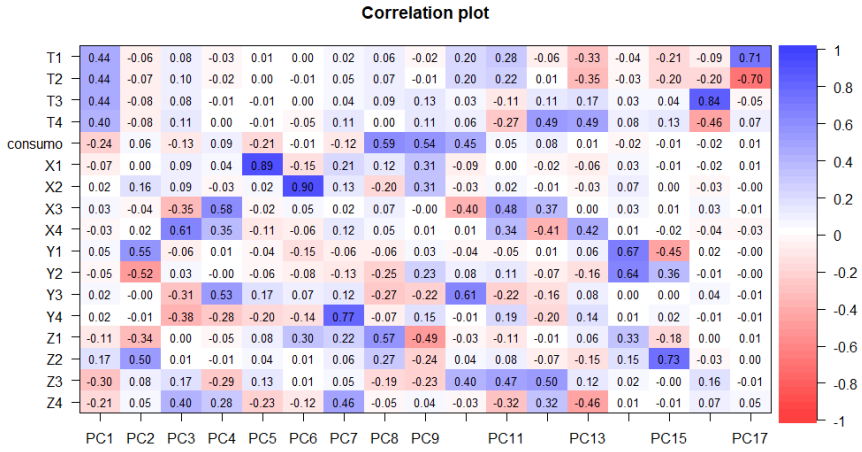


Figura 5.36: Gráfico de correlaciones. Importancia de cada variable en cada componente PCA Robusto

En este caso, para explicar el 90 % de la varianza, se necesitan las primeras 11 componentes principales. Con estas componentes conseguimos un modelo con sensibilidad 0.98, especificidad de 0.1, BA 0.54 y NPV de 0.1. En cambio, si utilizamos las componentes 1, 2, 3, 8, 9, 10, 11, 12 y 13, se tienen en cuenta todas las variables originales, se reduce el número de componentes a 9, se consigue una sensibilidad de 0.997, una especificidad de 0.85, una BA de 0.92 y un NPV de 0.825. Dado que sigue existiendo autocorrelación entre las observaciones, aun habiendo obtenido relativamente buenos resultados de detección de verdaderas anomalías y relativa ausencia de falsas anomalías, el siguiente paso es aplicar modelos de series de tiempo multivariantes, pasando a controlar los residuos de los mismos, con el objeto de que puedan cumplir las hipótesis de partida de los gráficos de control –independencia, normalidad y homocedasticidad de las variables críticas para la calidad del proceso–.

La representación de las series de tiempo de las 9 componentes seleccionadas puede verse en la Figura 5.37.

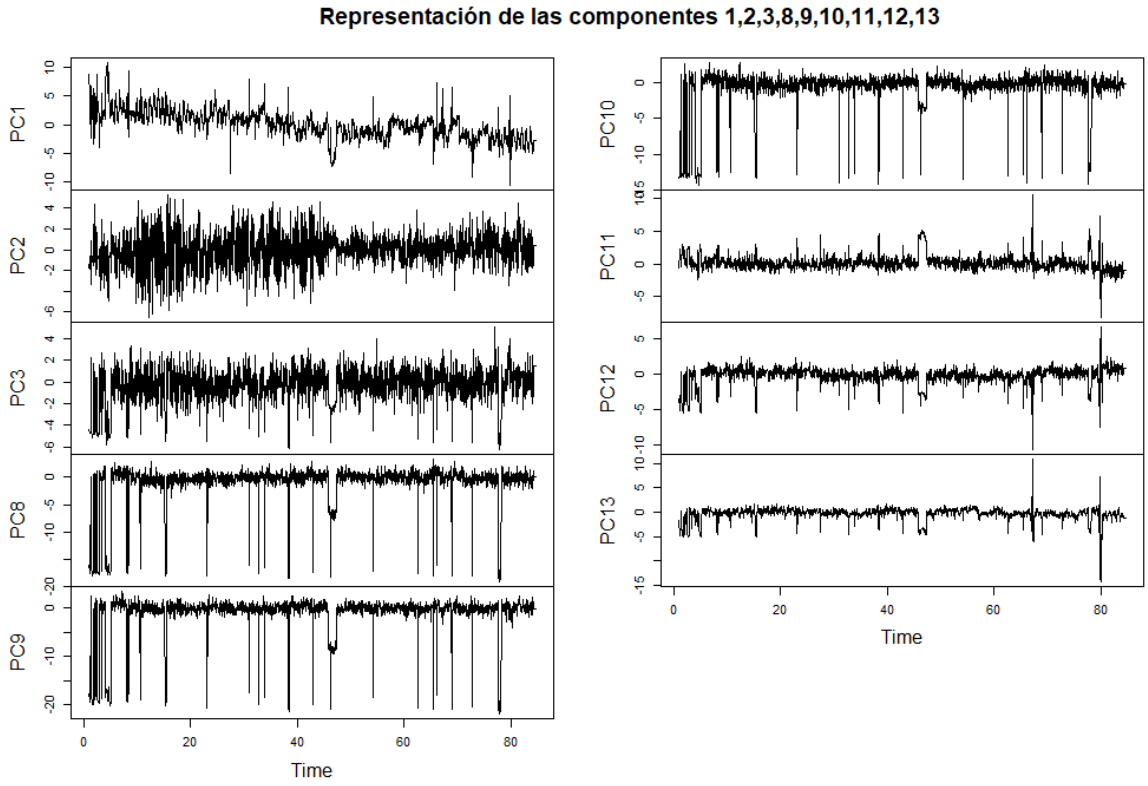


Figura 5.37: Representacion de las series de tiempo para las componentes principales seleccionadas

La componente 1 presenta tendencia y un aumento en el nivel de las vibraciones causadas posiblemente por la ruptura de la correa, a partir de este momento la componente 2 reduce la amplitud de las vibraciones. En la componente 3 es difícil identificar patrones. En las componentes 8, 9 y 10 hay muchos picos inferiores pero puede identificarse la ruptura de cadena. En las componentes 11 y 13 se puede identificar sutilmente la ruptura de correa -debido a un aumento en el nivel de las vibraciones- y la falta de grasa provocando un aumentando la amplitud de las vibraciones, patrón que se repite en la componente 12, además, en esta se identifica sutilmente lo que podría clasificarse como la ruptura de correa con una bajada en el nivel de las vibraciones.

Se divide la serie en dos: una serie de entrenamiento y una serie de test. A partir de este punto se trabaja con la serie de entrenamiento. Se comprueba la estacionariedad –como se ha indicado anteriormente–, esta vez es necesario diferenciar la serie 1 vez. La nueva representación de las series puede verse en la Figura 5.38, la componente 1 ha eliminado su tendencia y las series parecen más estables, con la media en torno al 0.

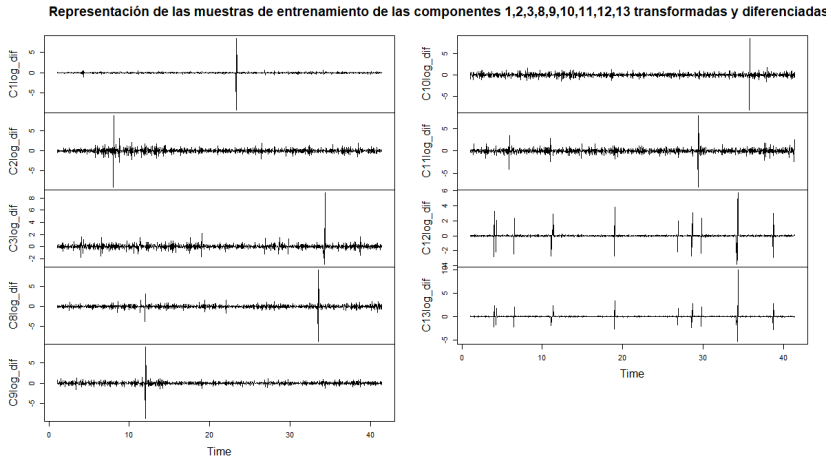


Figura 5.38: Representacion de las series de tiempo diferenciadas para las componentes principales seleccionadas

Tras esto se selecciona el orden del modelo (véase Figura 5.39) con una diferenciación y constante, en este caso el rezago óptimo es 8, ya que 2 de los criterios de selección obtienen este resultado. El modelo será un $VAR_9(8)$ con una diferenciación y constante.

```
> #Se modeliza la serie diferenciada
> vars::VARselect(serie_train75dif, type= "both",lag.max = 21) #VAR_9(8) (Rezago optimo=8)
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     8      4      2      8

$criteria
                1           2           3           4           5           6           7           8           9          10          11
AIC(n)  2.514923 1.727983 1.420683 1.168659 1.058648 0.9172846 0.9146922 0.8681299 0.8795894 0.872316 0.8714692
HQ(n)   2.707753 2.078582 1.929031 1.834797 1.882555 1.8989615 2.0541387 2.1653459 2.3345749 2.485071 2.6419937
SC(n)   3.021018 2.648155 2.754932 2.916985 3.221051 3.4937654 3.9052504 4.2727654 4.6983021 5.105106 5.5183365
FPE(n) 12.365774 5.629602 4.140653 3.218912 2.884569 2.5055696 2.5008220 2.3892418 2.4196169 2.405592 2.4078365
               12          13          14          15          16          17          18          19          20          21
AIC(n) 0.9028733 0.9081006 0.8952485 0.9212883 0.9393186 0.9916076 1.024383 1.054538 1.092627 1.145266
HQ(n)  2.8311673 2.9941641 3.1390815 3.3228908 3.4986906 3.7087491 3.899294 4.087219 4.283077 4.493486
SC(n)  5.9638179 6.3831225 6.7843476 7.2244648 7.6565723 8.1229386 8.569791 9.014024 9.466190 9.932906
FPE(n) 2.4899521 2.5093049 2.4845375 2.5587295 2.6153856 2.7679085 2.874316 2.978772 3.113568 3.304371
```

Figura 5.39: Selección del orden del modelo

Ahora se debe verificar si se cumplen las especificaciones del modelo: estacionariedad, independencia de los residuos, homocedasticidad de la varianza de los residuos y normalidad de los residuos. Como se ha explicado anteriormente los contrastes, librerías e hipótesis ahora únicamente se anunciarán los resultados:

- Hay estacionariedad, se ha especificado correctamente el orden del modelo.

```
> roots(modVAR75R)
 [1] 0.8980039 0.8980039 0.8448710 0.8448710 0.8429554 0.8429554 0.8403485 0.8403485 0.8370455 0.8370455
[11] 0.8351245 0.8351245 0.8314102 0.8314102 0.8282157 0.8282157 0.8173661 0.8173661 0.8171637 0.8171637
[21] 0.8095429 0.8095429 0.8063755 0.8063755 0.8040822 0.8040822 0.8035750 0.8035750 0.7956760 0.7956760
[31] 0.7894925 0.7894925 0.7864695 0.7864695 0.7851064 0.7851064 0.7792827 0.7792827 0.7782855 0.7782855
[41] 0.7775101 0.7775101 0.7773401 0.7773401 0.7706163 0.7706163 0.7699624 0.7699624 0.7645764 0.7645764
[51] 0.7605419 0.7605419 0.7580652 0.7580652 0.7557585 0.7557585 0.7403199 0.7403199 0.7379108 0.7379108
[61] 0.7255771 0.7255771 0.6950836 0.6950836 0.6797922 0.6797922 0.6651787 0.6651787 0.6448425 0.6448425
[71] 0.5916156 0.5916156
```

Figura 5.40: Comprobación estacionariedad

- No hay independencia de residuos.

```
> #Comprobamos autocorrelación de residuos
> #H0: Los residuos se distribuyen de forma independiente (No hay autocorrelación)
> serial.test(modVAR75R, lags.pt=8, type="PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object modVAR75R
Chi-squared = 239.35, df = 0, p-value < 2.2e-16
```

Figura 5.41: Comprobación independencia de los residuos

- La varianza de los residuos no es constante

```
> #Ho: la varianza de los residuos es homocedástica (cte)
> arch.test(modVAR75R, lags.multi = 8)

        ARCH (multivariate)

data:  Residuals of VAR object modVAR75R
Chi-squared = 17483, df = 16200, p-value = 1.811e-12
```

Figura 5.42: Comprobación homocedasticidad de la varianza de los residuos

- Los residuos no siguen una distribución normal.

```
> norm$jb.mul
$JB

        JB-Test (multivariate)

data:  Residuals of VAR object modVAR75R
Chi-squared = 14950, df = 18, p-value < 2.2e-16


$Skewness

        Skewness only (multivariate)

data:  Residuals of VAR object modVAR75R
Chi-squared = 159.88, df = 9, p-value < 2.2e-16


$Kurtosis

        Kurtosis only (multivariate)

data:  Residuals of VAR object modVAR75R
Chi-squared = 14790, df = 9, p-value < 2.2e-16
```

Figura 5.43: Comprobación normalidad de los residuos

Se comprueba si en este caso es necesario utilizar un modelo ARCH basándonos en el modelo $VAR_9(8)$ con una diferenciación y constante. El primer paso es calcular los residuos del modelo al cuadrado; al representarlos gráficamente (véase Figura 5.44) puede observarse que la varianza no es constante.

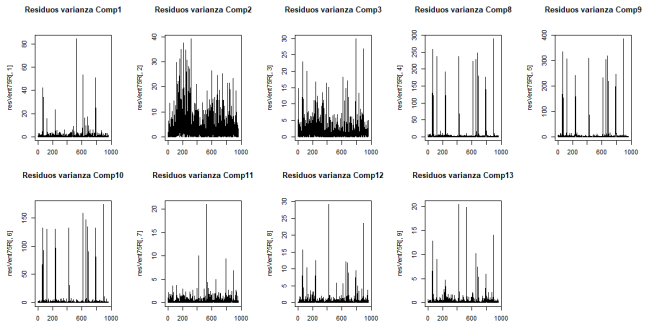


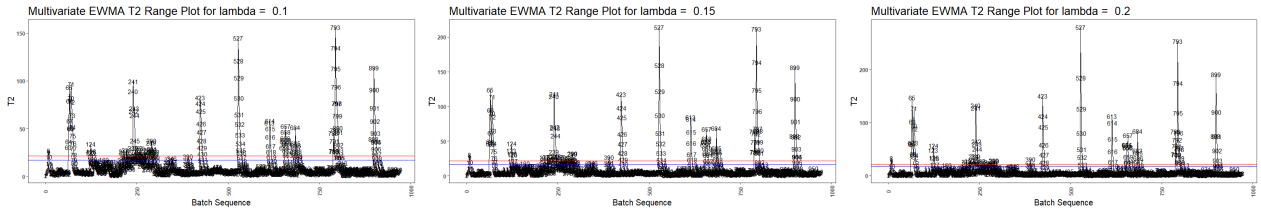
Figura 5.44: Representación de los residuos de las componentes principales seleccionadas

Si se comprueba el efecto ARCH a través de la prueba de Lagrange multivariante se acepta que hay efecto ARCH(1) (véase Figura 5.45).

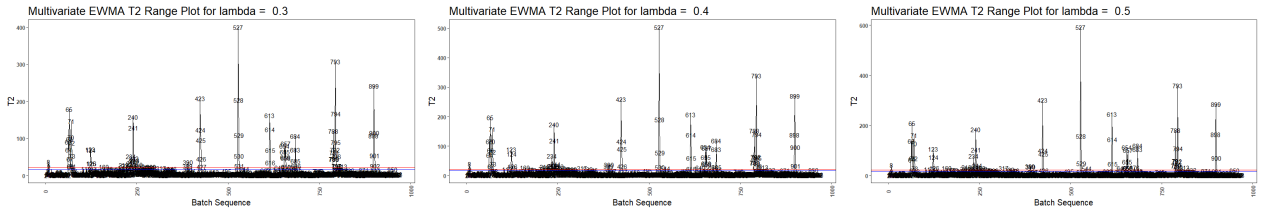
```
> #H0: No hay efectos ARCH p-value>0.05
> ArchTest(resVent75R) #Rechazamos Ho, hay efecto ARCH

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  resVent75R
Chi-squared = 146.63, df = 12, p-value < 2.2e-16
```

Figura 5.45: ArchTest

A partir del modelo $VAR_9(8)$ con 1 diferenciación y constante con efecto ARCH(1) se realiza el gráfico de control MEWMA, en este caso se ha probado valores de λ 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, siendo el más adecuado el valor $\lambda = 0{,}15$. De nuevo el proceso está fuera de control y no se pueden controlar porque hay demasiados puntos fuera de los límites.



(a) Gráfico MEWMA con $\lambda = 0{,}1$ (b) Gráfico MEWMA con $\lambda = 0{,}15$ (c) Gráfico MEWMA con $\lambda = 0{,}2$



(a) Gráfico MEWMA con $\lambda = 0{,}3$ (b) Gráfico MEWMA con $\lambda = 0{,}4$ (c) Gráfico MEWMA con $\lambda = 0{,}5$

Se ha realizado el mismo procedimiento para las medias de los datos cada 60 y 30 minutos pero únicamente con PCA Robusto porque tanto en los datos originales como para las medias de los datos cada 75 minutos parece que se obtienen NPV más elevados con este método que con PCA.

- **Medias de datos cada 60 minutos**

Al realizar la transformación de los datos a partir de calcular las medias de los datos en intervalos de 60 minutos, se ha obtenido una base de datos de 2510 observaciones de las cuales 46 están clasificadas como anomalías.

Los datos siguen una distribución similar al caso estudiado anteriormente, si distribución no es gaussiana (véase Figura 5.46), pero la transformación Box-Cox (véase Figura 5.47) no realiza mejoras significativas, por eso se seguirá trabajando con los datos normales.

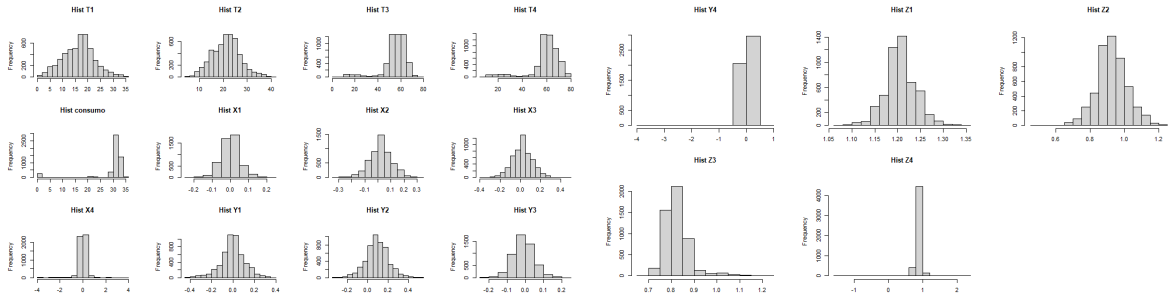


Figura 5.46: Histogramas de los datos transformados realizandoles la media cada 60 minutos

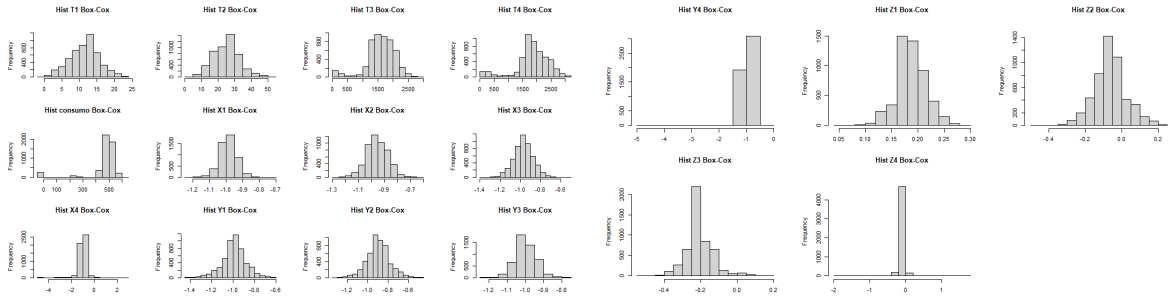


Figura 5.47: Histogramas de los datos transformados realizandoles la media cada 60 minutos y una transformación Box-Cox

Se realiza el PCA Robusto ya que los datos presentan combinaciones lineales pero hay variables que son independientes –$X1$, $X2$, $X3$, $Y1$, $Y2$–.

Si se observa la Figura 5.48, se observa que las variables más relevantes en cada componentes son:

- Componente 1: $T1$, $T2$, $T3$, $T4$, $Z3$.
- Componente 2: $Y1$, $Y2$, $Z2$, $Z1$.
- Componente 3: $X3$, $X4$, $Y3$, $Z4$, $Y4$.
- Componente 4: $X4$, $X3$, $Y3$, $Y4$, $Z4$.
- Componente 5: $X1$, $X2$.
- Componente 6: $X2$, $X1$, $Z1$.
- Componente 7: $Y4$, $Z4$.
- Componente 8: $Consumo$
- Componente 9: $Z1$, $Y2$, $Consumo$.
- Componente 10: $Y3$, $Z3$, $X3$, $Consumo$.
- Componente 11: $Z3$, $X3$, $X4$, $T1$.
- Componente 12: $T4$, $X4$, $Z3$, $X3$, $Z4$.
- Componente 13: $T4$, $Z4$, $X4$, $T2$, $T1$.
- Componente 14: $Y1$, $Y2$, $Z1$.
- Componente 15: $Z2$, $Y1$, $Y2$.
- Componente 16: $T3$, $T4$.
- Componente 17: $T1$, $T2$.

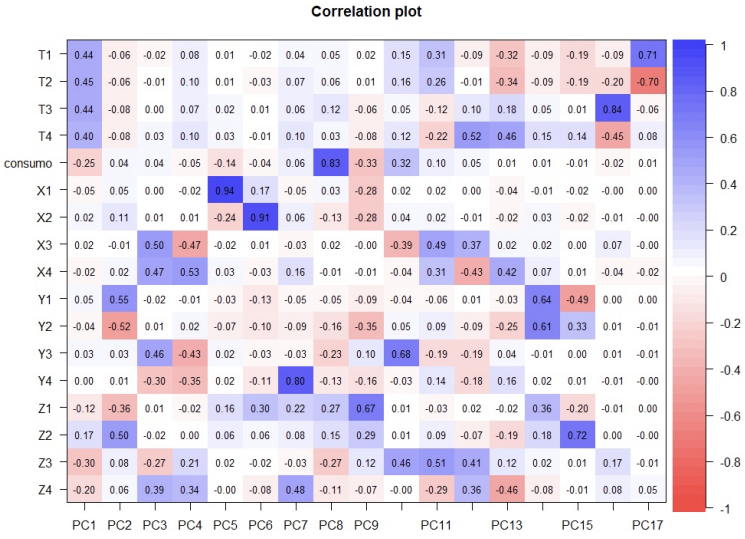


Figura 5.48: Gráfico de correlaciones

Para explicar el 90 % de la varianza se necesitan las primeras 11 componentes principales. Con estas componentes conseguimos un modelo con sensibilidad 0.99, especificidad de 0.06, precisión global 0.53 y NPV de 0.15. En cambio, si utilizamos las componentes 1,2,8,9,10,11,12,13, se tienen en cuenta todas las variables, se reduce el número de componentes a 8, se consigue una sensibilidad de 1, una especificidad de 0.43, una precisión global de 0.72 y un NPV de 1; por ese motivo se han seleccionado esas 8 componentes para ajustar un modelo de series de tiempo para posteriormente trabajar con los residuos a causa de la correlación existente entre las componentes.

En las series de tiempo de las 8 componentes (Figura 5.49) se observa que la componente 1 presenta tendencia y, puede detectarse lo que podría ser la ruptura de la correa por la disminución en el nivel de las vibraciones; a partir de ese punto la componente 2 reduce la amplitud de las vibraciones. Las componentes 8 y 10 tienen una estructura similar, presentan muchos picos inferiores pero se puede identificar la ruptura de la correa por ser un descenso inferior que se mantiene; en la componente 9 sucede el caso contrario a las componentes 8 y 10, los picos son superiores. En la componente 11 parece que puede detectarse la ruptura de cadena por un aumento en el nivel de las vibraciones y la falta de grasa por un aumento en la amplitud de las mismas. Por último, las componentes 12 y 13 tienen una estructura similar, identifican lo que parece la ruptura de cadena por una bajada en el nivel de las vibraciones y la falta de grasa por un aumento en la amplitud de las vibraciones.

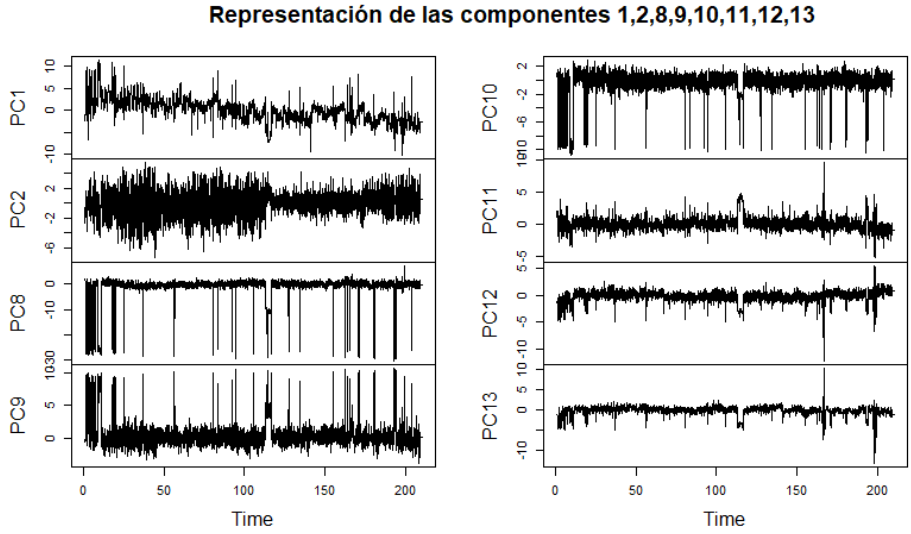


Figura 5.49: Representación series de tiempo de las componentes seleccionadas

La serie se divide en serie de entrenamiento –con la que se trabajará a partir de ahora– y muestra test. Es necesario diferenciar la serie, de este modo se consigue que desaparezca la tendencia de la primera componente y que las medias se estabilicen en torno al 0 (Figura 5.50) .

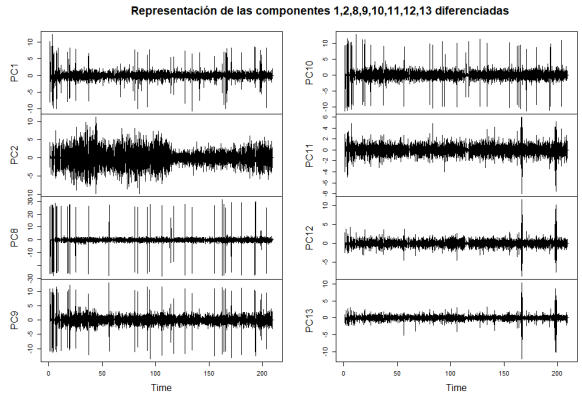


Figura 5.50: Representación series de tiempo diferenciadas de las componentes seleccionadas

Tras esto se debe ajustar el modelo

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     7      4      2      7

$criteria
              1         2         3         4          5            6           7
AIC(n) 1.175855 0.5572865 0.2706607 0.0975332 0.01241244 -0.009990897 -0.03149541
HQ(n)  1.330443 0.8355461 0.6725913 0.6231347 0.66168485  0.762952444  0.86511887
SC(n)  1.581769 1.2879313 1.3260365 1.4776400 1.71725019  2.019577852  2.32280434
FPE(n) 3.240932 1.7459901 1.3109697 1.1027103 1.01294403  0.990808719  0.97014187
               8           9         10         11         12
AIC(n) -0.02609754 -0.02779211 0.01144702 0.04505788 0.08228274
HQ(n)   0.99418767  1.11616403 1.27907410 1.43635589 1.59725169
SC(n)   2.65293321  2.97596964 3.33993977 3.69828163 4.06023749
FPE(n)  0.97594009  0.97498527 1.01490587 1.05073468 1.09199914
```

Figura 5.51: Selección del orden del modelo

En este caso un $VAR_8(7)$ con 1 diferenciación y constante, y comprobar si se cumplen las especificaciones del modelo:

- El modelo es estacionario y el número de rezagos está correctamente estimado

```
> #Comprobamos estacionariedad del modelo
> roots(modVAR60R)
 [1] 0.8219749 0.8219749 0.8120216 0.8097345 0.8097345 0.8045737 0.8045737 0.7998533
 [9] 0.7998533 0.7839003 0.7839003 0.7755356 0.7755356 0.7753223 0.7753223 0.7730419
[17] 0.7730419 0.7685322 0.7685322 0.7673163 0.7673163 0.7531528 0.7531528 0.7463466
[25] 0.7463466 0.7312960 0.7312960 0.7273394 0.7273394 0.7271535 0.7216218 0.7216218
[33] 0.7181370 0.7181370 0.7150879 0.7150879 0.7055034 0.7055034 0.7031491 0.7031491
[41] 0.7025777 0.7025777 0.6870150 0.6870150 0.6694791 0.6694791 0.6619288 0.6619288
[49] 0.6362241 0.6362241 0.6162300 0.6060832 0.4365534 0.4365534 0.3774280 0.3774280
```

Figura 5.52: Comprobación estacionariedad

- Los residuos no cumplen la condición de independencia.

```
> #Comprobamos autocorrelación de residuos
> #H0: Los residuos se distribuyen de forma independiente (No hay autocorrelación)
> serial.test(modVAR60R, lags.pt=7, type="PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object modVAR60R
Chi-squared = 160.29, df = 0, p-value < 2.2e-16
```

Figura 5.53: Comprobación independencia de los residuos

- La varianza de los residuos no es constante.

```
> #Comprobamos homocedasticidad de la varianza de los residuos
> #Ho: la varianza de los residuos es homocedástica (cte)
> arch.test(modVAR60R, lags.multi = 7)

        ARCH (multivariate)

data:  Residuals of VAR object modVAR60R
Chi-squared = 10774, df = 9072, p-value < 2.2e-16
```

Figura 5.54: Comprobación homocedasticidad de la varianza de los residuos

- Los residuos no siguen una distribución gaussiana.

```
> #Ho: los residuos se distribuyen como una dist. normal
> norm= normality.test(modVAR60R,multivariate.only = FALSE)
> norm$jb.mul
$JB

        JB-Test (multivariate)

data:  Residuals of VAR object modVAR60R
Chi-squared = 14618, df = 16, p-value < 2.2e-16


$Skewness

        Skewness only (multivariate)

data:  Residuals of VAR object modVAR60R
Chi-squared = 401.26, df = 8, p-value < 2.2e-16


$Kurtosis

        Kurtosis only (multivariate)

data:  Residuals of VAR object modVAR60R
Chi-squared = 14216, df = 8, p-value < 2.2e-16
```

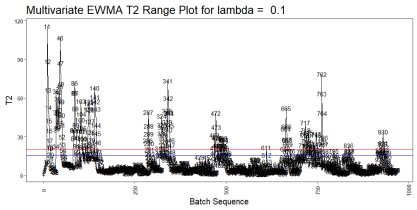
Figura 5.55: Comprobación normalidad de los residuos

Como de nuevo la varianza no es homocedástica se comprueba si sería correcto ajustar un modelo ARCH, a través de la prueba de Lagrange multivariante se acepta que hay efecto ARCH(1).

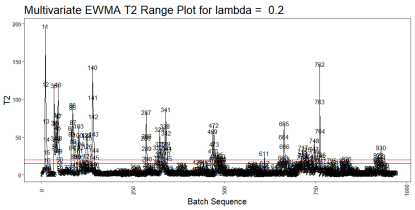
```
> #H0: No hay efectos ARCH p-value>0.05
> ArchTest(resVent60R) #Rechazamos Ho, hay efecto ARCH

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  resVent60R
Chi-squared = 194.21, df = 12, p-value < 2.2e-16
```

Figura 5.56: ArchTest

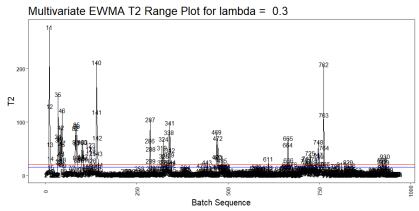
A partir del modelo $VAR_8(7)$ con 1 diferenciación y constante con efecto ARCH(1) se realiza los gráficos de control MEWMA, en este caso se ha probado los valores de λ 0.1, 0.2, 0.3, 0.4, 0.5, siendo el más adecuado el valor $\lambda = 0{,}2$.



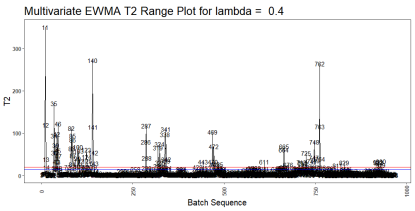
(a) Gráfico MEWMA con $\lambda = 0{,}1$



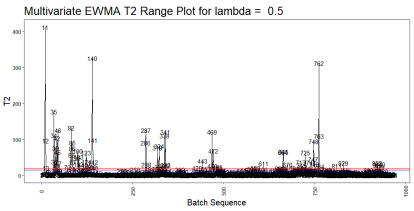
(b) Gráfico MEWMA con $\lambda = 0{,}2$



(a) Gráfico MEWMA con $\lambda = 0{,}3$



(b) Gráfico MEWMA con $\lambda = 0{,}4$



(c) Gráfico MEWMA con $\lambda = 0{,}5$

De nuevo el proceso está fuera de control y no se pueden controlar porque hay demasiados puntos fuera de los límites.

- **Medias de datos cada 30 minutos**

Por último, se ha realizado este procedimiento con las medias de los datos cada 30 minutos, obteniendo así una base de datos con 5021 observaciones, de las cuales 93 están clasificadas como anomalías.

En este caso, al igual que en los anteriores, los datos no siguen una distribución gaussiana (Figura 5.57) y la transformación Box-Cox no realiza mejoras significativas (Figura 5.58), con lo cual se trabajará con las medias de los datos cada 30 minutos sin transformar.

Se realiza el PCA Robusto porque existen combinaciones lineales entre las variables y se conoces que existen variables que son independientes –$X1$, $X2$, $X3$, $Y1$, $Y2$–.

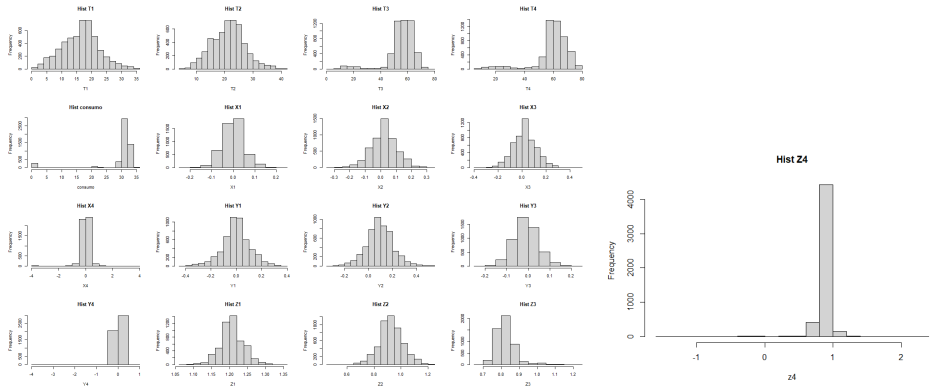


Figura 5.57: Histogramas de los datos transformados realizandoles la media cada 30 minutos

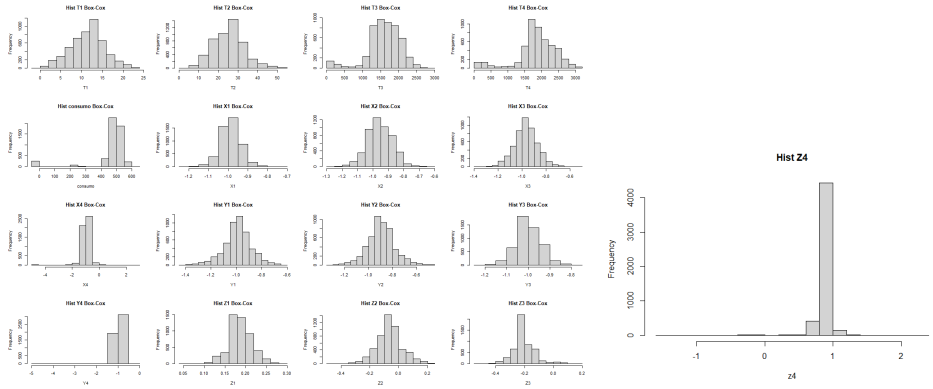


Figura 5.58: Histogramas de los datos transformados realizandoles la media cada 30 minutos y una transformación Box-Cox

Observando la Figura 5.59, se repara en que las variables más relevantes en cada componentes son:

- Componente 1: $T1,\ T2,\ T3,\ T4,\ Z3$.
- Componente 2: $Y1,\ Y2,\ Z2,\ Z1$.
- Componente 3: $X3,\ Y3,\ X4,\ Z4,\ Y4$.
- Componente 4: $X4,\ X3,\ Y3,\ Y4,\ Z4$.
- Componente 5: $X1$.
- Componente 6: $X2,\ Z1$.
- Componente 7: $Y4,\ Z4$.
- Componente 8: $Consumo$.
- Componente 9: $Z1,\ Y2,\ Z2$.
- Componente 10: $Y3,\ Z3,\ X3,\ Consumo$.
- Componente 11: $Z3,\ X3,\ X4,\ T1$.

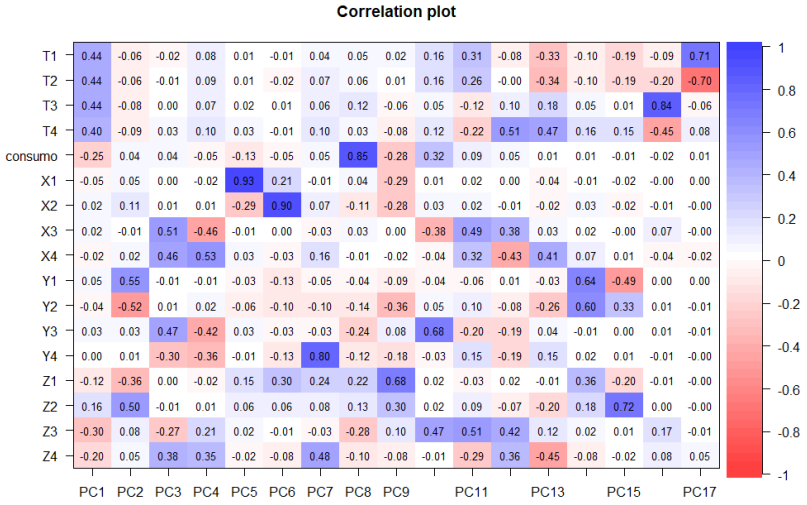


Figura 5.59: Gráfico de correlaciones. Importancia de cada variable en cada componente

- Componente 12: $T4,\ X4,\ Z3,\ X3,\ Z4$.
- Componente 13: $T4,\ Z4,\ X4,\ T2,\ T1$.
- Componente 14: $Y2,\ Y1,\ Z1$.
- Componente 15: $Z2,\ Y1,\ Y2$.
- Componente 16: $T3,\ T4$.
- Componente 17: $T1,\ T2$.

Para explicar el 90 % de la varianza se necesitan las primeras 11 componentes principales. Con estas componentes conseguimos un modelo con sensibilidad de 0.99, una especificidad y un NPV con valor 0 y una precisión global de 0.5; en cambio, si utilizamos un modelo con las componentes 1, 2, 8, 9, 10, 11, 12 y 13 se reduce el número de componentes de 11 hasta 8, y, además, se mejora el modelo en términos de conseguir el objetivo de la investigación, obteniendo una sensibilidad de 1, una especificidad de 0.43, una precisión global de 0.71 y un NPV de 1. Por ese motivo se ha utilizado este segundo modelo para ajustar la serie de tiempo.

En la Figura 5.60 puede verse la representación gráfica de la serie de tiempo para cada componente. En él se observa que la componente 1 presenta tendencia y sufre una disminución en el nivel de las vibraciones en el momento que podría ser la ruptura de la correa, a partir de ese momento la componente 2 reduce la amplitud de sus vibraciones. Las componentes 8 y 10 tienen un patrón similar, muestran muchos picos inferiores pero, a pesar de esto, puede identificarse lo que se intuye es la ruptura de la cadena, ya que hay una pequeña estabilización de los valores de las vibraciones por debajo de los niveles normales; a la componente nueve le sucede un patrón similar pero contrario, los picos son superiores a los niveles habituales. En la componente 11 puede identificarse sutilmente lo que coincidiría con la ruptura de correa ya que el nivel de las vibraciones aumenta y se mantiene constante un periodo de tiempo corto, además, pueden identificarse también los momentos de falta de grasa ya que hay un aumento de la amplitud de las vibraciones. Las componentes 12 y 13, siguen un patrón muy similar e inverso al de la componente 11, ya que en su caso el nivel de las vibraciones al detectar la ruptura de

correa disminuyen, pero la falta de grasa la detectan de forma similar -aumento en la amplitud de las vibraciones.

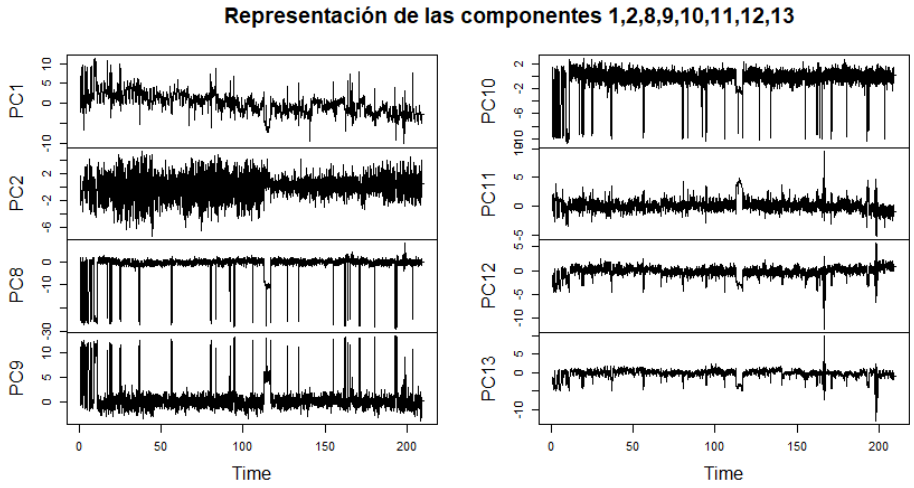


Figura 5.60: Representación series de tiempo de las componentes seleccionadas

El siguiente paso es comprobar la estacionariedad de las series para saber si es necesario diferenciarlas, en este caso se concluye que no es necesario y se ajusta la serie con un modelo $VAR_8(4)$ con constante, ¿cumplirá con las especificaciones?

```
$selection
AIC(n)  HQ(n)  SC(n) FPE(n)
     4      2      1      4

$criteria
                1           2           3           4           5
AIC(n) -0.6083724 -0.8572816 -0.8701492 -0.8747489 -0.8437938
HQ(n)  -0.4677588 -0.5916780 -0.4795557 -0.3591654 -0.2032204
SC(n)  -0.2393782 -0.1602925  0.1548347  0.4782300  0.8371799
FPE(n)  0.5442384  0.4243271  0.4189302  0.4170609  0.4302622
                6           7           8           9          10
AIC(n) -0.80958062 -0.7482561 -0.7121138 -0.6604751 -0.6016292
HQ(n)  -0.04401732  0.1422972  0.3034294  0.4800580  0.6638938
SC(n)   1.19938796  1.5887074  1.9528446  2.3324781  2.7193189
FPE(n)  0.44537413  0.4737429  0.4914553  0.5178740  0.5497580
               11          12          13          14          15
AIC(n) -0.5280697 -0.4534134 -0.4037572 -0.3508975 -0.2826474
HQ(n)   0.8624432  1.0620894  1.2367355  1.4145852  1.6078252
SC(n)   3.1208732  3.5235244  3.9011754  4.2820300  4.6782750
FPE(n)  0.5923729  0.6391311  0.6727151  0.7105200  0.7623023
               16          17          18          19          20
AIC(n) -0.2208735 -0.1534588 -0.08642798 -0.01634744 0.03948351
HQ(n)   1.7945891  1.9869937  2.17901442  2.37408489 2.55490577
SC(n)   5.0680438  5.4634534  5.85847904  6.25655445 6.64038027
FPE(n)  0.8128242  0.8718785  0.93518660  1.00652445 1.06839247
               21          22          23          24
AIC(n) 0.1072905 0.1447463 0.2098296 0.2726433
HQ(n)  2.7477027 2.9101484 3.1002216 3.2880252
SC(n)  7.0361821 7.4016328 7.7947110 8.1855195
FPE(n) 1.1481992 1.1975983 1.2847026 1.3756812
```

Figura 5.61: Selección del orden del modelo

Pues bien, sucede lo mismo que en los modelos anteriores: es estacionario, no se cumple la condición de independencia de residuos ni la de varianza de residuos constante, además, los residuos tampoco siguen una distribución gaussiana.

```
> #Comprobamos estacionariedad del modelo
> roots(modVAR3OR)
 [1] 0.945741507 0.937198406 0.937198406 0.666219583 0.641894021 0.615860962
 [7] 0.615860962 0.611253051 0.611253051 0.553801172 0.553801172 0.547130734
[13] 0.547130734 0.545521908 0.545521908 0.544487504 0.544487504 0.540632789
[19] 0.540632789 0.538186921 0.538186921 0.517206072 0.517206072 0.502162020
[25] 0.441798422 0.441798422 0.396973171 0.396973171 0.285409430 0.285409430
[31] 0.098322034 0.007094845
```

(a) Comprobación estacionariedad

```
> #H0: Los residuos se distribuyen de forma independiente (No hay autocorrelación)
> serial.test(modVAR3OR, lags.pt=4, type="PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object modVAR3OR
Chi-squared = 58.23, df = 0, p-value < 2.2e-16
```

(b) Comprobación independencia de los residuos

```
> #Comprobamos homocedasticidad de la varianza de los residuos
> #Ho: la varianza de los residuos es homocedástica (cte)
> arch.test(modVAR3OR, lags.multi = 4)

        ARCH (multivariate)

data:  Residuals of VAR object modVAR3OR
Chi-squared = 7242, df = 5184, p-value
< 2.2e-16
```

(a) Comprobación homocedasticidad de la varianza de los residuos

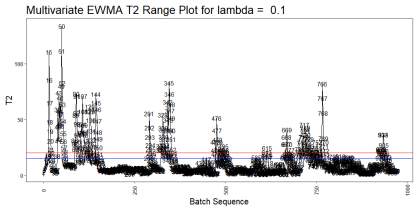
```
> #Comprobamos normalidad
> #Ho: los residuos se distribuyen como una dist. normal
> norm= normality.test(modVAR3OR,multivariate.only = FALSE)
> norm$jb.mul
$JB

        JB-Test (multivariate)

data:  Residuals of VAR object modVAR3OR
Chi-squared = 13854, df = 16, p-value < 2.2e-16


$Skewness

        Skewness only (multivariate)

data:  Residuals of VAR object modVAR3OR
Chi-squared = 366.02, df = 8, p-value < 2.2e-16


$Kurtosis

        Kurtosis only (multivariate)

data:  Residuals of VAR object modVAR3OR
Chi-squared = 13488, df = 8, p-value < 2.2e-16
```

(b) Comprobación normalidad de los residuos

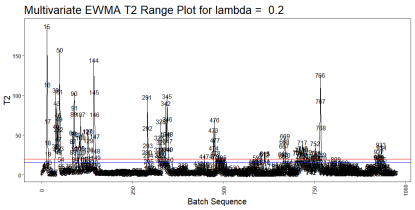
Se ha ajustado el modelo mediante un modelo ARCH (1) (Figura 5.62), con lo cual nos encontramos ante un modelo $VAR_8(4)$ con constante y efecto ARCH(1), con este modelo se ha intentado ajustado los gráficos de control MEWMA con diferentes λ (0.1, 0.2, 0.3, 0.4, 0.5), el más adecuado es $\lambda = 0{,}5$. En todos los casos hay demasiados puntos por encima de los límites de control, el proceso está fuera de control.

```
> #H0: No hay efectos ARCH p-value>0.05
> ArchTest(resVent3OR) #Rechazamos Ho, hay efecto ARCH

        ARCH LM-test; Null hypothesis: no ARCH effects

data:  resVent3OR
Chi-squared = 361.98, df = 12, p-value < 2.2e-16
```

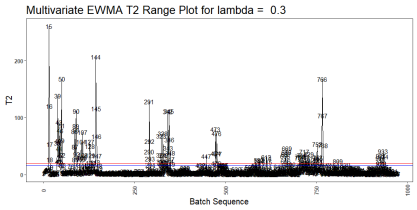
Figura 5.62: ArchTest



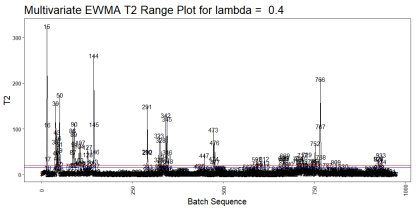
(a) Gráfico MEWMA con $\lambda = 0{,}1$



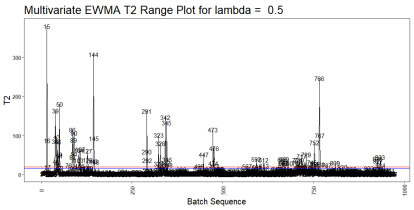
(b) Gráfico MEWMA con $\lambda = 0{,}2$



(a) Gráfico MEWMA con $\lambda = 0{,}3$



(b) Gráfico MEWMA con $\lambda = 0{,}4$



(c) Gráfico MEWMA con $\lambda = 0{,}5$

Para los datos de estudio, los gráficos de control no dan buenos resultados ya que en ningún caso se logra controlar los procesos para las muestras de entrenamiento.

En el Cuadro 5.5, se muestra un resumen de la precisión de los modelos estudiados en este apartado.

	TPR	TNR	BA	NPV	Modelo
Datos Originales (ACP)	0.999	0	0.5	0	-
Datos Originales (ACP Robusto)	1	0.00049	0.5	0.2	-
Medias 75 min (ACP)	0.93	0.9	0.91	0.2	$Var_{10}(1)$ con efectos ARCH(1)
Medias 75 min (ACP Robusto)	0.997	0.85	0.92	0.825	$Var_{9}(8)$ con efectos ARCH(1)
Medias 60 min (ACP Robusto)	1	0.43	0.71	1	$Var_{8}(7)$ con efectos ARCH(1)
Medias 30 min (ACP Robusto)	1	0.43	0.71	1	$Var_{8}(4)$ con efectos ARCH(1)

Cuadro 5.5: Tabla resumen del estudio de gráficos de control

5.3.2. Machine Learning

Una alternativa al uso de gráficos de control es la aplicación de modelos de clasificación supervisada. En esta sección se utilizarán este tipo de procedimientos, representativos en el marco del *Machine Learning*.

Aprendizaje Supervisado

Se seguirá trabajando con las medias de los datos cada 30 minutos, ya que en el apartado anterior obtuvo el modelo más simple dentro de los ACP Robusto, y, además, es la base de datos con mayor tamaño y mayor número de anomalías clasificadas por calcularse las medias de los datos en intervalos de tiempo más cortos.

Como se adelantaba en el Apartado 3.3, cuando se aplican técnicas de *machine learning*, es habitual utilizar todo el conjunto de datos que se tiene disponible para construir un modelo que refleje lo que ocurre en la máquina de la forma más real posible –modelo válido–. En este caso, se ha optado por dividir la muestra de estudio en dos, pero poder incluir en la muestra de entrenamiento los dos tipos de anomalías identificadas se ha tenido que romper la secuencia temporal, ya que las anomalías de la base de datos se detectan en 3 momentos separados y si no se incluye alguna de las anomalías en la muestra de entrenamiento, al utilizar métodos de aprendizaje supervisado, los resultados empeoran. Por ese motivo, para la muestra de entrenamiento, se incluyen los meses de septiembre, octubre, noviembre –hasta mitad del fallo–, además de, diciembre y enero hasta el primer fallo, conformando una muestra con 3542 observaciones en las cuales se detectan 45 anomalías. El resto de la base de datos forma la muestra test, integrada por 1479 observaciones entre las que se registraron 49 anomalías.

En primer lugar se ha ajustado un modelo lineal generalizado con la muestra de entrenamiento utilizando la variable *anomala* como variable respuesta y las variables $T1$, $T2$, $T3$, $T4$, $Consumo$, $X1$, $X2$, $X3$, $X4$, $Y1$, $Y2$, $Y3$, $Y4$, $Z1$, $Z2$, $Z3$, $Z4$ como variables explicativas, aplicando el modelo de clasificación SVM –máquinas de soporte vectorial–. Para ello se han probado utilizando diferentes semillas, valores del parámetro C [3] y de *sigma*[4]. Finalmente se ha fijado la semilla a 15, y se ajusta el modelo en la muestra de entrenamiento empleando un kernel lineal, el parámetro C con valor 1 y sigma con valor 0.05.

```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0518803493069236

Number of Support Vectors : 187

Objective Function Value : -43.011
Training error : 0.001976
Probability model included.
```

Figura 5.63: Resumen de la aplicación de SVM

Para evaluar la precisión del modelo, se realiza la predicción de la variable respuesta en la muestra test. En este caso se obtiene que la sensibilidad del modelo es 0.984, la especificidad es 1, la BA es 0.992 y el NPV es 0.53.

También se ha han aplicado modelos de clasificación supervisada Random Forest (RF) sobre la muestra de entrenamiento, fijando la semilla a 20 [5]. En este caso, se trata de un RF de clasificación con 500 árboles, en el que se toman 4 variables en cada decisión. La tasa de error es del 0.2 % y las anomalías mal clasificadas son el 0.16 % .

[3] El parámetro C es el parámetro que especifica el coste de la violación de las restricciones.
[4] El parámetro sigma es el parámetro que hace referencia a γ, es decir, el inverso del parámetro ventana.
[5] Al igual que en SVM, se han probado varias semillas y esta es la que obtiene mejores resultados.

```
Call:
 randomForest(formula = anomalia ~ T1 + T2 + T3 + T4 + consumo +
X1 + X2 + X3 + X4 + Y1 + Y2 + Y3 + Y4 + Z1 + Z2 + Z3 + Z4,
data = muestra)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0.2%
Confusion matrix:
     0  1 class.error
0 3497  0   0.0000000
1    7 38   0.1555556
```

Figura 5.64: Resumen de la aplicación de RF

Se puede examinar la convergencia del error OOB [6] en las muestras de entrenamiento según el número de árboles utilizados en la Figura 5.65. Como puede observarse, los errores se estabilizan. Esto podría hacernos pensar que aparentemente hay convergencia, aunque en situaciones de alta dependencia entre los árboles su interpretación se dificulta.

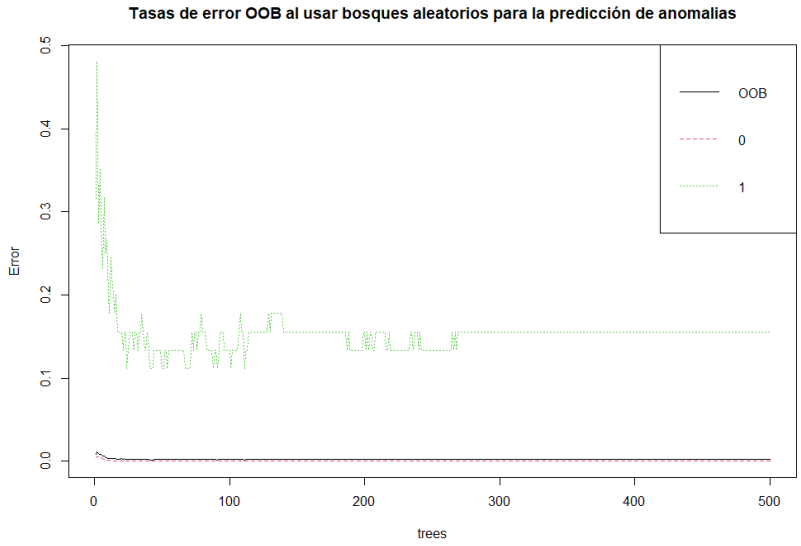


Figura 5.65: Tasas de error OOB utilizando RF

La evaluación de la muestra de test en este modelo, nos indica que se obtiene una sensibilidad de 0.9965, una especificación de 0.9565, una BA de 0.9765 y un NPV de 0.898. Estos son los mejores resultados conseguidos hasta el momento, por eso resulta interesante conocer las variables más importantes y conocer la relación entre las variables.

Las variables que tienen una mayor importancia (véase Figura 5.66) son $Z4, X4, T4, T3, Y4$.

[6] OOB: Out Of Bag. El OOB es una técnica utilizada para medir el error de predicción en modelos de aprendizaje automático que utilizan el método boostrap.

	MeanDecreaseGini
T1	3.0469755
T2	2.4549669
T3	10.6117507
T4	11.6095302
consumo	2.7243977
X1	0.3868281
X2	0.3960664
X3	1.1115100
X4	19.4698966
Y1	0.9480568
Y2	0.5034712
Y3	1.0023166
Y4	8.4380453
Z1	0.4108021
Z2	0.6967202
Z3	3.8277977
Z4	20.9181766

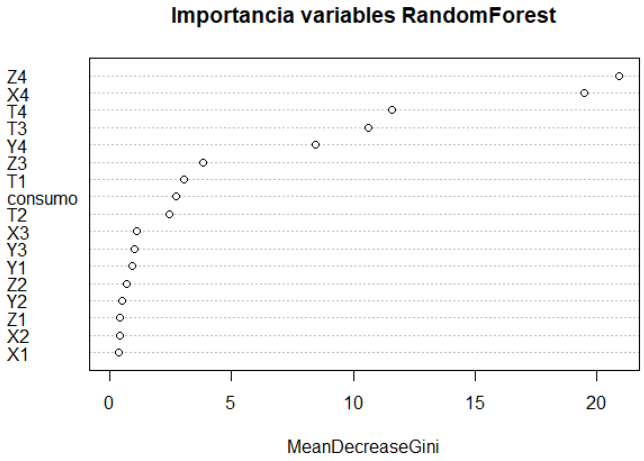
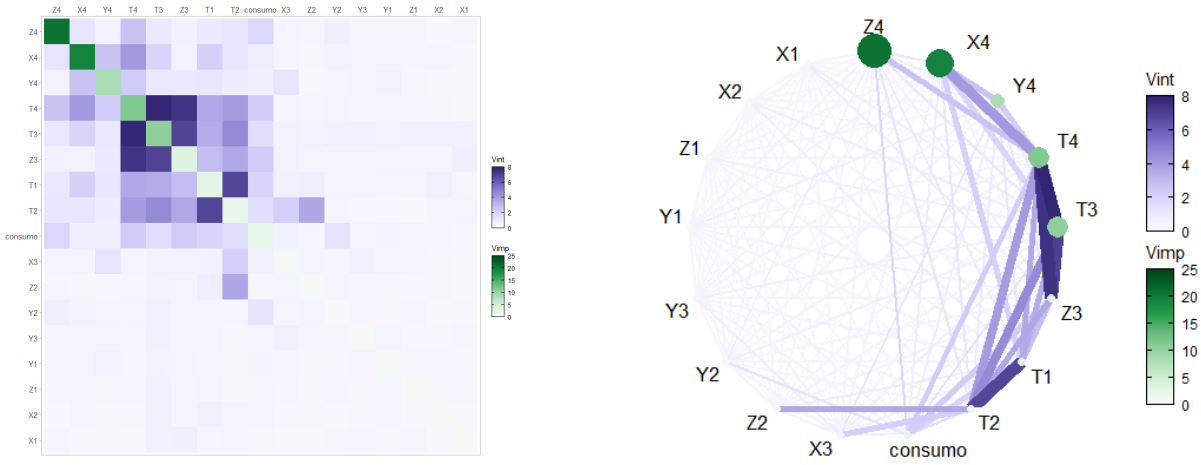


Figura 5.66: Importancia de las variables explicativas al emplear RF

La relación entre variables puede observarse en la Figura 5.67 [7], en la que las *Vint* –señaladas en violeta– hacen referencia a la interacción entre las variables 2 a 2 y las *Vimp* –mostradas en verde– señalan la importancia de los predictores –variables explicativas–.



(a) Mapa de color

(b) Gráfico de red

Figura 5.67: Importancia de las variables explicativas al emplear RF

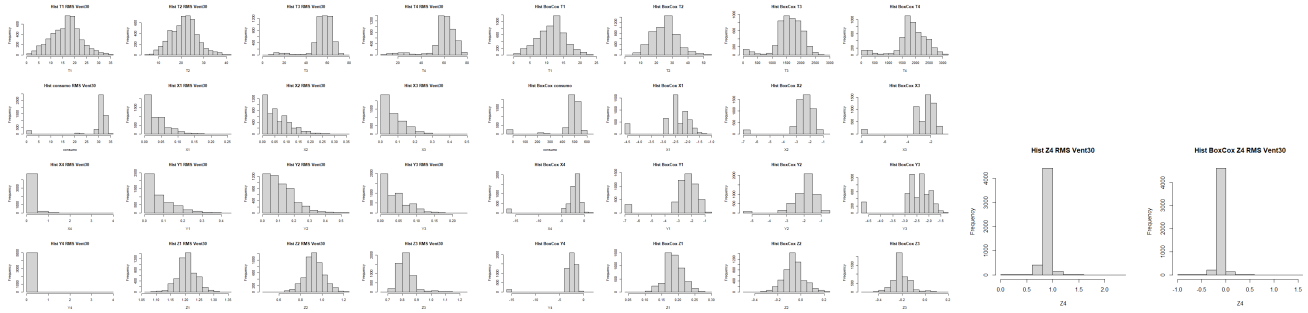
En este caso puede verse que las variables explicativas más importantes son $Z4$, $X4$, $Y4$, $T4$ y $T3$ y, además, existen correlaciones fuertes entre $T4-T3$, $T4-Z3$, $T3-Z3$ y $T2-T1$, correlaciones medias entre $Z4-T4$, $Z4-T4$, $Z4-Y4$, $X4-T3$, $X4-T4$, $X4-T1$, $Y4-T4$, $T4-T1$, $T4-T2$,

[7] Realizada con la librería *vivid* de RStudio.

$T4 - Consumo$, $T3 - T1$, $T3 - T2$, $Z3 - T1$, $Z3 - T2$, $Z3 - Consumo$ y $T2 - Z2$, y correlaciones débiles entre $Z4 - X4$, $Z4 - Consumo$, $Z4 - T4$, $Z4 - T1$, $Z4 - T2$, $X4 - T2$, $T3 - Z4$, $T3 - X4$, $T3 - Consumo$, $T1 - Y4$, $T1 - Consumo$, $T2 - Consumo$, $T2 - X3$, $Consumo - Y1$, $X3 - Y4$.

Además de la media, existen diversas formas de suavizar los datos, una de las más utilizadas en la práctica es a través del RMS. En este caso, se ha realizado el RMS de los datos en intervalos de 30 minutos obteniendo de este modo una BBDD de 5021 observaciones de las cuales 93 son anomalías, es decir, el 1.85 % del conjunto de datos.

Al realizar el RMS, algunas variables se aproximan en mayor medida a una distribución gaussiana pero otras se alejan; si se les realiza la transformación Box-Cox se producen cambios significativos. En la teoría sería beneficioso (Figura 5.68) decantarse por los datos transformados pero en la práctica estos datos están sobresuavizados, es por eso que se continúa trabajando con los datos del RMS sin la transformación Box-Cox.



(a) Histogramas de las variables RMS

(b) Histogramas de las variables RMS + transformación BoxCox

(c) Histogramas de $Z4$

Figura 5.68: Histogramas de las variables a las que se les ha realizado el RMS cada 30 minutos y de las que se les realiza la transformación Box-Cox además del RMS

Al igual que en el caso anterior, la muestra de datos se divide en 2 submuestras para conseguir una muestra de entrenamiento y una muestra test y se rompe la secuencia temporal por el mismo motivo; aunque en este caso la muestra de entrenamiento incluye los meses de septiembre, octubre, noviembre –hasta la mitad del fallo– y enero hasta el primer fallo, estableciendo una muestra de entrenamiento de 2760 observaciones de las cuales 44 están clasificadas como anomalías. Por otro lado, la muestra de test está conformada por el resto de la base de datos –noviembre desde la 2ª mitad el fallo, diciembre y enero desde después del 1er fallo–, en este caso cuenta con 2261 observaciones de las cuales 49 están clasificadas como anomalías.

Se han realizado los mismos análisis que en el caso de la media, el SVM y el RF.

El SVM se ha ajustado con los siguientes parámetros: la semilla 20, el parámetro C con valor 1 y sigma con valor 0.06. A partir del modelo obtenido, se ha verificado sobre la muestra test una sensibilidad de 0.996, una especificidad de 0.947, una BA 0.97 y un NPV de 0.73.

El RF de clasificación, al igual que en el caso anterior, se ha realizado sobre la muestra de entrenamiento con 500 árboles en el que se toman 4 variables en cada decisión. En este caso, la tasa de error es del 0.47 % y las anomalías mal clasificadas son el 0.14 %

```
Call:
 randomForest(formula = anomalia ~ T1 + T2 + T3 + T4 + consumo +
X1 + X2 + X3 + X4 + Y1 + Y2 + Y3 + Y4 + Z1 + Z2 + Z3 + Z4,
data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0.47%
Confusion matrix:
     0  1  class.error
0 1444  1 0.0006920415
1    6 38 0.1363636364
```

Figura 5.69: Resumen de la aplicación de RF

En la Figura 5.70 se puede explorar la convergencia del error en las muestras OOB según el número de árboles utilizados, al igual que en el caso de las medias los errores se estabilizan.

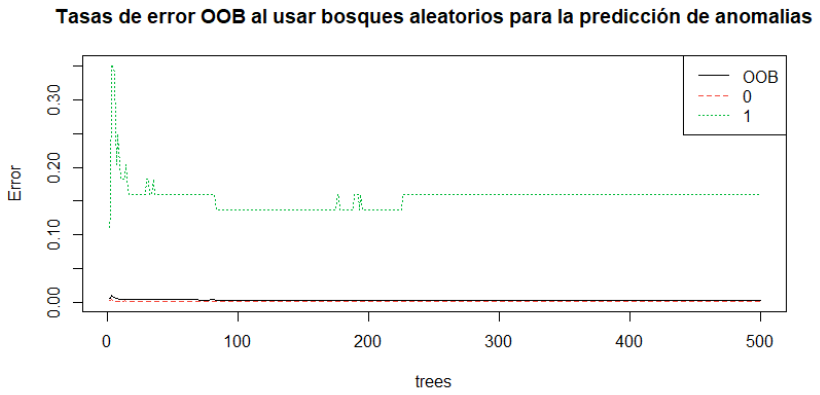


Figura 5.70: Tasas de error OOB utilizando RF

Este modelo se ha verificado sobre la muestra test obteniendo una sensibilidad del 0.9986, una especificidad de 0.9458, una BA 0.9785 y un NPV de 0.85, este es el mejor modelo conseguido.

Se ha estudiado la relación entre las variables y la importancia de las mismas en el modelo.

Como puede verse en la Figura 5.71, las variables más importantes son $X4$, $Z4$, $T4$ y $T3$. En la Figura 5.72, podemos ver que existe una fuerte correlación entre $T4 - T3$, $T4 - Z3$, $T3 - Z3$, una correlación media entre $X4-Z4$, $X4-T4$, $X4-T3$, $Z4-T4$, $Z4-T3$, $T4-Consumo$, $Z3-Consumo$, $Consumo-T3$, $T1-T3$, $T1-T2$, $T1-T4$ y existe correlación débil entre $X4-T1$, $X4-T2$, $Z4-T1$, $Z4-T2$, $T4-T1$, $T4-T2$, $T4-Y1$, $T3-T2$, $Z3-T1$, $Z3-T2$, $Z3-Y1$, $Consumo-T1$, $Consumo-Y1$, $T2 - Y1$, $T2 - X1$, $T3 - Y1$, $T1 - Y1$, $T2 - Y1$, $Y1 - Z1$

	MeanDecreaseGini
T1	2.4409287
T2	2.4769225
T3	8.7957781
T4	10.9817953
consumo	4.2789389
X1	0.4379782
X2	0.3616201
X3	1.1326937
X4	25.0735613
Y1	1.5681697
Y2	0.6954472
Y3	0.8297905
Y4	5.8723186
Z1	0.4467637
Z2	0.8983044
Z3	2.7203284
Z4	17.8971448

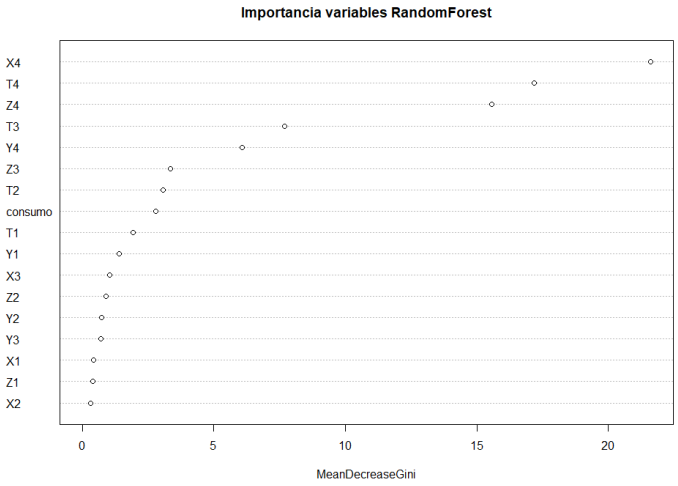
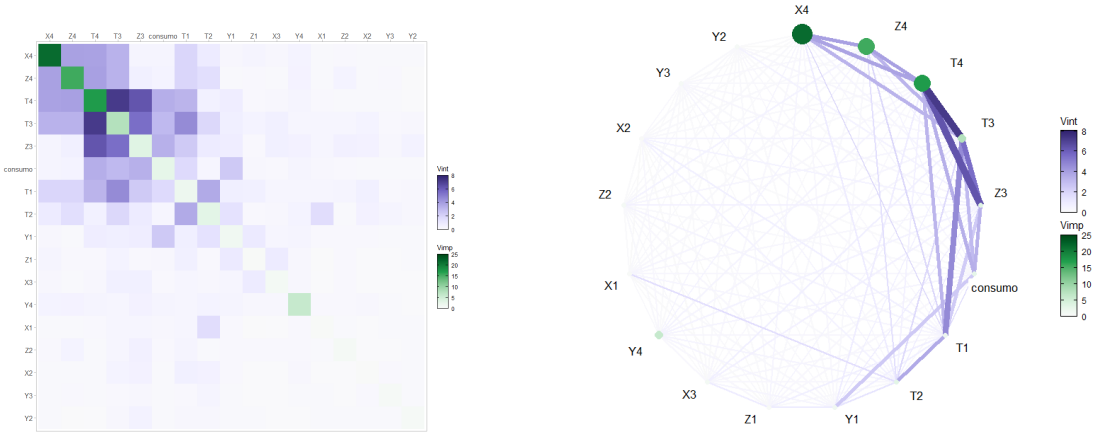


Figura 5.71: Importancia de las variables explicativas al emplear RF



(a) Mapa de color

(b) Gráfico de red

Figura 5.72: Importancia de las variables explicativas al emplear RF

Detección de anomalías

Los métodos de aprendizaje supervisado no son capaces de detectar observaciones que se alejen de los valores habituales sino están previamente identificadas como anomalía. Por ese motivo se ha intentado dar un paso más y utilizar métodos de detección de anomalías. Estos son métodos que se utilizan cuando el número de anomalías registradas es muy pequeño en comparación con el tamaño de la muestra o cuando nos interesa identificar anomalías no catalogadas en la base de datos (Ng et al. 2012). Todo esto los convierte en métodos que podrían ser muy útiles para el problema de estudio, ya

que nuestra muestra tiene pocas anomalías identificadas y, además, nos interesa identificar anomalías estén o no identificadas como tal.

Se ha empleado el algoritmo ALSO, para ello se ha utilizado el código visto en Amat (2020), que puede encontrarse en el Apéndice C. Se ha trabajado con toda la base de datos de RMS en intervalos de 30 minutos, estandarizando las variables ya que están medidas en diferentes unidades.

En la Figura 5.73 puede verse la importancia de cada variable en el modelo; en este caso la variable más importante son $T3, T2, T1, T4, Consumo$ y $Z3$.

```
Modelo T1 --> Peso: 0.6458
Modelo T2 --> Peso: 0.6496
Modelo T3 --> Peso: 0.7099
Modelo T4 --> Peso: 0.6232
Modelo consumo --> Peso: 0.5426
Modelo X1 --> Peso: 0.0384
Modelo X2 --> Peso: 0.022
Modelo X3 --> Peso: 0.0512
Modelo X4 --> Peso: 0.2756
Modelo Y1 --> Peso: 0.2077
Modelo Y2 --> Peso: 0.2392
Modelo Y3 --> Peso: 0.0414
Modelo Y4 --> Peso: 0.0761
Modelo Z1 --> Peso: 0.0744
Modelo Z2 --> Peso: 0.2396
Modelo Z3 --> Peso: 0.4288
Modelo Z4 --> Peso: 0.0699
```

Figura 5.73: Importancia de las variables(peso) en el modelo ALSO

Seguidamente, se utilizan los score como criterio para detectar anomalías, y se representan mediante un gráfico de violín (véase Figura 5.74) en los que, a su vez, se muestra, en rojo, la distribución de las no anomalías y en azul la distribución de las anomalías; en este caso en valor promedio. El problema está en que no todas las anomalías tienen puntuaciones altas, existe solapamiento entre anomalías y no anomalías, lo que produce que si se clasifican las n observaciones con mayor score como anomalías, se estaría incurriendo en errores de falsos positivos.

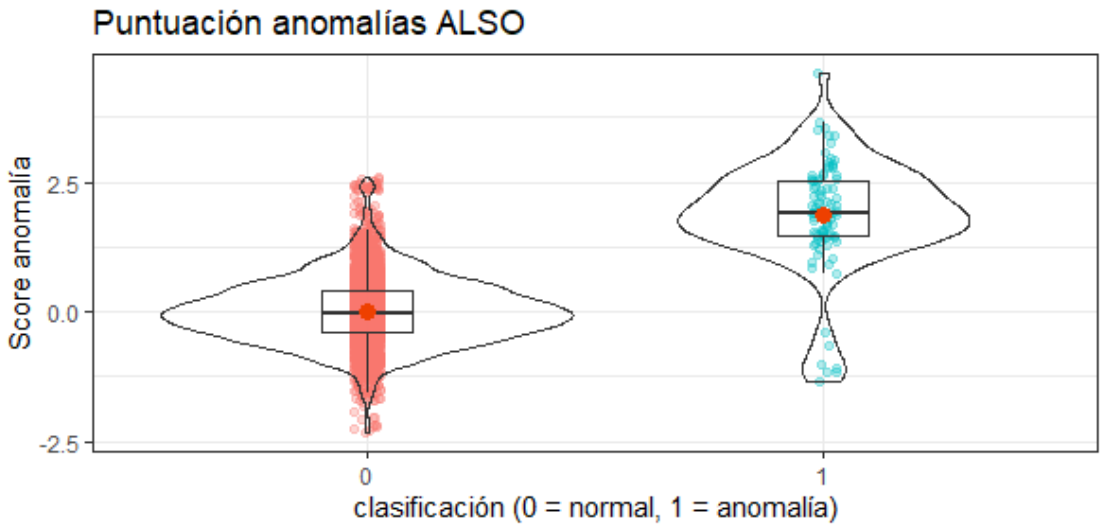


Figura 5.74: Gráfico de violín

Por ese motivo se ha validado sobre la muestra test obteniendo una sensibilidad 0.997, una especificidad de 0.279, una BA de 0.638 y un NPV de 0.65.

A continuación, en el Cuadro 5.6, se muestran las técnicas de *Machine Learning* utilizadas en este apartado, así como las medidas de bondad de clasificación que se ha obtenido en cada técnica.

Técnica	TPR	TNR	BA	NPV
SVM media 30 min	0.984	1	0.992	0.53
RF media 30 min	0.997	0.957	0.977	0.898
SVM RMS 30 min	0.996	0.947	0.97	0.73
RF RMS 30 min	0.999	0.946	0.979	0.85
ALSO	0.997	0.279	0.638	0.65

Cuadro 5.6: Tabla resumen con las técnicas de *Machine Learning* utilizadas y las medidas de bondad de la clasificación obtenidas.

Capítulo 6

Análisis en el dominio de la frecuencia

Los datos para el análisis en el dominio de la frecuencia son los datos recogidos por la FCM en batch. Se me han facilitado en ficheros csv en periodos de 5 segundos en los cuales había datos de los 4 sensores en los ejes X, Y y Z –12 variables–. Los datos se recogieron el día 2 de marzo –4 conjuntos de datos–, y el día 9 de marzo –2 conjuntos–.

Estos datos han tenido que limpiarse y organizarse para poder diferenciar el sensor y el eje al que corresponde cada observación y eliminar las lecturas incompletas. Una vez preparadas las bases de datos se ha comenzado a trabajar con ellas.

En primer lugar se han analizado los datos del día 2 de marzo en el dominio del tiempo, representándolos como en el contexto univariante –a través de gráficos secuenciales, en los que se establecen los límites de especificación fijados mediante la experiencia– y calculando los estadísticos utilizados en dicha sección: el RMS, la media, el FC, la kurtosis y el valor pico-pico.

Tras esto se les realiza la transformada de Fourier y se representa la frecuencia en hertzios para analizar los armónicos que se observan en las gráficas e intentar identificar patrones. Es importante prestar atención al momento donde se completa un ciclo –1x– es decir, al número de vueltas que dan en un segundo, para los sensores 1 y 2 es en los 25Hz, ya que el motor gira a 1500RPM, cada segundo dan 25 vueltas (1500/60=25); en cambio para los sensores 3 y 4 el 1x se sitúa en los 15 Hz, porque la turbina del ventilador gira a 900RPM (900/60=15). Como se mencionó anteriormente, el análisis de las frecuencias debe realizarse en diferentes momentos del tiempo, ya que lo interesante es poder observar la evolución de los armónicos para ver si se produce alguno de los efectos mencionados en el Apartado 2.4.2.

En la Figura 6.1 hace referencia a la variable $X1$ en los 4 periodos recogidos el día 2 de marzo. Puede observarse que en la primera gráfica –FCM031– y en la cuarta –FCM034– hay una observación que supera los límites establecidos, esto puede ser fruto del azar ya que es únicamente 1, si hubiese más deberíamos prestar una mayor atención. La moda para la media es el valor -0.01 y para el RMS el valor 0.07; además, el FC supera el valor 1.8 en los cuatro conjuntos, esto deja entreabierta la probabilidad de la existencia de anomalías no cíclicas; y en 3 de los 4 conjuntos la kurtosis supera el valor 3, lo que indica que la onda es el resultado de la interacción entre dos o más ondas con diferentes frecuencias; y la distancia pico-pico oscila entre 0.39 y 0.53.

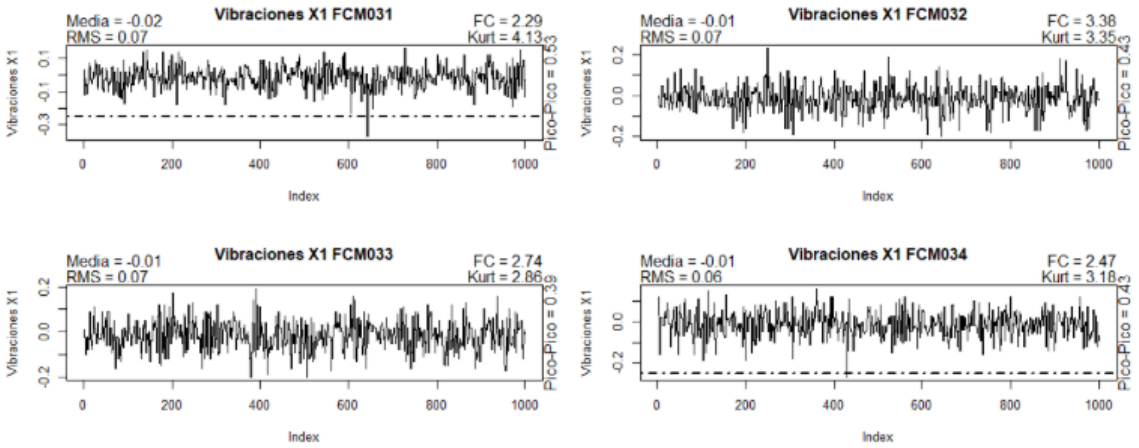


Figura 6.1: Variable $X1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

En el análisis para los sensores 1 y 2 no se hará demasiado hincapié en los picos inferiores a 12.5 Hz por consejo de mi tutor. Si se realiza la transformada de Fourier en las gráficas de la Figura 6.1, las gráficas se transforman en las gráficas de la Figura 6.2, es decir, en su representación en el dominio del espectro. En las 4 gráficas se distinguen picos en torno a los 28-29 Hz, esto es después de 1x; los armónicos no siguen una estructura concreta.

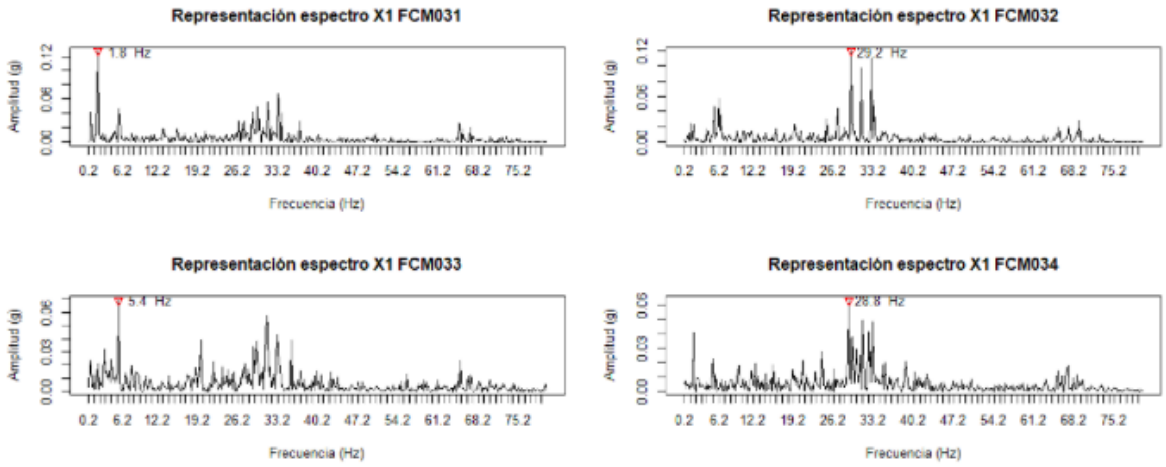


Figura 6.2: Variable $X1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $X2$ (Figura 6.3) mantiene todas las observaciones dentro de los límites de control. Su media oscila entre 0.04 y 0.05, la moda para el RMS es 0.09 y para el valor pico-pico es 0.54, el FC vuelve a ser mayor a 1.8 en todos los casos y la kurtosis toma valores superiores a 3.

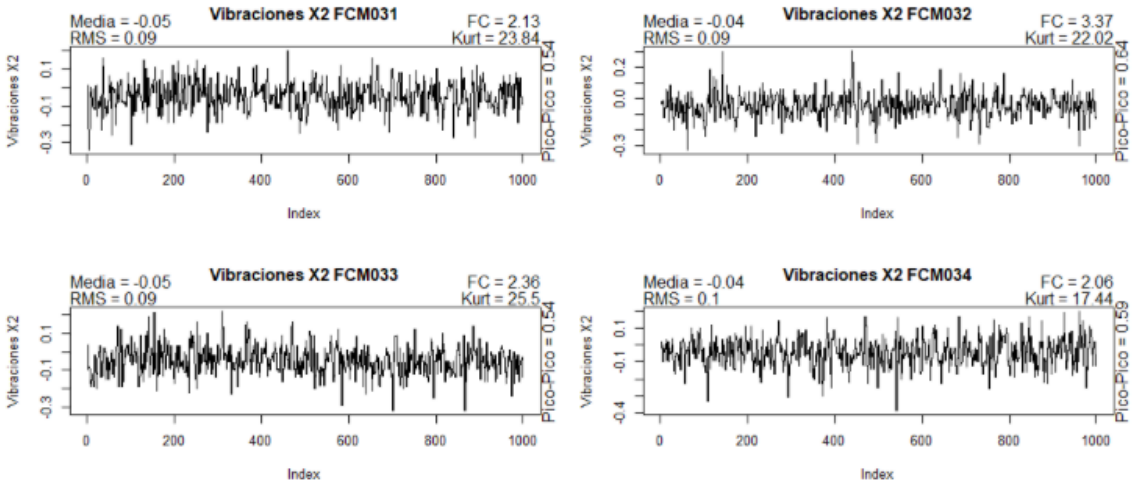


Figura 6.3: Variable $X2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (véase Figura 6.4) es difícil observar algún patrón claro ya que hay muchos picos.

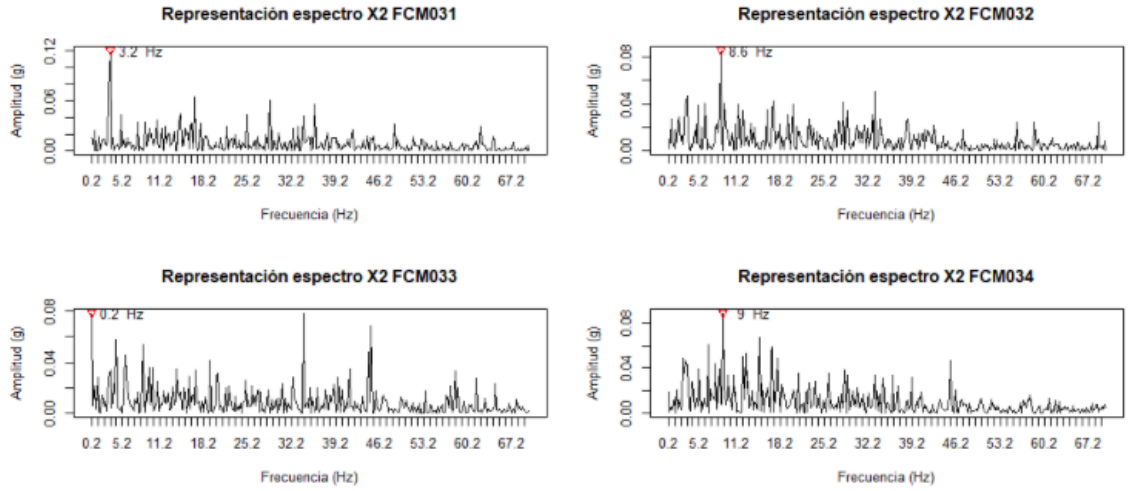


Figura 6.4: Variable $X2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $X3$ (Figura 6.5) mantiene todas las observaciones dentro de los límites de control. Su media oscila entre 0.03 y 0.04, el RMS entre 0.04 y 0.05 y el valor pico-pico entre 0.15 y 0.19, el FC vuelve a ser mayor a 1.8 en todos los casos y la kurtosis toma valores muy por encima de 3.

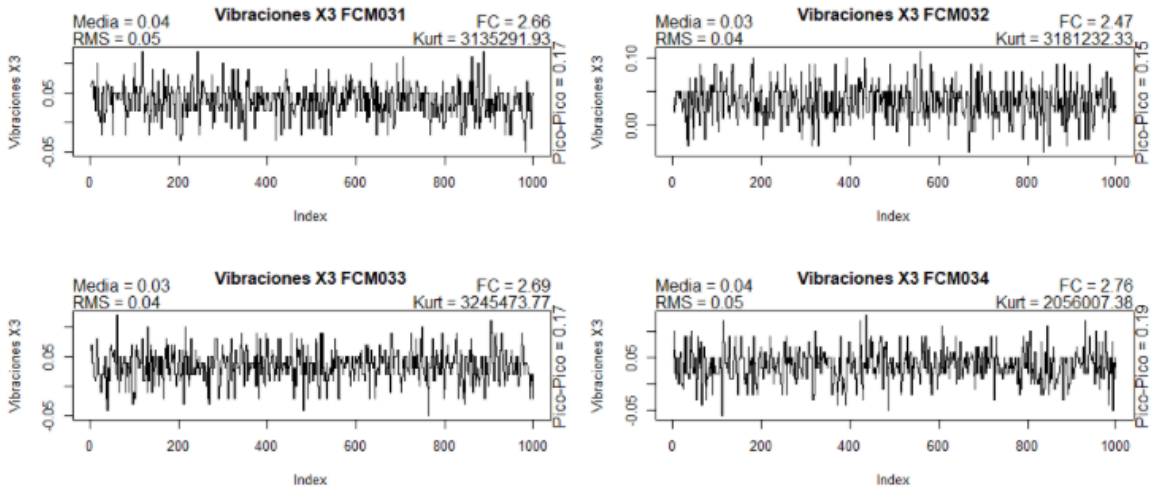


Figura 6.5: Variable $X3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Al realizar la transformada de Fourier –véase Figura 6.6– puede observarse dos armónicos, el primero que en general fue de mayor amplitud en 14.2Hz y otro de menor amplitud en el 17.2HZ, esto es cerca del 1x; este patrón se repite con menor amplitud en los 30Hz y en los 34Hz, cerca del 2x

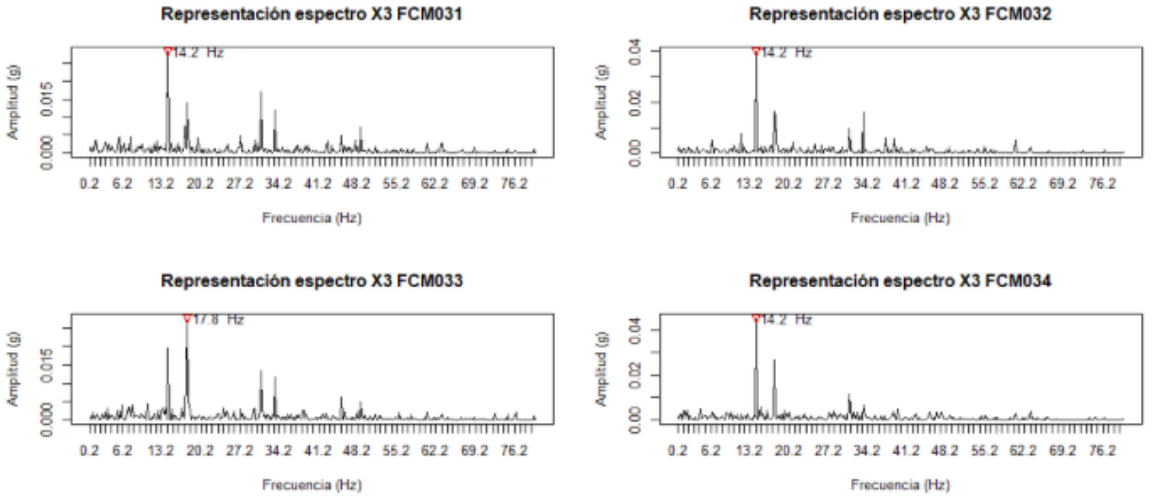


Figura 6.6: Variable $X3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $X4$ (Figura 6.7) mantiene todas las observaciones dentro de los límites de control. La moda para la media es 0.06, el RMS oscila entre 0.04 y 0.05 y el valor pico-pico entre 0.44 y 0.52, el FC vuelve a ser mayor a 1.8 en todos los casos y la kurtosis toma valores superiores a 3.

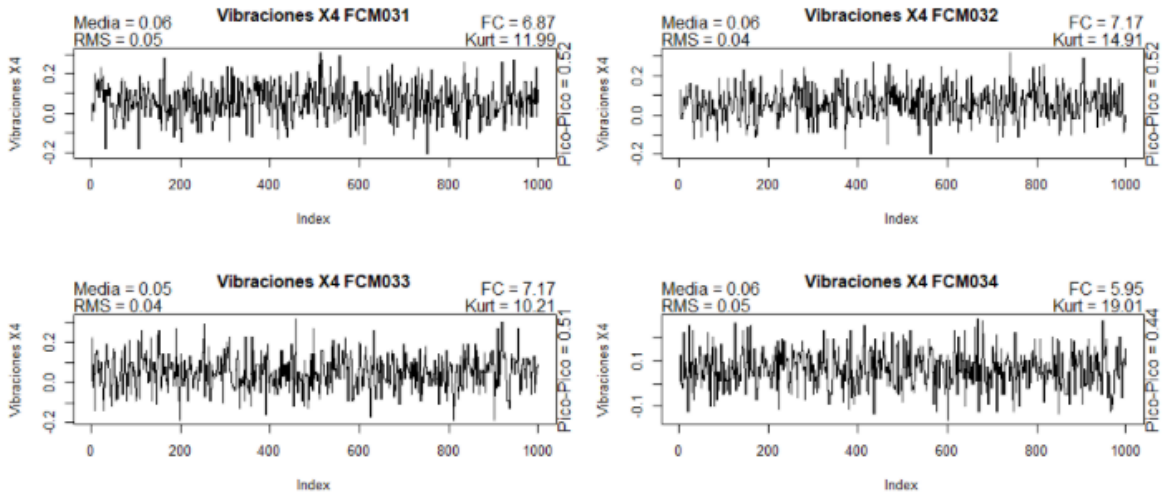


Figura 6.7: Variable $X4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (véase Figura 6.8) es difícil observar un patrón conjunto ya que hay muchos picos, pero parece que hay un armónico en torno al 1 x -15Hz- y otro en torno a 30Hz -2x-.

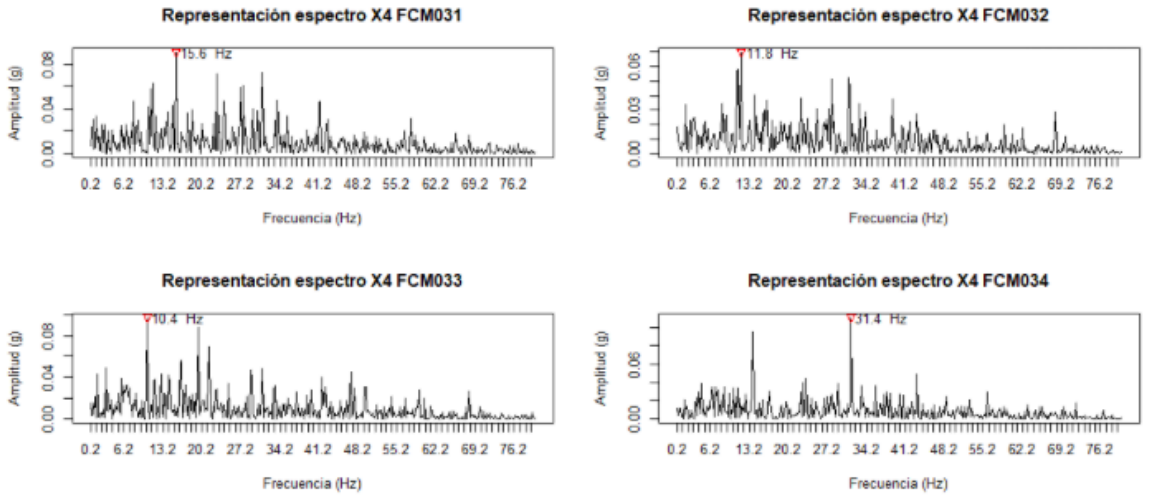


Figura 6.8: Variable $X4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Y1$ (Figura 6.9) mantiene todas las observaciones dentro de los límites de control. Su media es 0.01 y el RMS es 0.12, el FC vuelve a ser mayor a 1.8 en todos los casos, la kurtosis toma valores superiores a 3 y el valor pico-pico oscila entre 0.29 y 0.33.

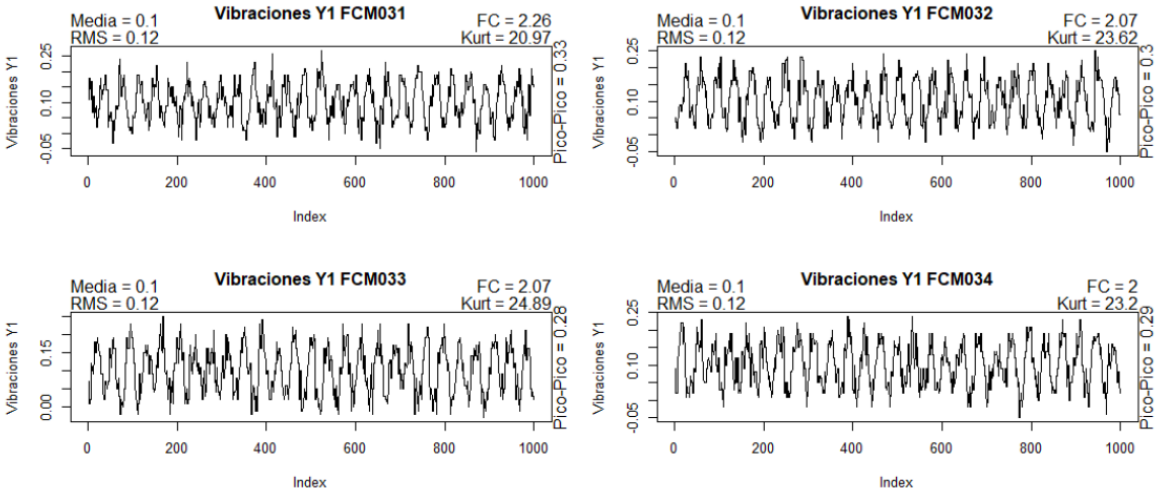


Figura 6.9: Variable $Y1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (Figura 6.10) se observa un único pico en 5.4Hz, pero a este no debe prestársele demasiada atención si se trata de un armónico que solo se produce esta vez, deben analizarse las gráficas en otros momentos para poder observar su comportamiento.

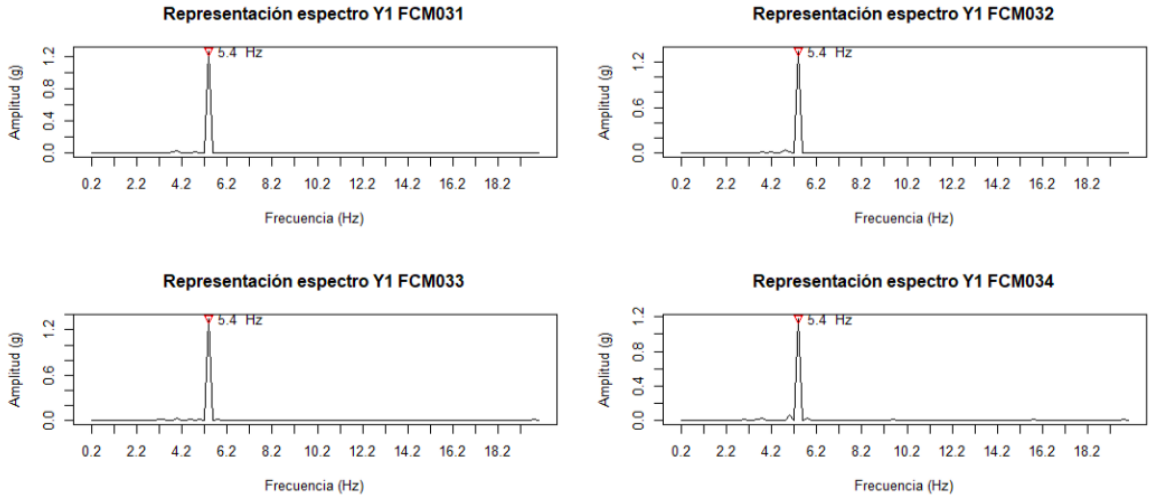


Figura 6.10: Variable $Y1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Y2$ (véase Figura 6.11) mantiene todas las observaciones dentro de los límites de control. Su media es 0.01, la moda para el RMS es 0.06, el FC es mayor a 1.8 en todos los casos, la kurtosis toma valores superiores a 3 y el valor pico-pico oscila entre 0.41 y 0.59.

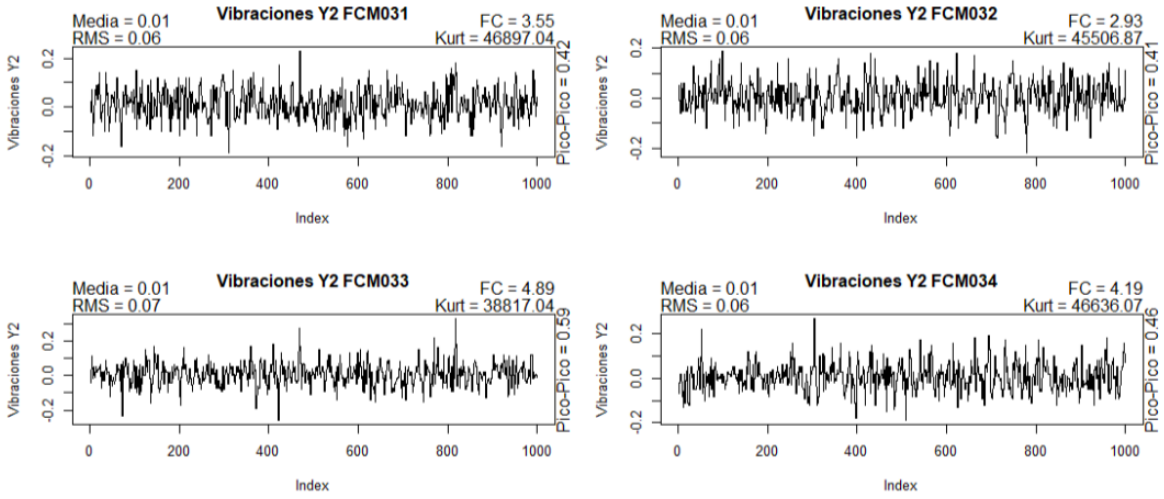


Figura 6.11: Variable $Y2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (Figura 6.12) pueden verse varios armónicos pero el de mayor amplitud y más claro en las cuatro gráficas se encuentra en 9.2Hz, este pico es inferior a 12.5Hz, por esa razón es interesante analizar si sigue apareciendo en el futuro.

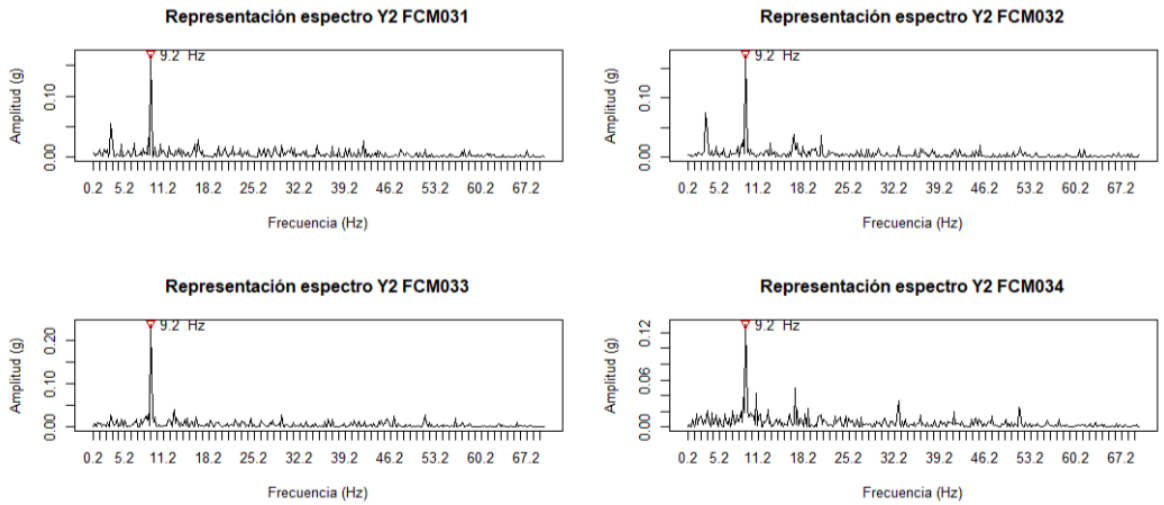


Figura 6.12: Variable $Y2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

En la variable $Y3$ (Figura 6.13) todas las observaciones se mantienen dentro de los límites de control. Su media es -0.03 y el RMS es 0.04, el valor pico-pico oscila entre 0.12 y 0.14, el FC es menor a 1.8 en todos los casos y la kurtosis es mayor a 3.

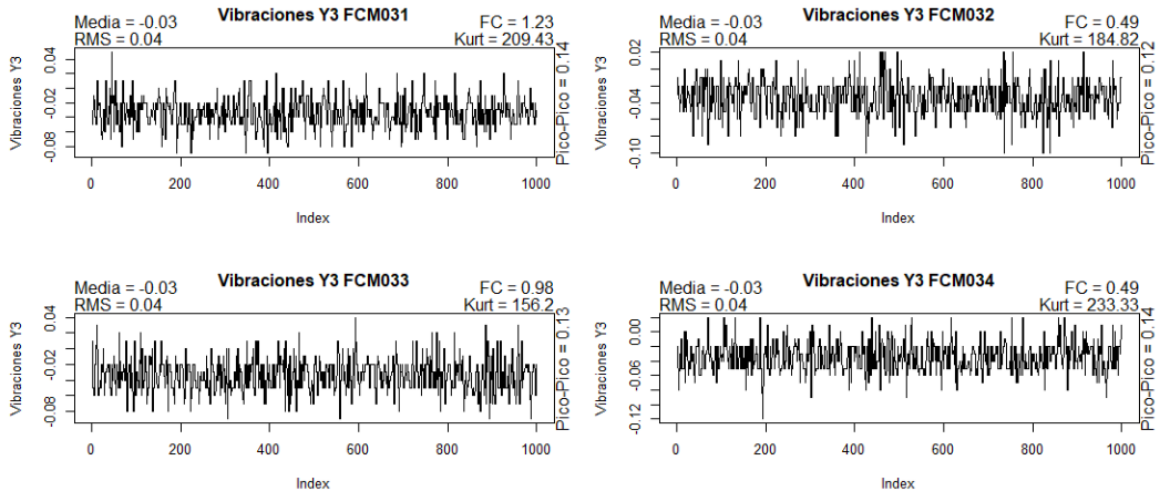


Figura 6.13: Variable $Y3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (Figura 6.14) en 3 de las 4 gráficas se produce un pico claro en torno a 1x en el 16.8Hz, y en las 4 gráficas se producen dos armónicos de diferente amplitud en 46.4Hz -mayor amplitud- y 47Hz -menor amplitud-, es decir, cerca de 3x.

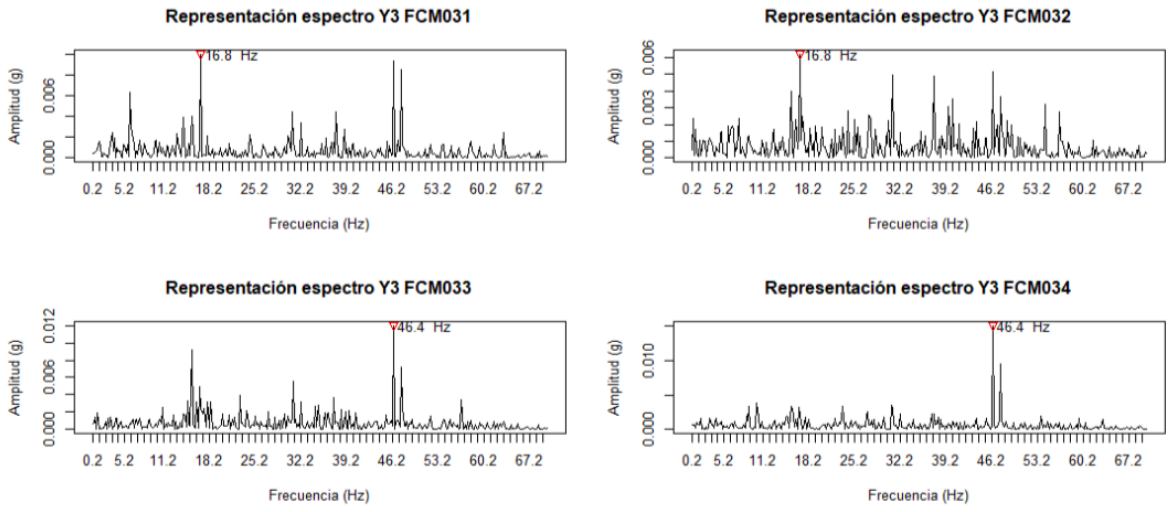


Figura 6.14: Variable $Y3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

En la variable $Y4$ (Figura 6.15) en los 4 conjuntos de datos recogidos el día 2 de marzo hay alguna observación que sobresale de los límites de control. Su media se encuentra entre 0 y 0.02, el RMS entre 0.24 y 0.26, el valor pico-pico entre 1.41 y 1.52, el FC es mayor que 1.8 y la kurtosis toma valores próximos a 0.

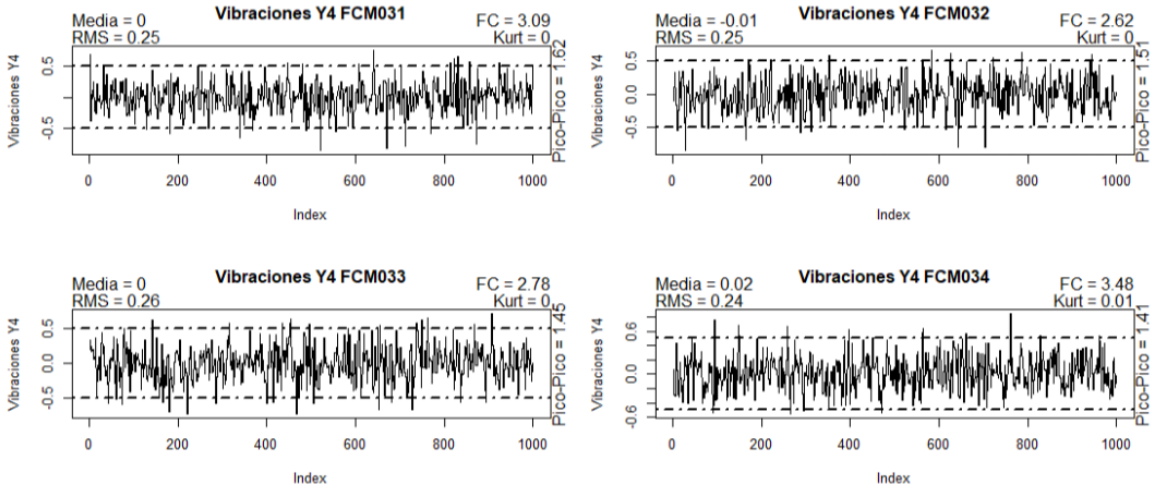


Figura 6.15: Variable $Y4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Al realizar la transformada de Fourier y representarla en el dominio del espectro (Figura 6.16) es difícil identificar picos claros ya que se producen muchos armónicos durante todo el periodo de estudio, lo que sí puede verse es la tendencia descendente esperada al aumentar los hertzios.

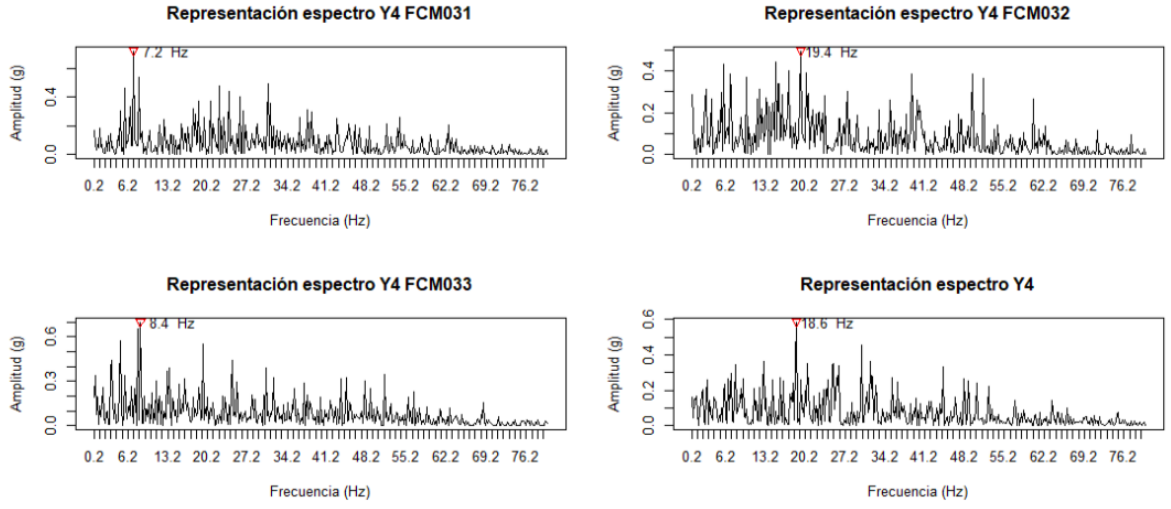


Figura 6.16: Variable $Y4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Z1$ (Figura 6.17) mantiene todas las observaciones dentro de los límites de control. Su media y su RMS es 1.22, la moda del FC es 1.05 y la del valor pico-pico 0.13, la kurtosis toma valores superiores a 3.

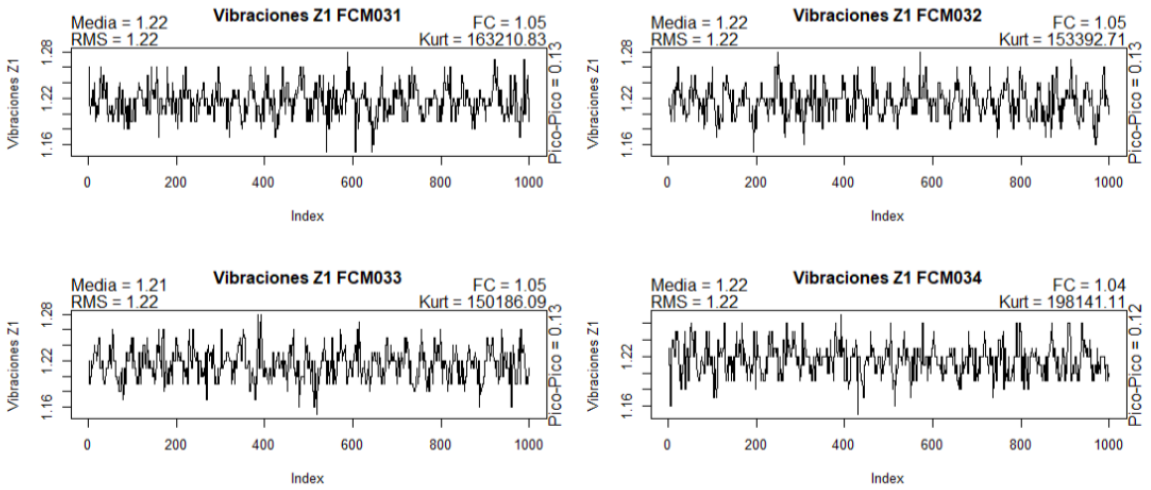


Figura 6.17: Variable $Z1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (Figura 6.18) sucede algo parecido al análisis de Y1, se observa un único pico en 5.4Hz, para poder tomar conclusiones sobre el deben analizarse las gráficas en otros momentos para poder observar su comportamiento.

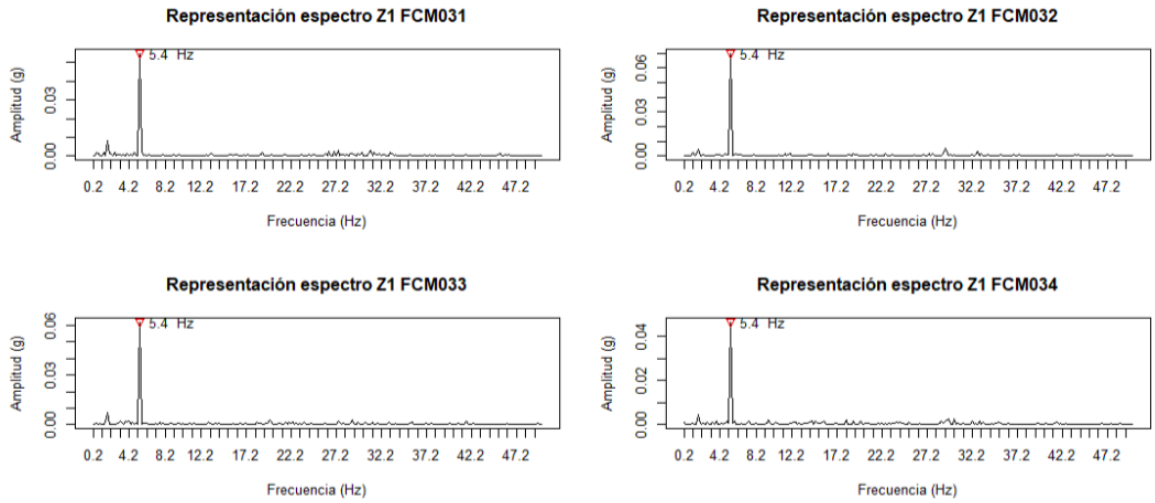


Figura 6.18: Variable $Z1$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Z2$ (véase Figura 6.19) mantiene todas las observaciones dentro de los límites de control. Su media y su RMS es 0.95, el FC toma valor 1.06, el valor pico-pico oscila entre 0.11 y 0.13 y la kurtosis toma valores muy superiores a 3.

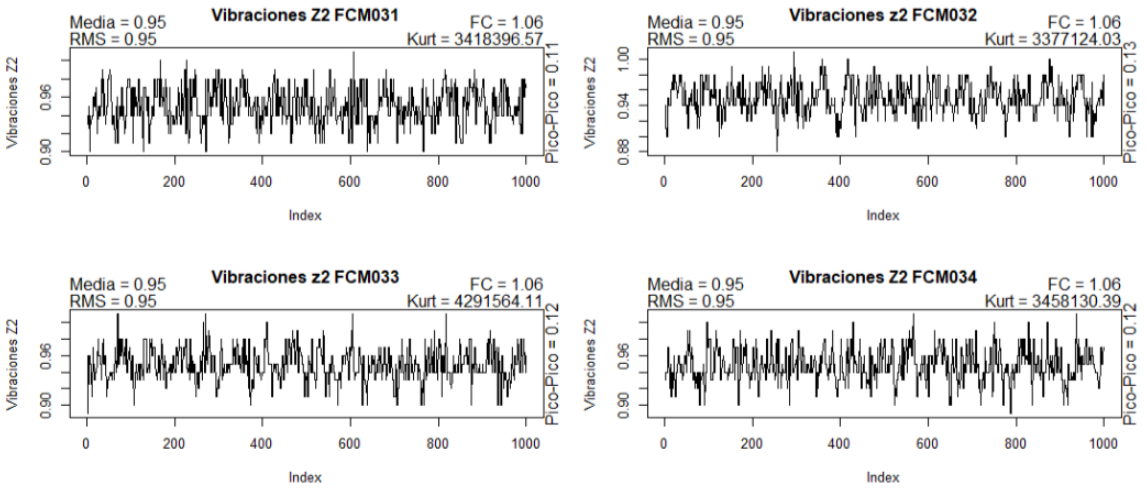


Figura 6.19: Variable $Z2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Si se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (Figura 6.20) se produce el pico de mayor amplitud en torno a 3 Hz aunque no se debe prestar demasiada atención hasta observar más gráficas porque está a frecuencias muy bajas.

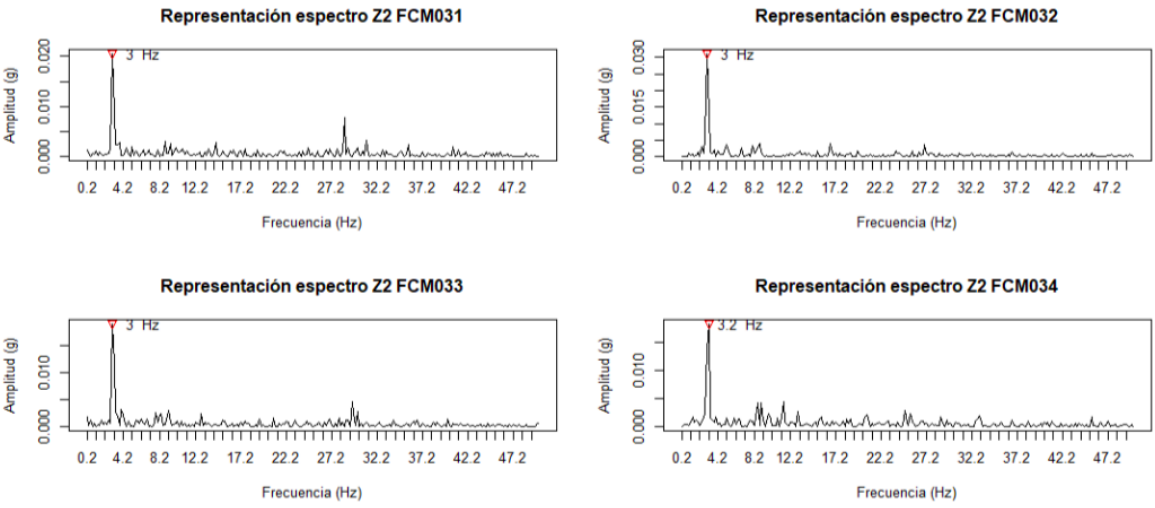


Figura 6.20: Variable $Z2$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Z3$ (Figura 6.21) mantiene todas las observaciones dentro de los límites de control. Su media y su RMS es 0.81, el valor pico-pico oscila entre 0.05 y 0.08 y el del FC entre 1.03 y 1.05, la kurtosis toma valores muy por encima de 3.

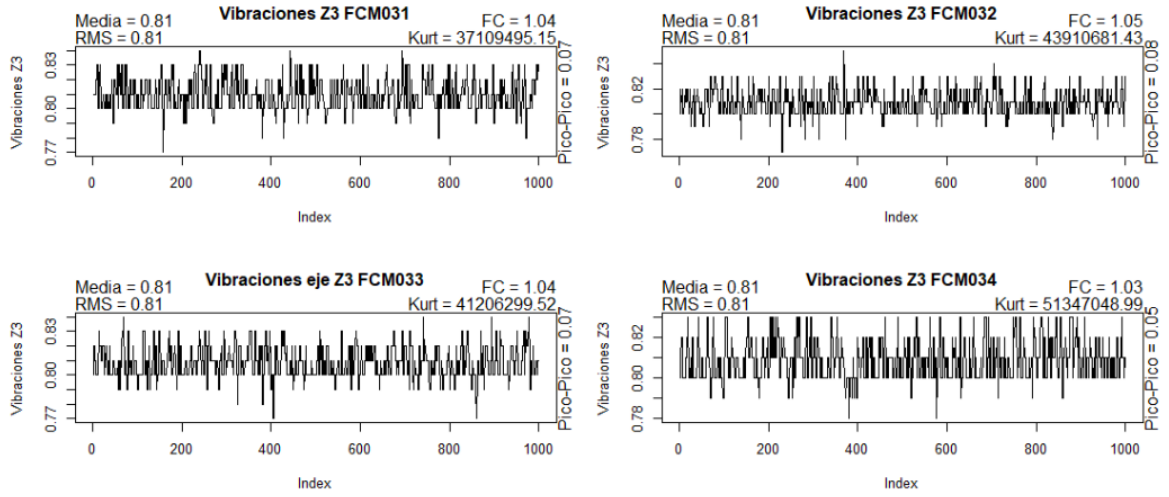


Figura 6.21: Variable $Z3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Al realizar la transformada de Fourier y representar las observaciones en el dominio de la frecuencia (véase Figura 6.22) el primer armónico que cabe resaltar es el de mayor amplitud y se encuentra en torno a 17Hz -poco después de 1x-, está rodeado por dos armónicos de menor amplitud. Además, se observa otro pico en torno a 33Hz, es decir, cerca de 2x.

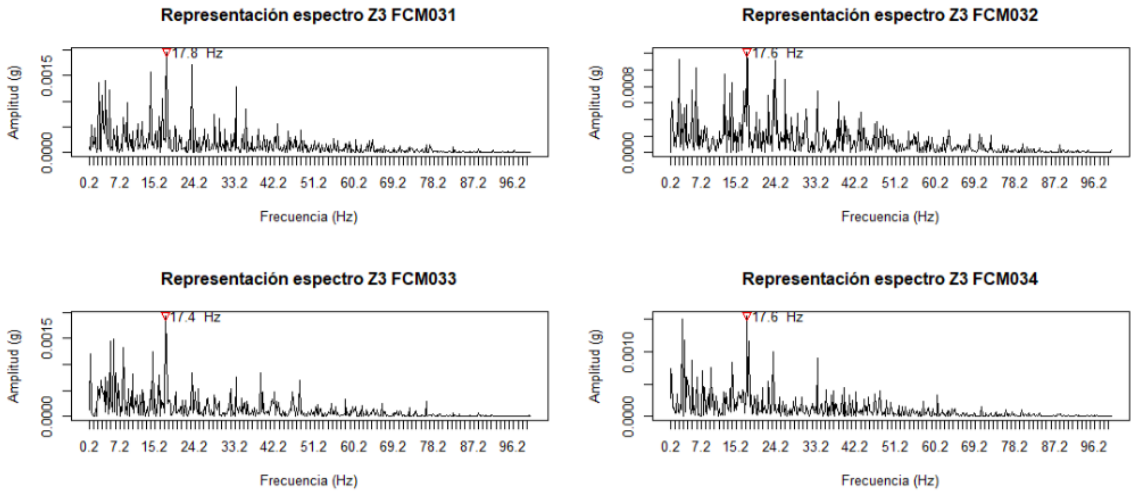


Figura 6.22: Variable $Z3$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

La variable $Z4$ (véase Figura 6.23) mantiene todas las observaciones dentro de los límites de control. Su media y su RMS es 0.86, el FC oscila entre 1.18 y 1.27 -siempre menor a 1.8-, el valor pico-pico entre 0.34 y 0.41 y la kurtosis toma valores próximos a 3.

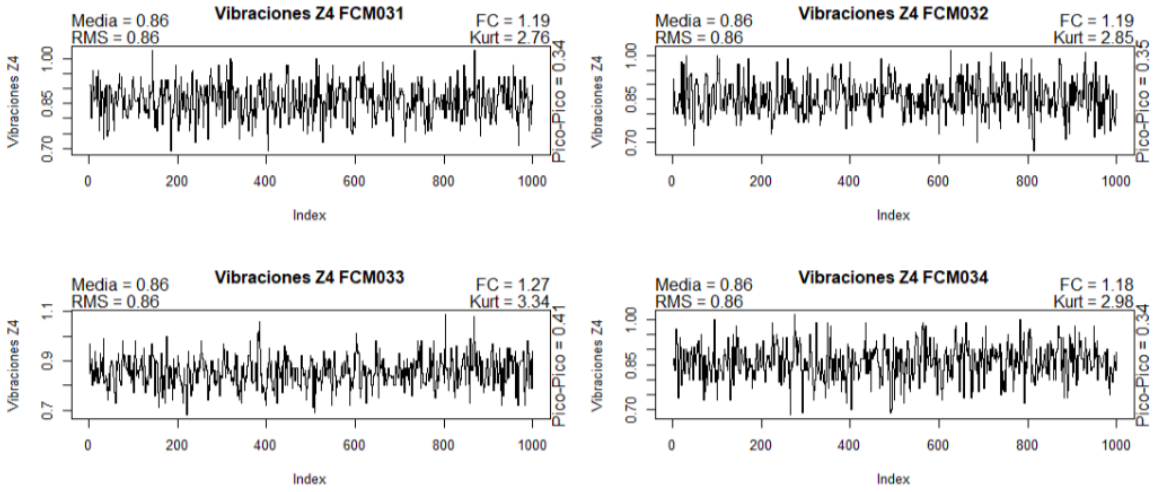


Figura 6.23: Variable $Z4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio del tiempo)

Cuando se realiza la transformada de Fourier y se representa en el dominio de la frecuencia (véase Figura 6.24) no existe un patrón claro.

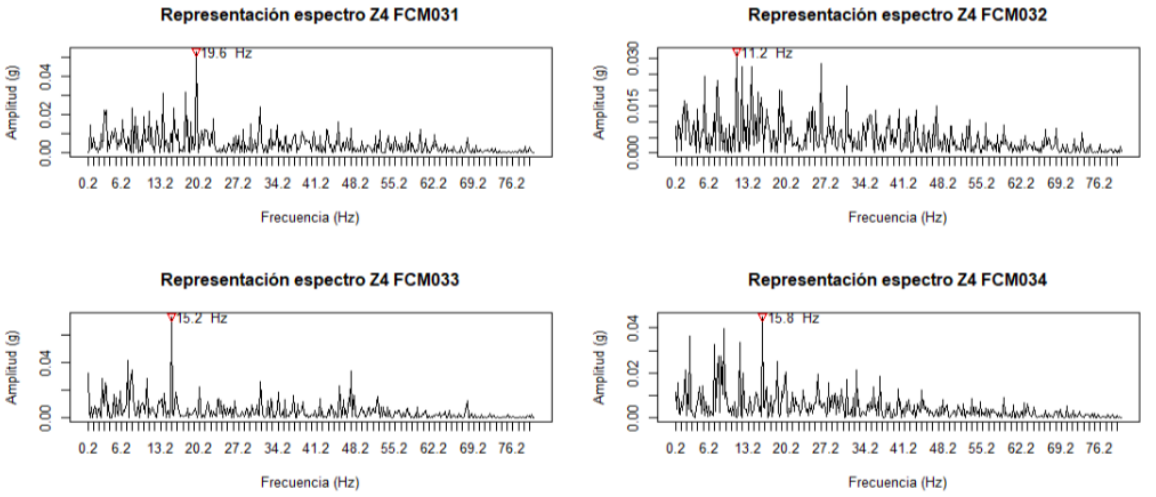


Figura 6.24: Variable $Z4$ en los 4 periodos recogidos el día 2 de marzo (medidas dominio de la frecuencia)

Al analizar los datos recolectados el día 9 de marzo puede verse, en la Figura 6.25, que si analizamos $X1$ en el dominio del tiempo, todos los valores se encuentran dentro de los límites habituales, la media es -0.01, el RMS oscila entre 0.22 y 0.38, el FC es menor a 1.8 en todos los casos, la kurtosis próxima a 3 y la distancia pico-pico está entre 0.45 y 0.51. Además, en el dominio del espectro pueden diferenciarse 3 armónicos, el primero puede obviarse porque se encuentra en los 3Hz, el segundo -mayor que el primero- se encuentra cerca de 0.5x, en los 15Hz, y el tercero es el de mayor amplitud -0.32-0.35g- y se encuentra en torno a los 19Hz.

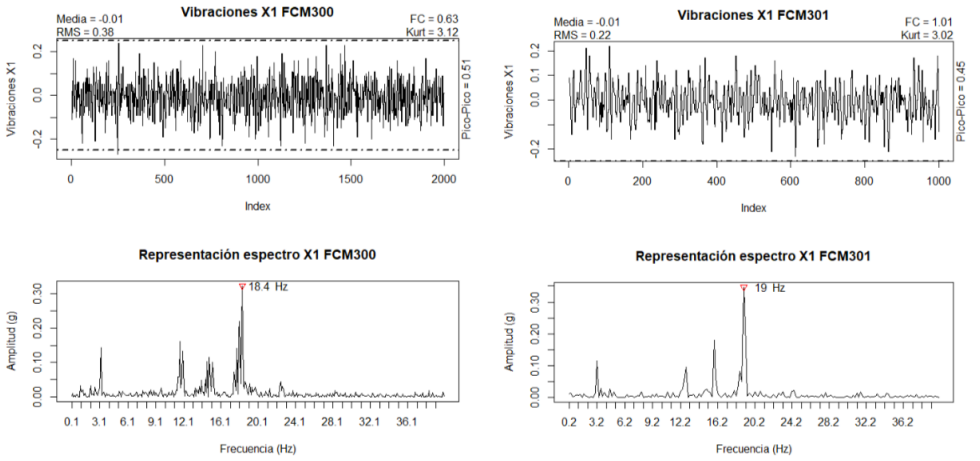


Figura 6.25: Variable $X1$ en los 2 periodos recogidos el día 9 de marzo

En la representación de $X2$ (obsérvese Figura 6.26) puede verse que todos los valores, en el dominio del tiempo, se encuentran dentro de los límites. La media es -0.03, el RMS 0.08, el FC es mayor a 1.8 -entre 2.62 y 2.89-, la kurtosis es superior a 3 y el valor pico-pico oscila entre 0.49 y 0.55. En cuanto a la representación del espectro, se aprecian 2 armónicos en torno a los 10Hz y otro armónico en torno a 25-29Hz.

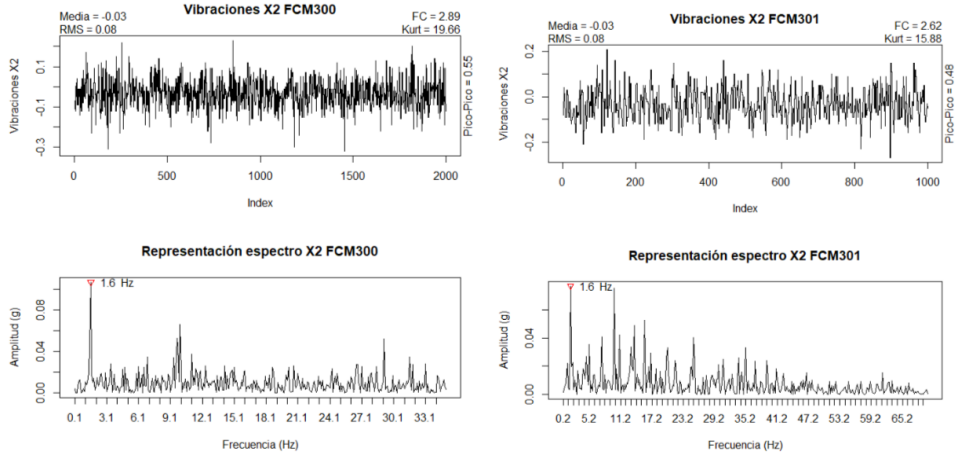


Figura 6.26: Variable $X2$ en los 2 periodos recogidos el día 9 de marzo

Si analizamos la representación de $X3$ (Figura 6.27) vemos que ningún valor sobresale de los límites, la media y el RMS toman valor 0.04, el valor pico-pico es 0.12, el valor del FC oscila entre 2.11 y 2.35 y la kurtosis es superior a 3. El análisis en el dominio de la frecuencia muestra varios armónicos de pequeña amplitud -como máximo 0.02g-, el primero cerca del 0.5x sobre los 6 Hz, el segundo y de mayor amplitud se encuentra cerca de 1x, sobre los 14 Hz; el tercero sobre los 19 Hz y el cuarto y último armónico se encuentra cerca de 2x, 31-32Hz.

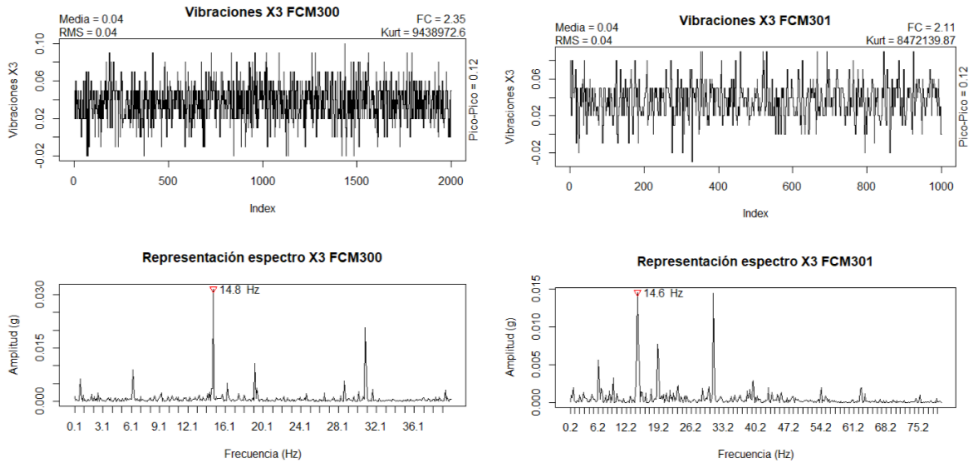


Figura 6.27: Variable $X3$ en los 2 periodos recogidos el día 9 de marzo

Al analizar la variable $X4$ (Figura 6.28) se observa que todos los valores se encuentran dentro de los límites; la media toma el valor 0.06, el RMS 0.04, el valor pico-pico oscila entre 0.14 y 0.17, el valor FC se encuentra en torno a 3, es decir, se queda entreabierta la posibilidad de existencia de anomalías no cíclicas; y la kurtosis toma un valor mucho mayor a 3. Si analizamos su representación del espectro podemos observar un armónico de amplitud 0.06g en torno a 2x, en 31Hz.

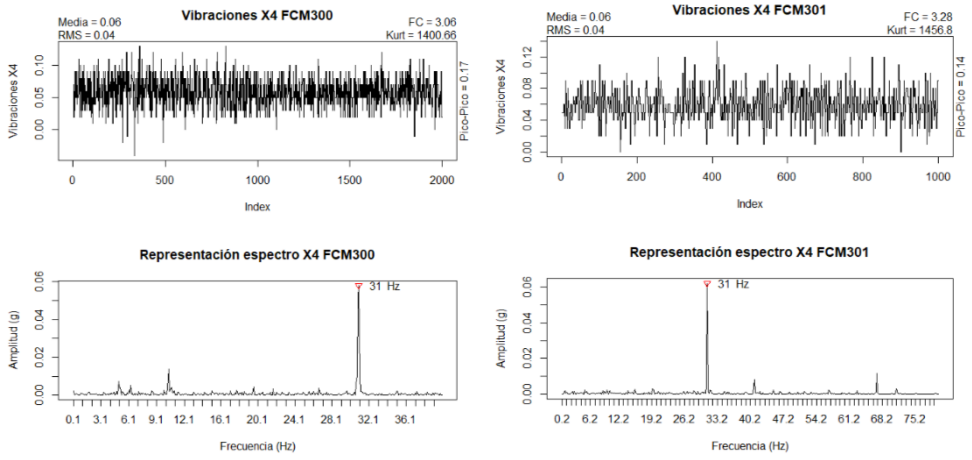


Figura 6.28: Variable $X4$ en los 2 periodos recogidos el día 9 de marzo

Si observamos el Figura 6.29, vemos que todas las observaciones de la variable $Y1$ se encuentran dentro de los límites, su media toma valor 0.1, el RMS valor 0.11, el valor de FC oscila entre 1.67 y 1.82 y el valor pico-pico entre 0.17-0.2, la kurtosis es superior a 3. En cuanto a su representación del espectro se observa claramente un armónico en 9.6Hz, pero este es inferior a 0.5x.

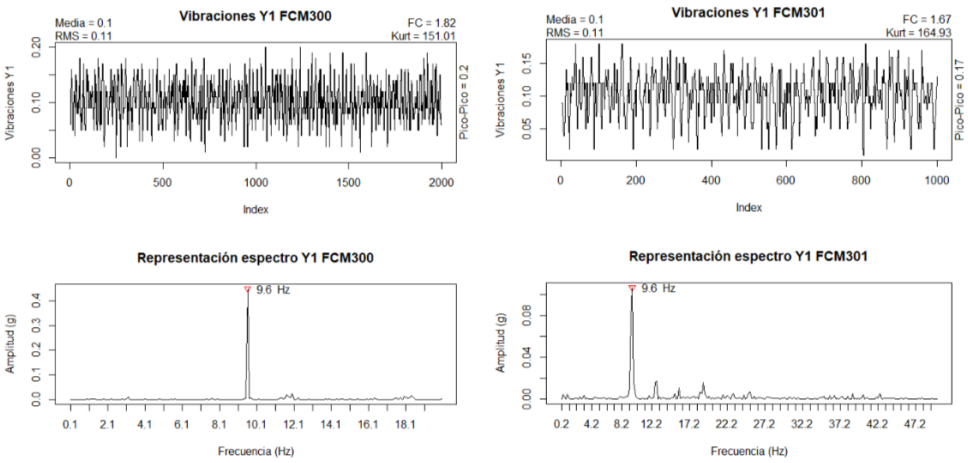


Figura 6.29: Variable $Y1$ en los 2 periodos recogidos el día 9 de marzo

La representación de $Y2$ (Figura 6.30) muestra que todos los valores se encuentran dentro de los límites, el valor de la media es 0.01, y el del RMS es 0.06, el valor pico-pico oscila entre 0.39 y 0.43, el valor del FC es superior a 1.8 y el de la kurtosis es superior a 3. En el dominio de la frecuencia es difícil encontrar un patrón.

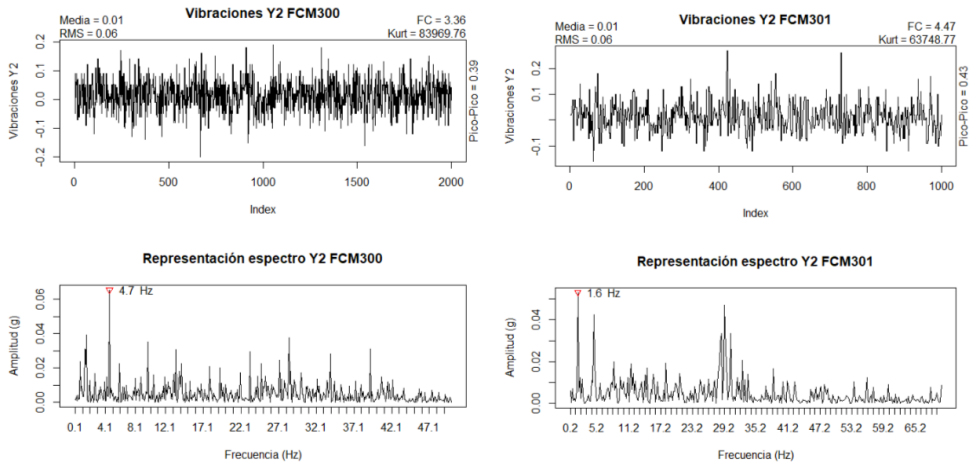


Figura 6.30: Variable $Y2$ en los 2 periodos recogidos el día 9 de marzo

Si analizamos la representación de $Y3$ (Figura 6.31) vemos que ningún valor sobresale de los límites, la media toma valor -0.03 y el RMS 0.04, el valor pico-pico oscila entre 0.12 y 0.13, el valor del FC oscila entre 0.49 y 0.99, valores inferiores a 1.8 y la kurtosis es superior a 3. El análisis en el dominio de la frecuencia muestra dos armónicos en torno a 3x, en 46 y 47Hz siendo el primero de mayor amplitud. En torno a 15 Hz -1x- y 30 Hz -2x- también se detectan armónicos de amplitudes muy bajas.

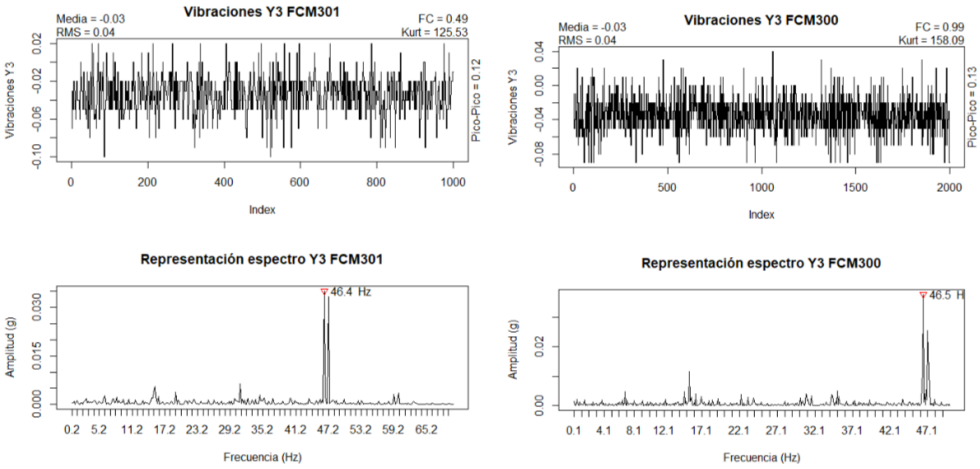


Figura 6.31: Variable $Y3$ en los 2 periodos recogidos el día 9 de marzo

Al analizar la variable $Y4$ (Figura 6.32) se observa que todos los valores se encuentran dentro de los límites; la media toma el valor 0.01, el RMS 0.04, el valor pico-pico oscila entre 0.22 y 0.33, el valor FC oscila entre 2.73 y 3.61 -es superior a 1.8-; y la kurtosis toma valores próximos a 3. Si analizamos su representación del espectro podemos observar en ambos conjuntos un pico en torno a 2x -31 Hz- y otro en torno a 3x -46 Hz-, según la gráfica se modifica la amplitud.

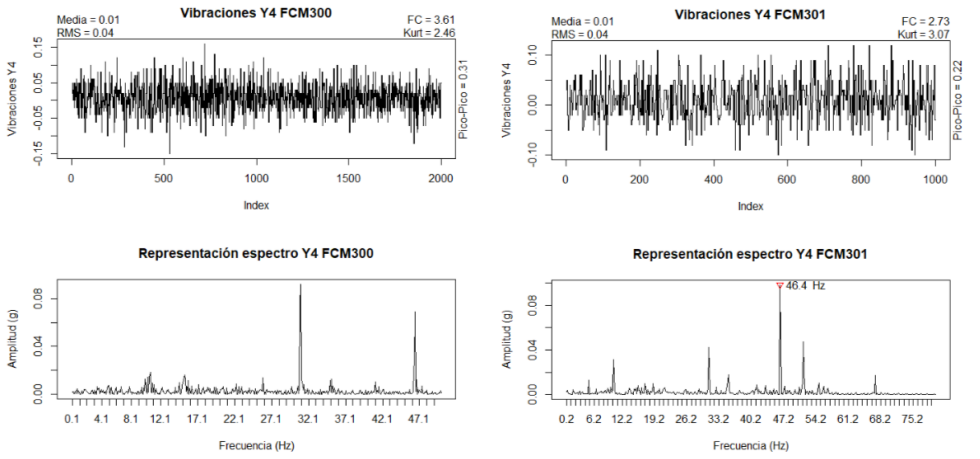


Figura 6.32: Variable $Y4$ en los 2 periodos recogidos el día 9 de marzo

Si observamos la Figura 6.33, vemos que todas las observaciones de la variable $Z1$ se encuentran dentro de los límites, la media y el RMS tienen valor 1.19, el valor de FC es 1.04 y el valor pico-pico está entre 0.11-0.12, la kurtosis es mucho mayor que 3. En cuanto a la representación del espectro los patrones más claros se observa en valores menores a 0.5x -un armónico cerca de 3.1 Hz y otro cerca de 10 Hz-.

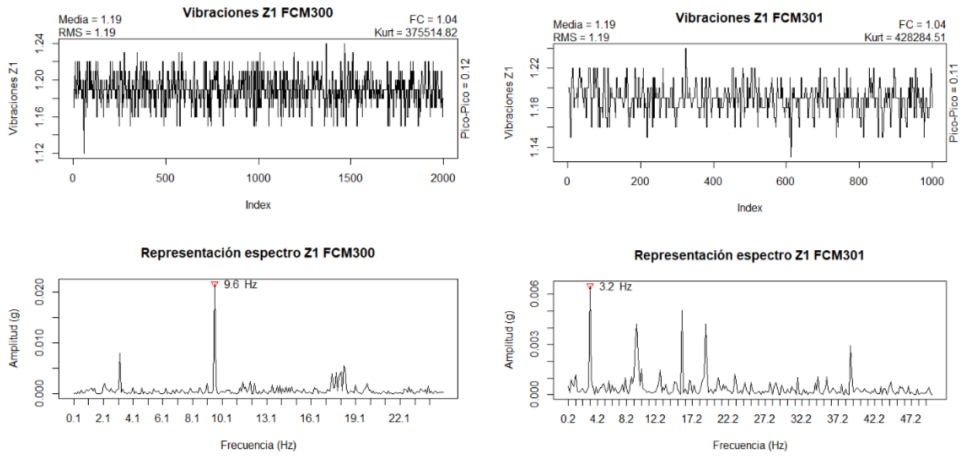


Figura 6.33: Variable $Z1$ en los 2 periodos recogidos el día 9 de marzo

La representación de $Z2$ (Figura 6.34) muestra que todos los valores se encuentran dentro de los límites, el valor de la media y el del RMS es 0.96, el valor pico-pico oscila entre 0.11 y 0.14, el valor del FC oscila entre 1.06 y 1.07 y la kurtosis toma valores muy superiores a 3. En el dominio de la frecuencia el pico de mayor amplitud se observa en 1.6Hz.

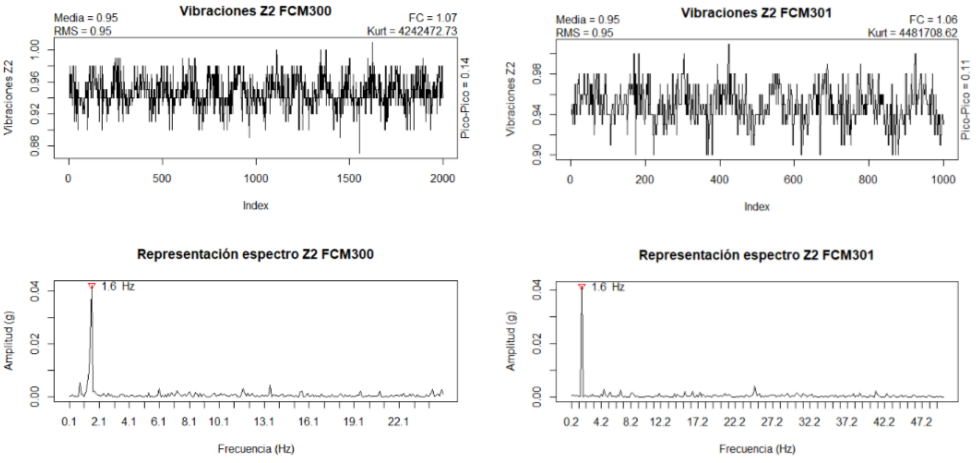


Figura 6.34: Variable $Z2$ en los 2 periodos recogidos el día 9 de marzo

Si analizamos la representación de $Z3$ (Figura 6.35) vemos que ningún valor sobresale de los límites, el valor de la media y del RMS oscilan entre 0.78 y 0.79, el valor pico-pico oscila entre 0.06 y 0.4, el valor del FC oscila entre 1.03 y 1.04, valores inferiores a 1.8 y la kurtosis toma valores muy superiores a 3. El análisis en el dominio de la frecuencia presenta muchos picos de baja amplitud.

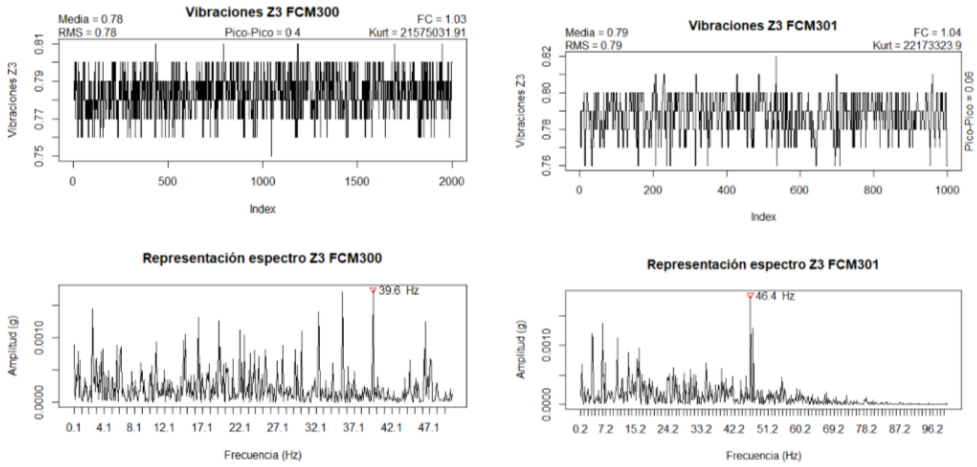


Figura 6.35: Variable $Z3$ en los 2 periodos recogidos el día 9 de marzo

Al analizar la variable $Z4$ (Figura 6.36) se observa que todos los valores se encuentran dentro de los límites; la media y el RMS toman valor 0.85, el RMS 0.04, el valor pico-pico oscila entre 0.08 y 0.11, el valor FC oscila entre 1.05 y 1.06 y la kurtosis toma valores entre 2.15 y 2.67. Si analizamos su representación del espectro no se puede identificar un patrón claro, pero se detecta un armónico en 3x -sobre 31/32 Hz- de amplitud en torno a 0.01g seguido por otro armónico de menos amplitud.

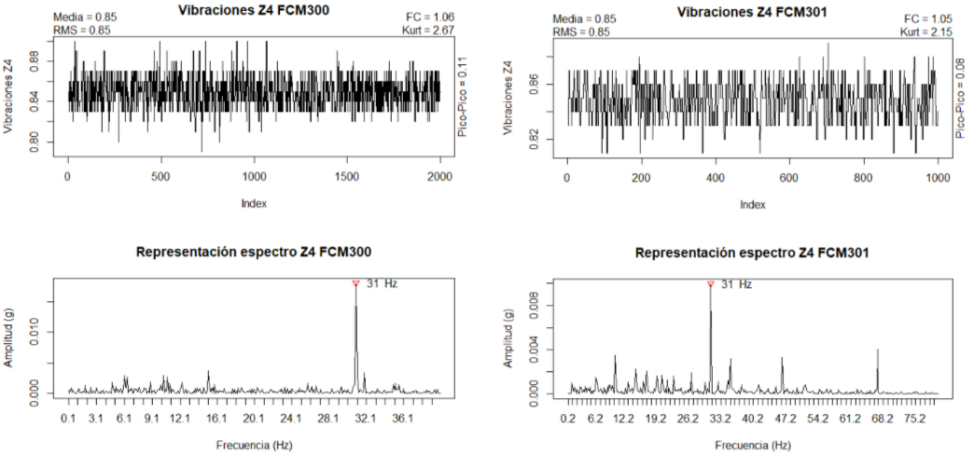


Figura 6.36: Variable $Z4$ en los 2 periodos recogidos el día 9 de marzo

Tras este análisis realizado por días, se van a señalar los patrones que se repiten en los conjuntos estudiados:

- La variable $X2$ el RMS se sitúa en torno a 0.08 y 0.09, el FC suele superar el valor 1.8 y la kurtosis el valor 3.
- La variable $X3$ obtiene un RMS de 0.04, un valor de FC superior a 1.8 -sobre 2- y un valor de kurtosis superior a 3. Además se detecta un armónico de mayor amplitud en torno a 1x y otro de menor amplitud en torno a 2x
- La variable $X4$ obtiene un RMS de 0.04, un valor de FC superior a 1.8 -sobre 2- y un valor de kurtosis superior a 3. Además se detecta un armónico de amplitud entre 0.06-0.1g en 3x -sobre 31 Hz-.
- La variable $Y1$ obtiene un RMS en torno a 0.11/0.12 y un valor para la kurtosis superior a 3.
- La variable $Y2$ tiene un RMS de 0.06, un FC superior a 1.8 y una kurtosis superior a 3. Además, se detecta un pico de diferentes amplitudes en torno a 9/10 Hz y otro de amplitud 0.02-0.04g en torno a 18 Hz
- La variable $Y3$ tiene un RMS de 0.04, un FC inferior a 1.8 y una kurtosis superior a 3. Además, se detectaron dos armónicos de diferentes amplitudes en torno a 3x.
- La variable $Y4$ tiene el FC superior a 1.8. Además se detecta un armónico de 0.04g en 2x.
- La variable $Z1$ obtiene un valor de RMS entre 1.19 y 1.22, el FC inferior a 1.8 y la kurtosis muy superior a 3.
- La variable $Z2$ tiene un RMS de valor 0.95, un FC inferior a 1.8 y una kurtosis muy superior a 3.
- La variable $Z3$ tiene un FC inferior a 1.8 y una kurtosis muy superior a 3.
- La variable $Z4$ el RMS se sitúa en torno a 0.85 y 0.86 y el FC es inferior a 1.8.

Para poder examinar si se produce alguno de los fallos descritos en el Apartado 2.4.2, deben analizarse los datos en más intervalos en diferentes momentos del tiempo, por ejemplo semanalmente, y observar si estos patrones se mantienen o si varían y cómo lo hacen.

Capítulo 7

Conclusiones

Este trabajo ha tenido por objeto la detección de anomalías en instalaciones de la empresa Finsa, más concretamente en el ventilador MW (microwave), situado en la fábrica de Ourense de FINSA (Orember), para poder avisar a los operarios, con antelación, de que algo extraño está sucediendo en la máquina y, de este modo, evitar paradas imprevistas por las pérdidas de producción que conllevan, reparaciones que puedan ocasionar mayores costes y tiempo de parada, etc.

En este trabajo se han utilizado técnicas univariantes para detectar anomalías, tanto en el dominio del tiempo como de la frecuencia, además de estudiar la aplicación de técnicas estadísticas multivariantes únicamente para el primer dominio mencionado.

Las técnicas univariantes en el dominio del tiempo han consistido en realizar un análisis descriptivo sobre los datos recogidos por la FCM en streaming, establecer valores aceptables para cada variables –a través de los histogramas y los cuantiles–, realizar un análisis del periodo de estudio de los valores y de ciertas medidas estadísticas. Además, se ha elaborado un código en python que detecta tanto los fallos identificados como nuevas anomalías. En este caso se han obtenido resultados satisfactorios.

El análisis en el dominio de la frecuencia ha consistido en transportar los datos recogidos por la FCM en batch del dominio del tiempo al dominio de la frecuencia a través de la transformada de Fourier y analizar los datos en ambos dominios en busca de patrones. En este caso se han detectado patrones no concluyentes, es necesario un análisis en un periodo de tiempo mayor para conseguir buenos resultados –analizar los valores habituales– y de este modo, posteriormente se podrán detectar las anomalías en este dominio.

Por último, se han analizado los datos en el dominio del tiempo en un contexto multivariante. Esto se ha llevado a cabo a través de la aplicación de gráficos de control y de técnicas de *machine learning*, concluyendo que los gráficos de control no son el método más adecuado para trabajar con los datos de estudio, ya que no se cumplen muchas de sus hipótesis de partida –normalidad e independencia entre las observaciones, entre otras– y, a pesar de todas las transformaciones realizadas para intentar llevarlos a cabo, incluyendo transformaciones box-cox, ajuste de modelos VAR y reducción de dimensión mediante PCA, los resultados obtenidos, en términos de detección de verdaderas anomalías y ausencia de falsas anomalías, no han sido satisfactorios, si bien los mejores resultados se alcanzan con el modelo en el que se realiza la media de los datos cada 75 minutos y se le aplica un ACP robusto, consiguiendo una sensibilidad = 0.997, una especificidad= 0.85, un balanced accuracy= 0.92 y un NPV= 0.825.

Dado que las técnicas de control estadístico de procesos no han dado el resultado esperado o, al menos, el demandado en FINSA, se ha optado por utilizar técnicas pertenecientes al ámbito del *Machine Learning*, en concreto modelos de clasificación supervisada SVM y RF. Atendiendo a los resultados la que mejor funciona es RF. Los mejores resultados obtenidos, en términos de sensibilidad, especificidad, balanced accuracy y NPV son 0.997, 0.957, 0.997 y 0.898, respectivamente. En todo caso, con respecto a los métodos multivariantes, los resultados obtenidos tampoco han sido, por el momento,

los deseados por la empresa. Por todo ello, atendiendo a su desempeño, se aconseja la utilización de algoritmos de detección de anomalías, como el ALSO.

Teniendo en cuenta la complejidad de los datos que definen el funcionamiento de las instalaciones de la empresa FINSA, se propone, para futuras líneas de investigación, la aplicación de los siguientes nuevos enfoques o algoritmos:

- **El algoritmo Wadjet** –consultado en Sadik et al. (2018)–, el cual se está ideado para trabajar simultáneamente con múltiples flujos de datos. Este algoritmo se basa en dos fases:
  - *FASE 1*: Detectar valores atípicos en flujos de datos individuales analizando las posibles correlaciones temporales dentro de cada flujo. Este se basa en distancias, a través de la idea de que los valores atípicos estarán rodeados de pocos puntos, con lo cual la distancia de un valor atípico al dato más próximo será grande.
  - *FASE 2*: Analiza las posibles correlaciones entre múltiples flujos en 4 etapas:
    - Calcula las correlaciones cruzadas, es decir, entre los diferentes flujos.
    - Agrupación por atributos mediante un análisis de conglomerados.
    - Estandarizar las medidas de los diferentes atributos.
    - Detección de valores atípicos mediante la técnica de k-vecinos más cercanos y diferentes pruebas de significación estadística.
- **El algoritmo de detección de anomalías propuesto por Andrew Ng**, mundialmente conocido como uno de los líderes de la inteligencia artificial y pionero en aprendizaje automático. La idea del algoritmo de Ng et al. (2012) consiste en utilizar la fórmula de la densidad normal multivariante de forma similar a las técnicas basadas en conceptos de profundidad, es decir, una vez se conoce la función de densidad se clasificarían como anomalías las observaciones que se encuentren en las colas de la distribución, o dicho de otro modo, alejadas de la media.

  Para asegurarse de que se está elaborando un modelo válido, como se explica en el Apartado 3.3, si la BBDD es grande se realiza la siguiente división:

  - Muestra de entrenamiento: es la de mayor tamaño, en ella no se pueden incluir anomalías. Se utiliza para definir la función de densidad normal.
  - Muestra de validación: es la segunda muestra más grande, en ella se introduce alguna anomalía, es decir, está formada tanto por anomalías como por no anomalías. Se utiliza para definir el umbral, ϵ, a partir del cual los valores de densidad se consideran poco probables y el tamaño de x_j.
  - Muestra test, formada tanto por anomalías como por no anomalías y utilizada para evaluar el modelo.

  Para llevar a cabo la detección de anomalías se debe:

  - Elegir n características x_i que sirvan como buen indicador de las anomalías.
  - Ajustar los parámetros (vector de medias y la matriz de varianzas-covarianzas) para definir la densidad normal multivariante.
  - Calcular el valor óptimo de ϵ
  - Calcular $p(x)$ para nuevos x siendo $p(x) = p(x_1; \mu_1, \sigma_1^2) \cdot p(x_2; \mu_2, \sigma_2^2) \cdot ... \cdot (x_n; \mu_n, \sigma_n^2) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$
  - Predecir si y_i $\begin{cases} 1 \text{ if } p(x) < \epsilon \ (y_i = \text{anomalía}) \\ 0 \text{ if } p(x) \geq \epsilon \ (y_i = \text{no anomalía}) \end{cases}$

- **Análisis de datos funcionales**, esta es otra posible rama por la que llevar el análisis de los datos estudiados, ya que las observaciones de estudio son puntos y para su análisis es conveniente conocer sus curvas de comportamiento; por eso se han graficado los datos en prácticamente todas las ocasiones. A través de las gráficas de las curvas se puede conocer el comportamiento habitual de los datos, observar si cambia la forma de la curva y detectar, de este modo, las anomalías.

Bibliografía

[Amat, 2020] Amat, J. (2020, Mayo). Detección de anomalías: Attribute wise learning for scoring outliers (ALSO). *cienciadedatos.net*. https://www.cienciadedatos.net/documentos/67_deteccion_anomalias_also#IntroducciÃ§n

[Aneiros, 2022] Aneiros Pérez, G. (2022). *Series de tiempo*[Diapositivas de PowerPoint]. Departamento de Matemáticas, Universidade da Coruña, A Coruña.

[Barbeito et al., 2017] Barbeito, I., Zaragoza, S., Tarrío, J., Naya, S. (2017). Assessing thermal comfort and energy efficiency in buildings by statistical quality control for autocorrelated data. *Applied energy*, 190, 1-17. http://dx.doi.org/10.1016/j.apenergy.2016.12.100

[Cabrera, 2012] Cabrera García, S. (2012). *Gráficos de Control*. http://hdl.handle.net/10251/16262

[Elvatron, 2022] Elvatron S.A. (23 de enero de 2022). *Webinar: Calidad de energía eléctrica*. [Vídeo]. Youtube. https://www.youtube.com/watch?v=IuUNEImEX24

[Fernández et al., 2021] Fernández Casal, R., Costa Bouzas, J., Oviedo de la Fuente, M. (2021). *Aprendizaje estadístico*. https://rubenfcasal.github.io/aprendizaje_estadistico

[Fernández, s.f.] Fernández Jauregui, A. (s.f.). Cómo programar un árbol de decisión en Python desde 0. *Blog Data Science*https://anderfernandez.com/blog/programar-arbol-decision-python-desde-0/

[Ferrero, 2020] Ferrero, R. (2020, mayo). Qué son los árboles de decisión y para qué sirven. *Máxima Formación*. https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

[Francisco, 2021] Francisco, M. (2021). Parte III. Distribuciones notables.

[Gempp, 2006] Gempp Fuentealba, R. (2006). El error estándar de medida y la puntuación verdadera de los tests psicológicos: Algunas recomendaciones prácticas. *Terapia psicológica, 24*(2), 117-129. https://www.redalyc.org/pdf/785/78524201.pdf

[Kim, 2016] Kim, N. (2016). A robustified Jarque–Bera test for multivariate normality. *Economics Letters, 140*, 48-52. https://doi.org/10.1016/j.econlet.2016.01.007

[López y Fernández, 2022] López Taboada G., Fernández Casal R. (2022) *Prácticas de Tecnologías de Gestión y Manipulación de Datos* https://gltaboada.github.io/tgdbook/

[Mauricio, 2007] Mauricio, J.A.(2007, marzo). Introducción al análisis de series temporales. *Universidad Complutense de Madrid*. https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf

[Metravib, s.f.] Metravib technologies (s.f.). *Vibration monitoring of rolling element bearings*. http://www.plant-maintenance.com/articles/bearing_vibration_monitoring.pdf

[Montgomeri, 2020] Montgomery,D.C. (2020). *Introduction to statistical quality control*. John Wiley & Sons. https://books.google.es/books?hl=es&lr=&id=oh7zDwAAQBAJ&oi=fnd&pg=PR3&dq=Introduction+to+statistical+quality+control&ots=DrExg6Kb8B&sig=m-ad5jd0Pt4eiz1f7SQA2wyWA3M#v=onepage&q=Introduction%20to%20statistical%20quality%20control&f=false

[Naya y Tarrío, 2021] Naya, S. y Tarrío, J. (2021). *Control Estadístico de la Calidad (CEC) Tema 3: Introducción a los gráficos de control*. Departamento de Matemáticas, Universidade da Coruña, A Coruña.

[Ng et al., 2012] Ng, A., Shyu, E., Bagul, A., Ladwig, G. (2012). *Unsupervised Learning, Recommenders, Reinforcement Learning*. Coursera. https://www.coursera.org/learn/unsupervised-learning-recommenders-reinforcement-learning/home

[Novales, 2017] Novales, A. (2017). Modelos vectoriales autoregresivos (VAR). *Universidad Complutense*, 1-26. https://www.ucm.es/data/cont/media/www/pag-41459/VAR.pdf

[Núñez, 2014] Nuñez Pérez, R.F. (2014). La tendencia del factor de cresta ayuda a detectar eventos nacientes; circuito electrónico, programas y aplicaciones a señales de diversos campos. *Ingeniería, Investigación y Tecnología*, XV(1), 63-81. https://doi.org/10.1016/S1405-7743(15)30007-X

[Palomio, 2008] Palomino Marín, E. (2008). *Curso de análisis de vibración. La medición y el análisis de vibración en el diagnóstico de máquinas rotativas*. Renovetec. https://www.academia.edu/31421088/Curso_de_an%C3%A1lisis_de_vibraci%C3%B3n

[Pateiro y Sánchez, 2022] Pateiro López, B. y Sánchez Sellero, C. (2022). *Análisis multivariante*. Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, Santiago de Compostela.

[Restrepo, 2020] Restrepo, J.B. [3Blue1Brown Español]. (2020, 21 de septiembre). *¿Qué es la Transformada de Fourier? Una introducción visual* [Vídeo]. Youtube. https://www.youtube.com/watch?v=h4PTucW3Rm0

[Roca, 2017] Roca Pardiñas, J. (2017). *Gráficos estadísticos con R*

[Royo et al., s.f.] Royo, J. , Rabanaque, G. , Torre, F. (s.f.). *Análisis de vibraciones e interpretación de datos*. http://www.todosensores.es/articulos/vibraciones.pdf

[Ruiz, 1994] Ruiz, E. (1994). *Modelos para series temporales heterocedásticas*. Departamento de Estadística y Econometría, Universidad Carlos III, Madrid. https://core.ac.uk/download/pdf/29428239.pdf

[Sadik et al., 2018] Sadik, S., Gruenwald, L., Leal, E. (2018, April). *Wadjet: Finding outliers in multiple multi-dimensional heterogeneous data streams*. In 2018 IEEE 34th International Conference on Data Engineering (ICDE) (pp. 1232-1235). IEEE. doi: 10.1109/ICDE.2018.00118

[Saéz y Pérez, 1994] Saéz Zafra, M y Pérez Rodríguez, J.V.(1994). Modelos autorregresivos para la varianza condicionada heteroscedastica (ARCH). *Estudios de Economia Aplicada, 2*(3), 71. https://www.researchgate.net/profile/Jorge-Perez-Rodriguez/publication/28088549_Modelos_autorregresivos_para_la_varianza_condicionada_heteroscedastica_ARCH/links/5523b3db0cf2b351d9c31a81/Modelos-autorregresivos-para-la-varianza-condicionada-heteroscedastica-ARCH.pdf

[Sebas, 2020] Sebas Pirón, N. (2020). *Técnicas para el control estadístico de calidad a partir de datos multivariantes* [Trabajo Final de Máster, Universidade de Santiago de Compostela]. http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1793.pdf

[Ssusera1e9de, 2022] Ssusera1e9de (2022, 15 de marzo). *Vibration diagnostic chart* [Diapositivas de PowerPoint]. SlideShare. https://www.slideshare.net/ssusera1e9de/vibration-diagnostic-chart

[Trocel, 2021] Trocel, D. (2021, 15 de abril). Análisis de vibraciones en ventiladores centrífugos. Power-MI Blog. https://power-mi.com/es/content/an%C3%A1lisis-de-vibraciones-en-ventiladores-centr%C3%ADfugos

[Universidad de Córdoba, 2018] Universidad de Córdoba (2018). *Parámetros que definen una vibración*. Laboratorio virtual de riesgos laborales. http://www.uco.es/RiesgosLaborales/fisicoyquimico/vibraciones/tutorials/view/5-Parametros-que-definen-una-vibracion

[Vaamonde, 2019] Vaamonde Rivas, M. (2019). *Sistema de Detección de Anomalías en Equipos Industriales*[Trabajo Final de Máster, Universidade de Santiago de Compostela].

[Vilar, 2021] Vilar, J.A.(2021). *Métodos no paramétricos*[Diapositivas de PowerPoint]. Departamento de Matemáticas, Universidade da Coruña, A Coruña.

[Wang et al., 2021] Wang, Q., Yan, B., Su, H., Zheng, H. (2021, March). *Anomaly detection for time series data stream*. In 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA) (pp. 118-122). IEEE. doi: 10.1109/ICBDA51983.2021.9402957

[White, 1990] White, G. (1990). *Introducción al Análisis de Vibraciones*. Azima DLI

[Zubia, 2021] Zubia Garea, A. (2021). *Detección de anomalías en redes IoT mediante Stream Machine Learning*. http://hdl.handle.net/10810/53304

Apéndice A

Tipos de tableros y sus acabados

A continuación se muestran los diferentes tipos de tableros -FDM, aglomerado y superpan-, los colores con los que se fabrican tableros, las formas que se pueden reproducir en los tableros añadiendo presión; así como los acabados con melamina y chapa.

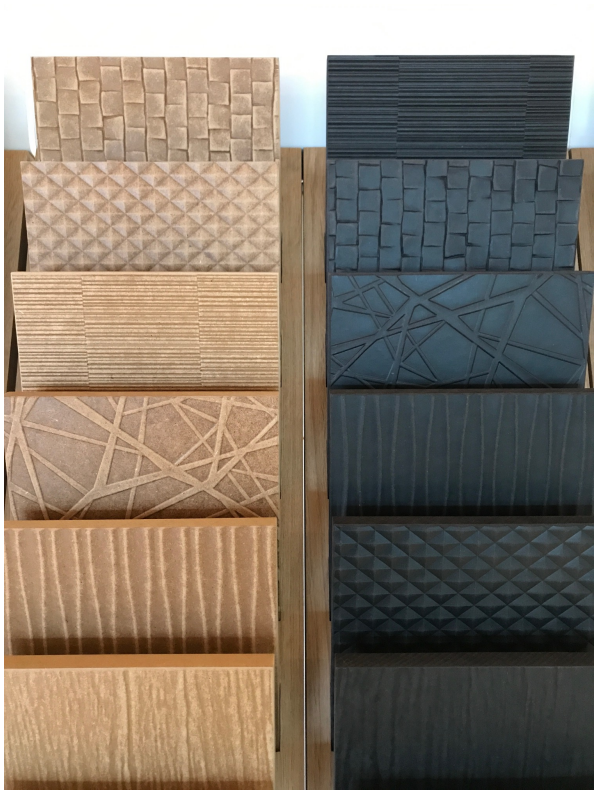
(a) FDM tamaño habitual (b) FDM grueso

Figura A.1: Tablero FDM

(a) Tablero de aglomerado

(b) Tablero Superpan

(a) Posibles colores para los tableros



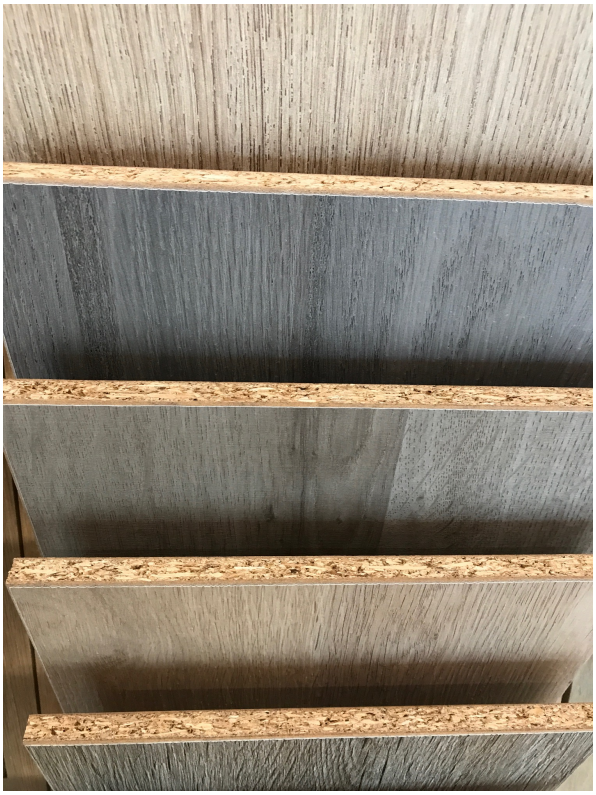
(b) Relieve en los tableros por presión

Figura A.2: Diferentes estilos de melamina: madera, piedra, color, etc

(a) Melaminas con diferentes acabados

(b) Acabado melamina madera

(a) Tablero con chapa



(b) Grosor de la chapa

Apéndice B

Demostraciones

Demostraciones premisas PCA para la base de datos original

```
> for (i in 2:18){
+   LB= Box.test(Vent[,i], type= "Ljung-Box")
+   p_value= c(p_value, LB$p.value)
+   if (LB$p.value < 0.05){
+     print(paste("Se rechaza H0 para la variable", names(Vent)[i] , ", presenta
autocorrelación" ))
+     i+1
+   }
+ }
[1] "Se rechaza H0 para la variable T1 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable T2 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable T3 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable T4 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable consumo , presenta autocorrelación"
[1] "Se rechaza H0 para la variable X1 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable X2 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable X3 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable X4 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Y1 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Y2 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Y3 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Y4 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Z1 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Z2 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Z3 , presenta autocorrelación"
[1] "Se rechaza H0 para la variable Z4 , presenta autocorrelación"
```

Figura B.1: BBDD Original. Demostración autocorrelación. Se acepta dependencia

```
Call:
princomp(x = Vent[2:18], cor = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9
1.9275213 1.6234360 1.3470404 1.2452843 1.1988950 1.0237000 1.0103037 0.9740415 0.9031981
  Comp.10   Comp.11   Comp.12   Comp.13   Comp.14   Comp.15   Comp.16   Comp.17
0.7559297 0.6011188 0.5650257 0.5208130 0.4815537 0.4250352 0.2433087 0.1352639

 17  variables and  4519632 observations.
```

Figura B.2: BBDD Original. Demostración linealidad, se rechaza

Demostración efectos ARCH en todas las componentes seleccionadas en la BBDD de los datos transformados realizando la media cada 75 minutos y aplicándo el PCA normal.

```
> summary(resC1) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 1] ~ L(resVent75[, 1]), data = vent75)

Residuals:
      Min        1Q    Median        3Q       Max
-6.493e-14  3.000e-17  6.600e-17  9.200e-17  2.711e-15

Coefficients:
                   Estimate Std. Error    t value Pr(>|t|)
(Intercept)       -7.985e-16  7.082e-17 -1.128e+01   <2e-16 ***
L(resVent75[, 1])  1.000e+00  2.306e-17  4.337e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.101e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.881e+33 on 1 and 968 DF,  p-value: < 2.2e-16
```

(a) Efecto ARCH componente 1

```
> summary(resC2) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 2] ~ L(resVent75[, 2]), data = vent75)

Residuals:
      Min        1Q    Median        3Q       Max
-6.290e-15 -3.300e-16 -2.200e-16 -9.100e-17  1.794e-13

Coefficients:
                   Estimate Std. Error   t value Pr(>|t|)
(Intercept)       4.563e-15  2.283e-16 1.999e+01   <2e-16 ***
L(resVent75[, 2]) 1.000e+00  3.400e-17 2.941e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.785e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 8.649e+32 on 1 and 968 DF,  p-value: < 2.2e-16
```

(b) Efecto ARCH componente 2

```
> summary(resC3) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 3] ~ L(resVent75[, 3]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-1.321e-15 -4.400e-17 -3.400e-17 -2.400e-17  3.334e-14

Coefficients:
                   Estimate Std. Error   t value Pr(>|t|)
(Intercept)       5.133e-16  3.779e-17 1.359e+01   <2e-16 ***
L(resVent75[, 3]) 1.000e+00  3.686e-17 2.713e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.075e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.361e+32 on 1 and 968 DF,  p-value: < 2.2e-16
```

(a) Efecto ARCH componente 3

```
> summary(resC4) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 4] ~ L(resVent75[, 4]), data = vent75)

Residuals:
      Min        1Q    Median        3Q       Max
-4.472e-14  3.400e-17  5.100e-17  6.800e-17  1.488e-15

Coefficients:
                   Estimate Std. Error    t value Pr(>|t|)
(Intercept)       -9.126e-16  5.472e-17 -1.668e+01   <2e-16 ***
L(resVent75[, 4])  1.000e+00  5.112e-17  1.956e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.441e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.826e+32 on 1 and 968 DF,  p-value: < 2.2e-16
```

(b) Efecto ARCH componente 4

```
> summary(resC5) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 5] ~ L(resVent75[, 5]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-2.574e-13  2.260e-16  2.920e-16  3.710e-16  8.363e-15

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)      -6.160e-15  3.278e-16 -1.879e+01   <2e-16 ***
L(resVent75[, 5])  1.000e+00  1.323e-16  7.557e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.285e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 5.711e+31 on 1 and 968 DF,  p-value: < 2.2e-16
```

(a) Efecto ARCH componente 5

```
> summary(resC6) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 6] ~ L(resVent75[, 6]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-2.390e-15 -1.160e-16 -8.100e-17 -4.800e-17  7.476e-14

Coefficients:
                  Estimate Std. Error   t value Pr(>|t|)
(Intercept)       1.711e-15  9.209e-17 1.858e+01   <2e-16 ***
L(resVent75[, 6]) 1.000e+00  5.087e-17 1.966e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.408e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.865e+32 on 1 and 968 DF,  p-value: < 2.2e-16
```

(b) Efecto ARCH componente 6

```
> summary(resC7) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 7] ~ L(resVent75[, 7]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-1.235e-15 -5.620e-17 -2.760e-17  5.930e-17  3.266e-15

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)      -2.281e-16  6.741e-18 -3.384e+01   <2e-16 ***
L(resVent75[, 7])  1.000e+00  2.935e-18  3.407e+17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.807e-16 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.161e+35 on 1 and 968 DF,  p-value: < 2.2e-16
```

(a) Efecto ARCH componente 7

```
> summary(resC8) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 8] ~ L(resVent75[, 8]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-2.079e-14 -1.000e-17  2.100e-17  4.970e-17  7.022e-16

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)      -4.563e-16  2.593e-17 -1.759e+01   <2e-16 ***
L(resVent75[, 8])  1.000e+00  1.565e-17  6.388e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.74e-16 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.081e+33 on 1 and 968 DF,  p-value: < 2.2e-16
```

(b) Efecto ARCH componente 8

```
> summary(resC9) #Hay efectos ARCH

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 9] ~ L(resVent75[, 9]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-6.601e-14  4.600e-17  7.400e-17  1.020e-16  2.104e-15

Coefficients:
                  Estimate Std. Error    t value Pr(>|t|)
(Intercept)      -1.369e-15  8.152e-17 -1.679e+01   <2e-16 ***
L(resVent75[, 9])  1.000e+00  5.247e-17  1.906e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.126e-15 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.633e+32 on 1 and 968 DF,  p-value: < 2.2e-16
```

(a) Efecto ARCH componente 9

```
> summary(resC10) #Hay efectos ARCH, intercepto no significativo

Time series regression with "numeric" data:
Start = 1, End = 970

Call:
dynlm(formula = resVent75[, 10] ~ L(resVent75[, 10]), data = vent75)

Residuals:
       Min         1Q     Median         3Q        Max
-4.433e-15 -4.300e-18  8.700e-18  1.780e-17  6.880e-16

Coefficients:
                   Estimate Std. Error   t value Pr(>|t|)
(Intercept)        0.000e+00  5.894e-18 0.000e+00        1
L(resVent75[, 10]) 1.000e+00  5.671e-18 1.764e+17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.489e-16 on 968 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.11e+34 on 1 and 968 DF,  p-value: < 2.2e-16
```

(b) Efecto ARCH componente 10

Apéndice C

Código ALSO

Código de RStudio con el que se ha realizado el algoritmo ALSO, el código se ha visto en Amat (2020)

```
################################################################################
#####################                ALSO            ###########################
#--------------------------------------------------------------------------------
###############################       RMS30        ###############################
################################################################################
################################################################################
#Attribute wise learning for scoring outliers (ALSO) es un algoritmo no supervisado
#para la deteccion de anomalias. Para cada variable disponible en el set de datos,
#se entrena un modelo de regresion que predice dicho atributo en funcion del resto
#de variables. El grado de anomalia de cada observacion se obtiene a partir el
#agregado del error cuadratico de los modelos al tratar de predecirla. De esta
#forma, ALSO reformula el problema, inicialmente no supervisado, en una combinacion
#de problemas supervisados, permitiendo asi hacer uso de cualquiera de los
#algoritmos supervisados disponibles (Amat, 2020)

###ALSO CON RANDOM FOREST
detectar_anomalias_also <- function(datos, standarize=TRUE, verbose=TRUE) {

  # Escalado de las variables
  # ------------------------------------------------------------------------------
  if (standarize) {
    datos <- purrr::map_if(.x = datos, .f = scale, .p = is.numeric) %>%
      as.data.frame()
  }

  # Identificacion de variables numericas
  # ------------------------------------------------------------------------------
  numeric_columns <- which(purrr::map_lgl(.x = datos, .f = is.numeric))

  # Iteracion para predecir cada variable numerica
  # ------------------------------------------------------------------------------
  mat_errores <- matrix(
    data = NA,
    ncol = length(numeric_columns),
    nrow = nrow(datos)
  )

  for (i in seq_along(numeric_columns)) {

    columna <- numeric_columns[i]

    # Modelo random forest
    # ------------------------------------------------------------------------------
```

```
    modelo_k <- ranger::ranger(
      x = datos[-columna],
      y = datos[[columna]],
      num.trees = 1000,
      max.depth = 4,
      oob.error = TRUE
    )

    # Predicciones out of bag
    # ------------------------------------------------------------------------
    predicciones <- modelo_k$predictions

    # Peso del modelo
    # ------------------------------------------------------------------------
    # El modelo ranger almacena el out of bag mean squared error
    rmse <- sqrt(modelo_k$prediction.error)
    peso_modelo <- 1 - min(1, rmse)

    # Error
    # ------------------------------------------------------------------------
    error <- (datos[[columna]] - predicciones)^2

    # Se pondera por el peso del modelo
    error <- peso_modelo * error
    mat_errores[, i] <- error

    if (verbose) {
      cat(
        paste(
          "Modelo", colnames(datos)[columna], "-->",
          "Peso:", round(peso_modelo, 4), "\n"
        )
      )
    }
  }

  # Agregado del error de todos los modelos
  # ----------------------------------------------------------------------------
  score_anomalia <- apply(mat_errores, 1, sum)

  return(score_anomalia)
}

Vent_RMS30$anomalia= as.character(Vent_RMS30$anomalia)

#Prediccion (lo que aporta cada variable)
score_anomalia <- detectar_anomalias_also(
  datos = Vent_RMS30[,2:18],
  standarize = TRUE
)

#Deteccion de anomalias
Vent_RMS30$score_anomalia <- score_anomalia

#Se utiliza la puntuacion (Score) como criterio para detectar anomalias
dev.off()
ggplot(data = Vent_RMS30, aes(x = score_anomalia)) + geom_histogram() +
  labs(title = "Distribuci n de la puntuaci n de anomal a ALSO",
       x = "clasificaci n (0 = normal, 1 = anomal a)") +
  theme_bw()

ggplot(data = Vent_RMS30,
       aes(x = anomalia, y = log(score_anomalia))) +
  geom_jitter(aes(color = anomalia), width = 0.03, alpha = 0.3) +
  geom_violin(alpha = 0) +
```

```
  geom_boxplot(width = 0.2, outlier.shape = NA, alpha = 0) +
  stat_summary(fun = "mean", colour = "orangered2", size = 3, geom = "point") +
  labs(title = "Puntuaci n anomal as ALSO",
       x = "clasificaci n (0 = normal, 1 = anomal a)",
       y = "Score anomal a") +
  theme_bw() +
  theme(legend.position = "none")
#El valor promedio en el grupo de anomalias es mayor al promedio en el grupo de no
    anomalias
#Pero no todas las anomalias tienen altas puntuaciones, existe solapamiento, si se
    clasifican
#las n observaciones con mayor score como anomalias, se incurriria en errores de
    falsos positivos.

resultados <- Vent_RMS30[,19:20]%>%
  arrange(desc(score_anomalia)) %>%
  mutate(clasificacion = if_else(row_number() <= 40, "1", "0"))

mat_confusion <- MLmetrics::ConfusionMatrix(
  y_pred = resultados$clasificacion,
  y_true = resultados$anomalia
)
mat_confusion
```

Código C.1: Código de Amat (2020) para realizar el algoritmo ALSO con RF