Universidad de Oviedo

ESCUELA POLITÉCNICA DE INGENIERÍA DE GIJÓN.

MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

ÁREA DE MATEMÁTICA APLICADA

TRABAJO FIN DE MÁSTER

Caracterización automática de las averías en escaleras mecánicas en el área Asia-Pacífico

D. CABEZAS RODRÍGUEZ, RAÚL
TUTOR: Dña. SAN LUIS FERNÁNDEZ, ANA Mª
D. CARLEOS ARTIME, CARLOS

FECHA: junio de 2019

Contenido

1.- Hipótesis de partida y alcance

Una de las fuentes de información más valiosas para comprender el funcionamiento de máquinas y equipos es el análisis de sus fallos y el consiguiente proceso de reparación. Este estudio permite mejorar el diseño, la fabricación y el mantenimiento de estos equipos. En la mayoría de los casos, estas reparaciones —también llamadas medidas de mantenimiento correctivo— solo quedan registradas mediante informes escritos por el propio personal de mantenimiento. El nivel de detalle con el que se realizan estos informes permite clasificarlos según dos grandes patrones que presentan tanto ventajas como inconvenientes.

En primer lugar, se consideran los informes en los que se seleccionan opciones cerradas para cada uno de los campos de interés del estudio: causa de fallo, elemento de fallo y medidas tomadas. Esta configuración presenta la ventaja de producir datos estructurados, los cuales son fáciles de procesar con los métodos actuales. Sin embargo, al restringir demasiado la respuesta, no se pueden reflejar todos los valores posibles de cada campo. Esto supone que alternativas muy diversas acaben en una categoría generalista “Otros” que en ocasiones se convierte en una amalgama de difícil aprovechamiento. Además, se pierden los matices y la subjetividad que aporta la valoración libre del operario.

El otro método para elaborar informes de reparación consiste en emplear campos con respuesta abierta en los que el trabajador describe toda la actuación llevada a cabo. En este caso se dispone de una gran cantidad de información, pero está desestructurada y es difícil de aprovechar con los sistemas de gestión de datos actuales.

Normalmente se utiliza una combinación de ambos patrones: se emplean opciones cerradas para aquellas características más estandarizadas (modelo, características mecánicas particulares, momento del fallo) y una descripción más amplia cuando son interesantes los matices u otro tipo de información (descripción del fallo, medidas tomadas).

Esta variabilidad de los informes hace que, en muchas ocasiones, estos documentos queden solo como testigo y elemento de trazabilidad del mantenimiento sin que se explote todo su potencial.

La aplicación de técnicas de análisis de datos sobre los reportes puede permitir detectar tendencias relacionadas con el desempeño de equipos y operarios, tanto en planta, durante su construcción, como en campo durante su uso, que pueden servir para mejorar la toma de decisiones y aumentar la calidad, productividad y eficiencia de la empresa. Por este motivo, se considera que puede ser muy beneficioso conseguir estructurar la información contenida en dichos informes, así como aplicar diversos métodos estadísticos al campo del mantenimiento de equipos.

2.- Objetivos

En el presente Trabajo Fin de Máster (TFM) se realiza el estudio de los informes de fallo de una muestra muy amplia de escaleras mecánicas y pasillos rodantes en el área Asia-Pacífico cuyo mantenimiento corre a cargo de la empresa Thyssenkrupp.

Los principales objetivos de este TFM son:

- Aprender nociones sobre distintas herramientas y lenguajes de programación de análisis de datos, así como técnicas estadísticas para trabajar con dichos datos.
- Creación de una base de datos homogénea a partir de los datos desestructurados de los informes de reparación de los operarios de mantenimiento.
- Detectar tendencias que relacionen los parámetros del mantenimiento correctivo con los distintos productos, mercados y fábricas de escaleras y pasillos rodantes, con el fin de ayudar a mejorar la toma de decisiones a nivel de gestión.
- Generar modelos predictivos empleando datos conocidos de los equipos y los fallos para mejorar la planificación de los programas actuales de mantenimiento.

Para alcanzar estos objetivos se han llevado a cabo los siguientes pasos:

- Estudio de la estructura de los informes de fallo, así como las características definitorias de cada modelo instalado y la idiosincrasia propia de cada mercado nacional que determinan y ayudan a explicar la naturaleza de los datos.
- Programación de distintos algoritmos de limpieza, procesamiento del lenguaje natural de los informes y clasificación de los datos brutos disponibles.
- Presentación de los resultados y conclusiones.

3.- Antecedentes

3.1.- IMPORTANCIA DEL MANTENIMIENTO

De acuerdo con la federación europea de asociaciones nacionales de mantenimiento (EFNMS) [1], se define el mantenimiento como el conjunto de acciones técnicas, administrativas y de gerencia cuyo objetivo principal es conservar un elemento durante todo su ciclo de vida en buen estado o repararlo para evitar su degradación y que pueda seguir cumpliendo las funciones requeridas.

El mantenimiento es una actividad fundamental en todos los sectores económicos. Un documento elaborado por la iniciativa MORE4CORE en 2015 [2] se cifra en 1200 millones €/año el gasto europeo en mantenimiento —incluyendo el doméstico— y se calcula en más de 6 millones el número de trabajadores que se dedican en exclusiva al mantenimiento de equipos en el continente. Más importante, si cabe, es el valor de los daños, en tiempo de parada y coste de reposición de equipos, evitado gracias al mantenimiento y que esta misma organización cifra en más de 10000 millones €/año solo en la UE [3].

Adicionalmente, hay que tener en cuenta la disminución del impacto ambiental que implica un buen mantenimiento al alargar la vida útil de los equipos, y evitar así el derroche de recursos y la sobreproducción [4].

Si se centra el análisis en el sector de las escaleras mecánicas y pasillos rodantes, se encuentran una serie de requisitos que exigen los clientes y que solo se pueden conseguir gracias a un mantenimiento adecuado [5]. Debido a su condición de servicio de cara al público (metros, aeropuertos, centros comerciales) un adecuado programa de mantenimiento supone las siguientes ventajas:

- Se reduce el tiempo de parada, con lo que aumenta la disponibilidad del equipo y mejora la impresión que se lleva el usuario del establecimiento.
- Disminuye el riesgo de accidentes y evita aglomeraciones de personas en hora punta, que podrían crear situaciones peligrosas.
- Mejora la accesibilidad para las personas con movilidad reducida y facilita la evacuación en caso necesario.

3.2.- PERSPECTIVA DE LA CIENCIA DE DATOS

Se define como Ciencia de Datos el empleo de técnicas propias de la estadística y la informática, aplicadas a grandes cantidades de datos, para entender su estructura y obtener información relevante a partir de ellos [6]. Es decir, la ciencia de datos es la intersección de distintas disciplinas tal y como se puede ver en figura- 3.1.

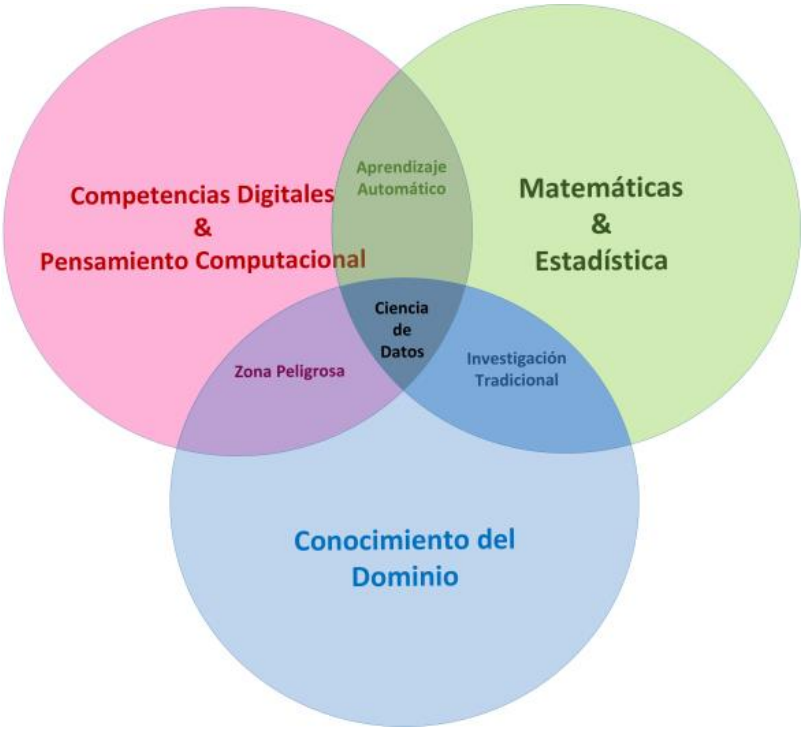


Figura- 3.1 Diagrama de Venn de la ciencia de datos

Este campo ha cobrado especial relevancia en los últimos años gracias a dos tendencias globales: por un lado, el aumento exponencial de la capacidad de cálculo de los ordenadores y el uso de la computación en la nube que ha abaratado enormemente el hardware necesario. Por otro lado, el uso masivo de redes sociales, la digitalización de bases de datos de las empresas y la irrupción del internet de las cosas han favorecido el almacenamiento de ingentes cantidades de datos (*Big Data*) fácilmente disponibles para su análisis [7].

Un informe del McKinsey Global Institute [8] cifra en un ahorro de hasta el 25% en los costes de operación y mantenimiento, y una reducción de hasta un 50% en los costes de desarrollo de producto en el sector manufacturero si se aprovecha el potencial de estas nuevas tecnologías. Ese mismo documento también afirma que existen múltiples barreras que han impedido hasta la fecha el despliegue de estas técnicas. Entre dichos obstáculos destacan el escepticismo de los directivos respecto al impacto de estas tecnologías, la ausencia de un sistema interconectado que aproveche todos los datos que genera cada empresa y la falta

de personal especializado: se estima que el 50% de los puestos de científico de datos generados en 2020 no tendrán un candidato adecuado para ocuparlos.

Por tanto, se puede ver que ya existe una tecnología madura y un conocimiento técnico suficiente para aplicar estas técnicas en el sector del mantenimiento y son, principalmente, las trabas organizativas las que han impedido hasta la fecha su correcto desarrollo.

3.3.- THYSSENKRUPP ELEVATOR TECHNOLOGY

3.3.1.- Historia de la empresa y el sector

Thyssenkrupp es un conglomerado alemán fruto de la fusión en 1999 de dos compañías del sector del acero, Thyssen AG, creada en 1867 y Krupp en 1811, ambas creadas cerca de Essen, en Alemania.

Actualmente se halla en un proceso de división en dos compañías independientes, cada una centrada en un sector. Por un lado, Thyssenkrupp Materiales, con las áreas de negocio de *Steel Europe*, centrada en la producción de acero y productos derivados, y *Material Services*, en el que se incluyen las actividades de minería, gases industriales, otros metales y servicios energéticos. Por otro lado, Thyssenkrupp Bienes Industriales, compuesta a su vez por tres áreas de negocio:

Components Technology: su actividad se desarrolla en toda la cadena de valor, desde el diseño hasta la fabricación de piezas para dos sectores principales:

- En el sector del automóvil, donde realiza tareas que van desde el diseño hasta la fabricación de sistemas de suspensión, el mecanizado de árboles de levas, cigüeñales o el ensamblaje de chasis.
- En el sector de la maquinaria, donde diseña y fabrica componentes para equipos de construcción o turbinas eólicas entre otros.

Industrial Solutions: su trabajo se centra en el diseño, construcción y gestión de todo tipo de plantas industriales, desde complejos mineros hasta petroquímicos. También tiene una importante participación en el sector naval, siendo uno de los principales suministradores globales de submarinos y buques militares.

Elevator Technology: se ocupa del diseño, fabricación, instalación y mantenimiento de sistemas de transporte urbano de pasajeros. Esto incluye: ascensores de pasajeros, montacargas, escaleras mecánicas, pasillos rodantes, pasarelas de embarque a aeronaves y plataformas salvaescaleras

Este trabajo se enmarca en esta última subdivisión de la compañía, que además es una de las que más crecimiento está experimentando debido a la evolución demográfica global. De

acuerdo con la ONU [9], la población urbana se incrementará del 50% actual a más de un 68% en 2050 y el número de megaciudades (con más de 10 millones de habitantes) pasará de las 33 actuales a casi 50. Por otro lado, la población mundial está experimentando un envejecimiento acelerado [10] y se espera que la población con más de 60 años pase del 10% actual a representar casi un 25% a nivel global, superando el 30% en regiones desarrolladas como Norteamérica o Europa.

En este panorama de concentración de la población en las ciudades y de envejecimiento generalizado, se harán más necesarios que nunca sistemas de transporte ecológicos, accesibles, y que permitan mover gran volumen de pasajeros. Por ello, se espera que en el periodo 2017-2020 el sector de ascensores, escaleras mecánicas y pasillos rodantes registre un crecimiento anual cercano al 5%, alcanzando unas ventas, solo en China, de más de medio millón de equipos al año [11]. Además, el mantenimiento de todos estos equipos representará un mercado de 47200 millones de dólares en este país asiático.

3.3.2.- Productos

Este trabajo se centra, dentro de los productos desarrollados por Thyssenkrupp Elevator Technologies, en las escaleras mecánicas y pasillos rodantes. Desde que el primer pasillo rodante se mostrara en 1893 en la Exposición Mundial de Chicago y de que la primera escalera mecánica fuera instalada en 1895 por Jesse Reno en Nueva York, su tecnología ha mejorado y ha ido incluyendo nuevos elementos para hacer más eficiente y seguro su funcionamiento. Sin embargo, su estructura y principio de acción no ha variado desde su invención a finales del siglo XIX.

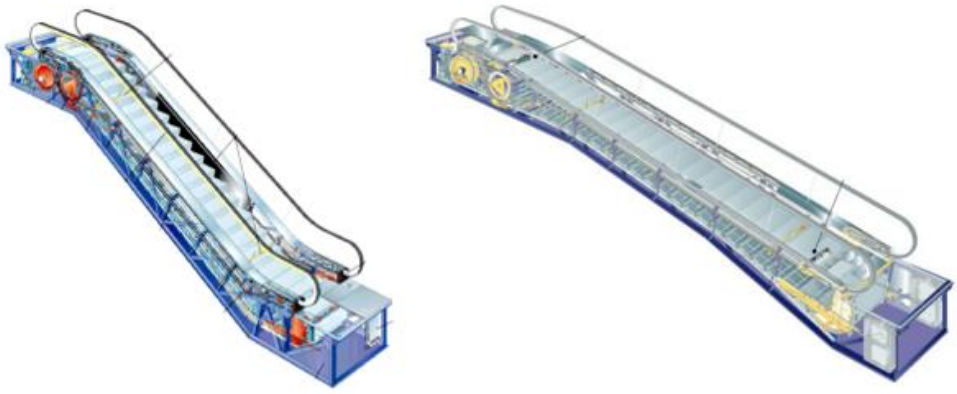


Figura- 3.2 Escalera mecánica y pasillo rodante de Thyssenkrupp

Tanto las escaleras mecánicas como los pasillos rodantes se basan en el movimiento continuo de una serie de peldaños cuyo fin último consiste en el transporte de personas de un lugar a otro. Ambos equipos se componen de las siguientes partes principales [12]:

Estructura portante: bastidor soldado con perfiles de acero y unido a los laterales por medio de travesaños y chapa soldada:

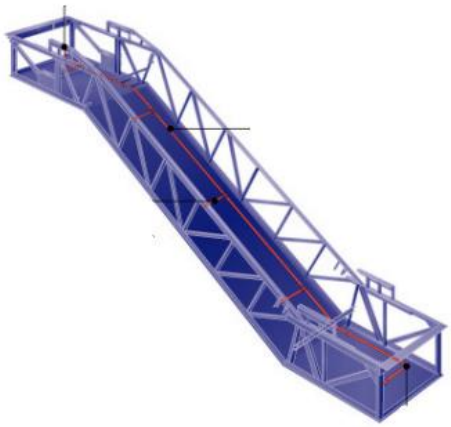


Figura- 3.3 Estructura portante

Sirve de elemento de soporte del resto de componentes y va apoyado en ambos extremos al suelo del edificio mediante tornillos ajustables y un sistema de amortiguación para evitar vibraciones.

Conjunto motor: motor de corriente alterna con freno electromagnético de zapata que transmite el movimiento mediante una caja reductora de engranajes y una cadena doble al eje principal. Este eje transmite potencia tanto a las cadenas de rodillos de peldaños (a ambos lados del eje principal) como a la cadena de accionamiento del pasamanos. Siempre se sitúa en la cabeza superior del equipo.

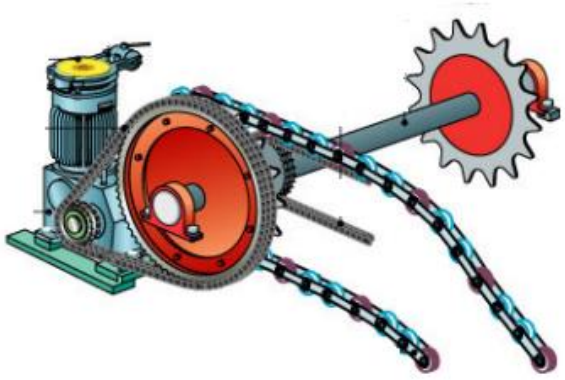


Figura- 3.4 Conjunto motor

Guías y estación tensora: estos componentes facilitan el movimiento suave y sincronizado de todos los peldaños. Las guías son perfiles de chapa planos o con forma de U que actúa de carril para que los rodillos de los peldaños se muevan alineados y que la cadena no se desenganche. La estación tensora se compone de los engranajes que permiten el retorno de la cadena de peldaños, que van situados sobre dos guías móviles con dos muelles de

compresión de tal manera que empujan el engranaje y con él la cadena para que mantenga siempre su tensión constante.

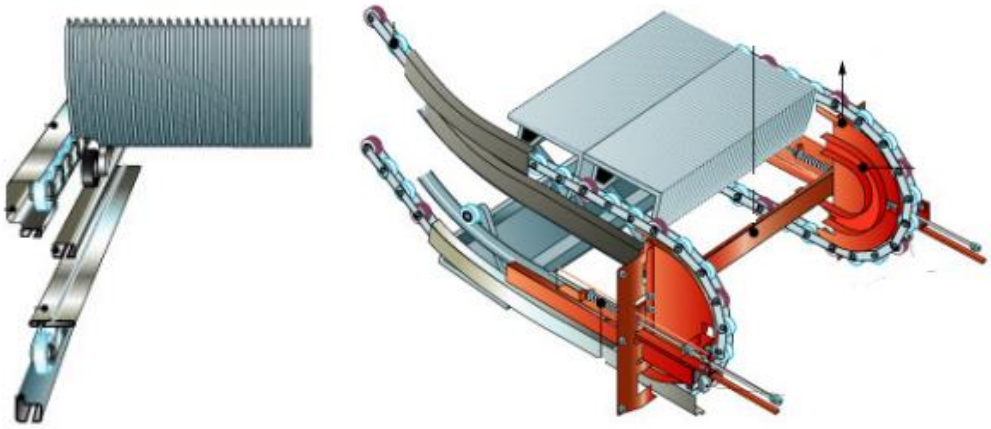


Figura- 3.5 Guías (izquierda) y estación tensora de retorno (derecha)

Peldaños: elementos fabricados por fundición de aluminio, cuentan con nervaduras y ranuras en su superficie para mejorar el agarre y evitar resbalones. Tienen en su parte inferior dos pares de rodillos de poliuretano con un cojinete en su interior. Los superiores van unidos por medio de bulones a la cadena de peldaños y son los que transmiten el movimiento del motor. Por el contrario, los inferiores giran arrastrados por los primeros y, gracias a la distancia entre la guía de unos y otros, se consigue dar la forma adecuada al peldaño y que no queden huecos entre un peldaño y el siguiente.

Figura- 3.6 Peldaño con línea de demarcación amarilla

Balaustrada: elemento lateral situado a ambos lados de los peldaños, cuya misión principal es servir de apoyo al pasamanos para evitar caídas por fuera de la escalera. Puede ser de cristal o metal y en su parte inferior tiene un zócalo de chapa de acero galvanizada. Lleva unido un cepillo y el conjunto tiene como misión evitar la entrada de cuerpos extraños entre peldaño y balaustrada.

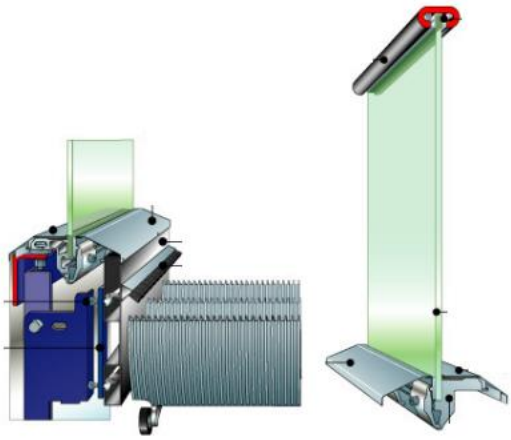


Figura- 3.7 Balaustrada con faldón de protección

Pasamanos: elemento formado por varias capas textiles, con cubierta de goma negra, que se mueve en sincronismo con los peldaños para facilitar a los usuarios la sujeción y estabilidad sobre la escalera. Se mueve sobre unas guías de chapa de acero en la parte visible por los usuarios de la escalera y mediante rodillos en el retorno. Recibe la potencia por medio de una polea movida gracias a la cadena que conecta directamente con el eje principal. Una correa ajustable presiona al pasamanos contra la polea para garantizar un movimiento adecuado.

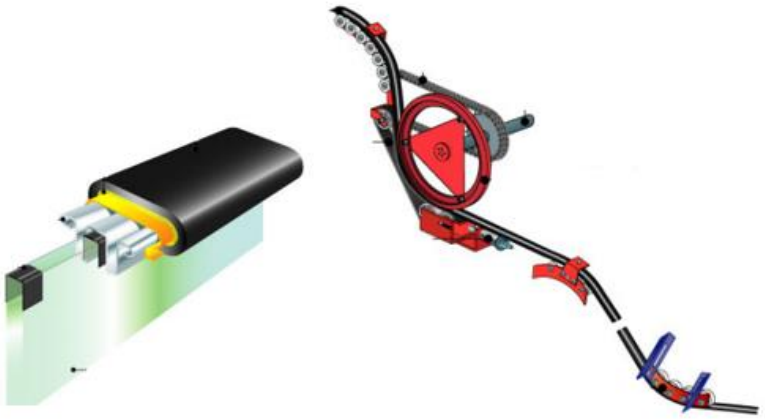


Figura- 3.8 Pasamanos (izquierda) y mecanismo de movimiento del pasamanos (derecha)

Plataformas de llegada: perfiles de aluminio fundido con ranuras para que la superficie sea antideslizante. Tienen como fin servir de tapa para todo el mecanismo interior e impedir que queden atrapados cuerpos extraños en el retorno de los peldaños.

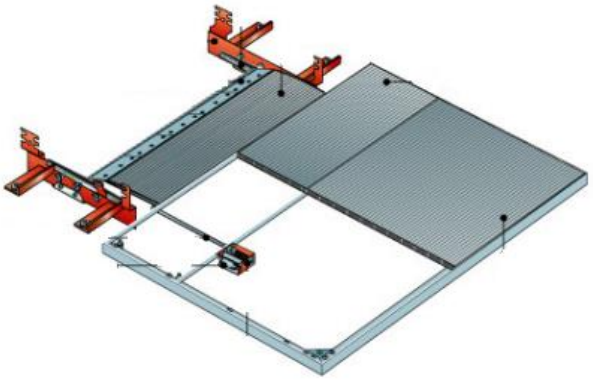


Figura- 3.9 Plataforma de llegada

Control: armario eléctrico cerrado situado en la cabeza superior de la escalera. Contiene los contactores principales, las protecciones del motor, variadores de frecuencia, el cableado de la línea de seguridad y otros elementos de control (relés, plc).

Debido a su posible peligrosidad y amplio uso en todos los países estas máquinas se rigen principalmente por las normas ASTM A 17.1-2012 y EN115-1:2008+A1:2010 que establecen una serie de seguridades mínimas, pudiendo existir otras adicionales, para garantizar un uso adecuado de los equipos y minimizar el riesgo de accidente. Entre estas seguridades cabe destacar:

- Pulsadores de emergencia en ambas plataformas de llegada.
- Protector de entrada del pasamanos, que impide que un objeto quede atrapado entre balaustrada y pasamanos.
- La monitorización del estado de la cadena principal, control de la velocidad del pasamanos para garantizar el sincronismo, medida del estado de los frenos y de la velocidad del motor.
- Freno doble de emergencia.
- Seguridad de placa de peines para evitar objetos atrapados entre peldaño y plataforma de llegada.
- Dispositivos de seguridad de peldaños, que avisan en caso de daño en el peldaño o huecos entre peldaños.
- Dispositivos de tensión de la cadena de peldaños.
- Interruptores de apertura de los pozos de las plataformas de llegada.

4.- Estado del arte

4.1.- MANTENIMIENTO

Una vez analizada la importancia del mantenimiento cabe preguntarse cómo debe ser ese mantenimiento para lograr dar un servicio de calidad, maximizando la seguridad de uso y minimizando el tiempo que no está disponible la escalera mecánica o el pasillo.

4.1.1.- Modalidades del mantenimiento

Para ello, en primer lugar, hay que determinar qué tipos de mantenimiento existen. En función del momento en que se realiza esta tarea se puede clasificar el mantenimiento en [13]:

- **Mantenimiento correctivo:** es el conjunto de acciones de sustitución y reparación de elementos deteriorados que se lleva a cabo cuando aparece un fallo en dicho elemento.
  Este tipo de mantenimiento siempre se intenta evitar, ya que el fallo puede producirse en un momento inoportuno. Además, la rotura de un elemento puede provocar daños en otros componentes aumentando el tiempo de reparación y el coste de los recambios.
  El principal método para disminuir el uso del mantenimiento correctivo es lo que se conoce como mantenimiento modificativo: aumentar la calidad de los componentes más difíciles de cambiar o tener en cuenta durante el diseño qué pieza es más probable que falle, modificando su construcción de tal manera que su sustitución sea sencilla y no dañe otras partes, reduciendo con ello la duración del mantenimiento correctivo.
  Otra técnica habitual consiste en el mantenimiento de oportunidad: aprovechar que se ha producido el fallo para realizar también una revisión de otros elementos. Esto es habitual en equipos de funcionamiento continuo y con alto coste de parada.
- **Mantenimiento preventivo:** es el conjunto de acciones programadas con antelación con el objetivo de reducir la frecuencia y el impacto de los fallos.
  Este es el tipo de mantenimiento más habitual, puesto que permite revisar todos los elementos de la máquina de una sola visita en lugar de esperar a que fallen, y establecer una planificación para las inspecciones, de manera que puedan realizarse en los momentos en los que interfieran menos con la actividad del equipo (por la noche o en fin de semana en el caso de las escaleras mecánicas).
  Sin embargo, este tipo de mantenimiento suele tener asociado un alto coste por varios motivos. En primer lugar, se basa en hacer visitas periódicas, con mayor

frecuencia cuanto mayor sea la disponibilidad deseada del equipo. Esto implica una utilización intensiva de mano de obra, ya que en la mayor parte de visitas solo se comprobará que el equipo funciona correctamente sin identificar ningún fallo y, por tanto, no serían realmente necesarias. En segundo lugar, suele implicar un alto uso de repuestos, al cambiar por precaución piezas cuya vida útil aún podría alargarse. Finalmente, este uso de repuestos suele obligar a tener grandes stocks de piezas en almacenes cerca de los distintos lugares donde están instalados los equipos, lo que también conlleva un coste a tener en cuenta.

El documento que define como se va a llevar a cabo el mantenimiento preventivo es el plan de mantenimiento, En él se recogen las tareas a realizar en cada visita y la política de repuestos. Existen dos estrategias principales, el cambio a intervalo fijo —en el que independientemente del estado del elemento se sustituye tras ciertas horas de funcionamiento— y el cambio según condición, en el que, mediante algún análisis *in situ* o por la inspección visual del operario, se decide si la pieza necesita ser sustituida o puede continuar funcionando.

- **Mantenimiento predictivo:** es el conjunto de acciones basadas en el diagnóstico continuo de los parámetros del equipo con el fin de detectar síntomas de fallo antes de que se produzcan daños y así poder llevar a cabo medidas correctivas menos invasivas.

  Este tipo de mantenimiento permite realizar solo intervenciones cuando sea necesario, reduciendo así las paradas del mantenimiento preventivo lo que aumenta la disponibilidad del equipo y reduce la mano de obra necesaria. Además, permite tener un seguimiento en tiempo real del estado de las piezas críticas, pudiendo anticipar cuándo se va a producir un fallo y así pedir las piezas que se vayan a necesitar para que lleguen en el momento adecuado, reduciendo también los stocks en los almacenes.

  Todos los métodos de mantenimiento predictivo se basan en obtener datos de funcionamiento de los equipos bajo supervisión [14]. Para ello, es necesario contar con sensores que monitoricen las variables de interés. Estos sensores dependen del componente de estudio; algunos de los más comunes son los acelerómetros y sensores piezoeléctricos para medir vibraciones, o los sensores térmicos para detectar calentamientos por desgaste, rozamientos indebidos o fallos en componentes eléctricos.

  A partir de los datos recopilados se aplican distintos métodos matemáticos (redes neuronales, arboles de decisión, sistemas estadísticos) que permiten predecir cuánto tiempo de funcionamiento queda antes de que se produzca un fallo y cuándo se recomienda realizar el recambio de piezas.

En general, se combinan siempre los 3 tipos de mantenimiento anteriormente expuestos. En la figura- 4.1 se puede observar una comparación del coste y tiempo de reparación con cada uno de los tipos de mantenimiento. Nótese que el mantenimiento correctivo es el que implica mayor tiempo de parada debido a situaciones no planeadas, mientras que el mantenimiento preventivo es el de mayor coste anual por el uso intensivo de mano de obra y repuestos. En contraste, el mantenimiento predictivo es el que permite minimizar tanto el tiempo de parada por fallos como el coste asociado al servicio.

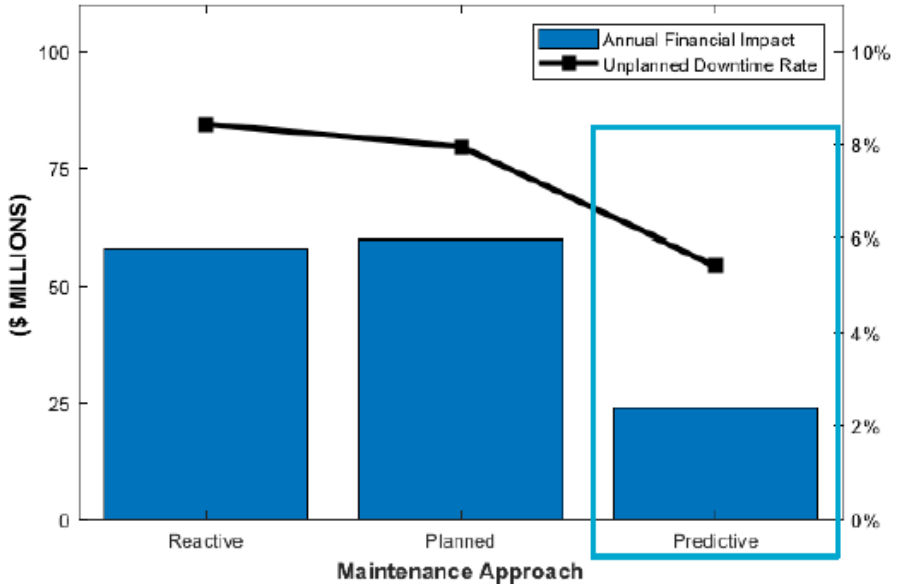


Figura- 4.1 Coste (azul) y tiempo de fallo (negro) en una instalación de GE Oil&Gas en función del tipo de mantenimiento

4.1.2.- Parámetros del mantenimiento

Después de decidir qué mantenimiento se va a llevar a cabo es conveniente establecer una serie de parámetros que permitan comprobar si las medidas tomadas tienen el efecto deseado o está lejos el objetivo establecido. Las principales variables para medir el desempeño del mantenimiento son:

- **Fiabilidad:** se entiende por fiabilidad la probabilidad *P* de que un equipo funcione de manera correcta durante un periodo específico *T* bajo condiciones operativas determinadas. Se expresa matemáticamente mediante la función de supervivencia en función del tiempo *S*(*t*) [15]:

$$S(t) = P(T > t) = 1 - F(t) \qquad (3.1)$$

Siendo *F*(*t*) la función de distribución de fallo acumulada que se define como:

$$\int_{-\infty}^{t} F(t) = f(t)dt \tag{3.2}$$

Con *f*(*t*) la función de densidad de probabilidad.

A partir de estas funciones se define la tasa de fallo λ como la probabilidad de fallo en un tiempo infinitamente pequeño, estando en funcionamiento en el instante anterior.

$$\lambda(\mathrm{t}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} * \frac{F(t+\Delta t) - F(t)}{S(t)} = \frac{f(t)}{S(t)} \tag{3.3}$$

La tasa de fallos es un concepto de gran relevancia en el mantenimiento ya que permite estimar la duración de un componente y comprobar cómo le afecta el envejecimiento. Si λ es constante, la probabilidad de fallo es independiente del tiempo, mientras que si λ es creciente indica que a mayor tiempo de funcionamiento mayor probabilidad de fallo, es decir, el componente presenta envejecimiento. El caso de λ decreciente, finalmente, se da en la vida inicial de muchos productos, e indica que hay muchas unidades defectuosas que no sobreviven a las primeras horas de funcionamiento. No obstante, si sobrevive a esa etapa inicial, la unidad no forma parte de ese grupo con taras y debería funcionar durante mucho tiempo sin presentar fallos.

Otra manera habitual de medir la fiabilidad de un equipo es mediante el tiempo medio entre fallos MTBF que se define como:

$$\mathrm{MTBF} = \frac{intervalo\ de\ tiempo\ estudiado}{número\ de\ fallos\ en\ dicho\ intervalo} = \frac{T}{n} \tag{3.4}$$

La relación entre MTBF y λ(*t*) depende de la variación temporal de la tasa de fallos. En general, si λ es constante y se disponen de todos los datos de fallo de los equipos se tiene que:

$$\mathrm{MTBF} = \frac{1}{\lambda} \tag{3.5}$$

- **Mantenibilidad:** se define como la capacidad inherente de un sistema para ser recuperado para el servicio cuando se realizan sobre él las tareas necesarias de mantenimiento. Puede hacerse un tratamiento estadístico similar al de la fiabilidad, pero es habitual caracterizarla mediante el tiempo medio de reparación MTTR:

$$\mathrm{MTTR} = \frac{tiempo\ de\ reparación\ en\ el\ intervalo\ estudiado}{número\ de\ fallos\ en\ dicho\ intervalo} = \frac{T_r}{n} \tag{3.6}$$

- **Disponibilidad:** la disponibilidad *A* es la probabilidad de que un equipo esté preparado para funcionar en el momento requerido. Se puede calcular, a partir de las medidas anteriores como:

$$A = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \tag{3.7}$$

  Esta variable permite estimar el porcentaje de tiempo que estará operativo un equipo para el cliente. Por tanto, si se conoce la disponibilidad que desea el cliente se puede establecer cómo de fiable y reparable debe ser un sistema para poder proporcionar dicho servicio.

4.2.- CIENCIA DE DATOS

En los últimos años, términos como *Big Data*, Inteligencia Artificial, *Machine Learning* o *deep Learning* se han puesto muy de moda, muchas veces sin tener muy claro que se quiere expresar con ellos; por ello, empezaremos definiendo con precisión qué significado tendrán en nuestro trabajo cada una de las siguientes expresiones:

***Big Data*:** este término tiene una doble acepción. Por un lado se refiere a los conjuntos de datos cuyo tamaño, complejidad y velocidad de crecimiento son tan grandes que son muy dificiles de almacenar, procesar y analizar de forma eficiente mediante herramientas tradicionales como las bases de datos relacionales. Por otra parte, se refiere también al conjunto de técnicas que se emplean para tratar estos conjuntos de datos, tales como la computación en la nube o la Inteligencia Artificial.

Inteligencia artificial (IA): es una disciplina del campo de la Informática que busca la creación de máquinas que puedan imitar comportamientos inteligentes propios de los humanos tales como el aprendizaje o la resolución de problemas. Una definición más precisa es la que emplean Andreas Kaplan y Michael Hainlein que define la IA como "la capacidad de un sistema para interpretar correctamente datos externos, para aprender de dichos datos y emplear esos conocimientos para lograr tareas y metas concretas a través de la adaptación flexible" [16]. Dentro de la Inteligencia Artificial se distingue entre IA débil, que es aquellas que solo puede realizar un conjunto limitado de tareas, e IA fuerte, que puede enfrentarse a gran variedad de problemas distintos. Actualmente todas las técnicas de IA que existen son de las denominadas débiles.

La Inteligencia Artificial, a su vez, se subdivide en numerosos campos; algunos de los que más desarrollo han tenido en los últimos años son los siguientes:

- **Procesamiento de Voz:** capacidad para procesar la señal sonora emitida por la voz humana, reconociendo la información que contiene y convirtiendola en texto u órdenes para una máquina. Algunos de los ejemplos más conocidos son los asistentes

de voz de los teléfonos móviles —como *Siri* de Apple— o los asistentes domésticos de los altavoces *Echo* de Amazon.

- **Procesamiento de lenguaje natural (NLP):** capacidad para establecer relaciones entre palabras y extraer la intención y el significado último de textos producidos por humanos. Algunos de los ejemplos más conocidos son los predictores y correctores de texto de los teléfonos móviles o la aplicación de traducción de Google.
- **Visión artificial:** capacidad para adquirir, procesar, analizar y comprender imágenes del mundo real y extraer información útil de dichas imágenes. Algunos de los ejemplos más conocidos que incluyen visión artificial son los sistemas de conducción autónoma como el *Autopilot* de Tesla o el sistema de compra automática de los supermercados *Whole Foods* de Amazon.
- **Robótica:** capacidad de moverse y adaptarse al entorno pudiendo realizar acciones que lo modifiquen. Algunos de los ejemplos más conocidos son los robots de rescate de Boston Dynamics o los cobots de KUKA.

Machine Learning (Aprendizaje automático): es la especialidad de la Inteligencia Artificial que busca cómo dotar a las máquinas de capacidad de procesar datos brutos y convertirlos en conocimiento útil. Se basa en técnicas (algoritmos) de aprendizaje mediante ejemplos en lugar de reglas escritas por un programador. Estas técnicas de *Machine Learning* son transversales y se aplican a todos los subcampos de la inteligencia artificial.

Se distingue, en función del tipo de aprendizaje empleado, entre las siguientes estrategias de *Machine Learning*:

- **Aprendizaje supervisado:** algoritmos que se centran en, descubrir la relación existente entre una variable de salida —lamada etiqueta— y las variables de entrada de las que depende, a partir de muchos valores conocidos de las entradas y sus correspondientes salidas. Es decir, es capaz de generalizar, a partir de casos concretos conocidos, la relación entre distintas variables. El aprendizaje supervisado es el enfoque más usado actualmente, si bien se necesitan bases de datos con suficientes ejemplos fiables para entrenar a los modelos, lo que constituye su mayor desventaja.

  Algunas de las técnicas de aprendizaje supervisado más empleadas son los árboles de clasificación, los métodos de regresión, las redes neuronales y las máquinas vector soporte.
- **Aprendizaje no supervisado**: algoritmos que se centran en buscar patrones de similitud a partir únicamente de la información de entrada, esto es, sin introducir qué información de salida se está buscando. La principal ventaja de estas técnicas es que no se necesita conocer el valor de la variable de salida, por lo que los datos son

mucho más sencillos de conseguir, siendo su mayor inconveniente el menor grado de desarrollo actual frente a los de aprendizaje supervisado
Algunos de los algoritmos más empleados son la clusterización o el análisis de componentes principales.

- **Aprendizaje por refuerzo:** algoritmos que se basan en retroalimentar la entrada con la información de salida para afinar la respuesta. En resumen, se establece un sistema de recompensas y penalizaciones para que el algoritmo (agente inteligente) trabaje por ensayo-error dentro de un entorno (problema a resolver), reforzando su comportamiento con aquellas acciones positivas que le acercan a la respuesta deseada y corrigiéndolo cuando realiza acciones negativas que lo alejan del objetivo [17]. Este enfoque es el más parecido al modo de aprendizaje natural y el que tiene más potencial para aplicaciones de robótica. Actualmente solo está desarrollado en aplicaciones muy específicas, como el ajedrez, algunos videojuegos o AlphaGo, la inteligencia artificial de Google que en 2017 consiguió vencer al campeón mundial del juego de estrategia go. Algunos de los algoritmos más empleados son el Q-Learning o el SARSA.

***Deep Learning*:** este concepto hace referencia a redes neuronales con un gran número de capas, en las que la información se codifica de manera jerárquica. De esta manera, los primeros niveles aprenden los patrones más sencillos y los niveles posteriores emplean los patrones deducidos en las capas iniciales para detectar patrones más complejos y abstractos. Por tanto, el *Deep Learning* es solo un tipo concreto de *Machine Learning*. Un ejemplo son las redes neuronales usadas en visión artificial. Las primeras capas se especializan en detectar líneas y otros patrones sencillos, mientras las siguientes emplean esas líneas detectadas para inferir figuras más complejas cómo círculos o rectángulos, y en las capas finales se combinan dichas figuras para deducir información más abstracta, como caras humanas o la forma de un perro o un gato.

4.2.1.- Aplicaciones del *Machine Learning*

El *Machine Learning* ha supuesto un gran avance en la inteligencia artificial y ha permitido dar solución a muchos problemas que se creían que una máquina nunca podría resolver. No obstante, el aprendizaje automático no realiza cualquier tarea y hay que identificar en qué casos es adecuado su uso y en cuáles es mejor emplear otras técnicas.

Como norma general, no se recomienda emplear *Machine Learning* si se puede programar mediante reglas o cálculos sencillos una receta para llegar desde los datos iniciales a la solución buscada. Este es el caso de las ciencias y gran parte de los problemas de ingeniería en los que son bien conocidas las leyes naturales que se aplican y es relativamente facil relacionar entre sí los datos disponibles. En cambio, el uso de técnicas de *Machine Learning*

es conveniente cuando el problema no se puede codificar mediante una serie de pasos sencillos.

Esto suele deberse a que la solución depende de un gran número de factores o a que existen relaciones o diferencias muy sutiles entre las entradas que cambian por completo la salida. En estos casos, aunque una persona pueda hallar fácilmente la solución (detectar si un correo es o no *spam*, identificar si hay una cara en una foto o recomendar una película a partir de otra que hayas visto), puede ser dificil explicar qué reglas ha usado para llegar a dicho resultado y además siempre hay riesgo de equivocarse.

Los algoritmos de *Machine learning* son la mejor alternativa [18] para este tipo de problemas, y también para realizar tareas repetitivas y de poco valor para las que no interesa tener empleada a una persona.

La mayoría de problemas que se quieren resolver en este TFM entran dentro de la categoría que se acaba de describir ya que, en algunos casos, existen muchas variables cuya influencia en el resultado final no es sencilla de prever (cómo afecta el modelo o el operario al tiempo de reparación o predecir cuándo va a volver a fallar un equipo) y en otros, una persona podría resolver de manera sencilla la tarea (decidir si un fallo es accidente o avería a partir de la descripción del informe, establecer qué elemento causó el fallo) es preferible automatizarlo para evitar errores y uniformizar el criterio usado. Así pues, este TFM se centrará en el *Machine Learning*, que emplea los mismos modelos que el *Deep Learning* y el *Big Data* con la ventaja de no necesitar servicios de computación en la nube al manejar conjuntos de datos menos extensos.

4.2.2.- Algoritmos de *Machine Learning*

Árbol de decisión: técnica basada en una representación gráfica con nodos (óvalos), donde se dividen los datos en función de un test (dato mayor o menor que cierto valor de alguna de las variables de entrada), y se envían, en función de si cumple o no la condición del test, a cada uno de los subgrupos de datos formados las ramas del arbol. Se repite el proceso de subdivisión hasta llegar a las hojas (rectángulos) donde se asigna a cada subconjunto un valor de la variable salida (etiqueta)

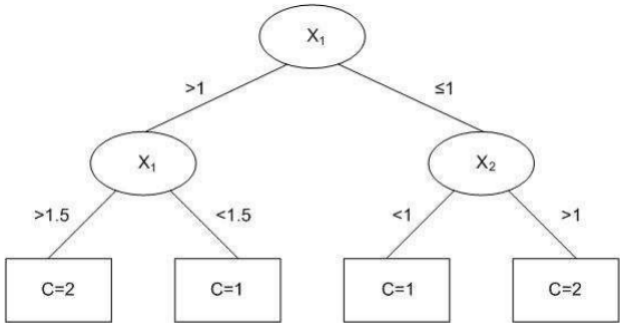


Figura- 4.2 Árbol de decisión

Teniendo en cuenta esta estructura, hay que decidir cómo se escoge el test en cada nodo, tanto qué variable como qué valor de esa variable, y cuántas veces (cuántos nodos) se realiza la subdivisión en subgrupos. Después de crear un primer modelo de árbol se realiza el proceso de poda: se estudia si todas las ramas creadas son útiles o si es necesario eliminar algunas.

En función de cómo se hagan dichas elecciones se hablará de un algoritmo u otro. Actualmente, el más usado en árboles de decisión es el C4.5 desarrollado por Quinlan en 1993 a partir del algoritmo ID3 ideado también por él en 1986. El C4.5 incorpora mejoras como la posibilidad de trabajar con datos continuos, reduce el sesgo hacia grupos grandes del ID3, y un mecanismo de poda, basado en un test de hipótesis que comprueba si seguir subdividiendo el árbol disminuye significativamente el error.

Estos algoritmos escogen la variable de test para el nodo a partir de dos criterios basados en la teoría de la información. Primero, el criterio de ganancia de información, que escoge aquel atributo que más reduce la incertidumbre de saber qué etiqueta les corresponde a los miembros de un subgrupo. Este es la base del ID3, pero da preferencia a las categorías con mayor número de individuos, por lo que en C4.5 se añadió otro criterio que corrige este sesgo, el criterio de ratio de ganancia. Este criterio calcula la ratio de reducción de incertidumbre, entendida como la reducción de incertidumbre calculada por el primer criterio entre la incerditumbre total que aporta cada uno de los subconjuntos en los que se dividiría. Se puede ver cómo se calcula esta incertidumbre o entropía y los criterios del árbol en la figura- 4.3.

$$g_1 = -\sum_{j=1}^{k} \frac{freq(C_j,T)}{|T|} \times \log_2\left(\frac{freq(C_j,T)}{|T|}\right) + \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \sum_{j=1}^{k} \frac{freq(C_j,T_i)}{|T_i|} + \log_2\left(\frac{freq(C_j,T_i)}{|T_i|}\right)$$

$$g_2 = \frac{g_1}{-\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times log_2\left(\frac{|T_i|}{|T|}\right)}$$

Donde:

n: número de categorías de la variable escogida para dividir el árbol.

k: número de categorías de la variable predictora

$|T|$: número de casos en el conjunto

$Freq(C_j, T)$: número de casos de la categoría C_j en T

$|T_i|$: número de casos en el subconjunto T_i (casos de T eliminando los vacíos de la variable considerada)

$Freq(C_j, T_i)$: número de casos de la clase C_j en el subconjunto T_i

Figura- 4.3 Cálculo de los criterios de selección de categoría del árbol

En función de la variable de salida se dividen en: arboles de clasificación, que son aquellos en los que la variable resultado es categórica, y árboles de regresión, en los que la variable resultado es continua [19].

Red neuronal: Este modelo computacional se ha convertido en uno de los más populares en los últimos años debido al gran número de aplicaciones en las que ha logrado altas tasas de éxito. Su funcionamiento se basa en la neurona artificial que se muestra en figura- 4.4.

Una neurona artificial no es más que una regresión lineal, es decir, una suma ponderada de las variables de entrada x_i y una variable de sesgo o bias, cuyo resultado se pasa como argumento a una función de activación φ, que transforma dicha función lineal en una no lineal. La salida de la neurona y_j clasifica en una clase u otra en función de si el resultado de la función de activación supera o no un umbral establecido.

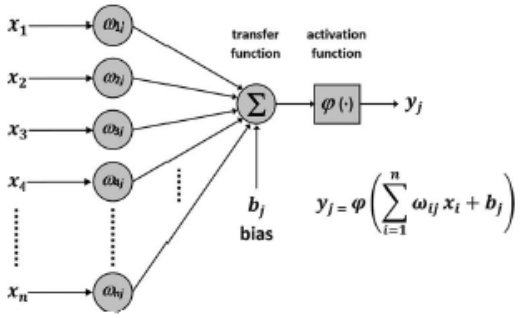


Figura- 4.4 Estructura de una neurona artificial

Las principales funciones de activación se muestran en la figura- 4.5, siendo las principales la función sigmoide $\sigma(x)$, cuya imagen de salida es [0,1], la tangente hiperbólica tanh, cuya

imagen es [-1, 1] y la función lineal rectificada ReLU, que es una función definida a trozos, cuya salida es nula cuando la entrada es negativa y el propio valor de la entrada cuando se trata de una entrada positiva. En general se suele emplear la sigmoide, ya que su salida se puede interpretar fácilmente como una probabilidad al estar en el intervalo [0,1]. La ReLU ha mostrado muy buenos resultados en redes neuronales de *Deep Learning*, ya que es muy eficiente cuando el número de capas es muy alto [20].

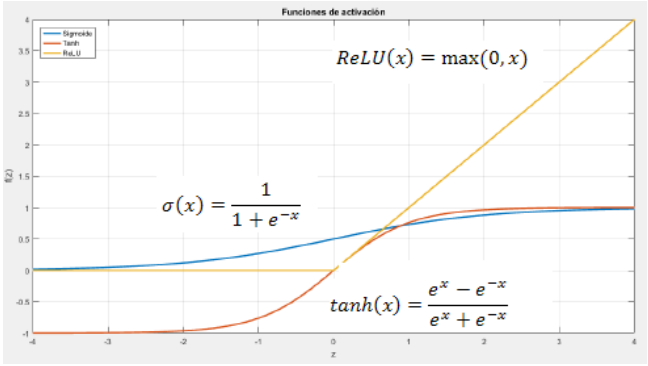


Figura- 4.5 Funciones de activación

Una vez conocido el principio de funcionamiento de una neurona se pasa a explicar cómo funciona la red neuronal. Este modelo se basa en dos principios que se desarrollan a continuación.

En primer lugar, el empleo de neuronas en paralelo (capas), permite obtener, al combinar el efecto de cada neurona, fronteras de clasificación curvas, que serían imposibles de obtener con una sola neurona por capa. Esto se puede ver en la figura- 4.6, en la que a partir de 4 sigmoides se define una superficie curva que, al intersecar con el umbral de 0,5, divide el dominio en la zona verde de puntos de una clase y la zona roja, perteneciente a los puntos de la otra clase. Cuánto más intrincado es el patrón a detectar, mayor número de neuronas por capa serán necesarias para poder definirlo correctamente.

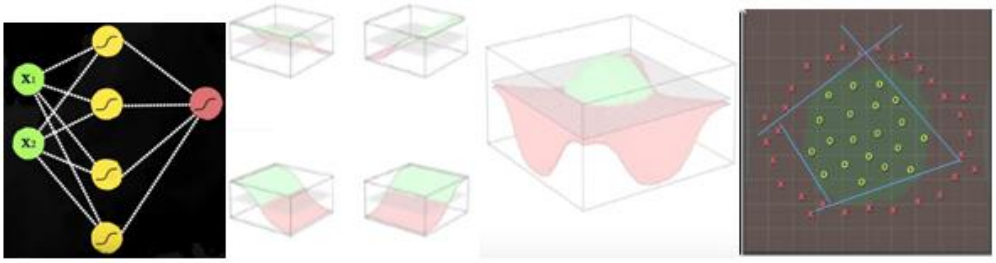


Figura- 4.6 Combinación de neuronas para formar una frontera de clasificación curva

El segundo principio (interconexión) se basa en conectar entre sí las capas utilizando la información de salida de cada capa como entrada para la siguiente. Esto permite jerarquizar la información, ya que cada capa se puede centrar en una tarea, de manera que se obteniene sucesivamente información máss elaborada y abstracta. Un ejemplo de esto son

las redes neuronales para el reconocimiento facial: en las primeras capas se identifican los rasgos generales, como los ojos o la boca, en las posteriores se unen dichos rasgos para identificar la cara, y en las siguientes se puede diferenciar a quién pertenece la cara y clasificarla según diferentes características.

En función de la posición de cada capa se habla de capa de entrada, que es aquella que recibe los datos, capa de salida, que suministra información al exterior, y capas ocultas, que son aquellas intermedias entre la entrada y salida.

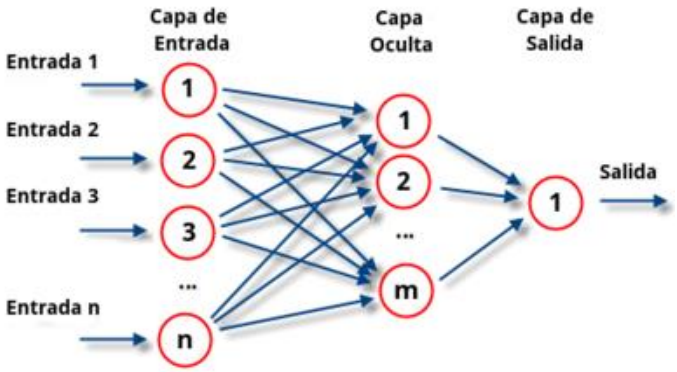


Figura- 4.7 Estructura de una red neuronal

En función de cómo se realizan las conexiones entre neuronas existen distintos tipos de redes neuronales. Las más empleadas son las siguientes:

Red neuronal profunda (DNN): todas las neuronas de cada capa están conectadas a todas las de la capa posterior. Es muy versátil, pero tienen un gran número de conexiones, por lo que solo se emplean cuando el número de datos es relativamente pequeño

Redes neuronales convolucionales (CNN): son empleadas principalmente en la clasificación de imágenes. Emplean la convolucion para agrupar información y disminuir así el volumen de información a tratar.

Redes neuronales recurrentes (RNN): son empleadas para datos secuenciales, en los que el valor anterior de un dato influye en el siguiente como la evolución de la bolsa o la predicción de texto. Emplean capas ocultas LSTM que tienen una conexión de recurrencia, es decir, su salida retroalimenta a su propia entrada.

Una vez descritos los elementos y funciones de cada parte de la red, se explicará brevemente cómo se consigue que estos elementos actúen de manera coordinada y se ajusten a la información disponible.

1. Se inicializan aleatoriamente los parámetros w_{ij} de las sumas ponderadas de las neuronas.
2. Se calcula el valor de la función de coste, es decir, el error cometido por la red neuronal.

3. Se calcula, mediante el algoritmo de *backpropagation*, el gradiente del coste respecto a cada capa. Este gradiente es una medida de la responsabilidad de cada neurona en el resultado obtenido, es decir, cuanto cambia la salida al variar un poco los parámetros de cada neurona.
4. A partir del gradiente, se emplea el algoritmo de descenso del gradiente para recalcular los pesos w_{ij} de las neuronas.
5. Se repite el proceso desde 2 hasta que el valor de la función de coste converja o se considere aceptable.

***Clustering*:** son algoritmos de aprendizaje no supervisado que se basan en agrupar datos en clústeres o grupos de tal manera que los individuos de un mismo clúster sean similares o estén cercanos entre sí y estén lejos o sean distintos de los pertenecientes a otros clústeres [21]. Esta idea de similitud o distancia es la que determina cómo se formarán los grupos. Existen muchas formas de definirla en función del tipo de atributo que se conozca de los datos y cómo se quiera hacer la agrupación. En variables numéricas se utilizan como distancias más habituales la euclidea, la distancia Manhattan o la de Minkowski. En variables categóricas las similitudes más empleadas son la binaria, la de atributos comunes o la de atributos no compartidos, que son una especie de ratio entre las características que comparten dos registros y el total de características posibles de cada registro [22]. Una vez definido cómo se va a medir cuánto se parecen dos registros entre sí se puede proceder al agrupamiento. En función del método empleado para realizar el *clústering* se habla de los siguientes algoritmos:

- **Jerárquicos:** no se determina previamente el número de grupos que se quieren conseguir. Se define un ranking de qué registros son los más cercanos entre sí y se van agrupando consecutivamente aquellos más similares. Pueden ser *bottom-up* (aglomerativo) si cada registro es inicialmente un clúster y se van uniendo entre sí formando clústeres más grandes hasta que se decida parar o se llega a un clúster único (algoritmos AGNES o CURE). También puede ser *top-down* (divisorio) si se empieza con un grupo único y se divide en clústeres más pequeños hasta llegar a tantos grupos como objetos o cumplir un criterio de parada (algoritmos DIANA o MONA)
- **K-means:** forma parte de una familia de algoritmos más grandes denominados de particiones que especifican el número de grupos *k* que se quieren conseguir con el clústering. El procedimiento es el siguiente:
    1. Se asigna aleatoriamente a *k* elementos como centros, también llamados medias, de los clústeres.
    2. Se asigna al resto de registros a los clústeres cuyo centro tienen más cerca.
    3. Se calcula, en cada clúster, su nuevo centro o media a partir de todos los elementos que lo forman.

4. Se reasignan de nuevo los elementos a los clústeres.
5. Se repite el proceso hasta que ya ningún elemento cambie de clúster.

Existen otros métodos, como los basados en densidades (DBSCAN) o los que emplean rejillas o grafos.

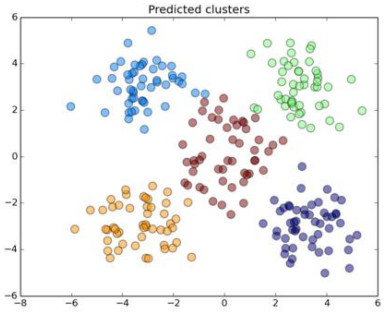


Figura- 4.8 Clústeres formados en un proceso de clústering

Análisis de correspondencia y análisis de componentes principales: son técnicas estadísticas para reducir la dimensión de un conjunto de datos. El análisis de componentes principales intenta reducir el número de variables numéricas continuas mientras el análisis de correspondencia reduce las variables cualitativas. Se basan en calcular la descomposición en autovalores de la matriz de covarianza o de coeficientes de correlación y escoger aquellos componentes que explican en mayor medida los datos, despreciando los demás. Normalmente se hace una representación bidimensional o tridimensional de dichas componentes principales para facilitar su interpretación [23].

Se describen otras técnicas no empleadas finalmente en ANEXO I.

5.- Metodología empleada

5.1.- DESCRIPCIÓN DE LA BASE DE DATOS

5.1.1.- Origen de los datos

Todos los informes de fallo son registrados por los propios operarios de mantenimiento de manera digital mediante tablets o, en ciertos países, en formato papel y luego son digitalizados en sus respectivas delegaciones. En total, de noviembre de 2017 a noviembre de 2018 se han reportado 68793 intervenciones no programadas (mantenimiento correctivo) en escaleras mecánicas y pasillos rodantes en el área Asia-Pacífico, que es en la que se centrará este trabajo. Estos datos son proporcionados en formato tabla de Excel.

El hecho de que la única información para el trabajo sean los informes de fallo implica una gran limitación. No se conoce en ningún momento el número total de equipos instalados; de hecho, la población de equipos instalados es variable ya que se instalan nuevos equipos a lo largo del periodo estudiado. Solo se conoce la existencia de aquellos que han informado de algún tipo de incidencia durante el periodo considerado. Por tanto, no se puede saber en en principio la población de equipos ni calcular ratios respecto a ella puesto que la muestra la constituyen solo los equipos que han presentado fallos. Así pues, los resultados que se obtengan se referirán solo a la muestra disponible y será necesario ampliar el muestreo a más años en trabajos futuros para que puedan aparecer la totalidad de los equipos.

5.1.2.- Características generales de los datos

Hay que tener en cuenta que estos informes presentan una serie de características propias debido al método de recogida de datos, basado, en unos pocos casos, en la selección de ítems de una lista, y, en la mayoría de campos, en la escritura manual por parte del operario de mantenimiento. Esto conlleva una serie de ventajas y limitaciones que se presentan a continuación:

- **Ventajas:**
    1. **Información fácil de obtener:** no necesita sensores, conexión a internet en cada equipo ni ningún otro equipo especial que aumente el coste de instalación.
    2. **Datos imposibles de obtener con otros métodos:** la información que se puede obtener mediante sensores, aunque útil, es limitada, ya que cada variable que se mida incrementa el coste en sensorización, y muchas de las variables que podrían ser de interés no son medibles, tales como la

explicación de las medidas tomadas, opinión del operario sobre la causa del fallo, etc.

3. **Información de situaciones clave:** en relación con lo anterior, la información que aquí se trata es obtenida en momentos de fallo del equipo, situaciones críticas que conviene conocer en detalle y donde muchos de los sensores podrían no captar la naturaleza del fenómeno al estar calibrados para medir las condiciones de funcionamiento normales del equipo.
4. **Gran abundancia de registros:** todos los informes de fallo quedan registrados, independientemente del año de fabricación, modelo o fabricante del equipo, mientras que el uso equipos conectados y sensores (proyectos de Thyssenkrupp MAX 1.0 Y 2.0 respectivamente) queda restringido a ciertos modelos de fabricación muy reciente.

- **Limitaciones:**
  1. **Registros incompletos:** los operarios de mantenimiento no tienen la obligación de conocer todas las características de los equipos y, salvo en ciertos países donde se usa información de ERP u otras fuentes para completar los reportes, variables como el año de fabricación, la fábrica o la distancia entre apoyos están vacias en muchos registros.
  2. **Información poco precisa:** ligado a la naturaleza humana de los reportes, es habitual que los registros tengan errores como fallos ortográficos, registros de ascensores (solo se analizan escaleras mecánicas y pasillos rodantes), empleo de distintas unidades de medida (mm, m, cm) o datos claramente erroneos (año de fabricación posterior a 2018, escaleras de tamaño superior al máximo fabricado, etc). Esto exige un trabajo exhaustivo y preciso de limpieza de datos para lograr una estructura homogénea, una labor imprescindible pero que consume gran cantidad del tiempo de trabajo en cualquier tarea de análisis de datos.
  3. **Uso de distintos idiomas y argots:** los datos provienen de 19 delegaciones distintas y en cada una de ellas existe un argot diferente con abreviaturas propias cuyo significado en principio se desconocía. Esto ha obligado a contactar con los responsables de las distintas delegaciones con el fin de preguntar el significado de dichos términos. Por otra parte, aunque las listas desplegables vienen por defecto en inglés, los campos están escritos en 6 idiomas distintos (inglés, árabe, coreano, chino, vietnamita y tailandés) lo que ha obligado a emplear distintos métodos de traducción automática para poder analizar todos los registros en inglés.
  4. **Distintos criterios de valoración:** en función del contrato con el cliente (mantenimiento total o por horas) y las indicaciones de la delegación, cada

operario emite un juicio subjetivo con su propio criterio sobre algunas de las variables registradas (debido a usuario/avería, motivo del fallo, etc). Esto complica enormemente la comparación entre los distintos países por lo que es necesario crear un criterio global y de reglas conocidas para analizar estas variables.

5.1.3.- Estructura de los datos

En cada uno de los informes se registran 59 variables que se dividen en los distintos elementos particulares de la intervención:

- Identificación y motivo de la llamada.
- Identificación y características del equipo.
- Identificación del operario de mantenimiento.
- Identificación y características del edificio y cliente.

Se presenta a continuación la lista de variables que incluye cada bloque, junto a una breve descripción de cada una de ellas. No se incluyen las descripciones de las siguientes variables al ser información clasificada: Name, SWP, SWPonarrive, SWPonreleaseddate, CommisioningDate.

- **Variables del incidente:** son aquellas que describen cómo se produjo el fallo, sus características más relevantes y las medidas tomadas para subsanarlo. Se enumeran a continuación:
    1. **Callnumber:** número de la llamada, sirve de identificador de cada incidencia.
    2. **AfterHoursNumber:** identificador de aquellas incidencias realizadas fuera del horario de servicio al cliente.
    3. **LoggedDate:** fecha y hora a la que llama el cliente informando de la incidencia.
    4. **ConfirmDate:** fecha y hora a la que confirman desde la delegación que se va a enviar a un técnico a revisar la incidencia.
    5. **ArriveDate:** fecha y hora a la que el operario llega a la instalación del cliente y empieza con la inspección.
    6. **OutDate:** fecha y hora a la que el operario se va de la instalación del cliente y cierra el procedimiento de inspección.
    7. **Problem_en:** descripción por parte del cliente de la incidencia.
    8. **Action_en:** descripción por parte del operario de la intervención realizada.
    9. **Activitycomment:** información logística sobre la incidencia,
    10. **AfterHours:** variable binaria. Si su valor es 1 indica que la incidencia se solucionó fuera del horario de servicio al cliente.

11. **Chargeable:** variable binaria. SI su valor es 1 indica que la incidencia no la cubre el contrato de mantenimiento que tiene el cliente y que debe abonar su coste.
12. **ChargeableNumber:** número de cuenta al que se debe pasar la factura de las incidencias no cubiertas por el seguro del cliente.
13. **Chargeableamount:** cantidad que se abona en las incidencias no cubiertas por el seguro.
14. **Interference:** variable binaria. Si su valor es 1 indica la presencia de un objeto externo en la zona del fallo o un fallo ajeno al equipo (alarma de incendios, corte de suministro eléctrico o inundación)
15. **Cancelled:** variable binaria. Si su valor es 1 indica que la llamada fue cancelada antes de que el técnico pudiera realizar la inspección.
16. **NoFault:** variable binaria. Si su valor es 1 indica que el incidente no fue causado por un uso indebido de la escalera o pasillo, sino por un fallo del propio equipo.
17. **OrderNumber:** identificador de la inspección a efectos internos de la delegación en el caso de que no se use el identificador por Docket.
18. **Docket:** identificador de la inspección a efectos internos de la delegación.
19. **FaultAreaName_en:** elemento en el que se produjo o causó el fallo.
20. **Incidentcomment_en:** recoge la misma información que Action_en
21. **PassengerInjured:** variable binaria. Si su valor es 1 indica que el incidente causó heridas a pasajeros.
22. **EmployeeInjured:** variable binaria. Si su valor es 1 indica que el incidente causó heridas a trabajadores del cliente.
23. **LastModified:** última fecha de consulta y modificación del registro.
24. **CalltoCallCentre:** variable binaria. Si su valor es 1 indica que el cliente contactó para informar de la incidencia con el centro de llamadas de Thyssen.

- **Variables del equipo:** son aquellas que describen las características geométricas, mecánicas, eléctricas, de mantenimiento y origen de la escalera mecánica o pasillo rodante. Las componentes de este grupo son:
    1. **Unitnumber:** identificador del equipo.
    2. **Yearmanufactured:** año de fabricación del equipo
    3. **Manufacturer_en:** compañía fabricante del equipo
    4. **DriveType_en:** tipo de control realizado en el motor.
    5. **UnitType_en:** variable que indica si el equipo es una escalera mecánica, un ascensor o un pasillo rodante.
    6. **Factory_en:** fábrica donde se ensambló el equipo.
    7. **Controller_en:** tipo de controlador que lleva instalado el equipo.

8. **Model_en:** denominación comercial del modelo.
9. **Floors:** número de plantas que salva el equipo.
10. **Speed:** velocidad a la que se mueven los peldaños, a partir de ahora se citará como velocidad del equipo.
11. **RisePerMetre:** desnivel que supera el equipo.
12. **ESClength:** distancia entre apoyos medida en horizontal.
13. **ESEsetwidth:** ancho de los peldaños.
14. **HandOverDate:** fecha en la que Thyssenkrupp empezó a llevar el mantenimiento de dicho equipo.
15. **UnitScheduledHours:** tiempo que se dedica al mantenimiento preventivo en cada visita programada.
16. **VisitPerYear:** número de visitas programadas al año para mantenimiento preventivo.
17. **EscalatorDrive:** variable binaria que indica si el equipo tiene dos motores (Dual) o solo uno (Single).
18. **EscalatorOutdoorUnit:** variable binaria. Si su valor es 1 indica que el equipo se encuentra a la intemperie.
19. **EscalatorAutoStartDevice_en:** variable binaria que indica si el equipo tiene dispositivo de arranque remoto (Autostart) o no (None).

- **Variables del operario: son** aquellas que identifican al trabajador que realizó la operación. Los campos incluidos en este bloque son:
    1. **PrimaryEmployeeNumber:** número identificador del operario de mantenimiento.
    2. **PrimaryEmployeeFirstName_en:** nombre del operario de mantenimiento.
    3. **PrimaryEmployeeSurname_en:** apellido del operario de mantenimiento.
- **Variables del cliente:** son aquellas que identifican el lugar del incidente y al propietario de la instalación. Se incluyen dentro de esta definición a las siguientes variables:
    1. **CountryName_en:** país donde está situado el equipo.
    2. **Caller_en:** nombre y apellidos del cliente que realizó la llamada.
    3. **ContactNumber:** número de teléfono al que llamar para contactar con el cliente.
    4. **BuildingName_en:** nombre del edificio donde está el equipo relacionado con la llamada.
    5. **CustomerName_en:** nombre de la empresa propietaria del equipo o para la que trabaja la persona que realizó la llamada.
    6. **EmployeeNumber:** identificador del empleado del cliente que realizó la llamada.

7. **SegmentType_en:** tipo de establecimiento donde está situado el equipo (aeropuerto, metro, centro comercial, residencial…).
8. **BuildingTypeSub_en:** subtipo de establecimiento donde está situado el equipo (tamaño del centro comercial, residencial privado o público, etc).

Se ha valorado la posibilidad de unir todos los campos de texto (Problem_en, Action_en, Activitycomment, Incidentcomment_en) pero se ha descartado esta posibilidad ya que el aumento del número de caracteres que se traducen implica un aumento del tiempo y recursos necesarios para la traducción que no compensan la mayor información que se obtiene. Esto es debido a que el campo Problem_en solo indica, en la mayor parte de los casos, qué unidad se ha detenido, sin aportar más detalles, el campo Activitycomment e Incidentcomment_en están mayoritariamente vacíos y solo se usan para indicar a quién se informó de las medidas tomadas, aspecto no relevante en este trabajo, por lo que solo se utilizará la variable Action_en para el análisis de texto.

Se puede ver en ANEXO II el porcentaje de registros vacíos (not available (NA)) inicialmente en cada variable y aquellas variables que no se considerarán en el análisis posterior por presentar información confidencial o poco relevante para el objetivo de este trabajo

No se incluirán 17 de las variables disponibles, además de las 5 confidenciales. Los 37 campos restantes serán los que constituyan la base de datos con la que se empezará a trabajar. No obstante, esto no implica que todas vayan a formar parte de los resultados presentados, ya que dependerá de si se considera relevante la información que se obtenga de su análisis.

5.2.- HERRAMIENTRAS DE ANÁLISIS DE DATOS

Ahora que se ha definido claramente el objetivo a conseguir y los pasos que se pretenden dar es necesario escoger qué herramienta informática se usará para conseguirlo. Existen una gran variedad de lenguajes y programas empleados para el análisis de datos y el *Machine Learning*. A continuación, se mostrarán algunos de los más populares indicando sus principales características, ventajas e inconvenientes para este trabajo.

- **Power BI:** herramienta de análisis de datos de Microsoft pensada especialmente para realizar *business intelligence*, es decir, para analizar resultados económicos de empresas y facilitar la toma de decisiones. Por tanto, su uso habitual es la realización de informes financieros más que el análisis de datos en general. Sus principales ventajas son la gran variedad de orígenes de datos que puede utilizar, desde tablas de Excel a bases de datos SQL o páginas web, la facilidad para la creación de gráficos dinámicos, el gran número de visualizaciones distintas con las que cuenta y su buena

conectividad, ya que se pueden compartir reportes online desde la nube o desde su propia app.

Sin embargo, cuenta con grandes limitaciones: en primer lugar, usa lenguaje M o DAX, por lo que el acondicionamiento de los datos es complejo, y no cuenta con una gran comunidad detrás a la que se puedan consultar dudas lo que complica aún más la tarea. Además, no cuenta con herramientas propias para el *Machine Learning*, aunque esto en parte se soluciona usando AzureML de Microsoft.

- **R:** lenguaje de programación *open source* específicamente diseñado para trabajos con métodos estadísticos. Es uno de los lenguajes junto con Python más empleados en ciencia de datos. Cuenta con una gran comunidad detrás, interfaces de uso intuitivas como RStudio e innumerables paquetes para realizar la inmensa mayoría de tareas que se necesitan en análisis de datos y *Machine Learning*. Su principal ventaja es su sencillez para el acondicionamiento y tratamiento de datos y la gran variedad y versatilidad a la hora de presentar resultados con gráficos.
- **Python:** lenguaje de programación *open source* multipropósito, con amplio uso en ciencia de datos en los últimos años. Al igual que en R, existe un gran número de usuarios dispuestos a ofrecer su ayuda en foros. Además, también cuenta como ventaja con su buena integración con servicios en la nube y la posibilidad de emplearlo para gran variedad de rutinas de programación. Así pues, es el lenguaje elegido mayoritariamente para implementar aplicaciones de Ciencia de datos en el mundo real, es decir, en productos que van a usar los clientes. Además, algunas de las librerías de código abierto más usadas para el *Machine Learning* y *Deep Learning* como Tensorflow o Keras fueron programadas inicialmente con Python, razón por la cuál ha alcanzado gran popularidad.
- **BigML, Amazon ML, Microsoft Azure ML Studio, IBM Watson ML:** estas son herramientas de *Machine Learning* ofrecidas por grandes empresas (Amazon, Microsoft, Google...) junto a servicios de computación y almacenamiento en la nube. Están pensadas para aplicar directamente modelos de *Machine Learning* a datos de la empresa de manera sencilla sin necesidad de tener conocimiento sobre Ciencia de datos ni saber programar. Su gran ventaja es su simplicidad, la gran potencia de cálculo que ofrecen y la facilidad para integrarlos en aplicaciones reales. Sin embargo, son suscripciones de pago y, en general, su flexibilidad es limitada ya que no ofrecen técnicas estadísticas exploratorias y su uso está limitado a las condiciones que impongan las compañías.

Inicialmente, desde Thyssenkrupp se planteó que se empleara Power BI para el trabajo, dado que era una herramienta conocida por la empresa y parecía que podía adaptarse a lo que se estaba buscando. Sin embargo, al intentar acondicionar los datos se observó que su uso no era nada eficiente y se descartó por su baja capacidad para transformar y filtrar los

registros. Considerando que este trabajo plantea un análisis inicial de los datos, cuyo fin es hallar correlaciones y casos de uso que se puedan explotar en profundidad posteriormente para mejorar el mantenimiento predictivo que se realiza en la empresa, se decide emplear como herramienta de trabajo el lenguaje R a través de la interfaz RStudio. Los principales motivos para esta decisión son el amplio número de recursos didácticos para aprender a manejarlo y la facilidad para acondicionar los datos, implementar modelos de *Machine Learning* y, sobre todo, la variedad de gráficos y visualizaciones, indispensables en la fase exploratoria para conocer bien la naturaleza de los datos y entender qué provecho se puede sacar de ellos.

En una fase posterior, cuando ya se haya establecido qué aplicación práctica se puede dar a estos datos puede ser necesario cambiar la herramienta usada para adaptarse a la infraestructura ya existente en la empresa.

Si se decide que la principal utilidad es la mejora de la toma de decisiones en la cadena de mando se podría combinar R con Power BI para facilitar la comunicación de los datos a mandos intermedios. En cambio, si se llega a la conclusión de que la principal aplicación sea una app de ayuda al operario de mantenimiento o en el *call center* de la delegación se debería emplear Python o algún servicio en la nube que garantice la disponibilidad del servicio.

5.3.- PROCESO DE ACONDICIONAMIENTO

Una vez recopilados los datos, la primera tarea de cualquier proceso de análisis consiste en lograr una estructura homogénea y normalizada que facilite la posterior construcción de modelos y la extracción de gráficos explicativos [24]. Esta etapa es la que más tiempo suele necesitar debido a que es un proceso fundamentalmente manual, sin algoritmos que faciliten la tarea, y exige profundizar en la estructura y contenido de la base de datos para conocer exactamente cómo organizarla y qué información se podrá obtener posteriormente de estos datos.

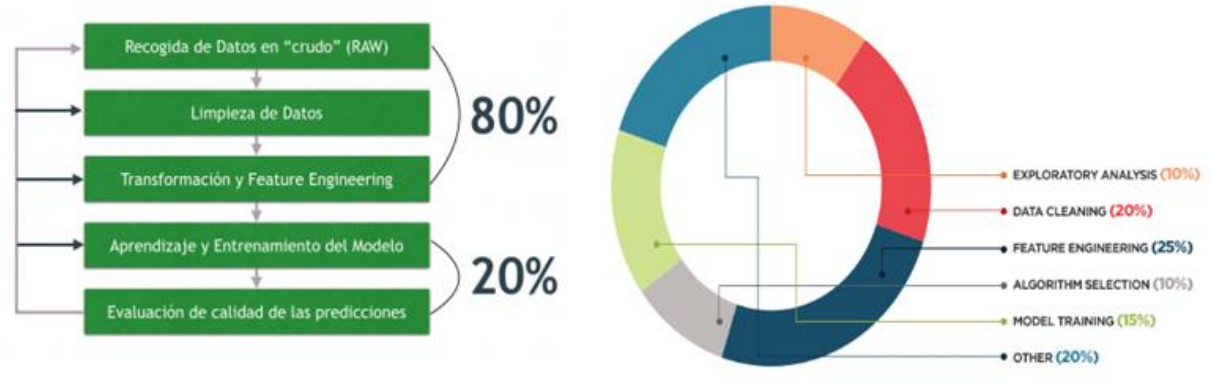


Figura- 5.1 Planificación temporal típica de un proceso de análisis de datos y Machine Learning [30] [31]

La fase de limpieza y transformación cobra aún más importancia en un caso como este, en el que los datos son introducidos de forma manual y la posibilidad de cometer errores es muy alta. Para llevarlo a cabo de manera ordenada, se divide el proceso de acondicionamiento en dos partes bien diferenciadas que se explican en 5.3.2.- y 5.3.3.-.

5.3.1.- Hardware y software empleado

El hardware utilizado es un ordenador portatil con sistema operativo Windows 10 y 4GB de RAM. En cuanto a software, se empleará la versión 3.5.2 de R y como entorno de desarrollo la versión 1.2.1335 de RStudio para Windows. Se ha valorado la utilización del paquete dplyr de R para facilitar el acondicionamiento, pero se ha considerado que no era necesario. Así pues, se han intentado usar solo aquellos paquetes imprescindibles, ya sea porque incorporan funcionalidades con las que no cuenta la versión básica de R o porque implementa métodos mucho más eficientes para realizar determinadas funciones. Se indican a continuación, en tabla 5.1, las librerías instaladas junto a un breve resumen del uso dado:

Librería	Versión	Algunas funciones usadas	Uso
readxl	1.2.0	readxlsl()	Lectura de datos iniciales, debido a la configuración de la hoja excel
ca	0.7.1	ca()	Análisis de correspondencia
tidytext	0.2.0	unnes_tokens()	Extracción de bigramas y trigramas
tm	0.7-6	DocumentTermMatrix(),Vcorpus	Creación de Bag of Words
NLP	0.2-0		requerida por tm
rpart	4.1-13	rpart(),plotcp(),printcp()	Creación de arboles de decisión
readr	1.3.1	write_excel_csv	Exporta en csv separado por comas con codificación UTF-8
Informationvalue	1.2.3	optCutoff(),ConfusionMatrix()	Ayuda en la creación de regresiones logísticas
tidyr	0.8.2	separate()	Facilita la creación del histórico de fallos
VCA	1.3.4	anovaMM()	Análisis ANOVA
nnet	7.3-12	nnet()	Creación de redes neuronales

Tabla 5.1 Librerías de R empleadas en el código

Una vez expuesto el hardware y el software, se explicará el proceso de acondicionamiento.

5.3.2.- Limpieza de datos

Lectura y definición de NA: en primer lugar, se importan los datos y se definen los valores que indican que el dato no está disponible (*Not available* (NA)). Debido al formato de fecha y

hora de los registros fue necesario emplear dos funciones distintas, una para leer los datos como archivo csv y otra como archivo xlsx que después se unieron, aprovechando las columnas tipo Date del fichero xlsx para sustituir a las del fichero csv. Se muestra el código necesario para llevar a cabo esta tarea en figura- 5.2.

```
d = read_xlsx("AP Esc&MW Call data 12mths to 05112018.xlsx",na=c("","*","N/A"))D = read.csv2("
AP Esc&MW Call data 12mths to 05112018.csv",encoding="UTF-8",na.strings=c("","","*","N/A"))
D[grep("Date",names(d))] = d[grep("Date",names(d))]
```

Figura- 5.2 Lectura de los datos

Filtrado: a continuación, hay que eliminar aquellos registros que no formarán parte del análisis. Serán excluidos los reportes que fueron cancelados, ya que no se llevaron a cabo y por tanto no contienen información de interés. También serán excluidos los reportes de ascensores que quedan fuera del alcance de este trabajo, que se centrará únicamente en escaleras mecánicas y pasillos rodantes, cuya mecánica y construcción es similar.

A diferencia de los reportes cancelados, que son fácilmente identificables, los reportes de ascensores son más difíciles de reconocer. En algunos casos son sencillos de encontrar ya que en UnitType vienen indicados como “Elevator”, pero se ha comprobado que existen otros registros de ascensores camuflados como “Escalator”, ya que es la opción que viene por defecto al rellenar dicho campo. Para poder extraer estos casos se emplearán las variables FaultAreaName_en y Action_en. Gracias a la información proporcionada por Thyssenkrupp se ha podido ver que existen una serie de fallos exclusivos de ascensores (relativos al sensor de cortina de luz, las puertas o el hueco de ascensor), por lo que aquellos reportes cuya causa de fallo sea de este tipo se puede asignar inequívocamente a un ascensor. Además, gracias al análisis semántico también se asignan a ascensores aquellos registros que describan la unidad reparada como *Elevator* o *Lift*. Se muestra el código para llevarlo a cabo en la figura- 5.3.

```
#%%%%%%%%%%%%% FILTRADO PARA ELIMINAR LOS CANCELLED
D=D[!D$Cancelled==1,]#N=736

#%%%%%%%%%%%%% FILTRADO PARA ELIMINAR ASCENSORES    %%%%%%%%%%%%%%%%%%%%%%

D=D[!D$UnitType_en=="Passenger Elevator",]#N=261
D=D[grep("elevator",invert=TRUE,tolower(D$Action_en)),]#N=57
D=D[grep("elevator",invert=TRUE,tolower(D$Problem_en)),]#N=11
D=D[grep("^Lift",invert=TRUE,D$Problem_en),]#N=233 #al coger que empiece es mas probable que s
ea lift que se suele iniciar el comentario por Lift/ESC/MW not working
D=D[grep("^Lift",invert=TRUE,D$Action_en),]#N=23
D=D[grep("^lift",invert=TRUE,D$Action_en),]#N=11
D=D[grep("^lift",invert=TRUE,D$Problem_en),]#N=55
D=D[grep("Cab|Rope|Shaft|Door|Dispatch|Destination|\\bCar\\b|Cradle|Voice|Hoist|Intercom",
invert=TRUE,D$FaultAreaName_en),]#N=774
D=D[grep("inspect secur elev|inspect elev|speaker|voice|voic|interphone|light curtain",
invert=TRUE,tolower(D$Action_en)),]#N=125
D=D[grep("^elev",invert=TRUE,D$Action_en),]#N=30
D=D[grep("^technician site elev",invert=TRUE,D$Action_en),]#N=500
```

Figura- 5.3 Filtrado de registros para eliminar los cancelados y pertenecientes a ascensores

Este método ha permitido eliminar unos 2000 registros, garantizando que los más de 65000 reportes restantes cumplen las condiciones necesarias para ser analizados conjuntamente.

Ordenación: posteriormente, se ordenaron los reportes en función de su fecha de registro (Logged Data). Esto es imprescindible para facilitar el cálculo del tiempo medio entre fallos (MTBF), variable fundamental del mantenimiento que ya fue descrita en 4.1.2.-. Además, se pone un identificador cronológico de los registros llamado "id" y se corrigen algunos registros temporales que se habían mostrado erróneos. Se muestra el código en figura- 5.4.

```
#%%%%%%%%%%%%%%%%%%%%%%%%%% ORDENAR CRONOLOGICAMENTE (PARA PODER ANALIZAR TIEMPO ENTRE FALLOS)
D=D[order(D$LoggedDate),]#PASA DE 2487 A 0 CON EL MINTB NEGATIVO
D$IntermediateDate=D$OutDate#Cambio de arrivedate y outdate en aquellas columnas que dan durac
iones negativas (damos por hecho que estan swapped esas columnas en esas filas)
D$IntermediateDate[which(D$ArriveDate>D$OutDate)]=D$ArriveDate[which(D$ArriveDate>D$OutDate)]
D$ArriveDate[which(D$ArriveDate>D$OutDate)]=D$OutDate[which(D$ArriveDate>D$OutDate)]
D$OutDate[which(D$ArriveDate==D$OutDate)]= D$IntermediateDate[which(D$ArriveDate==D$OutDate)]
D <- D[ ,!colnames(D)=="IntermediateDate"]
D$OutDate[D$CallNumber==1503687]="2018-01-02 16:15:00 UTC"
D$OutDate[D$CallNumber==158907]="2018-03-04 08:00:00  UTC"
D$OutDate[D$CallNumber==158908]="2018-03-04 08:40:00  UTC"

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% CREACION COLUMNA IDENTIFICADORA %%%%%%%%%
D$id=1:nrow(D)
```

Figura- 5.4 Ordenación cronológica de los registros

Corrección de variables numéricas: Las variables geométricas, de mantenimiento y año de origen del equipo (distancia entre apoyos, desnivel, ancho de peldaños y velocidad, número de visitas anuales programadas, duración de cada visita, año de fabricación y año de inicio de mantenimiento) son las únicas numéricas del archivo csv y su tratamiento requiere un formato especial. Para ello, hay que quitar las unidades, para que solo queden los números, cambiar la codificación de Thyssenkrupp (EK) por sus valores reales y unificar el separador decimal empleado. Se muestra el código en figura- 5.5.

```
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% ELIMINACION DE M y MM PARA PASAR A NUMERO %%%%%%%%%%%%%%%%%
D$SEsetpwidth = gsub("5EK","1000",as.character(D$ESEsetpwidth))
#as.character necesario al ser un factor
D$ESEsetpwidth = gsub("4EK","800",as.character(D$ESEsetpwidth))
D$ESEsetpwidth = gsub("3EK","600",as.character(D$ESEsetpwidth))
D$ESEsetpwidth = gsub("[^0-9,.]","",as.character(D$ESEsetpwidth))
#elimina todo lo que no sean numeros o separador digital
D$ESEsetpwidth = as.numeric(gsub(",",".",D$ESEsetpwidth))
D$Speed = gsub("[^0-9,.]","",as.character(D$Speed))
#elimina todo lo que no sean numeros o separador digital
D$Speed = as.numeric(gsub(",",".",D$Speed))
#cambia el separador digital en los casos que sea coma a punto y lo pasa a numero
D$RisePerMetre = gsub("[^0-9,.]","",as.character(D$RisePerMetre))
#elimina todo lo que no sean numeros o separador digital
D$RisePerMetre = as.numeric(gsub(",",".",D$RisePerMetre))
#cambia el separador digital en los casos que sea coma a punto y lo pasa a numero
D$ESClength = gsub("[^0-9,.]","",as.character(D$ESClength))
#elimina todo lo que no sean numeros o separador digital
D$ESClength = as.numeric(gsub(",",".",D$ESClength))
```

Figura- 5.5 Paso a formato numérico de las variables geométricas

Una vez que el formato es homogéneo hay que garantizar que los valores son correctos. Esto se conseguirá mediante las siguientes modificaciones en los registros:

1. Cambio de unidades para que todos los registros sean comparables (paso de m a mm y viceversa).
2. Intercambio de valores entre variables cuando se considera que el operario las ha introducido incorrectamente (ángulos en el ancho de peldaños, valores inasumibles de distancia entre apoyos que se corresponden con valores estándar de anchos o desnivel, etc).
3. Establecimiento de un rango de valores admisible para cada variable, de tal manera que si algún registro queda fuera se considerará que es un *outlier* (valor atípico) y se pasará a NA.
4. Rellenar registros vacíos a partir de otras variables (ángulo nulo para pasillos rodantes de aeropuerto y ángulo de 12° para centros comerciales, ancho de peldaños en modelo KLR 1000 de valor 1000 mm, año de fabricación como año anterior al inicio de mantenimiento si son de Thyssenkrupp, cálculo de distancia entre apoyos a partir del resto de variables geométricas)

Se puede observar cómo se han programado todas estas modificaciones de los registros en http://bellman.ciencias.uniovi.es/~raul/Acondicionamiento.html. La variable Ángulo aparece en el código, y también se ha mencionado en el texto, pero debido a que ha sido creada a partir de los datos y no viene por defecto en el fichero se comentará posteriormente cómo se ha definido, y solo se indica que sigue un tratamiento de corrección de registros análogo al resto de variables numéricas.

Todas las decisiones respecto a intervalos válidos e intercambio de variables han sido consultadas con expertos de Thyssenkrupp para garantizar que tiene sentido realizar dichos cambios y que se eliminan solo valores aberrantes. Se puede observar cómo quedan definidas las variables numéricas en tabla 5.2.

Variable	Unidad	Intervalo válido
Distancia entre apoyos	m	[4,70]
Desnivel	m	[0,50]
Velocidad	m/s	0,5; 0,63; 0,75
Ancho de peldaños	mm	[600,1400]
Ángulo	grados (°)	[0,35]
Número de visitas al año		Sin limites
Duración de las visitas	Horas	Sin limites
Año de fabricación		[1990,2018]
Año de inicio de mantenimiento		[1990,2018]

Tabla 5.2 Definición de las variables numéricas y sus intervalos admisibles

Corrección de variables factor: tras corregir las variables numéricas, queda repetir el proceso para las variables categóricas, llamadas factor en R. Estas variables solo pueden adquirir un valor concreto dentro una serie de niveles. El método para garantizar que el formato y contenido de estas variables es el correcto consta de los siguientes pasos:

1. Corrección formal, pasando el texto a mayúsculas, eliminando espacios en blanco al inicio o fin de palabra, signos de puntuación y otros caracteres.
2. Rellenar registros vacíos en función de otras variables o con valores por defecto (controlador en función de modelo, escalera a cubierto por defecto).
3. Reducción del número de niveles en cada variable, agrupando aquellos que sean equivalentes en función de criterio técnico y consultando a expertos de la empresa. Se asigna a un nivel "otros" a aquellos minoritarios en cada variable o que no tengan interés para el análisis posterior.

En el caso concreto de la variable FaultAreaName_en, cuyo número de niveles asciende a más de 150 inicialmente, se aplican técnicas estadísticas para facilitar la labor de agrupar esos niveles o categorías y reducir así su número. En concreto, se emplea el análisis de correspondencia ya explicado en algoritmos de *machine learning*, utilizando como variables los lexemas importantes del reporte de fallo, los cuales serán descritos posteriormente en análisis exploratorio. En resumen, este método mostrará cercanos entre sí a aquellos niveles de FaultAreaName_en cuyos registros empleen las mismas palabras, lo que debería indicar que podrían reducirse a un nivel único ya que se están refiriendo al mismo tipo de fallo. Se muestra a continuación la representación de las dos primeras dimensiones (los autovalores más explicativos) del análisis de correspondencia en figura- 5.6.

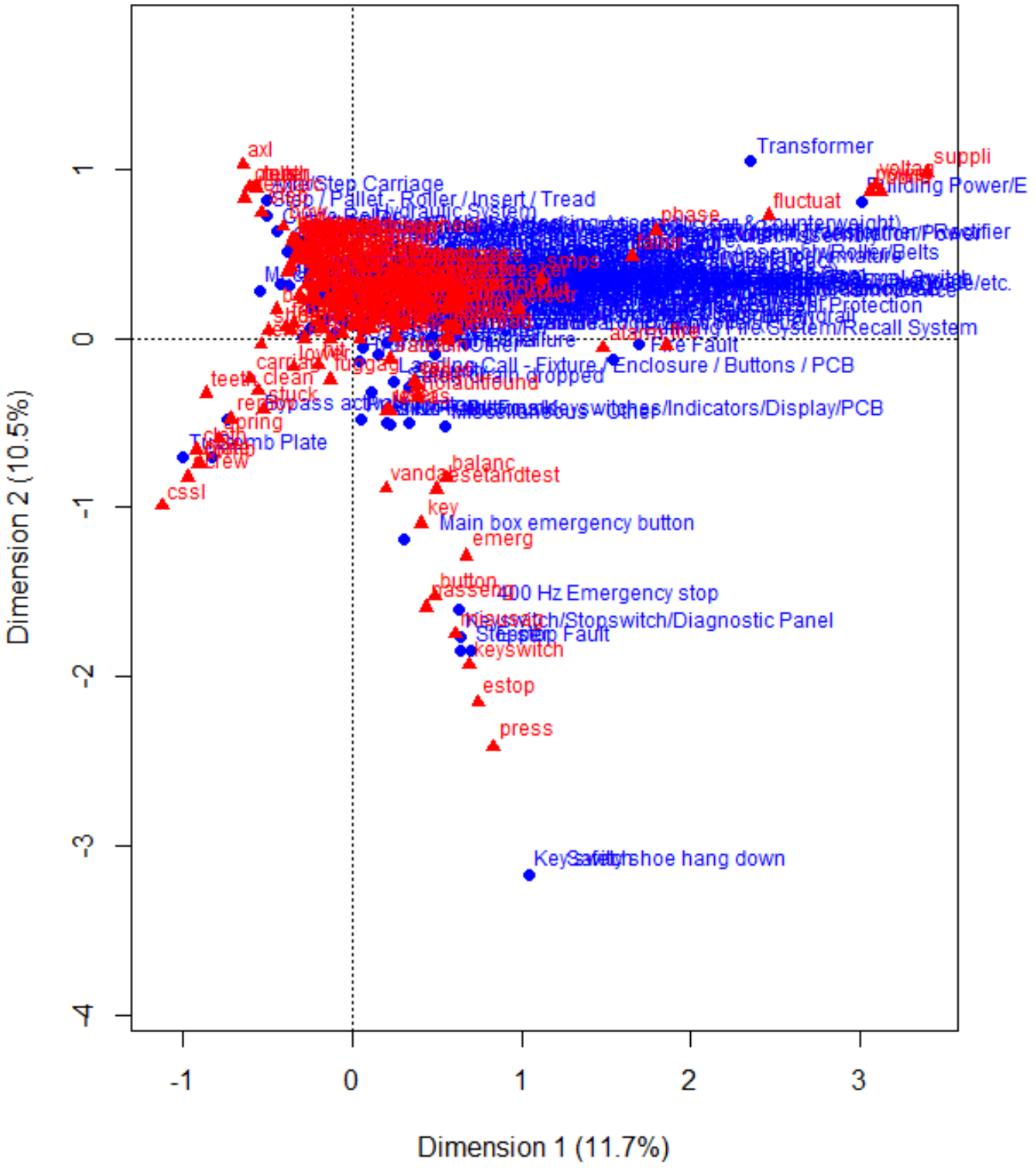


Figura- 5.6 Representación del análisis de correspondencia completo

Dado que la imagen es poco clara por el gran volumen de elementos a representar, se realizan ampliaciones en algunas zonas de interés. Se muestran en figura- 5.7 y figura- 5.8.

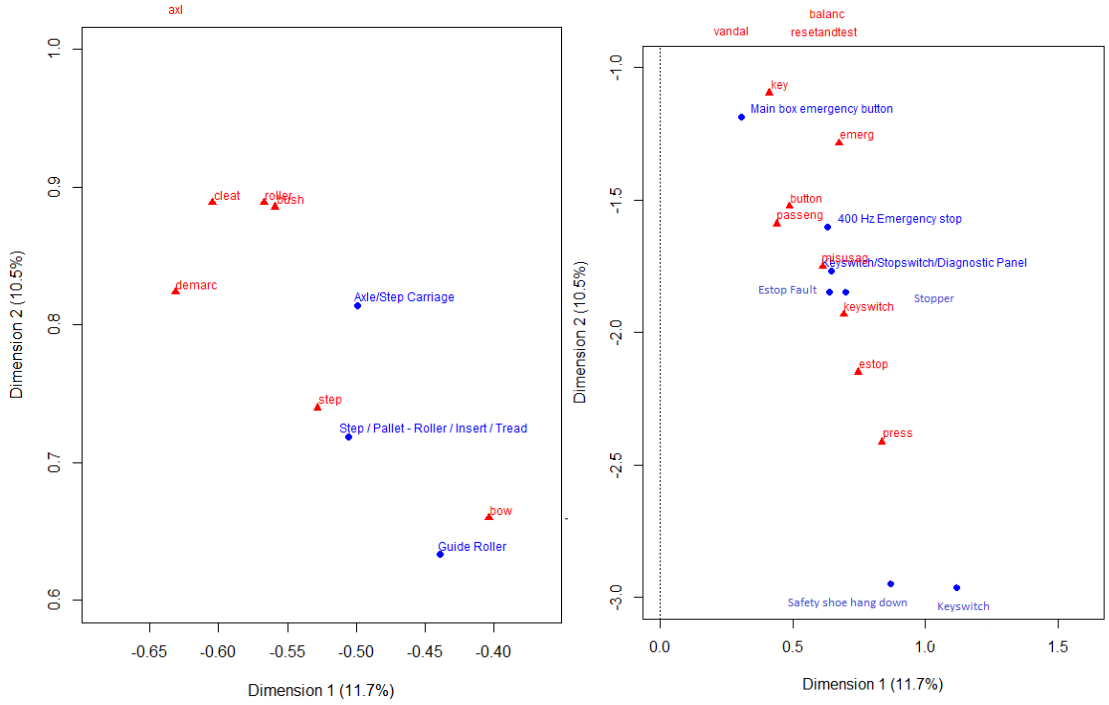


Figura- 5.7 Ampliación en la zona de fallo de peldaños (izquierda) y de fallo de emergencia (derecha)

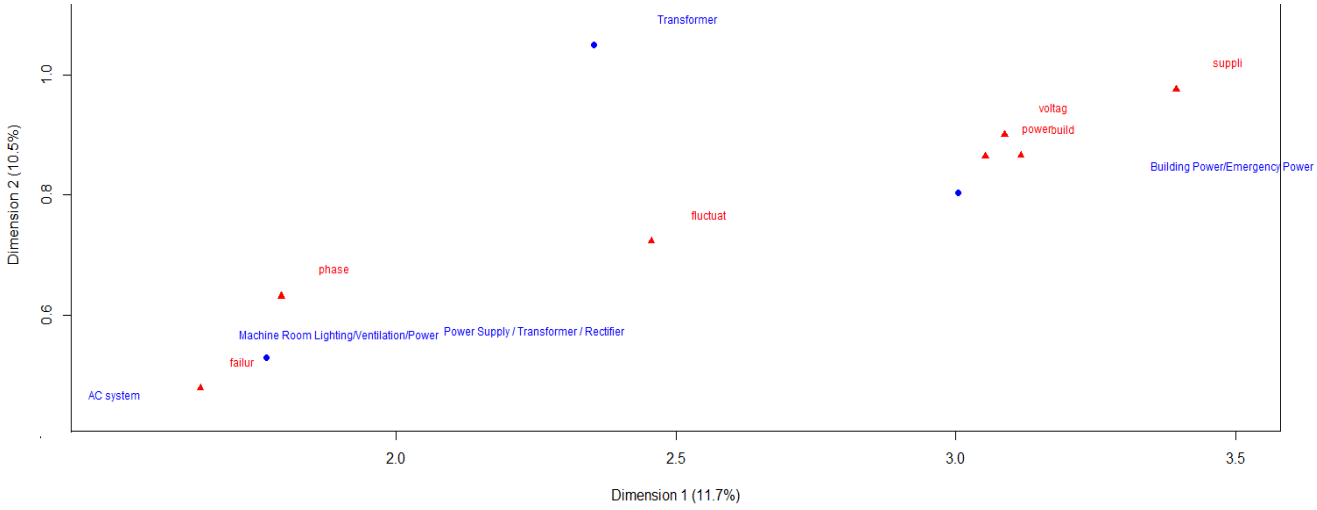


Figura- 5.8 Ampliación en la zona de fallo del suministro eléctrico

En estas imágenes se observa la aparición en la misma zona (mismas coordenadas) de niveles del FaultAreaName_en que se puedan considerar equivalentes (*Transformer, Power supply/Transformer/Rectifier, Building Power Emergency/Power* en figura- 5.8. Esto indica que se emplean las mismas palabras en sus registros y, por tanto, se pueden agrupar en un solo nivel. También se puede apreciar en la figura- 5.7 el principal problema de la variable FaultAreaName_en como clasificadora de la causa del fallo. Algunos niveles agrupan fallos muy distintos y diferentes entre sí como *Step/Pallet-Roller/Insert/Tread* que incluye fallos del rodillo de cadena de peldaños, de las demarcaciones y de la estructura del peldaño. Estos

elementos se averían por motivos muy distintos y su reparación es radicalmente distinta tanto en tiempo necesario como en repuestos o medidas tomadas. Por esta razón entre otras, desde la empresa se pidió que se definiera una nueva variable FaultAreaName_en, llamada FaultAreaName_en2, cuyos niveles serán fallos perfectamente caracterizados en cuanto a duración estimada, repuestos y medidas correctoras. El método de obtención de dicha variable se describirá posteriormente en clasificación del área de fallo.

Se muestra el análisis de correspondencia y la corrección realizada por medio de código en http://bellman.ciencias.uniovi.es/~raul/Acondicionamiento.html para algunas de las variables y en formato tabla para las demás en la tabla 5.3. La reducción de categorías conseguida se puede ver en tabla 5.4.

Pais	Nivel antiguo	Nuevo nivel
	UAEDIA	UAE

Fabricante	Nivel antiguo	Nuevo nivel
	FUJI	FUJITEC
	XIZI OTIS	OTIS
	LGSIGMA Y SIGMA	LG
	ORONA,HYUNDAI, BORAL/J&W, JOHNSON, 3RD PARTY SUPPLIERS, LOCAL SUPPLIER, CNIM, YIDA, TOSHIBA, HITACHI, DONGYANG, LG	OTHER

Segmento	Nivel antiguo	Nuevo nivel
	Entertainment/Leisure	Retail
	Education/Religion, Hospital/Healthcare,Industrial, Office, Parking garage, Private residential, Public residential, Urban mobility (footbridge)	Others

Unidad motriz	Nivel antiguo	Nuevo nivel
	Single Speed A.C, AC-1	Geared 1-Speed A.C
	Two speed A.C, AC-2 1&2SAC	Geared 2-Speed A.C
	Geared VVVF	VVVF
	VFR,VVVF,ACVV	VF
	Hydraulic, unknown, gearless, dc-motor generator	UNKNOWN

Fábrica	Nivel antiguo	Nuevo nivel
	TKE-CHN-ELE,TKE-ELE,TKE-IND-ELE	TKE-CHN-ZS-ELE
	EMFRANCE, TKE-FRA-ELE, TKE-JOR-ELE	TKE-GER-HH-ESC
	OTHERS, RDPARTYFACTORY,TURKLIFT,CNIM, HYUNDAI, KONE, MITSUBISHI, ORONA, SCHINDLER, YIDA, OTIS	UNKNOWN

Tabla 5.3 Agrupación de niveles en algunas de las variables factor

Variable	Niveles originales	Niveles para análisis
Pais	19	18
Fabricante	30	8
Unidad motriz	14	5
Controller	70	4
Modelo	76	48
Fabrica	35	8
Area de fallo	162	22
Segmento	13	4

Tabla 5.4 Reducción de categorías en las variables tipo factor

5.3.3.- Creación de variables (*feature engineering*)

El siguiente paso en el proceso de análisis de datos es la creación de variables características, también llamado *feature engineering*. Esta etapa consiste en crear, ya sea por medio de

transformaciones matemáticas o por combinación de distintas variables originales, nuevas variables que faciliten la interpretación de los datos y la creación posterior de modelos [25].

Esta fase es fundamental ya que permite extraer la información útil contenida en los campos primigenios, disminuyendo en gran medida el tiempo que será necesario para comprender qué variables afectan a los datos y qué modelos desarrollar. A continuación, se detallarán las variables creadas para este trabajo:

Variables continuas: son definidas a partir de las variables numéricas corregidas en 5.3.2.- y las variables tipo fecha. Se dividen en variables geométricas (ángulo y tiempo de viaje), de mantenimiento (tiempo de mantenimiento), de fallo (duración, tiempo de reacción y tiempo no disponible) y de mantenimiento correctivo del equipo (tiempo en reparación, tasa de fallos y MTBF). Estas variables se definen de la siguiente manera:

- **Variables de fallo:** son aquellas relacionadas con cada incidente individual. Son:

  1.Duración de reparación: tiempo que permanece el operario revisando el equipo para solucionar el incidente. Se mide como la diferencia de fechas entre el momento en el que completa el reporte de fallo y el momento en el que llega al lugar del incidente. Se calcula en minutos.

  $$Duración\ de\ reparación = Momento\ salida - Momento\ \ llegada \quad (5.1)$$

  2.Tiempo de reacción: tiempo que tarda el operario en llegar al lugar del incidente. Se mide como la diferencia de fechas entre el momento en que llega al lugar del incidente y el momento en el que se registra la llamada. Se calcula en minutos. Esta variable tiene outliers por lo que se establece un intervalo válido de 0 a 500 minutos.

  $$Tiempo\ de\ reacción = Momento\ \ llegada - Momento\ llamada \quad (5.2)$$

  3.Tiempo no disponible (tiempo parado): tiempo en el que el equipo no está disponible para su uso debido al incidente. Es la base para calcular la disponibilidad expuesta en 4.1.-. Se computa como la suma de la duración de reparación y el tiempo de reacción. Se mide en minutos. Se muestra a continuación el código con el que se calculan estas tres variables en figura- 5.9.

  $$Tiempo\ no\ disponible = Duración + Tiempo\ de\ reacción \quad (5.3)$$

```
D$TravelDistance=sqrt(D$RisePerMetre^2+D$ESClength^2)
D$TravelTime=sqrt(D$RisePerMetre^2+D$ESClength^2)/D$Speed

#Tiempo de reaccion

D$reactiontime=as.numeric(D$ArriveDate-D$LoggedDate,unit="secs")/60
D$reactiontime[D$reactiontime<0]=20
D$reactiontime[D$reactiontime>=1000]=500

#Tiempo de mantenimiento

D$MaintenanceTime=D$VisitPerYear*D$UnitScheduleHours

#Duracion de la tarea de mantenimiento

D$Duration=D$OutDate-D$ArriveDate#da la diferencia en segundos
D$Duration2=as.numeric(D$Duration, units="secs")/60#Duracion del mantenimiento en minuto
D$tiempoparado=D$Duration2+D$reactiontime
```

Figura- 5.9 Cálculo de variables de fallo, geométricas y de mantenimiento

- **Variables de mantenimiento:** el tiempo anual de mantenimiento preventivo es el producto entre el número de visitas periódicas y el tiempo establecido para cada visita. El código se muestra en figura- 5.9

$$Tiempo\ anual\ de\ mantenimiento = Nº\ visitas \cdot Tiempo\ de\ visita \qquad (5.4)$$

- **Variables geométricas:** son aquellas relacionadas con las características constructivas del equipo:

  1-Ángulo de inclinación (α): se muestra su definición gráficamente en la figura- 5.10. Cabe destacar que existen dos zonas horizontales en las áreas de llegada superior e inferior que habrá que incluir en el cálculo del ángulo. Esta longitud (Luh y Llh) depende del uso de la escalera, pero su influencia debería ser mínima en el ángulo por lo que se coge un valor estándar de 4,7 m.

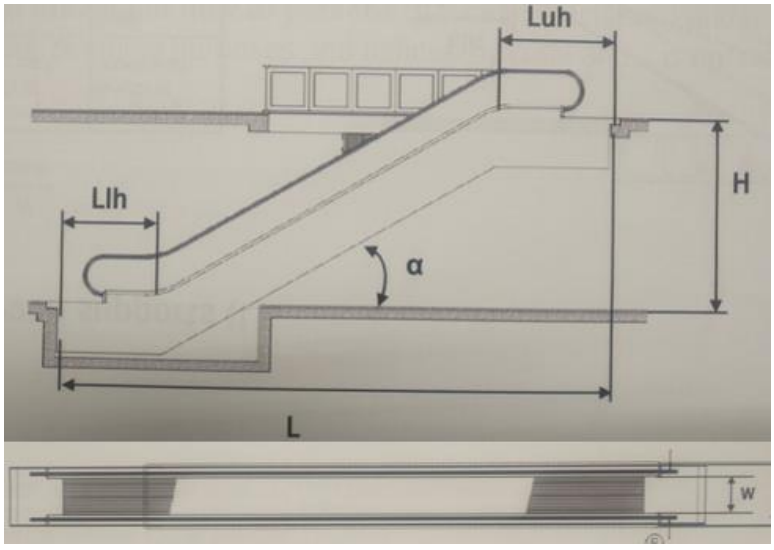


Figura- 5.10 Relación entre las variables geométricas de una escalera mecánica

Si se representa el desnivel en función de la distancia entre apoyos como en la figura- 5.11, se puede observar que existen unos ángulos preferentes, que son las pendientes de las alineaciones de puntos. Estos deberían ser cercanos a 0°; 12°; 27,5°; 30° y 35° ya que son los valores estándar de inclinación. Por tanto, sería interesante ver cómo de frecuente es cada uno y si tiene alguna relación la inclinación con el tipo de fallo más habitual. Este es el motivo por el que se considera necesario calcular el ángulo de inclinación

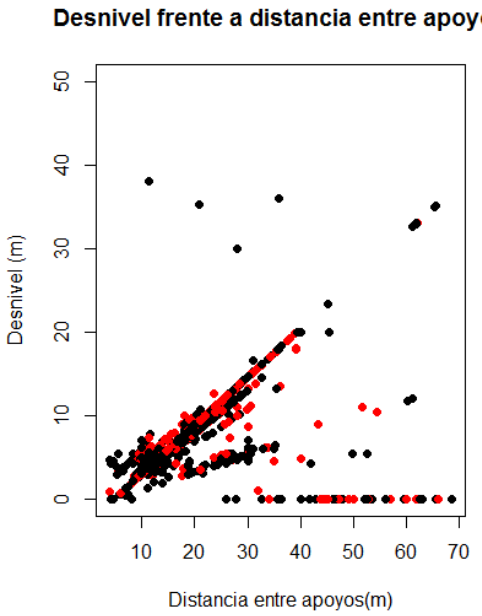


Figura- 5.11 Desnivel en función de la distancia entre apoyos

$$Ángulo(\alpha) = \text{atan}\frac{Desnivel\ (H)}{Distancia\ entre\ apoyos\ (L) - Luh - Llh} \qquad (5.5)$$

2-Tiempo de viaje: conviene contar con una variable que represente cómo de grande es la escalera. Se ha decidido que esta variable sea el tiempo de viaje ya que además incluye datos sobre la velocidad. Se define el tiempo de viaje cómo el tiempo que tarda un usuario que permanece inmóvil sobre el escalón en ir de una zona de llegada a la otra.

$$Tiempo\ de\ viaje = \frac{\sqrt{Desnivel^2 + Distancia\ entre\ apoyos^2}}{Velocidad} \qquad (5.6)$$

La definición que se muestra en la ecuación 5.9 sería más precisa si se incluyeran los tramos horizontales, no obstante, dado que solo se va a emplear esta variable para comparar unas escaleras con otras sin ser relevante su valor concreto, se opta por la

definición más sencilla posible. Se muestra en figura- 5.9 el código necesario para calcular las variables geométricas.

- **Variables de mantenimiento correctivo de equipo:** aquellas relacionadas con la fiabilidad, mantenimiento y disponibilidad de cada equipo. Solo tienen sentido hablando de equipos y no de cada uno de los incidentes, por lo que se calculan como variables independientes de la base de datos inicial y se integran en una tabla con las variables de equipo definidas en 5.1.3.- y las creadas en este apartado derivadas de ellas. Cabe destacar que están referidas todas al periodo que va de noviembre de 2017 a noviembre de 2018 ya que es el intervalo estudiado en este trabajo. Son:

  1-Tiempo en reparación: suma de las duraciones de reparación registradas para el equipo *n* durante el periodo considerado.

$$Tiempo\ en\ reparación = \sum_{equipo\ n} Duración_n \tag{5.7}$$

  2-Tasa de fallos: número de incidentes que ha sufrido el equipo *n* en el periodo considerado. Esta a su vez se puede dividir entre tasa de accidentes y tasa de averías, en función de si los fallos son causados por un mal uso por parte del pasajero o por un mal funcionamiento del equipo.

$$Tasa\ de\ fallos = \sum_{equipo\ n} Presencia\ fallo_n \tag{5.8}$$

  3-Tiempo medio entre fallos (MTBF): se emplea la definición proporcionada en 4.1.2.-. El MTBF es el cociente entre el tiempo desde que se produce el primer fallo hasta el final del periodo estudiado y el número de fallos que se dan en dicho periodo. Se verá un tratamiento más exhaustivo del significado y cálculo de esta variable en 6.3.2.-.

$$\mathrm{MTBF} = \frac{intervalo\ de\ tiempo\ estudiado}{número\ de\ fallos\ en\ dicho\ intervalo} = \frac{T}{n} \tag{5.9}$$

  Se muestra en la figura- 5.12 el código necesario para crear dichas variables e integrarlas en la tabla de equipos. En él se muestran otras variables internas que se precisan para el cálculo, pero que no son necesarias para el análisis posterior de los datos, por lo que no se definen en este apartado.

```
#%%%%%%%%%%%%%%%%%%%%%%%%% CREACION TABLA DE EQUIPOS CON VARIABLES DE FALLOS %%%%%%%%%%%%%  MinDa
te=min(D$LoggedDate)
MaxDate=max(D$LoggedDate)#ULTIMA FECHA DE LA QUE SE TIENEN DATOS

#DEFINICION DE LAS VARIABLES DE FALLOS

RealFaults_Tasafallo=sort(by(D$accident,D$UnitNumber,function(x)sum(x=="averia",na.rm = TRUE
FakeFaults=sort(by(D$accident,D$UnitNumber,function(x)sum(x=="accidente",na.rm = TRUE
Faults=sort(by(D$accident,D$UnitNumber,function(x)length(x)))#FALLOS TOTALES (REALES+MAL USO)

Falibilidad=sort(by(D$accident,D$UnitNumber,function(x)
sum(x=="averia",na.rm = TRUE)/length(x)))#PROPORCION DE TODOS LOS FALLOS QUE SON REALES
Stoppedtime=sort(by(D$Duration2,D$UnitNumber,function(x)sum(x,na.rm = TRUE)))#TIEMPO PARADO
Stoppedtime2=sort(by(D$Duration2,D$UnitNumber,function(x)sum(x,na.rm = TRUE)/length(x)))#MEDIA
Stoppedtimereal=sort(by(D$Duration2[D$accident=="averia"],D$UnitNumber[D$accident=="averia"],
function(x)sum(x,na.rm = TRUE)))#TIEMPO QUE SE HA PASADO LA MAQUINA PARADA POR FALLO

Stoppedtimeaccident=sort(by(D$Duration2[D$accident=="accidente"],
D$UnitNumber[D$accident=="accidente"],function(x)sum(x,na.rm = TRUE)))#TIEMPO QUE SE HA PASADO
LA MAQUINA PARADA POR FALLO
Reactiontime=sort(by(D$reactiontime,D$UnitNumber,function(x)sum(x,na.rm = TRUE)))#TIEMPO QUE S
E HA PASADO LA MAQUINA PARADA POR FALLO

#TIEMPO MEDIO ENTRE FALLOS

MTBF=sort(by(D$LoggedDate,D$UnitNumber,function(x)sum(as.numeric(diff(c(MinDate,x,MaxDate)),un
its="days"))/length(x)))
MTBF[order(names(MTBF))]=
MTBF[order(names(MTBF))]-Stoppedtime2[order(names(Stoppedtime2))]/(60*24)
MAXTBF=sort(by(D$LoggedDate,D$UnitNumber,function(x)max(as.numeric(diff(c(MinDate,x,MaxDate)),
units="days"))))
MINTBF=sort(by(D$LoggedDate,D$UnitNumber,function(x)min(as.numeric(diff(c(MinDate,x,MaxDate)),
units="days"))))#EL MIN PERMITE VER OUTLIERS NEGATIVOS
TODOStBF=(by(D$LoggedDate[D$NoFault==0],D$UnitNumber[D$NoFault==0],function(x)(as.numeric(diff
(c(x)),units="days"))))#EL MIN PERMITE VER OUTLIERS NEGATIVOS

#CREACION DE TABLA DE EQUIPOS

Deq=D
Deq=Deq[ ,colnames(D)=="X.U.FEFF.CountryName_en"|colnames(D)=="newyear"|colnames(D)=="Building
Name en"|colnames(D)=="UnitNumber"|colnames(D)=="YearManufactured"|colnames(D)=="NewLustro"|co
lnames(D)=="NewDecade"|colnames(D)=="UnitType_en"|colnames(D)=="factorycountry"|colnames(D)=="
newmanufacturer"|colnames(D)=="Manufacturer_en"|colnames(D)=="DriveType_en"|colnames(D)=="Fact
ory_en"|colnames(D)=="Controller_en"|colnames(D)=="Model_en"|colnames(D)=="Speed"|colnames(D)=
="RisePerMetre"|colnames(D)=="BuildingTypeSub_en"|colnames(Deq)=="SegmentType_en"|colnames(D)=
="ESClength"|colnames(D)=="ESEsetpwidth"|colnames(D)=="EscalatorDrive"|colnames(D)=="Escalator
OutdoorUnit"|colnames(D)=="Angle"|colnames(D)=="MaintenanceTime"|colnames(D)=="Heavyduty"|coln
ames(D)=="Heavypunctual2"|colnames(D)=="Balaustrade"|colnames(D)=="Handrailperfil"|colnames(D)
=="ModeloBase"|colnames(D)=="ExpectedUse"|colnames(D)=="NumSerie"|colnames(D)=="Lustro"|colnam
es(D)=="Decade"|colnames(D)=="Day"|colnames(D)=="TravelTime"]

Deq=Deq[!duplicated(Deq$UnitNumber),]

#ASIGNACION DE LAS VARIABLES DE FALLO A LA TABLA DE EQUIPOS

Deq$Faults[order(Deq$UnitNumber)]=as.numeric(Faults[order(names(Faults))])
Deq$RealFaults_Tasafallo[order(Deq$UnitNumber)]=RealFaults_Tasafallo[order(names(RealFaults_Ta
safallo))]
Deq$FakeFaults[order(Deq$UnitNumber)]=FakeFaults[order(names(FakeFaults))]
Deq$Falibilidad[order(Deq$UnitNumber)]=Falibilidad[order(names(Falibilidad))]
Deq$Stoppedtime[order(Deq$UnitNumber)]=Stoppedtime[order(names(Stoppedtime))]
Deq$Stoppedtimereal=0
Deq$Stoppedtimereal[match(names(Stoppedtimereal),Deq$UnitNumber)]=Stoppedtimereal
Deq$Timereaction[order(Deq$UnitNumber)]=Reactiontime[order(names(Reactiontime))]
Deq$MTBF[order(Deq$UnitNumber)]=MTBF[order(names(MTBF))]
Deq$MAXTBF[order(Deq$UnitNumber)]=MAXTBF[order(names(MAXTBF))]
Deq$MINTBF[order(Deq$UnitNumber)]=MINTBF[order(names(MINTBF))]
Deq$Timenoworking=Deq$Stoppedtime+Deq$Timereaction #SUMA DE TIEMPOS DETENIDOS
Deq$disponilibidadmant=1-Deq$Timenoworking/(365*24*60)*100#Porcentaje del tiempo usable
```

Figura- 5.12 Código para calcular las variables de mantenimiento correctivo del equipo

Variables factor: definidas a partir de las variables factor corregidas en 5.3.2.- y las variables tipo fecha. Se dividen en variables de fecha de fabricación (lustro, década), variables de fecha de fallo (día semanal, mes y día anual), variables comerciales (modelo base, uso esperado, fabricante y país de fabricación) y constructivas (balaustrada y perfil bajo pasamanos y número de serie). Estas variables se definen de la siguiente manera:

- **Variables de fecha de fabricación:** el año de fabricación, incluso empleando el año de inicio de mantenimiento para completar registros vacíos, sigue presentando un gran número de registros no disponibles. Por ello, se decide agrupar en lustros y décadas los años de fabricación de los equipos para que sean más representativas las series temporales. Se muestra el código en figura- 5.14.
- **Variables de fecha de fallo:** a partir de la fecha de registro de incidente (Logged Date) se extrae el día de la semana, el mes y el día del año en el que se produce este fallo. Estas variables permitirán encontrar patrones periódicos que no se hallarían de otra manera. Se muestra el código en figura- 5.14.
- **Variables comerciales:** el número de modelos existentes es excesivo (más de 40) y no es necesario analizarlos por separado ya que se puede agrupar por una serie de rasgos comúnes más sencillos de analizar. En primer lugar, solo se incluyen los modelos de Thyssenkrupp ya que no interesa analizar los de la competencia ahora mismo. Todos los modelos de Thyssenkrupp siguen la misma denominación que se muestra en la figura- 5.13:

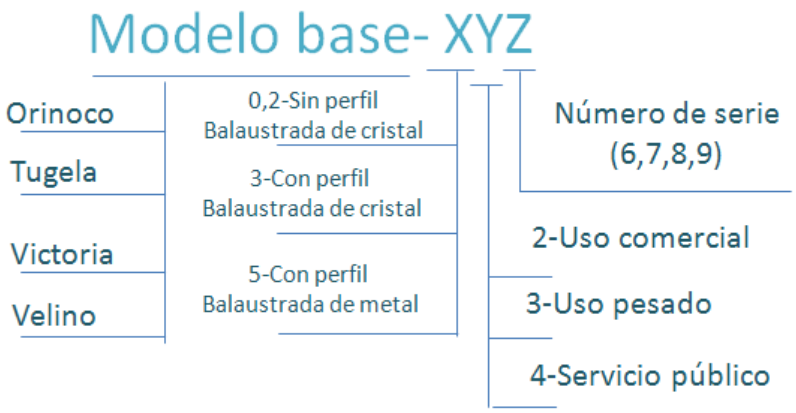


Figura- 5.13 Denominación de los modelos de Thyssenkrupp

Por medio de expresiones regulares (regex) se puede extraer cada una de las partes de la denominación, creando con ellas distintas variables (Modelo base, tipo de balaustrada, presencia de perfil bajo pasamanos y número de serie) que codifiquen esta información.

Finalmente, también se crea una variable de fabricante origen, combinando la de fabricante y la de factoría. Esta fue una petición de Thyssenkrupp para tratar como si fueran fabricantes

independientes sus tres fábricas principales (China, España y Alemania). Se muestra el código necesario para obtener estas variables en la figura- 5.14.

```
#Tipo de uso para el que fue fabricado (2 numero de FT XXX)

D$ExpectedUse=substr(D$Lastnumber2,1,1)
D$ExpectedUse=gsub("[^2-4]",NA,D$ExpectedUse)
D$ExpectedUse=gsub("2","Comercial",D$ExpectedUse)
D$ExpectedUse=gsub("3","Heavy duty",D$ExpectedUse)
D$ExpectedUse=gsub("4","Public use",D$ExpectedUse)
D$ExpectedUse=factor(D$ExpectedUse)

#Numero de serie (indicador de antiguedad y lugar de procedencia) (1 numero de FT XXX)

D$NumSerie=substr(D$Lastnumber3,1,1)
D$NumSerie=gsub("[^6-9]",NA,D$NumSerie)
D$NumSerie=factor(D$NumSerie)
D$Model_en=factor(D$Model_en)#No se si hace falta volverlo a pasar a factor
D <- D[ ,!colnames(D)=="Lastnumber2"]
D <- D[ ,!colnames(D)=="Lastnumber3"]

#Dia y mes del incidente (con logged date)

ordenday=c("lunes","martes","miércoles","jueves","viernes","sábado","domingo")
D$Day=factor(weekdays(D$LoggedDate),levels=ordenday)
ordenmes=c("enero","febrero","marzo","abril","mayo","junio","julio","agosto","septiembre",
"octubre","noviembre","diciembre")
D$Month=factor(months(D$LoggedDate),levels=ordenmes)
D$Hour=as.integer(strftime(D$LoggedDate, format = "%H"))
D$Dayyear=as.integer(strftime(D$LoggedDate, format = "%j"))
D$TravelDistance=sqrt(D$RisePerMetre^2+D$ESClength^2)
D$TravelTime=sqrt(D$RisePerMetre^2+D$ESClength^2)/D$Speed
#Year de inicio de mantenimiento por parte de Thyssen
D$maintenanceyearstart=as.integer(strftime(D$HandOverDate, format = "%Y"))
D$newyear=D$YearManufactured
D$maintenanceyearstart2=D$maintenanceyearstart
D$maintenanceyearstart2[D$maintenanceyearstart2<=1980|D$maintenanceyearstart2>=2019]=NA
D$newyear[which(is.na(D$newyear)&D$Manufacturer_en=="THYSSENKRUPP")]=
D$maintenanceyearstart2[which(is.na(D$newyear)&D$Manufacturer_en=="THYSSENKRUPP")]-1
D$newyear[D$newyear<2000|D$newyear>2018]=NA
D <- D[ ,!colnames(D)=="maintenanceyearstart2"]

D$NewLustro=D$newyear

D$NewLustro=gsub("1980|1981|1982|1983|1984","1980-1984",D$NewLustro)
D$NewLustro=gsub("1985|1986|1987|1988|1989","1985-1989",D$NewLustro)
D$NewLustro=gsub("1990|1991|1992|1993|1994","1990-1994",D$NewLustro)
D$NewLustro=gsub("1995|1996|1997|1998|1999","1995-1999",D$NewLustro)
D$NewLustro=gsub("2000|2001|2002|2003|2004","2000-2004",D$NewLustro)
D$NewLustro=gsub("2005|2006|2007|2008|2009","2005-2009",D$NewLustro)
D$NewLustro=gsub("2010|2011|2012|2013|2014","2010-2014",D$NewLustro)
D$NewLustro=gsub("2015|2016|2017|2018|2019","2015-2019",D$NewLustro)
D$NewLustro=factor(D$NewLustro)
D$NewDecade=D$NewLustro

D$NewDecade=gsub("1980-1984|1985-1989","1980",D$NewDecade)
D$NewDecade=gsub("1990-1994|1995-1999","1990",D$NewDecade)
D$NewDecade=gsub("2000-2004|2005-2009","2000",D$NewDecade)
D$NewDecade=gsub("2010-2014|2015-2019","2010",D$NewDecade)
D$NewDecade=factor(D$NewDecade)
```

Figura- 5.14 Código para la creación de variables categóricas

5.4.- PROCESAMIENTO DEL LENGUAJE NATURAL

Una de las variables que exige un tratamiento más particular, debido a la información que contiene, es la descripción por parte del operario de la causa del fallo y las medidas tomadas, llamada Action_en. El contenido de esta variable, al contrario que todas las vistas hasta ahora, que eran numéricas o de una categoría concreta dentro de una lista de opciones, es abierta. Es decir, el operario puede escribir lo que desee en cuanto a extensión y variedad de terminos. Por tanto, cada registro será distinto y tendrá pocas cosas en común con los demás. Además, se emplean distintos idiomas lo que complica aún más la tárea a realizar.

Sin embargo, como ya se ha mencionado en otras ocasiones, es necesario contar con una estructura estándar y normalizada para poder realizar un correcto análisis de los datos. Es aquí donde entra en juego el procesamiento del lenguaje natural (NLP), que es un conjunto de técnicas que intentan extraer el significado de textos escritos, despreciando toda la información no esencial. El NLP se ha desarrollado ampliamente en los últimos años logrando muy buenos rendimientos en tareas como la generación de texto (algoritmos para sustituir labores repetitivas de periodistas), creación de resúmenes, formulación de preguntas, análisis de sentimientos, traducción o, la que se va a emplear en este trabajo, la categorización de textos en función de su contenido.

Así pues, se aplicarán una serie de procedimientos específicos para conseguir acondicionar la variable Action_en. Los principales pasos que se han seguido se detallan a continuación.

5.4.1.- Traducción automática

Los registros, como ya se ha comentado anteriormente, provienen de 18 países distintos, 9 con reportes escritos en inglés y que no será necesario traducir (Australia, Bahréin, India, Kuwait, Malasia, Qatar, Arabia Saudi, Singapur y Dubái) y otros 9 (China, Egipto, Hong Kong, Indonesia, Jordania, Corea, Taiwán, Tailandia y Vietnam) con reportes escritos en 6 idiomas distintos (árabe, chino (tradicional y simplificado), tailandés, vietnamita, coreano e indonesio).

Por tanto, el primer paso consiste en traducir todos los registros al inglés para que los procesos posteriores, que implican transformaciones léxicas y semánticas, puedan realizarse de manera más sencilla. Esta labor constituye en sí misma uno de los campos más complejos del procesamiento del lenguaje natural y está totalmente fuera del alcance de este trabajo crear un algoritmo de traducción automática. Por tanto, se empleará alguno de los distintos motores de traducción, con acceso más o menos libre, que hay actualmente online.

Tras realizar un proceso de búsqueda se ha visto que existen tres grandes alternativas:

1. Pagar una licencia para usar la API (interfaz de programación de aplicación) de Google (*Translate*) o Microsoft (*Bing*). Esta opción tiene la ventaja de su sencillez, al estar perfectamente integrada con el entorno de R. Sin embargo, debido al volumen de texto con el que se trabaja, para el cuál se estima un coste de traducción de unos 500$, se decide descartar inicialmente el uso de estas APIs inicialmente hasta explorar las otras alternativas.
2. Emplear APIs o versiones web gratuitas, como Yandex, DeepL o Apertium. En este caso, se evita el coste monetario, pero existe una gran limitación: estos traductores, en su mayoría, solo trabajan con lenguas latinas o anglosajonas. Este hándicap obliga a descartar estas aplicaciones, ya que algunos registros están escritos en árabe y otras lenguas asiáticas, para las cuales estos motores no ofrecen traducciones. Además, la conexión con la versión web, al contrario que con la API, es compleja y no sería sencillo traducir un gran volumen de datos.
3. Emplear un traductor por línea de comandos, como *Translate Shell*, que permite emplear las versiones web (gratuitas) de distintos motores de traducción, pero sin la dificultad para trabajar con grandes volúmenes de datos que implicaban en principio estas páginas web. Es decir, esta opción permitirá usar *Google Translate* o *Bing*, que son los únicos válidos para los idiomas que se están manejando y, al emplear su página web en vez de su API, la traducción debería ser gratuita. Como se verá a continuación, esto no es totalmente cierto. No obstante, el empleo de *Translate Shell* sigue representando la mejor opción de todas las encontradas.

Requerimientos para *Translate Shell*: este traductor por línea de comandos es en realidad una aplicación en GNU Awk. En consecuencia, ha sido necesario contar con un equipo con Ubuntu o algún otro sistema operativo basado en Linux. En este caso, se ha empleado un ordenador portátil con Windows 10 en el que se ha instalado Ubuntu 18.04 desde la tienda de Microsoft, ya que una de las últimas actualizaciones de Windows 10 ha permitido la compatibilidad entre ambos sistemas, y una *Raspberry Pi* que se ha empleado como apoyo y para realizar algunas pruebas previas.

Procedimiento de uso y limitaciones de *Translate Shell*: la funcionalidad básica de *Shell* permite llamar a las versiones web de los traductores desde la consola del ordenador y obtener el texto traducido también por línea de comandos, lo que facilita enormemente el manejo de los datos. El flujo de trabajo que se empleará se puede ver en figura- 5.15.

Figura- 5.15 Flujo de datos para la traducción automática de registros

Antes de detallar este procedimiento, es imprescindible comentar alguno de los problemas que se han encontrado durante su programación y las soluciones que se han aplicado, puesto que han condicionado en gran medida la manera final en la que se realizará la traducción.

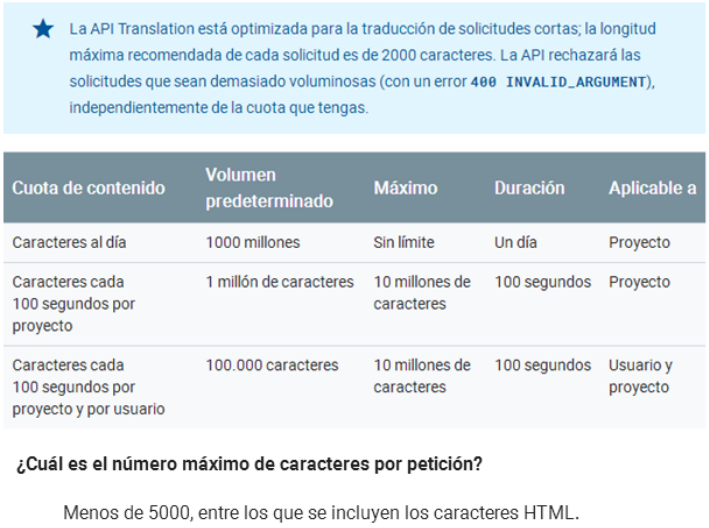
Todas estas dificultades derivan del motor de traducción escogido. Tal y como se ha comentado previamente, los únicos válidos —por incluir los idiomas necesarios— eran *Google Translate* y *Bing Translate*. Inicialmente, se llevó a cabo una prueba con unos registros tipo para ambos, cuyo resultado se muestra en tabla 5.5. Se comprobó que *Google Translate* ofrecía una traducción más precisa y completa, criterio por el que fue elegido para realizar todas las traducciones.

Id registro	Google translate	Bing
7205	when you arrive at the site, the stairs are found to be working, inspected and in service	on reaching the site, he found the stairs working, and he was lost in service.
1245	"KTV check ladder with mechanical error, processed, normal operation ladder."	"KTV Test ladder failed to win, processed, normal operation ladder. "
2119	"Check the ladder with a sound because there are strange objects stuck in the tooth, causing rubbing with the step, removed the strange object, the normal operation ladder."	"The Ladder is a cry from a strange material to a comb that is rubbed with a rank, has removed foreign objects, ladders normal operation. "

Tabla 5.5 Comparación de traducciones entre Google translate y Bing

Sin embargo, al profundizar en su uso se observó que existían limitaciones en la cantidad de texto que se podía introducir y en la frecuencia con la que se podían pedir traducciones. Después de distintas pruebas se comprobó que *Google Translate* bloqueaba las peticiones muy voluminosas o reiterativas, identificando y censurando la dirección IP durante un periodo estimado de una hora, en el cual cualquier petición de traducción, por mínima que fuera, era rechazada. Esto era algo esperable, ya que se está pidiendo a la aplicación web gratuita de *Google Translate* que traduzca volúmenes de datos para los que no está

diseñada. Para esta tarea se diseñó la API de pago, que también cuenta con limitaciones, pero mucho menos estrictas, mostradas en figura- 5.16.



★ La API Translation está optimizada para la traducción de solicitudes cortas; la longitud máxima recomendada de cada solicitud es de 2000 caracteres. La API rechazará las solicitudes que sean demasiado voluminosas (con un error `400 INVALID_ARGUMENT`), independientemente de la cuota que tengas.

Cuota de contenido	Volumen predeterminado	Máximo	Duración	Aplicable a
Caracteres al día	1000 millones	Sin límite	Un día	Proyecto
Caracteres cada 100 segundos por proyecto	1 millón de caracteres	10 millones de caracteres	100 segundos	Proyecto
Caracteres cada 100 segundos por proyecto y por usuario	100.000 caracteres	10 millones de caracteres	100 segundos	Usuario y proyecto

¿Cuál es el número máximo de caracteres por petición?

Menos de 5000, entre los que se incluyen los caracteres HTML.

Figura- 5.16 Información sobre los límites de Google Translate obtenida de su apartado de FAQ

Está limitación ha obligado a tomar dos medidas para que la traducción tuviera lugar de manera continua sin intervención humana. Por un lado, los archivos csv iniciales de cada país, que podían sumar entre 10000 y 300000 caracteres cada uno (en total sumaban en torno a medio millón de caracteres repartidos en unos 40000 registros), han tenido que ser divididos en porciones más pequeñas, con unos mil caracteres cada uno (menos de los 5000 que acepta la API ya que la versión web gratuita se ha comprobado más estricta tras varios ensayos). Por otro lado, para que las peticiones no fueran reiterativas, se ha establecido un tiempo de espera entre dos peticiones consecutivas. Este tiempo fue ajustado también por ensayo y error, intentando que fuera el mínimo posible para que la traducción no llevara mucho tiempo, pero que evitara que *Google Translate* bloqueara la IP.

Ahora que se conocen los condicionantes existentes se procede a detallar el flujo de información y cómo se ha programado.

- Se crea un bucle en R para exportar en csv los registros de aquellos países cuyos textos no están en inglés. También se exporta un identificador que facilite posteriormente su incorporación a la base de datos. Se ha decidido usar csv, en vez de la función *system* de R, que podría haber llamado directamente a *Translate Shell*, con el objetivo de poder tener los archivos en equipos que no tengan R instalado, como la *Raspberry PI* y así es también más sencillo distribuir la traducción en varios ordenadores para agilizar el proceso.

```
#%%%%%%%%%%%%%%% EXPORTAR LOS DATOS A TRADUCIR  %%%%%%%%%%%%%%%
países atraducir=c("China","Egypt","Hong Kong","Indonesia","Jordan",
"Korea","Taiwan","Thailand","Vietnam")
for (i in países atraducir)
{Datraducir=D[D$X.U.FEFF.CountryName_en==i,c("id","Action_en")]
   write_excel_csv(as.data.frame(Datraducir),paste0(c(i,".csv"),colla
pse = ""))}
```

Figura- 5.17 Bucle para exportar los datos en R

- Se programa un bucle en la consola de comandos de Linux que realice las siguientes funciones para cada país (el bucle es independiente para cada país por lo que debe inicializarse manualmente en cada caso):
    1. Dividir el csv en porciones con unos mil caracteres cada una, lo que representa unos 10-30 registros cada uno. El número de registros de cada porción depende del país por lo que fue ajustado en función del idioma (una palabra en coreano o chino requiere muchos menos caracteres que en vietnamita o indonesio por lo que 1000 caracteres son más registros en chino, y no se podía emplear un límite inferior común a todos para poder optimizar el tamaño de cada porción y así disminuir el número de peticiones a realizar).
    2. Emplear *Translate Shell* para llamar a *Google translate* y escribir los registros traducidos en otro csv almacenado de manera local en el equipo.
    3. Esperar un tiempo aleatorio entre 2 y 3 minutos, para que el proceso no fuera determinista y fuera más difícil de detectar para *Google Translate* la reiteración de peticiones, y volver al paso 2.
    4. Unir todas las porciones de csv traducidas creando un csv traducido para cada país.
    5. Adaptar el formato de dichos csv (quitar comillas, cambiar comas por puntos) para facilitar su lectura en R.

```
--------------------INSTALACION-------------------------------------
wget git.io/trans
chmod +x ./trans
cd /mnt/c/Users/Usuario(como se llame tu usuario)/carpeta donde metas los archivos que te pase

---------------------BUCLE--------------------------------
split -l 20 DTaiwanid.csv;
for fichero in x??;
do trans -s zh-TW -e google -b -i $fichero -o ${fichero}_en.csv;
echo hecho ;
sleep $(shuf -i 180-220 -n 1);
 echo hecho $fichero;
done;
sed 's/,/:/g' Dtaiwantrans.csv >> Djordantrans2.csv;
sed 's/:/,/' Djordantrans2.csv >> Djordantrans3.csv
```

Figura- 5.18 Bucle en Linux para Translate Shell

- Se crea un bucle en R para leer los csv traducidos y añadirlos a la base de datos en la posición adecuada.

```
#################### UNIR LA TRADUCCION Y EL ORIGINAL ##########################################
archivostraducidos=c("Dvietnamtrans3.csv","Djordantrans3.csv","Dindonesiatrans3.csv","
Degypttrans3.csv","Dthailandtrans3.csv","Dhongkongtrans3.csv","Dtaiwantrans3.csv","Dko
reatrans3.csv","Dchinatrans3.csv")
D$Action_en=as.character(D$Action_en)
for (i in archivostraducidos){
  Dtraducido=read.csv(i,encoding="UTF-8")
  colnames(Dtraducido)=c("id","translated")
  D$Action_en[match(Dtraducido$id,D$id)]=as.character(Dtraducido$translated)}
```

Figura- 5.19 Bucle para importar datos en R

Debido a los tiempos de espera y el pequeño tamaño de las particiones la traducción automática se ha completado en un tiempo estimado de una semana, con un ordenador traduciendo a tiempo completo todo el día y la *Raspberry* en momentos puntuales. Por tanto, se puede concluir que este método ha permitido ahorrar el coste monetario, pero ha consumido una gran cantidad de tiempo. Por tanto, si se decide implementar el resultado de este trabajo para dar servicio a los clientes será necesario comprar una licencia de API de *Google Translate* para que, aunque siga sin ser en tiempo real, la traducción pueda tener lugar en un tiempo que se cuente por minutos en vez de por días.

Finalmente, se ha comprobado que el resultado de la traducción es bastante satisfactorio, si bien es necesario añadir un postprocesado en R para mejorarla, debido al uso de palabras técnicas que no son correctamente traducidas en determinados idiomas por *Google Translate*. Como ejemplo se mencionan las siguientes traducciones en tabla 5.6:

Pais	Texto original	Traducción original de Google	Traducción al español	Significado real en español
Hong Kong	紅蟑螂	Red cockroach	Cucaracha roja	Pulsador de emergencia
Taiwan	級聯	Cascade	Cascada	Peldaños
Korea	Laura/Lola	Laura/Lola	Laura/Lola	Rodillo
China	猪嘴	Pig mouth	Boca de cerdo	Flap de seguridad del pasamanos
Hong Kong	鼠洞	Mouse hole	Agujeron de ratón	Flap de seguridad del pasamanos
China	蜻蜓	Dragonfly	Libélula	Balaustrada

Tabla 5.6 Algunas traducciones conflictivas

Dado que inicialmente, cuando se leyeron por primera vez los registros traducidos, no se sabía qué podían significar dichas palabras fue necesario un gran trabajo de investigación y también creativo para poder darles sentido a dichos términos, lo que ha consumido una considerable cantidad de tiempo.

5.4.2.- Limpieza del texto

El siguiente paso, una vez que ya se tienen todos los registros en inglés, consiste en lograr una estructura homogénea y estándar. Al igual que se hizo con el resto de las variables, se procederá a limpiar los reportes eliminando aquellos elementos cuya información es irrelevante e impiden que se identifiquen como equivalentes palabras que para este análisis lo son.

Cabe destacar que en este caso se cuenta con una gran ventaja. El lenguaje que emplean los operarios de mantenimiento se caracteriza por ser breve, conciso y repleto de tecnicismos que tienen como fin reducir las ambigüedades que se generan al emplear el lenguaje natural. Por tanto, no se suelen emplear palabras con varios significados, ni mucho menos frases con doble sentido o estructuras complejas. Esto permitirá simplificar al máximo el texto sin un gran riesgo de pérdida del contexto y del significado conjunto, ya que cada palabra suele tener sentido por sí misma. Para ello, se siguen las siguientes transformaciones:

- **Uniformizar formato:** en primer lugar, se modificarán los registros para que todos empleen la misma estructura de término, espacio en blanco, término. Para ello:
    1. Se pasará todo el texto a minúsculas ya que R es caso sensitivo y se podrían identificar como distintas palabras que son iguales.
    2. Se eliminarán los espacios en blanco al inicio y fin de frase y los espacios en blanco múltiples.
    3. Se eliminan también los signos de puntuación.
    4. Se eliminan todos los números, ya que se ha comprobado que en la mayor parte son identificador de la escalera o del momento en el que el operario realiza alguna tarea, información ya codificada en otras variables. Sin embargo, hay una excepción: en algunas ocasiones el equipo cuenta con un plc con diagnóstico y el empleado anota el código estándar de error que aparece. Esta información es de enorme utilidad, puesto que en esos registros se conoce con seguridad la causa de fallo. Por esta razón, es importante conservar esos números y es relativamente sencillo hacerlo ya que vienen asociados a la palabra code, node o error. Se crea así una nueva variable “error” que indica, si se emplea un código estándar, cuál es el que aparece.

  Finalmente, cabe mencionar que estos códigos estándar no son habituales, y solo están presentes en unos 2000 de los 70000 registros existentes por lo que aún será necesario establecer algún método para obtener la causa de fallo a partir del texto. De todas formas, ya que se tiene esta información en 2000 registros, se asigna automáticamente el área de fallo en función del error indicado. Se muestra en figura- 5.20 el código necesario para uniformizar el formato y en la figura- 5.21 para la asignación de errores.

```
D$Action_en=gsub("\\bnode ","node",tolower(D$Action_en))
D$Action_en=gsub("\\berror ","error",tolower(D$Action_en))
D$Action_en=gsub("\\bcode ","code",tolower(D$Action_en))

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% VARIABLE PARA VER CUANTO SE USAN LOS ERRORES
D$error=0
D$error[grep("node[0-9]|error[0-9]|code[0-9]|f[0-9]",D$Action_en,ignore.case = TRUE)]=
grep("node[0-9]|error[0-9]|code[0-9]|f[0-9]",D$Action_en,ignore.case = TRUE,value=TRUE)
rexp2="^(.*)(node)([0-9]*)(.*)$"

D$error[grep("node[0-9]",D$Action_en,ignore.case = TRUE)]=gsub(rexp2,"\\2\\3",D$error[grep("no
de[0-9]",D$Action_en,ignore.case = TRUE)])
rexp3="^(.*)(error)([0-9]*)(.*)$"
D$error[grep("error[0-9]",D$Action_en,ignore.case = TRUE)]=gsub(rexp3,"\\2\\3",
D$error[grep("error[0-9]",D$Action_en,ignore.case = TRUE)])
rexp4="^(.*)(code)([0-9]*)(.*)$"

D$error[grep("code[0-9]",D$Action_en,ignore.case = TRUE)]=gsub(rexp4,"\\2\\3",
D$error[grep("code[0-9]",D$Action_en,ignore.case = TRUE)])
rexp4="^(.*)(f)([0-9]*)(.*)$"

D$error[grep("f[0-9]",D$Action_en,ignore.case = TRUE)]=gsub(rexp4,"\\2\\3",
D$error[grep("f[0-9]",D$Action_en,ignore.case = TRUE)])

D$Action_en=gsub("[0-9]","",D$Action_en)#eliminar números
D$Action_en=gsub("[[:punct:]]"," ",tolower(D$Action_en))#signos de puntuación
D$Action_en=gsub(" $|^ ","",tolower(D$Action_en))#espacio inicial y final
D$Action_en=gsub("\\s{1,}"," ",tolower(D$Action_en))#espacio blanco multiple
```

Figura- 5.20 Código para uniformizar formato de textos

```
#CORRECIONES DE ERRORESD$FaultAreaName_en2=as.character(D$FaultAreaName_en2)
D$FaultAreaName_en2[grep("code02|error02|code03|error03|code05|error05|code06|error06",
D$error)]="Combplate"
D$FaultAreaName_en2[grep("code01|error01|code04|error04|code47|error47|code41|error41",
D$error)]="Emergency stop"
D$FaultAreaName_en2[grep("code07|error07|code08|error08",D$error)]="Chains"
D$FaultAreaName_en2[grep("code09|error09|code10|error10|code19|error19|code18|error18|code45|e
rror45|code46|error46",D$error)]="Handrail inlet"
D$FaultAreaName_en2[grep("code11|error11|code17|error17|code36|error36|code44|error44",
D$error)]="Floor plate switch"
D$FaultAreaName_en2[grep("code13|error13|code14|error14",D$error)]="Sensor rotura handrail"

D$FaultAreaName en2[grep("code20|error20|code12|error12",D$error)]="Step upthrust"
D$FaultAreaName_en2[grep("code15|error15|code16|error16|code21|error21|code22|error22",
D$error)]="Skirt switch"
D$FaultAreaName_en2[grep("code20|error20|code12|error12",D$error)]="Step upthrust"
D$FaultAreaName_en2[grep("code26|error26|code74|error74",D$error)]="Relay and wires"
D$FaultAreaName_en2[grep("code28|error28",D$error)]="Chains"
D$FaultAreaName en2[grep("code29|error29",D$error)]="Motor over/underspeed"
D$FaultAreaName_en2[grep("code30|error30|code31|error31|code32|error32|code33|error33|code35|
error35|code37|error37|code38|error38|code70|error70|code71|error71|code72|error72|code73|
error73",D$error)]="Brake"
D$FaultAreaName_en2[grep("code34|error34",D$error)]="Motor over/underspeed"
D$FaultAreaName_en2[grep("code42|error42",D$error)]="Power supply"
D$FaultAreaName_en2[grep("code39|error39|code82|error82|code97|error97",D$error)]=
"PLC and inverter"
D$FaultAreaName_en2[grep("code34|error34|code81|error81",D$error)]="Lubrication"
D$FaultAreaName_en2[grep("code50|error50",D$error)]="Miss step sensor"
D$FaultAreaName_en2[grep("code52|error52",D$error)]="Buggy switch"
D$FaultAreaName_en2[grep("code55|error55",D$error)]="Other safety switches"
D$FaultAreaName_en2[grep("code56|error56",D$error)]="Fire alarm"
D$FaultAreaName_en2[grep("code95|error95|code75|error75",D$error)]="Keyswitch"
D$FaultAreaName_en2[grep("code100|error100",D$error)]="sinfault"
D$FaultAreaName_en2=factor(D$FaultAreaName_en2)
```

Figura- 5.21 Asignación de área de fallo en función del error estándar indicado

- **Eliminación de "stop words":** se denomina "stop words" o palabras vacías a aquellas que no aportan significado a una frase, pero son necesarias para que sea correcta gramaticalmente como "el", "la", "un", "y" o "como" entre otras muchas.
  Es conveniente eliminar estas palabras de los registros ya que ocupan un gran espacio; sin ellas la base de datos es mucho más compacta. Se sigue una doble estrategia para excluirlas, en primer lugar, se emplea una función estándar del paquete tm que proporciona una lista de *stop words* en inglés.
  Tras su aplicación se comprueba que aún existen muchas palabras que, en este contexto, se consideran *stop words*. Por este motivo, se crea una función propia, más flexible que la de tm, en la que se puedan incluir todas las palabras que se consideren vacías. Se muestra el código necesario para completer la tarea en la figura- 5.22.

```
stopwords=c("unit","and","in","the","thbe","we","to","back","at","by","too",,"this","with","so"
,"then","also","it","i","due","of","for","from","as","on","that","after","they","there") patte
rn=paste(stopwords, collapse = '\\b|\\b')
pattern= paste0('\\b',pattern,'\\b')
D$Action en=gsub(pattern,"",tolower(D$Action en))
commonwords=c("but","per","inform","arriv","upon","be","check","work","normal","check",
"observ","arrival","found","taken","action","operation","all","ok","service","running","call",
"parti a","a","d","oper","resum")
pattern2=paste(commonwords, collapse = '\\b|\\b')
pattern2= paste0('\\b',pattern2,'\\b')
D$Action en=gsub(pattern2,"",tolower(D$Action en))
pattern3=paste(stopwords("english"), collapse = '\\b|\\b')
pattern3= paste0('\\b',pattern3,'\\b')
D$Action_en=gsub(pattern2,"",tolower(D$Action_en))
D$Action_en=stemDocument(D$Action_en)
```

Figura- 5.22 Código para eliminar "stop words" y reducir a lexemas los textos

***Lemmatization* y *stemming*:** dos de los grandes métodos para reducir la variabilidad de las palabras y realizar agrupaciones en función de su significado son la *lemmatization* y el *stemming*. Ambos son parecidos, ya que tienen como fin eliminar las desinencias que aportan un complemento de significado (género, número, tiempo verbal), pero su funcionamiento es ligeramente distinto [26]. El *stemming* se basa en cortar el final o inicio de las palabras para eliminar prefijos y sufijos y dejar solo la raíz que, si bien no existe como palabra, muchas veces, contiene el significado del término. El *stemming* es fácil de programar ya que suele emplear una lista de prefijos y sufijos habituales que aplica a todas las palabras, eliminando así las terminaciones en –s o –ing al considerarlas plural o gerundio. Su simplicidad es a la vez su mayor inconveniente ya que puede suprimir finales de palabra que no son sufijos, por ejemplo, quitar "ing" en *ring*. A consecuencia de esto, es habitual combinar el *stemming* con la *lemmatization*. Esta, a diferencia del *stemming*, reduce las palabras en función de su información morfológica, para lo cual requiere diccionarios completos que recojan el significado y origen de cada palabra que se quiera incluir. Como es fácil suponer, esto es mucho más complejo de programar al requerir más información, pero

se obtienen resultados mucho más precisos. A veces, demasiado precisos para el análisis que se quiere realizar si no se necesita tanto nivel de detalle.

En resumen, es necesario conocer bien la naturaleza de los textos y saber qué nivel de detalle se quiere conservar en las palabras para poder elegir adecuadamente a que profundidad aplicar la *lemmatization* para conservar el significado fundamental y el *stemming* para reducir la variabilidad y agrupar palabras similares. Se muestra un ejemplo en la tabla 5.7. Si se trata de un texto relativo al mar puede ser de interés conservar como diferentes las palabras "pleamar" y "bajamar", mientras en un texto más general puede que solo interese "mar" y puedan considerarse iguales ambas palabras. Por ello, lo más habitual es combinar ambas técnicas.

Stemming			Lemmatization		
Palabra	Sufijo/Prefijo	Raiz	Palabra	Información morfológica	Lexema
studies	-es	studi	studies	3º persona singular presente del verbo study	study
studying	-ing	study	studying	Gerundio de study	study
niñas	-as	niñ	niñas	Femenino plural de niño	niño
niñez	-ez	niñ	niñez	Sustantivo singular de niñez	niñez
pleamar	plea-	mar	pleamar	Sustantivo singular de pleamar	pleamar
bajamar	baja-	mar	bajamar	Sustantivo singular de bajamar	bajamar

Tabla 5.7 Ejemplos de stemming y lemmatization en inglés y castellano

En el paquete tm existe una función llamada stemDocument que combina bien las funcionalidades del *stemming* y la *lemmatization*, por lo que será empleada tal y como se ve en figura- 5.22. Posteriormente, si se considera que alguna palabra queda reducida en exceso, se modificará por código para que recupere su forma original.

Corrección ortográfica: finalmente, el último paso, aunque podría ser llevado a cabo al principio también, consiste en corregir aquellas palabras que se hayan tecleado incorrectamente. Esto es relativamente habitual en estos textos, que han sido escritos por operarios de manera rápida y sin ninguna revisión para detectar errores. Este proceso es difícil de automatizar, y si bien existen librerías en R como Hunspell que proporcionan sugerencias de corrección y podría programarse un código que escoja la opción más probable, se ha decidido no emplearlo ya que se ha visto que no obtiene los resultados esperados. Se pone como ejemplo el término "demarcation", que indica la franja amarilla de los peldaños. Debido a su longitud, muchas veces fue introducida erróneamente como "demacration" o "damercation". Al aplicar Hunspell se ofrecen como correcciones "democracy" o "democratic", claramente alejadas del término original. Esto se ha detectado también en otros términos como "sensor" o "combplate" y se achaca a que los correctores no están especializados en el vocabulario técnico y alejado del lenguaje común usado en estos reportes. Por este motivo, y aunque conlleva una mayor inversión en tiempo, se ha decidido corregir por código, sin paquetes especializados, aquellos términos erróneos más

habituales. Estas palabras se identifican, bien por lectura directa de los registros, o bien con mecanismos exploratorios, como el ranking de palabras más usadas o la función *agrep,* la cual busca términos similares a aquellos suministrados.

Además, se aprovecha para incluir en esta sección de código las traducciones mal realizadas que ya se mostraron en la tabla 5.6. Se muestra el código en http://bellman.ciencias.uniovi.es/~raul/Acondicionamiento.html.

Este proceso de limpieza ha permitido pasar de 1036595 palabras a tan solo 556877 reduciendo la variabilidad de 39879 términos distintos a apenas 21270. Así se ha reducido a casi la mitad el volumen de información y se facilita su procesamiento posterior. También hay que destacar que con el procedimiento de agrupación que se llevará a cabo en 5.4.3.- se reduce aún más la variabilidad, pasando de las 21270 mencionadas anteriormente a 6689 términos diferentes.

5.4.3.- Análisis exploratorio

Pese a que ahora la cantidad de texto es significativamente menor, sigue siendo excesiva para poder llevar a cabo offline, en un ordenador sin gran capacidad, las técnicas más habituales para la clasificación de texto. Por este motivo, hay que continuar con la reducción de texto, si bien con un enfoque ligeramente distinto al anterior. El proceso de limpieza se basaba en eliminar palabras que no aportaban contenido y unir aquellas con la misma raíz o lexema. En cambio, ahora se unirán aquellas palabras, que, aun proviniendo de distinta raíz, tienen significados suficientemente parecidos en este contexto para agruparlas. Finalmente se seleccionarán únicamente aquellas palabras más importantes, que será con las que se realizará el proceso de clasificación de textos. A continuación se detalla cómo se llevan a cabo estas tareas.

Agrupación de sinónimos: decidir qué palabras pueden considerarse sinónimos es algo tremendamente subjetivo y depende en gran parte de la intención y sentido que tenga dicha palabra en el contexto. Por ello, el primer hito fue consultar con expertos de Thyssenkrupp para que, para los principales elementos eléctricos y mecánicos de un equipo, sugiriesen qué palabras podían ser equivalentes.

Al llevar a cabo este proceso, y tras consultar también la lista de palabras más empleadas, se hizo evidente que existía un argot propio de los operadores de mantenimiento, que dependía del país de origen, con abreviaturas, siglas y nombres de marca cuyo significado los propios expertos de Thyssenkrupp desconocían.

La solución adoptada, y que ralentizó en parte esta fase, fue contactar directamente con los responsables de cada delegación y suministrarles la lista de términos indescifrables de su argot, esperando que ellos pudieran facilitar su significado. Por ejemplo, el término “sls”, resultó ser “step link switch”, o el término “arp needl” se refería al “non reversal device”.

Como se puede comprobar, era muy difícil saber a qué hacen mención si no se es miembro de esa delegación.

Una vez obtenida la respuesta a las palabras de argot, y obtenida una lista de sinónimos por parte de los técnicos de Thyssen, se procedió a agrupar los términos.

Adicionalmente a este procedimiento basado en el conocimiento de expertos, se plantea utilizar modelos estadísticos y de *Machine Learning* para llevar aún más lejos la agrupación de sinónimos. Para ello, se emplearán técnicas basadas en el contexto y en la similitud. Dos palabras serán más similares si aparecen rodeadas de las mismas palabras (mismo contexto). Como se puede observar, este empleo del contexto fue el mismo que se realizó en 5.3.2.- para agrupar niveles en FaulAreaName_en, por lo que se empleará también en este caso el análisis de correspondencia. Se muestran algunas ampliaciones del espacio vectorial, similares a las de 5.3.2.-, en la figura- 5.23.

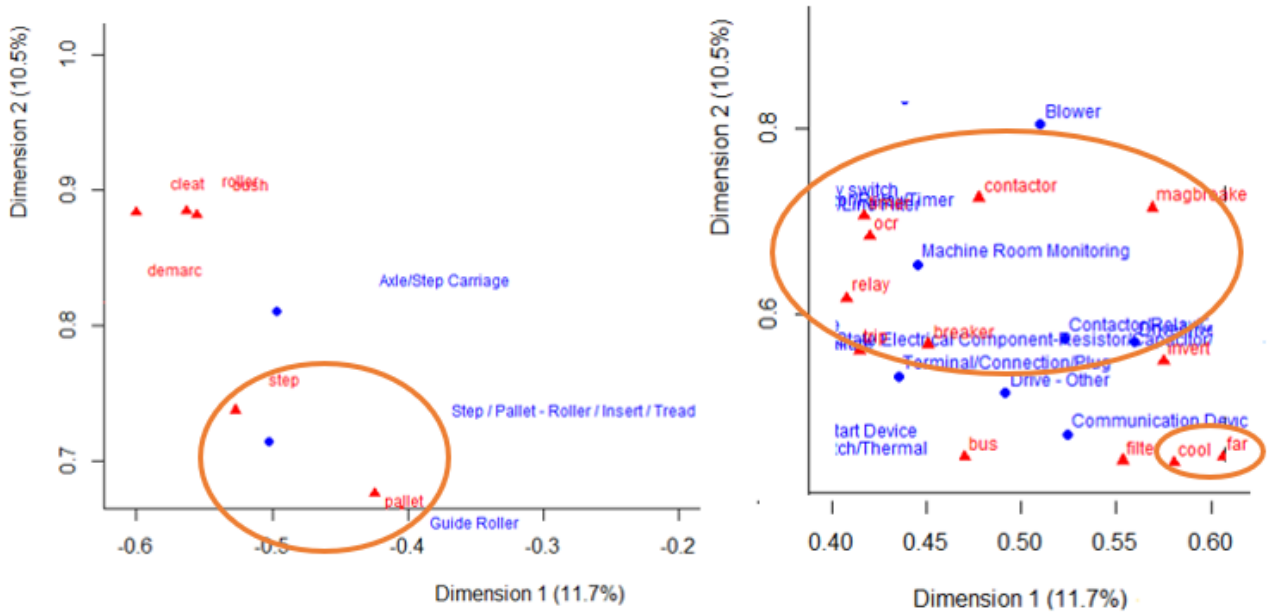


Figura- 5.23 Ampliaciones del análisis de correspondencia

Se puede observar que aparecen cercas palabras similares, como *step* y *pallet, relay, contactor, breaker* y *magbreaker o cool* y *fan.* Gracias a este método, se puede conocer qué palabras tienen un uso similar en este ámbito, y así poder agruparlas como sinónimos.

Por otro lado, se plantea la posibilidad de unir aquellas palabras que se escriben separadas, pero forman un término único ("frequency converter" es "inverter") y también aquellas que tienen sentido por separado, pero juntas cambian o concretan más su significado como "step" y "chain", que cuando se unen en "step chain" se refieren a la cadena de peldaños, un elemento completamente distinto al peldaño o a la cadena principal a la que se refieren por

separado. Para ello, además de consultar a empleados de Thyssenkrupp (es lo primero que se hizo) se emplea una técnica de *Machine Learning*, la clusterización jerárquica. Se emplea como distancia una medida basada en la correlación de términos, es decir, la proporción entre el número de veces que aparecen en el mismo registro y el número de veces que aparecen por separado. El árbol es excesivamente largo para incluirlo aquí, pero se muestra una sección en la figura- 5.24.

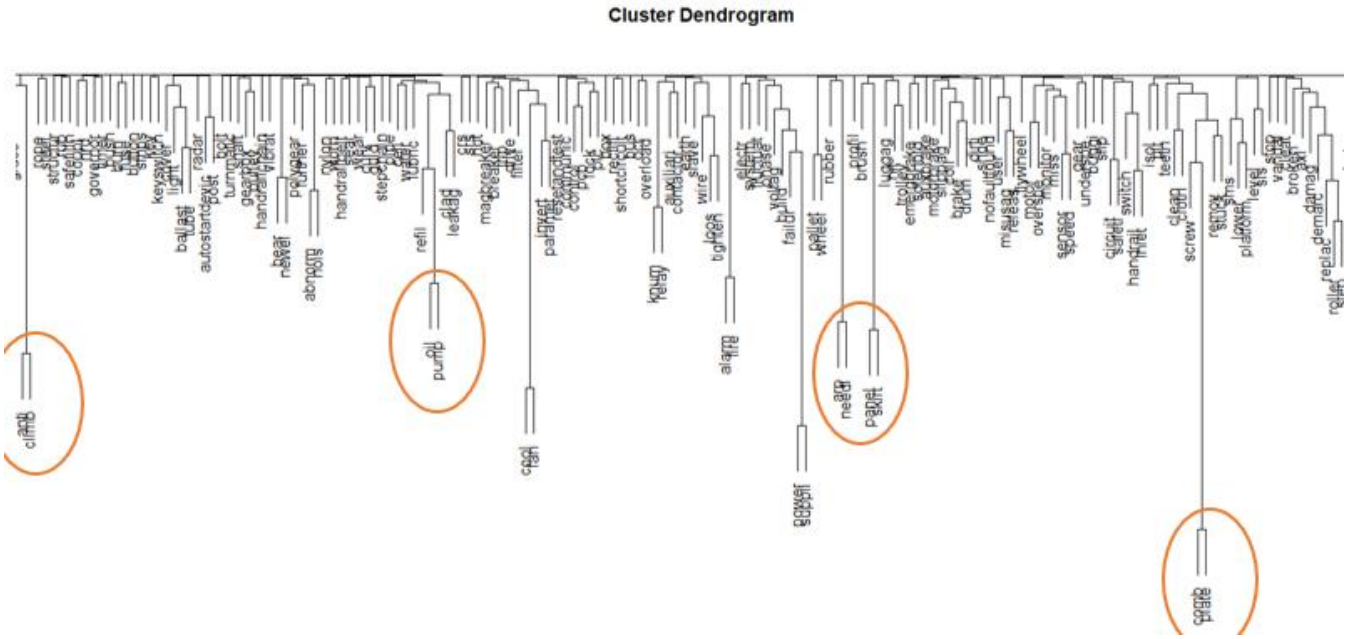


Figura- 5.24 Clusterización jerárquica de lexemas

En este árbol, cuanto más abajo están los términos unidos mayor correlación tienen entre sí, es decir, son candidatos más sólidos para unirse en una única palabra. Se destaca, por ejemplo, *anti* y *climb*, que se refieren al dispositivo para evitar que se trepe por los laterales de la escalera, *oil* y *pump*, ya que la única bomba que puede haber en la escalera tiene como misión bombear el aceite de lubricación y otros términos que solo tienen sentido en conjunto como “arp needl” (dispositivo para medir velocidad de motor), “comb plate” que es la placa de peines o “panel skirt”, que es el zócalo de la balaustrada. Esta información es de gran utilidad para agrupar términos y disminuir la variabilidad de palabras.

Selección de palabras importantes: en paralelo a la agrupación de sinónimos se escogen aquellas palabras que formarán parte del modelo de clasificación. Existen distintos criterios para definir la importancia de una palabra. Además del criterio subjetivo, de nuevo se ha consultado con Thyssenkrupp qué palabras se consideran que es imprescindible que se tengan en cuenta. Los criterios objetivos más usados se detallan a continuación:

Frecuencia absoluta: es la medida más usada por su sencillez. Indica el número de veces que aparece una palabra, bien sea en un registro, o bien en el conjunto de reportes. Esta frecuencia se puede calcular tanto para cada término individual como para conjuntos de

palabras, llamadas n-gramas. En este trabajo se empleará la frecuencia absoluta para ver qué palabras son más importantes. Se utiliza tanto para términos individuales como para bigramas (grupos de 2 palabras), para esto último se empleará el paquete tidytext que permite separar los bigramas de manera sencilla mediante unnest_tokens.

TF_IDF: la otra medida principal de la relevancia de una palabra se denomina TF_IDF, siendo TF la frecuencia del término en el registro, e IDF la inversa de la frecuencia del término en todo el documento, es decir, en el conjunto de registros.

$$TF\_IDF = TF \cdot IDF = \frac{Frecuencia\ de\ A\ en\ un\ registro}{Frecuencia\ de\ A\ en\ todos\ los\ registros} \quad (5.10)$$

Esta medida es de gran utilidad para encontrar términos que caractericen a un texto B y lo diferencien de otros textos con los que se compare. Los términos comunes tendrán gran frecuencia (TF) en el texto B, pero su frecuencia será similar en el resto de textos por lo que su valor de TF_IDF será próximo a 1/nº documentos. En cambio, si el término es característico de B y solo se emplea en dicho texto (como el nombre de un protagonista en una novela o un país inventado en una película), incluso aunque su frecuencia sea pequeña, tendrá un TF_IDF cercano a 1, siendo 1 si solo aparece en B. Esta medida es de gran utilidad en textos de gran longitud y con temáticas muy distintas, condiciones que no cumple el caso de estudio. Debido a este motivo, no se calcula el TF_IDF en este trabajo.

Tras seleccionar cómo se medirá la importancia de una palabra o bigrama se extrae una lista con los términos más usados. Se muestra el código para llevarlo a cabo en figura- 5.25.

```
do.call(c,strsplit(as.character(D$Action_en)," ")) -> palabras
head(rev(sort(table(palabras))),500)
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% PARTE DE TIDYTEXT %%%%%%%%%%%%%%%%%%%%%%%%   #u
nigrams=unnest_tokens(D,unigrams,Action_en, token="ngrams",n=1)
#bigrams=unnest_tokens(D,bigrams,Action_en, token="ngrams",n=2)
#head(rev(sort(table(bigrams$bigrams))),50)
#trigrams=unnest_tokens(D,trigrams,Action_en, token="ngrams",n=3)
#rev(sort(table(trigrams$trigrams)))
```

Figura- 5.25 Código para extraer el ranking de palabras por su frecuencia absoluta

Se pueden ver las palabras más frecuentes antes y después del proceso de limpieza inicial en la figura- 5.26 y la figura- 5.27.

and	the	unit	to
28662	24069	22006	19032
	back	normal	checked
16620	13517	11727	9421
found	is	working	escalator
9011	8890	8103	6803
normal.	then	step	switch
6177	5941	5672	5644
we	stop	observed	of
5476	5409	5388	4952
was	not	THE	reset
4735	4090	4017	3960
in	now	so	UNIT
3857	3789	3513	3454
operation	error	handrail	comb
3362	3304	3192	3192
running	operation.	check	roller
3107	3080	3030	3012
emergency	restart	replaced	fault
3011	3001	2996	2954
tested	per	side	adjusted
2942	2902	2877	2873

Figura- 5.26 Ranking de palabras más frecuentes antes del proceso de limpieza

escal	adjust	resum	replac	switch	no	step
16189	13936	13543	11814	11120	11007	9941
comb	reset	plate	handrail	test	return	have
9130	8671	8162	7841	7519	7088	6616
roller	stop	damag	clean	sensor	restart	brake
6075	6069	5896	5339	4824	4645	4269
run	upper	knum	power	fault	failur	caus
4141	3760	3613	3528	3252	3162	3119
emerg	now	lower	errorcod	chain	use	provid
3004	2896	2805	2786	2742	2716	2609
motor	passeng	side	screw	control	open	drive
2529	2474	2461	2449	2393	2371	2253

Figura- 5.27 Ranking de palabras más frecuentes después del proceso de limpieza

comb plate	emerg stop	roller step	plate have	stop switch	return replac	inspect escal	restart now	step roller
9830	4654	2777	2709	2676	2654	2578	2477	2445
reset test	right side	stop button	damag replac	power suppli	replac step	press emerg	switch activ	errorcod provid
2357	2121	2111	2085	2056	2001	1967	1768	1764
no errorcod	drive chain	scc no	handrail belt	inlet switch	step demarc	provid escal	damag return	step damag
1733	1711	1693	1654	1632	1533	1449	1387	1375
escal stop	inspect upper	upper comb	someon press	handrail inlet	lower land	escal run	upper land	have screw
1353	1324	1314	1299	1295	1282	1273	1257	1253
speed sensor	safeti switch	axl roller	adjust adjust	plate switch	left side	motor brake	power failur	inspect handrail
1235	1209	1202	1201	1141	1081	1076	1058	1026
skirt panel	no scope	replac new	test run	reset control				
986	966	949	941	941				

Figura- 5.28 Bigramas más frecuentes después de la limpieza y antes de la agrupación de sinónimos

En primer lugar, hay que destacar, en el ranking de palabras antes del proceso de limpieza en figura- 5.26, que las más usuales, como por otra parte era de esperar, son conectores y otras palabras sin significado como *and*, *the*, *to* o *back*. También estaría el espacio, lo que indica que hay espacios múltiples en blanco.

Por otro lado, en el ranking de palabras ya limpias (figura- 5.27), se observa que hay términos que parecen importantes y que se habrán de emplear en el modelo de clasificación como *step*, *replace*, *handrail*, *comb* o *plate*, pero también aparecen otros totalmente irrelevantes como *now*, *resum*, que indican que se están resumiendo las medidas tomadas, o *test*, que indica que se hizo un test de funcionamiento del equipo. Finalmente (figura- 5.28) los bigramas más frecuentes, viendo que algunos se pueden unir ya que forman entre los dos un solo término como *comb plate*, que pasará a ser *combplate* o *emer stop* que pasará a ser *estop*.

Por tanto, como se ha podido comprobar, si bien suele haber correspondencia entre la frecuencia de los términos y lo relevantes que serán para el modelo de clasificación, no es una relación exacta, y ha sido necesario realizar un proceso iterativo de prueba, consulta y toma de decisiones a la hora de programar el modelo para elegir cuántos términos y cuáles serían los escogidos.

Finalmente, se ha llegado a una lista de 193 lexemas, agrupados en función del elemento de la escalera o pasillo al que pertenecen, que serán los que se empleen para la clasificación de textos. Se muestra dicha lista en la figura- 5.29. Algunos registros no contarán con ninguno de los 193 pero se ha comprobado que constituyen menos del 0,5% del total por lo que son despreciables.

```
#%%%%%%%%%%%%%%%%%%%%%%%%%%%% TERMINOS RELEVANTES   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

posibles=c("safetilin","motorfault","bolt","crs","bush","broken","nrd","burn","keyswitch",
"vibrat","underspe","miss","vandal","anti","balanc","rectifi","needl","ceil","wear")
safetyswitches <- c ("switch", "sensor", "safeti","arp","sms","sls","sss","srs")
step=c("lower","buggi","teeth","step","roller","demarc","axl","guid","screw")
combplate=c("pit","platform","cssl","remov","cloth","screw","stuck","comb","plate","spring",
"tension","shoe","trolley")
estop=c("estop","emerg","stop","button","press","releas","passeng","misusag","user")
handrail=c("curv","nylon","seal","transit","profil","bow","slip","handrailbelt",
"handraildrivewheel","handrail","belt","rubber","rub","sprocket","inlet","newel","bear",
"pressur","polygear")
chaindrive=c("pulley","turnmatic","handrailchain","stepchain","drivechain","chain","drive",
"speed","shaft","flywheel","rope")#rope hace que steel no este con skirt, no convence
brake=c("slippag","latch","drum","auxbrake","emergbrake","motorbrake","brake","motor","emerg",
"auxiliari","solenoid","align","worn")
power=c("circuit","breaker","build","power","failur","suppli","loss","smps")
relay.wiring.controller=c("overload","earth","ocr","fuse","bus","pcb","timer","knum","electr",
"isol","shortcircuit","magbreaker","slave","relay","contactor","wire","control","panel",
"monitor","processor","plc","communic","fluctuat","lock")
motor.gearbox=c("wheel","governor","drive","coupl","gear","box","gearbox","invert","paramet","
overspe","underspe")
ventilation=c("cool","fan","filter")
ballustrade=c("skirt","brush","glass","bollard")
nofaultfound=c("key","keystart","resetandtest","reset","nofaultfound")
lightning=c("light","cover","tube","ballast")
autostartdevice=c("autostartdevic","radar","photocel","post")
lubrication=c("oil","pump","level","leakag","refil","clad","lubric","nois","abnorm","pipe")
builDfire=c("fire","alarm","system")
other=c("water","cleat","hit","clean","carriag","climb","dirti","phase","hot","luggag",
"voltag","steel","pit","structur")
auxiliares=c("adjust","replac","further","trip","switch","tighten","loos","damag")

lexemas.importantes=c(posibles,safetyswitches,step,combplate,estop,handrail,chaindrive, brake,
power,relay.wiring.controller,motor.gearbox,ventilation,ballustrade,nofaultfound,
lightning,autostartdevice,lubrication,builDfire,other,auxiliares)
lexemas.importantes=lexemas.importantes[!duplicated(lexemas.importantes)]
```

Figura- 5.29 Lexemas empleados en la clasificación de textos

Para concluir, se quiere destacar que, como ya se ha ido comentando a lo largo de este apartado, este proceso requiere mucho tiempo y conocimiento de la materia para seleccionar qué palabras se pueden descartar, cuáles son sinónimos y cuáles son las más relevantes. Este tiempo sería considerablemente menor si se contase con la capacidad para procesar los datos en la nube o un gran poder de cálculo offline en vez de los 4GB de RAM

disponibles, ya que se podrían incluir todos los términos que aparecen en los registros y no sería necesario tampoco agrupar sinónimos. Sería el posterior modelo de *Machine Learning* el que discriminaría qué términos son similares y cuáles son los más importantes sin necesitar casi intervención humana.

5.4.4.- Clasificación del área de fallo

Origen de la necesidad de clasificación: tal y como se comentó en el apartado 5.3.2.-, existe una variable llamada FaultAreaName_en que clasifica el motivo del fallo, pero no cumple adecuadamente su función por diversos motivos. En primer lugar, codifica dentro del mismo nivel fallos que en la empresa consideran muy distintos por el tiempo de reparación que implican, los repuestos empleados y las complicaciones que conllevan. En segundo lugar, se han encontrado registros en los que no hay relación entre el fallo que se describe en Action_en y su valor de FaultAreaName_en, lo que podría explicarse por un error del operario al elegir el FaultAreaName_en o por una mala definición de los niveles, que hace que sea difícil para el técnico elegir su valor. Finalmente, también existe un reducido número de registros que cuentan con descripción del fallo, pero tienen valor NA en FaultAreaName_en e interesa completarlos. Por estos motivos, se decide codificar, a partir de Action_en, una nueva variable de motivo de fallo llamada FaultAreaName_en2.

Estructura de datos de entrada: en este momento se cuenta con todos los registros con una estructura homogénea y reducida, puesto que solo pueden contener 193 lexemas distintos, en el orden y frecuencia oportuna en cada uno. Este formato cumple la condición que se había impuesto (al no tener gran poder de cálculo) de un bajo volumen de información, pero sigue siendo inadecuado al contener palabras, ya que todos los modelos de clasificación trabajan con variables numéricas. Por este motivo, es necesario transformar la variable Action_en, la cual contiene los registros, en un *Bag of Words*, que es uno de los formatos más usados en el análisis de texto [27]. El *Bag of Words* es una matriz en la que cada columna representa un término, cada fila un texto, y el componente x_{ij} expresa el número de veces que aparece el término *j* en el texto *i*. También puede emplearse una matriz binaria, en vez de frecuencias, donde el término x_{ij} es 1 si el término *j* aparece en el texto *i*, independientemente del número de veces que aparezca. Inicialmente se escoge la versión en frecuencia, pero tras unas pruebas iniciales con el algoritmo que se explicará más adelante se ha comprobado que se consiguen mejores resultados con la versión binaria.

Cabe destacar que este formato tiene una gran desventaja, ignora el orden de las palabras. Esto puede ser un inconveniente en algunos textos (no es lo mismo “Juan come un plátano” que “un plátano come a Juan”), pero en este caso no es una gran limitación, puesto que se ha evitado esa pérdida de información con la clusterización jerárquica, que ha permitido agrupar aquellos términos en los que es necesario saber que aparecen juntos y no separados en la frase.

Se emplean las funciones Vcorpus y Documenttermmatrix del paquete tm para extraer el "Bag of Words" y unirlo al resto de la base de datos. Se muestra el código empleado en la figura- 5.30.

```
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%   PARTE DE TM  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5

Corpus=VCorpus(VectorSource(D$Action_en))
dtm=DocumentTermMatrix(Corpus,list(dictionary=lexemas.importantes))
DTM=as.data.frame(as.matrix(dtm))
rm(dtm)
rm(Corpus)
D=cbind.data.frame(D,DTM)
```

Figura- 5.30 Código necesario para obtener el Bag of Words

Algoritmo de clasificación empleado: una vez que se tiene el "Bag of Words", se pasa a definir propiamente el problema que se quiere resolver. El objetivo es, a partir de las 193 variables numéricas que indican la frecuencia o presencia de cada lexema clave en cada texto, asignar dicho texto a una causa de fallo, la cual se desconoce en principio, dado que no se tiene una etiqueta que la identifique, al descartar como inadecuada la proporcionada por FaultAreaName_en. De acuerdo con este enunciado, se tiene un problema de aprendizaje no supervisado, puesto que se desconoce a qué grupo pertenece cada individuo. Por tanto, como técnicas de *Machine Learning* aplicables se tienen principalmente la clusterización, ya sea jerárquica o k-means, y los mapas de Kohonen.

No obstante, antes de aplicar dichas técnicas se desarrollará un algoritmo de clasificación propio basado en la asociación de palabras clave.

La hipótesis de partida para este algoritmo es que cada área de fallo se caracteriza por una serie de palabras únicas que la definen, y es condición necesaria que aparezcan mencionadas en el informe. No obstante, una única palabra no es, en general, condición suficiente, por lo que realmente lo que se necesita es valorar la presencia de varias palabras en conjunto. Además, se parte de que, al ser reportes meramente informativos del suceso, intentan ser escuetos en detalles no necesarios, y no emplearán casi palabras no relacionadas con la zona del fallo. Así pues, los pasos a seguir por el algoritmo serán los siguientes:

1. Puntuar la idoneidad de que un registro pertenezca a un área de fallo determinada. Para ello, se emplea como criterio la suma, ponderada por la importancia y exclusividad de cada término en esa área, de las palabras presentes en el reporte, puntuando en paralelo sinónimos circunstanciales, solo aplicables a ese fallo, e incluso restando puntuación para las palabras que pudieran causar confusión al usarse a menudo en otro tipo de fallos. Se muestra como ejemplo la puntuación de pertenencia a roller step en la figura- 5.31.

```
D$rollerstep=(D$roller>0)*1.5+(D$step>0)+(D$replac>0|D$broken>0)-(D$chain>0)-(D$switch>0)-(D$handrail)
```

Figura- 5.31 Puntuación de la categoría roller step

Se puede ver que se considera más probable que el fallo sea de *roller step* si aparece la palabra *roller* (1,5 puntos) que si aparece *step* (1 punto), puesto que la palabra *step* se usa en otras categorías como *step demarcation* o *step bolt* y, por tanto, no es exclusiva del fallo de *roller step*. Por otro lado, se consideran sinónimos para esta categoría los lexemas *broken* y *replac*, ya que se da por hecho que si está roto (*broken*) se sustituirá (*replac*) aunque no se diga explícitamente y viceversa, si se sustituye es porque se ha roto. Así pues, estás dos palabras no se podrían puntuar independientemente, dado que podría darse el caso de que un reporte mencione que algún elemento está roto y se reemplaza, y si ambos lexemas fueran independientes esto daría a ese registro dos puntos en la categoría *roller step*, sin que tuviera que aparecer siquiera la palabra *roller*. En cambio, al considerarlas en paralelo, aunque estuvieran presentes los dos términos (*broken* y *replac*), solo se otorgaría un punto en vez de dos tal y como se muestra en tabla 5.8.

Registro	Action_en	broken|replac	broken+replac
A	Sprocket broken, replaced	1	2
B	Replaced two step rollers	1	1

Tabla 5.8 Asignación de puntos en palabras en paralelo o sumadas

Finalmente, se penaliza que aparezcan las palabras *chain* y *switch*, ya que los *roller step* están unidos a la cadena de peldaños y muy cerca de varios *switch* (interruptores de seguridad). Por ello, es habitual que en los fallos por cadena de peldaños mencionen al *roller* o que mencionen que un *switch* se ha activado por el contacto de un *roller*. Además, también se llaman *roller* a los rodillos que guían al pasamanos, por lo que también se penaliza el término *handrail*.

Se puede apreciar que determinar la ponderación, las palabras en paralelo y las penalizadas es algo enormemente subjetivo, y ha precisado un largo periodo de consulta, prueba y ajuste para lograr puntuar adecuadamente los registros en cada caso. Además, hay que comentar que finalmente no se han utilizado los 193 lexemas en su totalidad en esta fase, pero se mantienen por si se necesitaran en métodos posteriores.

2. Definir un umbral, a partir del cual, si un registro lo supera para cierta área de fallo, se considera que podría pertenecer a ella. Cabe destacar que, dado que las puntuaciones son números concretos multiplicados por variables binarias, decidir el umbral equivale a decir qué combinaciones de palabras caracterizan a dicha área de fallo. Se continúa con el ejemplo de *roller step* en figura- 5.32.

```
D$FaultAreaName_en2[D$rollerstep>=2.5]="Step roller"#1665
```

Figura- 5.32 Código para programar el umbral de asignación a la categoría *roller step*

El umbral de 2,5 es equivalente a decir que un registro podría pertenecer a *roller step* si en él aparecen las palabras *roller* y *step*, o *roller* y *replaced* o *broken*, siempre y cuando no aparezcan la palabra *handrail*, *chain* o *switch*, puesto que en ese caso se considera más plausible que el fallo sea de alguno de esos elementos y, por tanto, no debería asignarse a *roller step*. En esta etapa la subjetividad también juega un papel muy importante, puesto que se podría haber elegido un umbral distinto, cambiando las combinaciones características, pero la experiencia y el ensayo y error han demostrado qué umbral es el adecuado en cada caso.

3. Definir una jerarquía de pertenencia y asignar los fallos. Un mismo registro puede contener palabras suficientes para pertenecer a dos áreas de fallo, una más general y otra más específica. Por ello, hay que decidir a cuál de las dos se le adjudica. Para ello, la asignación de fallos a áreas se realiza de tal manera que van primero las áreas de fallo más generales, aquellas con umbrales más bajos y palabras y combinaciones más comunes. De esta manera, si un registro posee un lexema o combinación característica de otro tipo de fallo más específico, será posteriormente asignado a esa área. Se muestra el orden de asignación de algunas áreas de fallo en figura- 5.33. De nuevo hay que destacar que establecer el orden es algo subjetivo que ha requerido múltiples pruebas y consultar a distintos técnicos de la empresa. Además, se menciona que inicialmente todos los registros son adjudicados al área *Others,* pero al acabar la asignación solo pertenecerán a *Others* aquellos reportes que no superen el umbral de ninguna otra área de fallo.

```
D$FaultAreaName_en2=as.character(D$FaultAreaName_en2)
D$FaultAreaName_en2="Others"#71
D$FaultAreaName_en2[D$step>=1]="stepjust"#970
D$FaultAreaName_en2[D$skirt>=1]="skirtonli"#970
D$FaultAreaName_en2[D$Resetandtest>=1]="Reset"#970
D$FaultAreaName_en2[D$otrosafeti>=2]="Other safety switches"#293
D$FaultAreaName_en2[D$handrailonli>=1]="Handrailonli"#76
D$FaultAreaName_en2[D$SensorRoturahandrail>=2]="Sensor rotura handrail"#10
D$FaultAreaName_en2[D$Motordrive>=1]="Motor drive"#68
D$FaultAreaName_en2[D$Combplate>=2.5]="Combplate"#996
D$FaultAreaName_en2[D$steponli>=2.5]="Step"#130
D$FaultAreaName_en2[D$Handrail>=3]="Handrail"#998
D$FaultAreaName_en2[D$firealarm>=3]="Fire alarm"#425
D$FaultAreaName_en2[D$controller>=1.5]="PLC and inverter"
D$FaultAreaName_en2[D$relaysandcontactorsandwires>=1.5]="Relay and wires"
D$FaultAreaName_en2[D$powersuppli>=3]="Power supply"#361
D$FaultAreaName_en2[D$lubrication>=3.5]="Lubrication"#422
D$FaultAreaName_en2[D$refrigeration>=2]="Refrigeration"#225
D$FaultAreaName_en2[D$sinfault>=1]="sinfault"#884
D$FaultAreaName_en2[D$balaustrada>=3]="Balustrade"#398
D$FaultAreaName_en2[D$light>=1]="Lighting"#371
D$FaultAreaName_en2[D$brakes>=3.5]="Brake"#729
D$FaultAreaName_en2[D$gearsandpulleys>=2]="Gears"#288
D$FaultAreaName_en2[D$guides>=1]="Guides"#376
D$FaultAreaName_en2[D$Circuitbreaker>=2]="Circuit breaker"#295
D$FaultAreaName_en2[D$Keyswitches>=1.5]="Keyswitch"#188
D$FaultAreaName_en2[D$Stopswitches>=1.5]="Emergency stop"#4455
D$FaultAreaName_en2[D$Passengerstart>=1]="Passenger detection"#229
D$FaultAreaName_en2[D$rollerstep>=2.5]="Step roller"#1665
D$FaultAreaName_en2[D$axistep>=3.5]="Step bolt"#1143
D$FaultAreaName_en2[D$demarcstep>=2.5]="Step demarcation"#2159
```

Figura- 5.33 Jerarquía de pertenencia a algunas áreas de fallo

Una buena muestra de cómo afecta el orden son las áreas de fallo *stepjust*, *Step* y *step roller*. La condición para ser asignado a *stepjust* consiste únicamente en incluir la palabra step en el reporte. En cambio, la condición para pertenecer a *Step* es contener la palabra *step* y algún calificativo de su condición mecánica como *broken*, *replac*, *align* o *adjust*, lo que es más especifico, ya que indicaría que el fallo se debe a que el peldaño ha sido dañado o desajustado y necesita una corrección.

Se desarrolla un ejemplo para mostrar la influencia de dicha jerarquía con 3 registros. El A se define como “Step dirty, Cleaned”, El B contiene “Noise in the step. Adjusted”. Y el C es “Replaced the broken roller step”. Los 3 contienen la palabra “step”, por lo que inicialmente serán asignados a “steponly”. Siguiendo con la jerarquía de asignación se encuentra “Step”, cuyo umbral cumplen los registros B y C pero no el A, ya que este no recoge ninguna condición mecánica del peldaño. Por tanto, B y C pasarán a pertenecer a “Step”, mientras el A sigue siendo “steponly”. Finalmente, el registro C también cumplirá el umbral de “step roller”, al contener la palabra *roller*, por lo que será asignado a esa área de fallo. Se consigue así que cada fallo pertenezca al área adecuada. En cambio, si la jerarquía hubiera sido la contraria, todos los fallos habrían acabado siendo asignados a “steponly”. El proceso seguido queda resumido en la tabla 5.9.

Registro	Action_en	> umbral steponly	Área de fallo inicial	> umbral Step	Nueva área de fallo	> umbral step	Área de fallo final
A	Step dirty, cleaned	SI	steponly	NO	steponly	NO	steponly
B	Noise in the step, adjusted	SI	steponly	SI	Step	NO	Step
C	Replaced the broken roller step	SI	steponly	SI	Step	SI	step roller

Tabla 5.9 Asignación secuencial a áreas de fallo

4. Posteriormente, los registros con errores normalizados, comentados en 5.4.2.- y cuyo código fue mostrado en figura- 5.21, y algunos con un texto muy específico se asignan directamente a un área de fallo sin importar las puntuaciones que obtuvieran. Esto tiene también su componente personal y subjetivo.
5. Finalmente, y a petición de Thyssenkrupp, se agrupan algunos de los fallos que se codificaron como niveles distintos, puesto que no se requiere tanto nivel de detalle. Se muestra el código en figura- 5.34.

```
#AGRUPACION DE AREA DE FALLO

D$FaultAreaName_en2=as.character(D$FaultAreaName_en2)
D$FaultAreaName_en2[grep("Sensor rotura handrail",D$FaultAreaName_en2)]="Handrail speed"
D$FaultAreaName_en2[grep("Handrailonli",D$FaultAreaName_en2)]="Handrail"
D$FaultAreaName_en2[grep("stepjust",D$FaultAreaName_en2)]="Step"
D$FaultAreaName_en2[grep("Not available|sinfault",D$FaultAreaName_en2)]="Others"
D$FaultAreaName_en2[grep("Combplate",D$FaultAreaName_en2)]="Combplate switch"
D$FaultAreaName_en2[grep("arpneedlswitch",D$FaultAreaName_en2)]="Motor over/underspeed"
D$FaultAreaName_en2[grep("Step chain tension",D$FaultAreaName_en2)]="Step chain"
D$FaultAreaName_en2[grep("skirtonli",D$FaultAreaName_en2)]="Balustrade"
D$FaultAreaName_en2[grep("Buggy switch",D$FaultAreaName_en2)]="Step upthrust"
D$FaultAreaName_en2=factor(D$FaultAreaName_en2)
```

Figura- 5.34 Código de agrupación de áreas de fallo

Se puede ver todo el código necesario para programar el algoritmo que se acaba de explicar en http://bellman.ciencias.uniovi.es/~raul/Acondicionamiento.html.

Cabe destacar que esto es una primera aproximación y ha sido posible emplear un algoritmo "clásico", sin aprendizaje automático, al haber un número relativamente reducido de categorías y atributos a valorar.

Se puede observar que los tres primeros pasos de este algoritmo son relativamente similares a muchos métodos de *Machine Learning* de aprendizaje supervisado. Se parece a las redes neuronales, en la medida en que se basa en asignar los coeficientes de la suma ponderada y el umbral de la "función de activación", que en este caso es directamente dicha suma, de cada "neurona" o área de fallo. También tiene ciertos componentes de los árboles de decisión, tanto en los umbrales como en la importancia de la jerarquía, pasando de fallos generales a ramas más especificas. La principal diferencia entre dichos métodos y el algoritmo creado es que todas las decisiones de coeficientes y orden han sido tomadas por humanos en vez de matemáticamente minimizando el error obtenido. Esto se debe a que no hay realmente un error a optimizar, dado que no se tiene originalmente la etiqueta de qué área es la correcta, por lo que no se podían emplear redes neuronales o árboles de decisión y hubo que optar por esta solución.

Además, la otra gran ventaja de este método es su sencillez, ya que se basa en obtener una puntuación para cada registro en cada área de fallo y asignarlo secuencialmente a los niveles

cuyo umbral supere, y esto requiere poco poder de cálculo. También cabe destacar que se estima que su fiabilidad es superior al 90%. Esta medida es bastante subjetiva, pero han sido leídos un gran número de muestras aleatorias por expertos de Thyssenkrupp y consideraron que al menos el 90% de los registros habían asignados correctamente a su área de fallo. Como inconveniente cabe mencionar que puede que se hayan despreciado palabras o combinaciones que pudieran asignar el área de fallo mejor que las aquí empleadas. Adicionalmente, este algoritmo ha requerido más tiempo para programarlo que uno de *Machine Learning* de los ya comentados. De todas formas, ha implicado mucho menos tiempo que si hubiera habido que asignar a cada registro una etiqueta manualmente, que era lo que hubiera sido necesario si se quisiera haber empleado aprendizaje supervisado desde el principio.

Esta primera asignación de fallo se empleará, a partir de ahora, como etiqueta para calibrar el grado de similitud de los resultados de clasificación obtenidos con *Machine Learning* y con el algoritmo propio. Esto permitirá ver lo robusto que es el algoritmo creado.

Se ha escogido como técnica de comprobación el árbol de clasificación. Esta elección se debe a la gran facilidad con que puede ser interpretado un árbol, en contraposición a otras técnicas como las redes neuronales, que son cajas negras. Adicionalmente, como comentario general, hay que indicar que en los dos árboles que se van a usar se han tenido que modificar ligeramente los parámetros del árbol para adaptarlo a este problema.

En primer lugar, se cogerán como variables predictores las puntuaciones de idoneidad creadas por el algoritmo propio. Este árbol permitirá ver qué influencia tienen los umbrales y el orden de asignación, que son las variables que elegirá el árbol para intentar minimizar el error. También podrá no utilizar todas las puntuaciones, lo que indicaría que esas idoneidades no utilizadas no son necesarias para la clasificación. Se pueden ver las variables usadas por el árbol y la evolución del error con las iteraciones en la figura- 5.35 y figura- 5.36. También se muestra en esa figura la estructura del árbol, no obstante, no se indican los niveles de las ramas debido a su gran longitud.

```
Variables actually used in tree construction:
 [1] arpneedl                 axlstep                  balaustrada              brakes                   Chains
 [6] chainstep                Circuitbreaker           Combplate                controller               demarcstep
[11] firealarm                gearsandpulleys          guides                   Handrail                 handrailinlet
[16] handrailspeed            lubrication              misssstepsensor          motorspeed               otrosafeti
[21] powersuppli              proximitichaindrive      relaysandcontactorsandwires Resetandtest          rollerstep
[26] skirtswitch              steponli                 Stepupthrust             Stopswitches
```

Figura- 5.35 Variables usadas en el árbol de clasificación

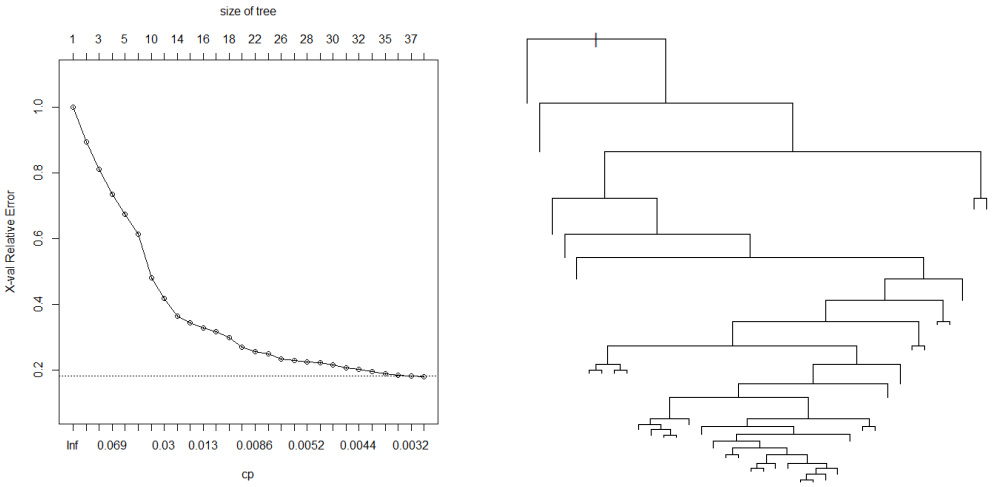


Figura- 5.36 Evolución del error con las iteraciones en el árbol 1 (izquierda) y estructura del árbol (derecha)

Tal y como muestra figura- 5.36, hay una discrepancia menor del 16% entre el algoritmo propio y este primer árbol de decisión. Es decir, parece que los umbrales y el orden de asignación del árbol es relativamente parecido al del algoritmo propio. Además, en la matriz de confusión, se observa que los registros atribuidos a distinta área por el algoritmo propio y el árbol son asignados por el árbol generalmente al nivel *Others* mientras el algoritmo, al considerar niveles más amplios y tener en cuenta más factores con los pasos 4 y 5, es capaz de encontrarles un nivel más adecuado que el *Others*. Esto se ve en la matriz de confusión de Figura- 5.37, donde las columnas son los niveles originales y las filas los niveles del árbol. El nivel *Step* y *Water leakage* son los peor clasificados por el árbol y son asignados en gran parte a "Others", como ya se había comentado.

arbol.af.pred	Relay and wires	Reset	Skirt switch	Step	Step bolt	Step chain	Step demarcation	Step roller	Step upthrust	Water leakage
Balustrade	0	0	1	0	0	8	4	0	0	2
Brake	0	0	6	0	1	11	6	0	2	30
Chains	0	0	0	0	0	17	0	1	0	3
Circuit breaker	0	0	2	0	0	0	0	0	0	10
Combplate switch	0	0	9	6	12	24	274	141	34	19
Emergency stop	0	0	4	0	4	4	7	9	18	21
Fire alarm	0	0	0	0	0	0	0	0	0	0
Floor plate switch	0	0	0	0	0	0	0	0	0	0
Gears	0	0	1	0	34	4	0	0	0	8
Guides	0	0	1	0	14	1	0	0	0	5
Handrail	0	5	5	0	154	18	1	0	1	225
Handrail inlet	0	0	11	0	0	3	0	0	0	0
Handrail speed	0	0	0	0	0	2	0	0	1	0
Keyswitch	0	0	0	0	0	0	0	0	0	0
Lighting	0	0	0	0	0	0	0	0	0	0
Lubrication	0	0	0	0	4	2	0	0	0	8
Main chain safety device	0	0	6	0	0	31	0	0	0	0
Miss step sensor	0	0	0	0	0	0	0	0	0	0
Motor drive	0	0	0	0	0	0	0	0	0	0
Motor over/underspeed	0	0	0	0	0	0	0	0	0	0
Other safety switches	0	0	0	0	17	8	1	0	16	37
Others	1	8	34	183	102	60	19	19	11	289

Figura- 5.37 Parte de la matriz de confusión del árbol 1

Así pues, con este primer árbol se ha visto que la mayor influencia para asignar los fallos son las palabras usadas para codificar cada área, por lo que cabe plantearse si es realmente necesario tener en cuenta las combinaciones de palabras, o serviría únicamente con los de manera independiente. Por ello, se plantea un nuevo árbol, cuyas variables predictoras serán los 193 lexemas, sin proporcionarle las más de 40 puntuaciones de idoneidad, para ver

si el propio árbol genera las mismas combinaciones de palabras para cada nivel (aunque sin ponderaciones, penalizaciones ni paralelas) con sus propias reglas. Esto permitirá comprobar si el criterio usado para elegir las palabras es robusto y se parece al que elige el árbol. Se muestra el resultado en figura- 5.38 y figura- 5.39.

```
 [1] arp          auxbrake    axl         brake      comb       contactor  demarc     drivechain estop      fire       guid       handrail
[13] handrailbelt inlet       invert      keyswitch  light      miss       motor      motorbrake nofaultfound oil        overspe    pcb
[25] phase        plc         power       relay      reset      resetandtest roller   sensor     skirt      sls        sms        speed
[37] step         switch      water
```

Figura- 5.38 Variables usadas en el árbol de clasificación 2

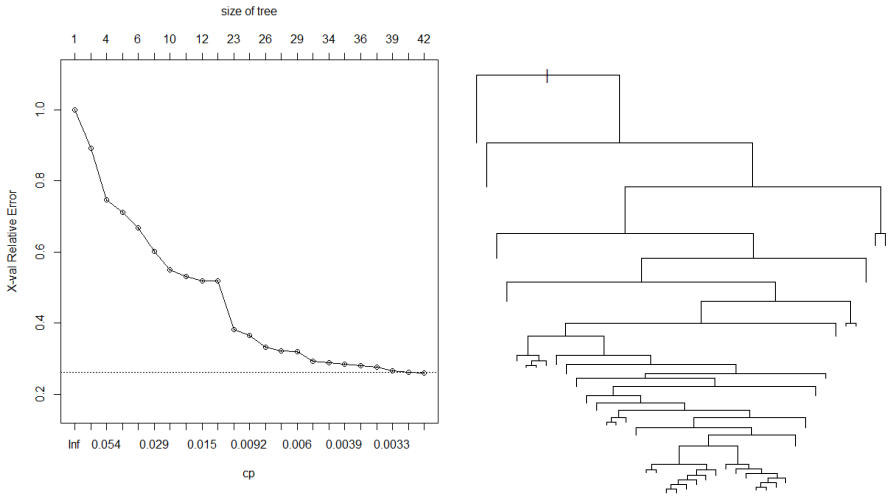


Figura- 5.39 Evolución del error con las iteraciones en el árbol 2 (izquierda) y estructura del árbol (derecha)

Se aprecia que este segundo árbol es menos preciso, con cerca de un 22% de discrepancia frente al algoritmo propio. Esto es lógico, porque aunque puede elegir las combinaciones de palabras, no puede ponderarlas, penalizar palabras o colocar otras en paralelo. Pese a ello, el resultado es bastante similar y se considera que estos dos árboles confirman que el algoritmo propio obtiene un resultado robusto y los criterios para fijar todos los coeficientes, pese a ser subjetivos, son bastante acertados.

Finalmente, hay que mencionar que se puede pensar que el algoritmo propio es mejor que los árboles debido a que está sobreajustado (*overfitted*), es decir, ha sido específicamente diseñado para este conjunto de datos. Esto es cierto, pero los árboles también han sido entrenados con la totalidad de datos y en este caso está justificado, puesto que todos estos métodos se han empleado para crear una etiqueta fiable que luego se utilizará en otros modelos. Por tanto, era necesario sobreajustar este algoritmo para garantizar que los datos que se emplearán en lo que queda de trabajo son correctos.

La otra posible estrategia que se seguirá como se comentó al principio, en vez de crear un algoritmo propio y ver su robustez, era emplear técnicas de aprendizaje no supervisado.

Dentro de estas, se descartan los mapas autoorganizados, ya que solo darían una estructura general de los registros, pero no servirían para esta clasificación de detalle que se pretende hacer. Por otro lado, la principal diferencia entre la clusterización jerárquica y el k-means es que este último método tiene como dato de entrada el número de clústeres que se quieren conocer. Este no es un gran impedimento, ya que con el algoritmo propio ya se han definido 36 áreas identificables. Así pues, se emplearán ambos métodos y se compararán los resultados obtenidos.

Se empieza por la clusterización jerárquica. En este caso se define la siguiente distancia de similitud:

$$dist_{ab} = \frac{lexemas\ iguales\ en\ registros\ A\ y\ B}{lexemas\ de\ A + lexemas\ de\ B} \quad (5.11)$$

El cálculo de esta distancia ha implicado una gran complicación. Se dispone de unos 65000 registros, por lo que hay que calcular 65000^2=4.2e+9 distancias comprendidas entre 0 y 1, la mayoría de ellas no nulas. Es inviable manejar esta cantidad de información con los recursos con los que se cuentan en este trabajo (un ordenador portatil con 4GB de RAM), por lo que se recurre a escalar el problema, escogiendo una muestra representativa cuyo coste de computación sea asumible. Se coge, tanto para k-means como para el clústering jerárquico, una muestra de 80 registros de cada una de las 36 áreas de fallo sumando un total de 2880 registros. Para garantizar que el resultado es independiente de la muestra escogida, se repite el experimento varias veces con distintas muestras y se comprueba que el resultado es siempre muy similar.

Una vez realizada la agrupación empleando el método “complete” de clusterización se corta el árbol jerárquico para formar 36 ramas o clúster y se comparan con los grupos formados por el algoritmo propio. Cabe destacar que esta comparación no forma parte del algoritmo de clústering, si no que se emplea para comprobar si el algoritmo propio puede sustituir, de manera mucho menos costosa computacionalmente, a estos métodos de *Machine Learning*. Se puede ver la comparación entre algunos clústeres (entre corchetes) y algunos niveles del algoritmo propio en figura- 5.40.

En el caso de k-means no es necesario definir una distancia ya que emplea todo el espacio vectorial como coordenadas para hacer la agrupación. Se muestra el resultado en la figura- 5.41.

	PLC and inverter	Power supply	Refrigeration	Relay and wires	Reset	Skirt switch	Step	Step bolt	Step chain
[1,]	0	0	0	1	0	1	0	0	13
[2,]	0	0	0	0	1	0	0	4	0
[3,]	1	0	0	0	0	0	0	0	0
[4,]	0	64	1	2	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	2	34	0	3
[7,]	0	4	2	5	0	0	0	0	3
[8,]	0	0	0	0	0	0	0	0	0
[9,]	1	0	48	0	1	2	0	0	0
[10,]	6	2	0	6	3	0	0	1	1
[11,]	1	0	0	0	0	0	0	0	2
[12,]	0	0	0	0	0	0	0	1	1
[13,]	1	0	0	3	2	20	0	0	5
[14,]	0	0	0	0	0	0	0	0	24
[15,]	0	0	0	0	0	0	0	0	0
[16,]	28	4	1	0	0	0	0	0	0
[17,]	3	1	7	16	0	3	0	42	4
[18,]	1	0	5	0	0	0	0	3	2
[19,]	0	0	0	0	1	0	0	0	1
[20,]	0	1	0	2	0	0	0	6	0

Guides	Relay and wires	Step upthrust	Brake	Passenger detection	Other safety switches
13	16	16	18	18	19
Motor over/underspeed	Circuit breaker	Handrail	Lighting	Balustrade	Step chain
20	21	21	22	23	24
Gears	Keyswitch	Lubrication	Main chain safety device	PLC and inverter	Water leakage
25	25	28	28	28	29
Miss step sensor	Handrail speed	Step	Others	Step roller	Step bolt
30	33	34	36	36	42
Step demarcation	Chains	Skirt switch	Motor drive	Refrigeration	Fire alarm
44	45	46	48	48	51
Combplate switch	Handrail inlet	Emergency stop	Power supply	Reset	Floor plate switch
54	54	55	64	67	74

Figura- 5.40 Matriz de comparación entre clústeres y niveles de FaultAreaName_en2 (superior) y máximo número de registros de cada nivel asignado a un único clúster (inferior) con clústering jerárquico

	PLC and inverter	Power supply	Refrigeration	Relay and wires	Reset	Skirt switch	Step	Step bolt	Step chain
[1,]	0	0	0	0	0	0	0	0	12
[2,]	0	0	0	0	0	0	0	0	24
[3,]	0	1	0	2	1	0	0	1	0
[4,]	0	0	0	0	0	0	6	15	0
[5,]	5	0	0	3	33	0	0	0	0
[6,]	3	0	2	5	0	0	0	0	3
[7,]	0	0	0	0	0	2	16	7	0
[8,]	1	0	0	0	0	1	0	0	0
[9,]	10	0	6	15	2	0	0	1	3
[10,]	0	3	0	2	0	0	0	0	0
[11,]	0	7	0	0	0	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0
[13,]	0	0	0	0	1	0	0	0	0
[14,]	0	0	0	0	0	0	0	0	0
[15,]	1	64	2	2	0	0	0	0	0
[16,]	2	0	0	7	2	67	0	1	2
[17,]	1	1	7	8	0	2	0	41	2
[18,]	1	0	0	0	0	0	0	0	1
[19,]	44	0	1	11	35	0	0	0	4
[20,]	1	0	0	0	2	0	0	0	0

Guides	Relay and wires	Motor over/underspeed	Step upthrust	Circuit breaker	Brake
14	15	16	17	20	21
Balustrade	Keyswitch	Passenger detection	Lighting	Step chain	Others
23	23	23	24	24	25
Handrail	Lubrication	Other safety switches	Miss step sensor	Water leakage	Chains
28	28	28	30	30	31
Step	Main chain safety device	Reset	Motor drive	Gears	Step bolt
34	35	35	36	38	41
Handrail speed	PLC and inverter	Step demarcation	Step roller	Refrigeration	Fire alarm
42	44	44	47	48	52
Handrail inlet	Combplate switch	Emergency stop	Power supply	Skirt switch	Floor plate switch
53	54	56	64	67	74

Figura- 5.41 Matriz de comparación entre clústeres y niveles de FaultAreaName_en2 (superior) y máximo número de registros de cada nivel asignado a un único clúster (inferior) con el método k-means

Si la correspondencia fuera exacta, cada uno de los niveles tendría un clúster propio (distinto al de los demás) con cerca de 80 registros asignados. Sin embargo, tras varios muestreos, se ve que en realidad el clúster que mejor clasifica cada nivel acierta de media con unos 35 de los 80 casos, mientras los otros 45 son asignados a otros clústeres de manera casi equitativa. Además, en el caso del k-means no siempre hay una correspondencia unívoca entre clúster y nivel y un mismo clúster puede ser el que más registros de dos áreas distintas recoge. El porcentaje puede parecer bajo (alrededor del 40%) pero hay que tener en cuenta que el número de clústeres que se quieren formar es muy alto (36) y el clústering debería mejorar al aumentar el tamaño de la muestra de registros considerados.

Así pues, se demuestra que el k-means y el *clústering* jerárquico obtienen resultados parecidos y relativamente similares a los conseguidos por el algoritmo propio. Se considera que esto valida aún más dicho algoritmo como método de clasificación y garantiza que su uso es lo más adecuado para las siguientes fases del trabajo.

5.4.5.- Otras variables léxicas

Aprovechando el proceso de acondicionamiento que se ha hecho con Action_en, se plantea crear otra serie de variables binarias, cuyo fin es indicar las condiciones del fallo y la reparación. Se considera que si aparecen ciertos términos en el registro se puede afirmar, (con poco riesgo de equivocarse), que el incidente cumple determinadas características. Cabe destacar que los términos empleados no tienen porqué restringirse a los 193 empleados en el "Bag of Words" y, de hecho, se emplean en torno a 200 términos —contando sinónimos asimilados a otras palabras— completamente distintos a los del Bag of Words. Se definen a continuación las variables creadas bajo esta hipótesis.

Fenómeno natural ("natural"): esta variable binaria adopta el valor 1 ("natural") si el incidente se debe a algún fenómeno natural que ha inutilizado el equipo como lluvia, arena del desierto, inundaciones o heladas. En caso contrario toma el valor 0 ("artificial").

Repuesto ("replaced"): esta variable binaria adopta el valor 1 ("replaced") si fue necesario emplear algún tipo de repuesto o remplazar alguna pieza durante la reparación; y vale 0 ("adjusted") si, por el contrario, solo se ajustó algún sensor y se arregló sin emplear nuevas piezas.

Incidente no solucionado ("shutdown"): esta variable binaria toma el valor 1 ("parada larga") si no fue posible eliminar la causa del incidente durante la visita del operario, ya sea porque el cliente prefiere que se lleve a cabo la reparación en otro momento, porque no se dispone de los repuestos necesarios, porque se ha acabo el turno del operario, o por algún otro motivo. Toma el valor 0 ("parada corta") en caso contrario.

Accidente ("accident"): esta variable binaria toma el valor 1 ("accident") si la causa primaria del incidente fue una persona o un objeto externo al equipo. Los accidentes más habituales suelen provocar la axctivación de algún diapositivo de seguridad para proteger al pasajero (flap de seguridad del pasamanos), pero también pueden implicar la rotura de algún elemento por un choque o impacto (golpes de maletas o carritos), el atasco de la unidad (llaves, monedas y otros pequeños objetos bloqueando el peine e impidiendo un movimiento fluido), que se hiera algún pasajero (caídas, zapato de tacón atascado en algún hueco) o se dañe algún objeto personal (ropa o paraguas atrapados en la placa de peines). Adopta el valor 0 ("avería") si no es un accidente.

Cabe destacar que las variables Nofault e Interference ya intentan codificar si el fallo fue causado por un accidente o una avería, sin embargo, la regla por la cual un operario decide si

es accidente o avería es subjetiva, y depende tanto de su propio criterio como de la pauta establecida por su delegación e incluso del contrato de mantenimiento que tenga el cliente. Por este motivo, es preferible definir una nueva variable con un criterio conocido, objetivo y global —para todas las delegaciones— que distinga entre avería y accidente.

Caída ("fall"): esta variable binaria toma el valor 1 si el incidente se debió a una caída del usuario mientras usaba el equipo. Toma el valor 0 en caso contrario

Fallo predecible ("max"): esta variable binaria toma el valor 1 ("Detectable") si hubo algún indicador que un humano pudiera percibir, ya sea antes del fallo o durante su reparación. Se entiende por indicador la información sensorial como el ruido, vibraciones, calor o exceso de temperatura u olor a quemado. También se cuenta como indicador a la información que sería fácilmente detectable por sensores como paradas pequeñas frecuentes o corriente fluctuante. Adopta el valor 0 ("Suddenly") en caso contrario.

Se puede apreciar que el método empleado para definir estas variables puede pecar de ser excesivamente sencillo al basarse únicamente en la presencia de unas pocas palabras. Es cierto que puede haber incidentes que cumplan la definición teórica de alguna de las variables, y sin embargo, al no aparecer ninguna de las palabras clave, ya sea porque el operario no informo correctamente o porque se expresara con palabras no contempladas, no sea codificada correctamente. No obstante, dado que no se tiene información suficiente para implementar alguna técnica más sofisticada relativa a estos aspectos, se considera que este enfoque es el más adecuado y que se obtienen resultados bastante ajustados a la realidad. Se muestra el código necesario para programar dichas variables en figura- 5.42.

```
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% COLUMNAS ESPECIALES DE TEXTO   %%%%%%%%%%%%%%%%%%%%%%%%%%
D$natural="artificial"#NATURAL DISASTER
D$natural[grep("water|storm|lightn|earthquake|\\bice\\b",D$Action_en)]="natural"
D$natural[grep("sand|gravel",D$sintrad)]="natural"
D$natural=factor(D$natural)
D$replaced="ajuste"#PIEZA CAMBIADA
D$replaced[grep("replac|pcs|nos|broken|new|instal|damag",D$Action_en)]="cambio"
D$replaced=factor(D$replaced)
D$shutdown="parada corta"#ESCALERA PARADA EN EL MOMENTO DEL REPORTE
D$shutdown[grep("onprogress|will|tomorrow|followup|workinprogress|further|shutdown|kept|keep|
futur|tempor|temp|temporari",D$Action_en)]="parada larga"
D$shutdown=factor(D$shutdown)
D$accident="averia"#FALLOS DEBIDO A USUARIO#añadir passeng en accidente
D$accident[grep("mistakenly|coin|ladi|woman|man|garbag|artifici|kick|umbrella|mous|injur|wheel
chair|accident|rat|nofaultfound|stone|vandal|passeng|balanc|cloth|shoe|stuck|somebod|misusag|u
ser|ball|object|foreign|hit|trolley|luggag",D$Action_en)]="accidente"
D$accident=factor(D$accident)
D$max="Suddenly"#FALLOS PREDECIBLES
D$max[grep("loud|smell|vibrat|unstabl|fluctuat|hot|frequent|nois",D$Action_en)]= "Detectable"
D$max=factor(D$max)
D$fall=0#CAIDAS
D$fall[grep("balanc",D$Action_en)]=1
D$fall=factor(D$fall)
```

Figura- 5.42 Código para variables léxicas binarias

6.- Resultados obtenidos

Una vez concluida la fase de acondicionamiento de la base de datos se procede a realizar distintas representaciones gráficas que permitan obtener más información relativa a tres grandes cuestiones.

1. Cómo es el parque de equipos cuyo mantenimiento lleva Thyssenkrupp y cuál es su desempeño habitual en cuanto al mantenimiento correctivo.
2. Qué fallos son los más habituales, bajo qué condiciones se producen y qué características de los equipos los hacen más proclives a sufrir incidentes.
3. Cómo es la actuación de los operarios a la hora de solucionar los incidentes y cuáles son las variables que tienen un mayor efecto en el tiempo de reparación.

La respuesta a estas preguntas permitirá ajustar el mantenimiento llevado a cabo para que se adapte a las necesidades que se hayan detectado y mejorar así el uso de recursos por parte de la empresa.

Puede parecer que está información —al menos la relativa al primer punto— debiera ser perfectamente conocida por la empresa previamente a la elaboración de este trabajo. Sin embargo, las delegaciones funcionan casi independientemente unas de otras por lo que la integración entre sus bases de datos no es muy alta y no se tienen datos globales fiables.

6.1.- DESCRIPCIÓN GENERAL DE LOS RESULTADOS

Antes de proceder a entrar en detalle en los resultados obtenidos, se ofrece un panorama general en figura- 6.1. Este da una primera impresión sobre el volumen de fallos en el periodo que va de noviembre de 2017 a noviembre de 2018. También permite ver el tamaño de la división de mantenimiento de Thyssenkrupp en el área de Asia-Pacífico.

Figura- 6.1 Totales de las principales variables de la base de datos

6.2.- CARACTERIZACIÓN DEL PARQUE DE EQUIPOS

6.2.1.- Caracterización geométrica

En primer lugar, es conveniente analizar las dimensiones más frecuentes de las escaleras y pasillos. Esto permitirá hacerse una idea de la homogeneidad del parque y ver así si existen equipos con dimensiones muy distintas a los demás, los cuales pueden tener un comportamiento diferente en cuanto a incidentes. Para ello, se construyen los histogramas o diagramas de barras en frecuencia de las siguientes variables:

- **Desnivel (Rise):** este es el parámetro que más condiciona el correcto funcionamiento del equipo, de acuerdo con los técnicos de Thyssenkrupp. Esto se debe a que un mayor desnivel implica un mayor esfuerzo para elevar la carga, al tener que vencer la gravedad. Por este motivo también se espera que los pasillos, cuya disposición es mayoritariamente horizontal, presenten un menor número de fallos y de menor gravedad que las escaleras. Se muestra el histograma en la figura- 6.2, donde cada barra representa un intervalo de 2 metros.

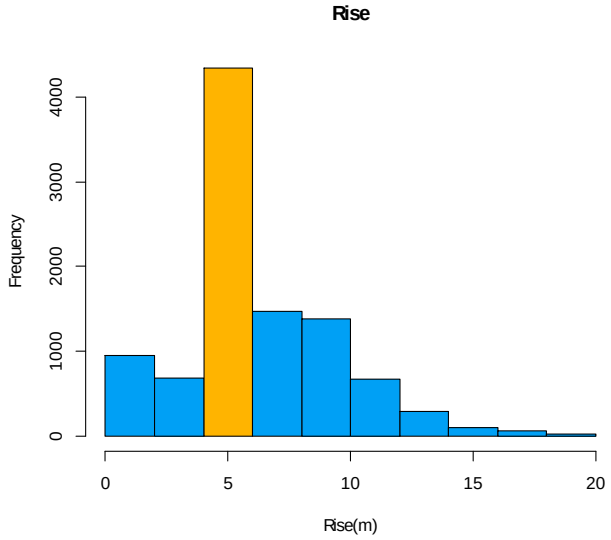


Figura- 6.2 Histograma del desnivel de los equipos instalados

Se puede ver que el valor más habitual —casi el 50% de los equipos— está entre 4-6 metros. Además, se observa que apenas hay equipos con un desnivel superior a 10 metros, por lo que se considera que, exceptuando los pasillos (equipos de 0-2m principalmente), esta variable no debería condicionar en exceso el desempeño en cuanto a fallos.

- **Distancia entre apoyos:** esta variable debería seguir un comportamiento casi idéntico al del desnivel, puesto que ambas se relacionan por el ángulo de inclinación y este suele adoptar valores normalizados. Se presenta el histograma, con barras de 5m de intervalo, en figura- 6.3. Se confirma que la distribución es muy similar a la

anterior, siendo el valor más habitual en este caso 10-15 metros y estando cerca del 90% de los equipos entre 10 y 30 metros.

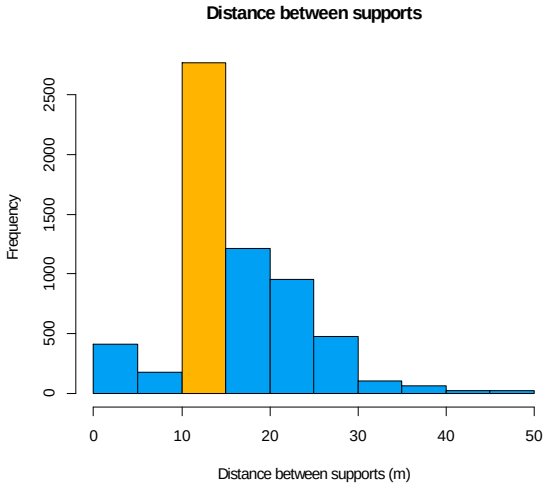


Figura- 6.3 Histograma de la distancia entre apoyos en los equipos

- **Ángulo de inclinación:** finalmente, para confirmar la prevalencia de esos ángulos normalizados se muestra la distribución de esta variable en figura- 6.4. Se observa que existen 4 intervalos distintivos, cada uno de los cuales se corresponde con una aplicación que se explica a continuación y se muesta también en figura- 6.4.
  0-5°: pasillos rodantes en aeropuertos, con disposición horizontal o muy ligeramente inclinada.
  10-15°: habitualmente 10° o 12°, pasillos rodantes en centros comerciales.
  25-30°: habitualmente se emplea 27, 5° en aeropuertos y 30° en metros y estaciones de tren. También son empleados en centros comerciales
  30-35 °: son característicos de las escaleras mecánicas en centros comerciales y otras aplicaciones de uso poco intensivo. Los de 35° se emplean casi exclusivamente en este segmento, aunque su uso es minoritario.

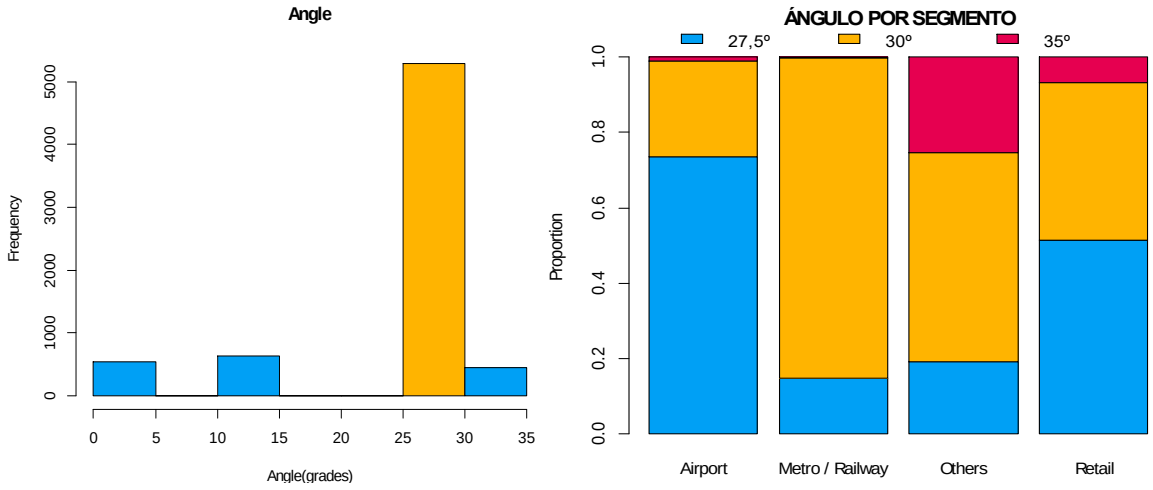


Figura- 6.4 Histograma del ángulo de inclinación en los equipos (izquierda) y ángulo por segmento (derecha)

Finalmente, para confirmar que efectivamente hay esa correspondencia entre ángulo y aplicación, se muestra en un grafico circular en la figura- 6.5 el número de equipos en función de si son escaleras o pasillos.

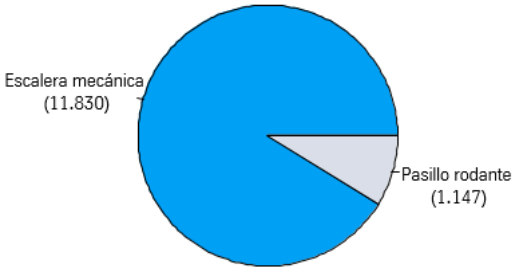


Figura- 6.5 Gráfico circular del tipo de unidad

Se puede ver que coincide perfectamente el número de pasillos con la suma de equipos con inclinaciones entre 0 y 15°. En cambio, el número de escaleras es mucho mayor que el número de equipos con inclinación entre 30° y 35°. Esto se explica por la incompletitud de los informes. Un operario sabe que el ángulo de un pasillo es 0 o muy cercano a la horizontal. En cambio, conocer la inclinación exacta de la escalera exige hacer más mediciones y son más complejas de realizar. Esto lleva a que en muchas ocasiones el técnico no introduzca ningún valor en el campo de distancia entre apoyos, impidiendo así obtener el ángulo para muchas de las escaleras.

Velocidad de los escalones: al igual que sucedía con los ángulos, las velocidades están fuertemente controladas por la norma y solo son posibles 3 valores distintos: 0.5, 0.63 y 0.75 m/s. Por tanto, en este caso se emplea en la figura- 6.6 un gráfico de barras en vez de un histograma. Se puede apreciar que el valor de 0,75 m/s casi no se emplea —apenas en 200 de los 12000 equipos— y, de hecho, está prohibido en la legislación nacional de muchos países por considerarlo excesivamente alto y aumentar el riesgo de accidentes.

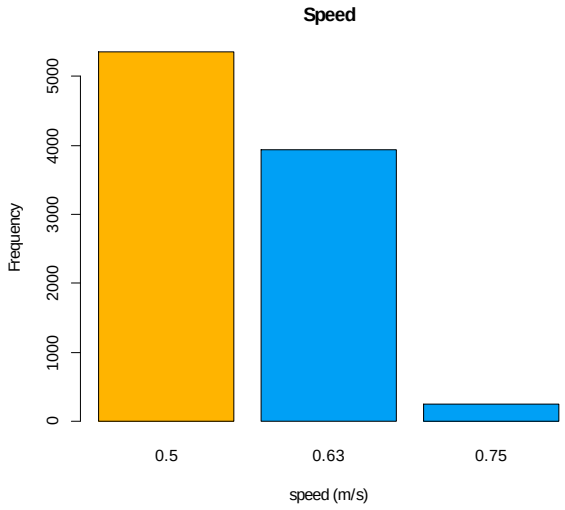


Figura- 6.6 Diagrama de barras de velocidades en equipos

- **Tiempo de viaje:** se combinan todas las variables anteriores para mostrar en la figura- 6.7 el histograma del tiempo de viaje de los equipos, es decir, el tiempo que tarda un escalón desde una plataforma de llegada a la otra.

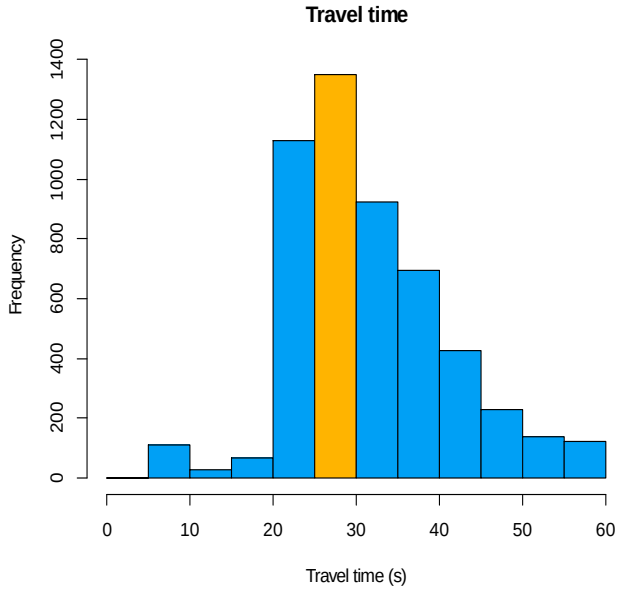


Figura- 6.7 Histograma del tiempo de viaje en los equipos

En este caso se observa que la moda, esto es, el valor más habitual, está entre 25 y 30 s, pero hay una mayor dispersión de los datos comparada con los gráficos anteriores al entrar más variables en juego. Así pues, al tener esta característica mayor variabilidad que las demás y definir mejor el comportamiento geométrico del equipo, será una de las empleadas para detectar tendencias dependientes de la geometría.

- **Ancho nominal de los escalones:** finalmente, se presenta en la figura- 6.8 el diagrama de barras —al tener valores normalizados— del ancho de peldaños. Nótese que la mayoría de equipos presentan un ancho de 1000mm, siendo muy poco frecuentes los anchos superiores y aún menos los inferiores.

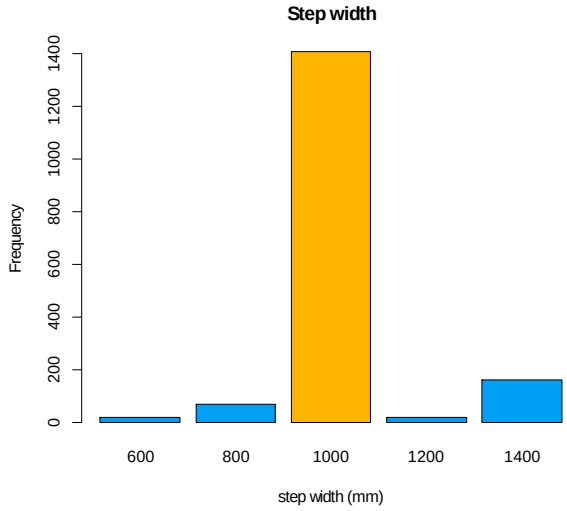


Figura- 6.8 Diagrama de barras del ancho de peldaños en los equipos

También se observa que entre todas las columnas no suman los 12000 equipos que se maneja. Esto se cree que es debido a que el ancho de 1000 mm es un estándar de facto y, por tanto, los operarios no se molestan en introducir el dato en muchas de las ocasiones, solo indicándolo en aquellos casos en los que el ancho es distinto al habitual.

6.2.2.- Caracterización comercial

El otro aspecto que más condiciona la construccion y por tanto el funcionamiento es el diseño del equipo, el cual depende en gran medida de la marca, el modelo y la factoría. Por este motivo, se presenta la distribución por fabricante de las unidades en mantenimiento, así como la distribución por modelo en aquellas pertenecientes a Thyssenkrupp.

- **Fabricante:** se muestra en figura- 6.9 el gráfico circular con el porcentaje de equipos producidos por cada fabricante. Destacar que para este gráfico se distingue entre las tres fábricas de Thyssenkrupp (Alemania (tkE-DE), España (tkE-ES) y China (tkE-CH) a petición de la empresa, ya que cada fábrica actúa casi como una empresa independiente y, por tanto, interesa conocer su cuota de mercado.

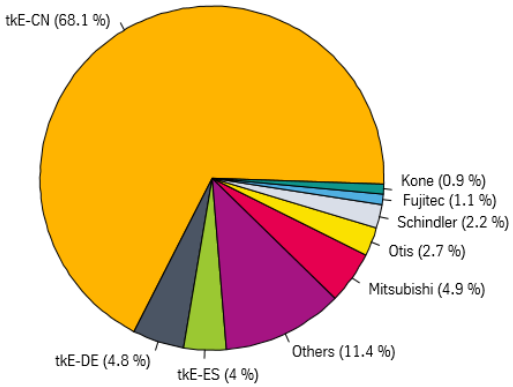


Figura- 6.9 Gráfico circular del fabricante de los equipos

Es notorio que la factoría de Thyssenkrupp en China es a la que pertenecen un mayor número de unidades. Esto no se debe tanto a que sea la fábrica de mayor tamaño en este sector, sino por los condicionantes de la base de datos manejada, centrada en Asia-Pacífico y en las escaleras mantenidas por Thyssenkrupp.

De esta manera, parece comprobarse que es habitual que la misma marca que suministra la escalera se encargue de su mantenimiento posterior, al menos es así en esta empresa. De todas formas, no hay que despreciar el hecho de que cerca del 20% de escaleras mantenidas por Thyssenkrupp fueron producidas por otras compañías.

Finalmente indicar que la categoría "Others" incluye tanto equipos cuya marca era desconocida, como aquellos cuyo fabricante era muy minoritario y se consideraba irrelevante para el análisis, como Hyundai, Toshiba, Orona o LG.

- **Modelo:** se centrará ahora el foco en Thyssenkrupp, puesto que es la compañía para la que se realiza este estudio y, además, es la mayoritaria en la base de datos. Se ofrece en figura- 6.10 el diagrama de barras con el número de equipos por modelo.

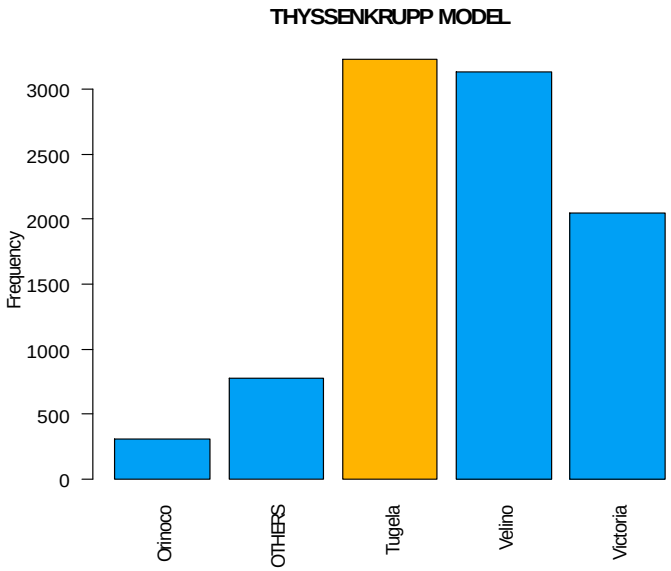


Figura- 6.10 Diagrama de barras de los modelos de los equipos

Los tres modelos que ocupan el pódium, en consonancia con lo visto en la caracterización geométrica, pertenecen a escaleras mecánicas. El modelo más habitual es el Tugela, que es el recomendado para estaciones de metro y aeropuertos por su diseño con mayores coeficientes de seguridad. Le sigue el Velino, destinado principalmente a centros comerciales, donde las cargas no son tan pesadas. En tercer lugar se sitúa el Victoria, específicamente diseñado para aplicaciones de servicio público con gran carga y, por tanto, empleado en metros de grandes ciudades, donde la carga en hora punta puede ser muy alta para un modelo Tugela.

El modelo Orinoco es el único que vende actualmente Thyssenkrupp en lo que a pasillos rodantes se refiere. Por eso, al igual que pasaba con el ancho de 1000 mm, el modelo Orinoco es el estándar en pasillos mecánicos y, por tanto, en muchas ocasiones no se introduce el dato al darlo por hecho. Así pues, se considera que la mayoría de los equipos cuyo modelo se desconoce son realmente Orinoco. De hecho, se puede observar que la suma de Orinoco y "Others" da prácticamente el número de pasillos rodantes en mantenimiento.

- **Segmento:** el uso que se le da a un modelo puede condicionar en gran medida su desempeño en cuanto al número y gravedad de fallos. Para comprobar que la aplicación a la que se destinan los equipos es la adecuada se muestra la distribución por segmento en figura- 6.11 y la distribución combinada de segmento y modelo en figura- 6.12.

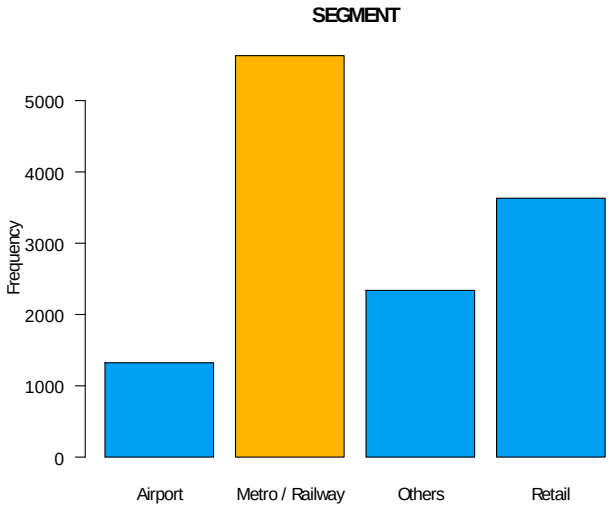


Figura- 6.11 Gráfico de barras del segmento al que pertenecen los equipos

Se aprecia que la aplicación más habitual de estos equipos es en el transporte público, lo que concuerda con que el Tugela sea el modelo más usado, al igual que el segundo puesto para el sector Retail y el Velino. Como apunte señalar que el segmento "Others" agrupa usos muy variados pero poco intensivos en cargas como hospitales, viviendas u oficinas.

Al combinar ambas distribuciones en figura- 6.12 se confirma que, mayoritariamente, cada modelo se emplea para la aplicación para la que fue diseñado: Victoria para metro, Tugela para metro y aeropuerto, Velino para comercial y Orinoco para centros comerciales y aeropuertos. Ademas, se observa que el segmento "Others", al ser de carga ligera, emplea preferentemente el modelo Velino de los centros comerciales. También se puede destacar ese pequeño porcentaje de escaleras Velino empleadas en segmentos donde las cargas suelen ser elevadas, como los aeropuertos (por las maletas) o el metro (por la hora punta). Por tanto, habrá que analizar si aparecen más problemas en este grupo.

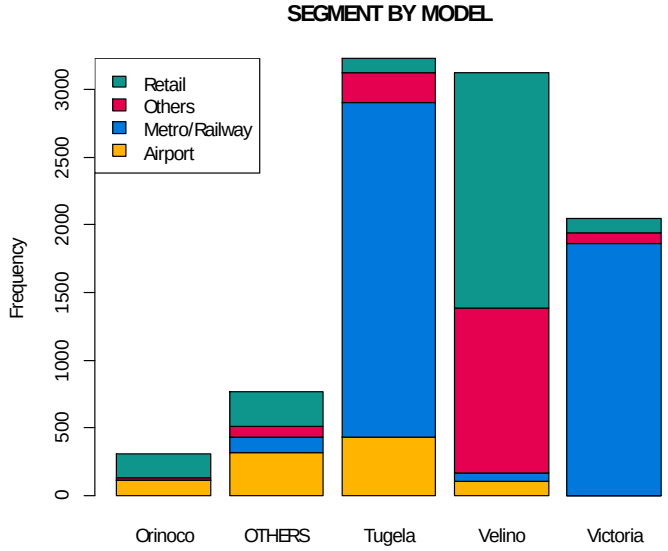


Figura- 6.12 Gráfico de barras apiladas del segmento para cada modelo

6.3.- CARACTERIZACIÓN DEL MANTENIMIENTO CORRECTIVO

6.3.1.- Histogramas de mantenimiento

El siguiente paso tras conocer los rasgos básicos que definen los equipos es analizar su comportamiento frente a fallos, es decir, cuáles son los valores de algunos de los parámetros ya introducidos en el apartado 4.1.2.-. Para ello, se ofrecen en figura- 6.13 los histogramas y una tabla resumen en tabla 6.1 con los estadísticos más significativos que dan una idea de la mantenibilidad (repair time), la disponibilidad (stopped time), el tiempo de respuesta desde que se conoce el fallo hasta que se empieza a trabajar en él (reaction time) y la tasa de fallos, dividida entre accidentes y averías.

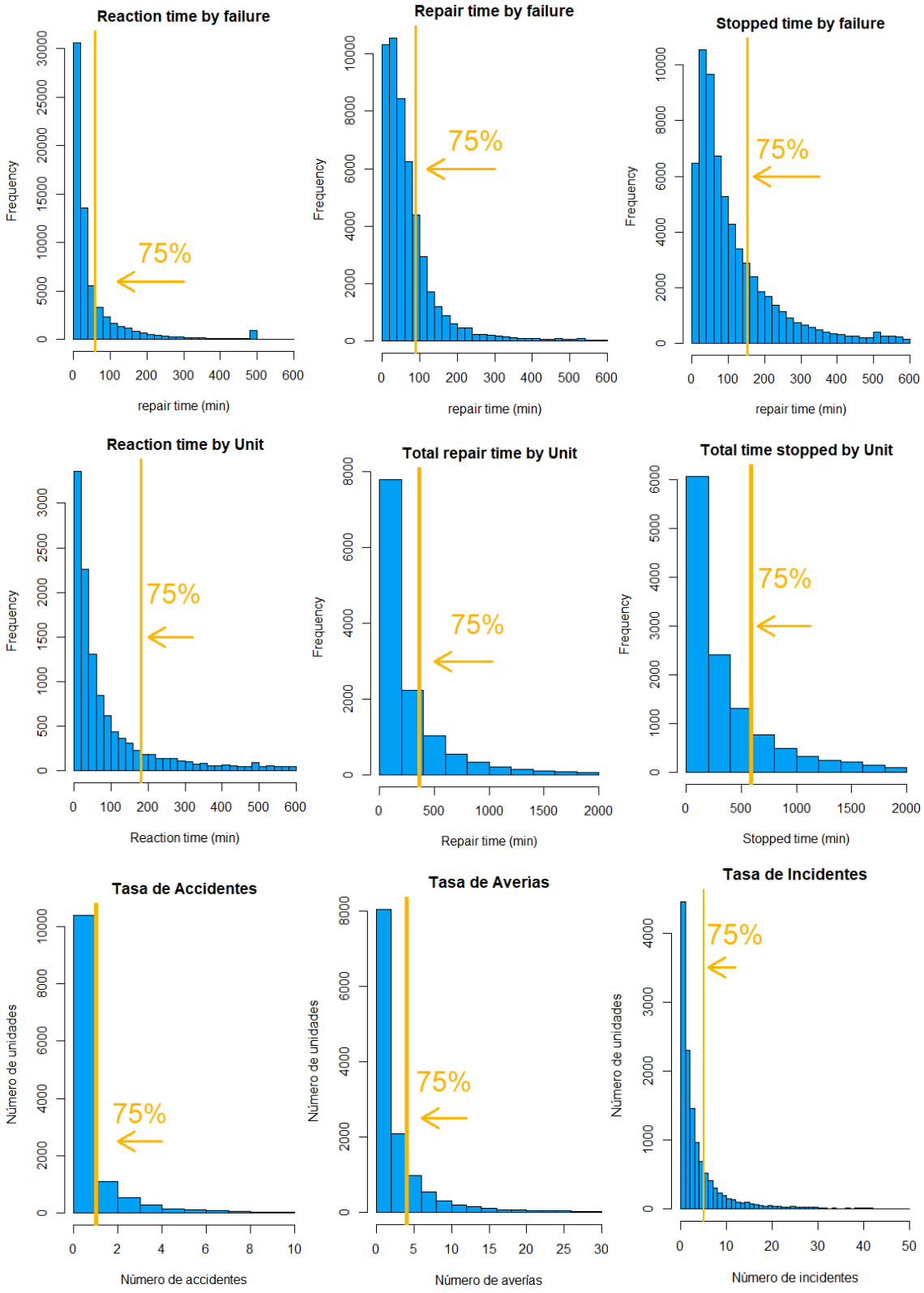


Figura- 6.13 Histogramas de las principales variables de mantenimiento correctivo

FALLO	Tiempo de reacción (min)	Tiempo de reparación (min)	Tiempo no disponible (min)
Mediana	23	42,5	76,9
3º cuartil	58	80	153
EQUIPO	Tiempo de reacción (min)	Tiempo de reparación (min)	Tiempo no disponible (min)
Mediana	53	138,2	224
3º cuartil	181	357,3	590
EQUIPO	Tasa de accidentes	Tasa de averías	Tasa de incidentes
Mediana	0	2	2
3º cuartil	1	4	5

Tabla 6.1 Estadísticos de las principales variables de mantenimiento correctivo

No se proporciona desde Thyssenkrupp el rango en el que deberían estar cada una de estas variables. Por tanto, no se puede saber si el mantenimiento cumple con los criterios esperados o hay que mejorar en alguna de las áreas. En principio, podría parecer que un tiempo no disponible de 10 horas al año por equipo (3º cuartil) es buen dato, ya que la disponibilidad del equipo es muy alta. De todas formas, se podrá realizar un análisis en mayor profundidad cuando se comparen entre sí las delegaciones en apartados posteriores.

6.3.2.- Función de supervivencia

Tal y como se vio en el apartado 4.1.2.-, el método principal para caracterizar el comportamiento de una máquina frente a fallos es su función de supervivencia *S(t).* En principio, esta función es desconocida, pero es posible estimar su valor mediante el método de máxima verosimilitud [15].

En primer lugar, para poder aplicar este método es necesario elegir una familia de funciones cuyos parámetros serán los que después se estimen mediante la aplicación de la verosimilitud. La función de supervivencia más común es la exponencial, que se define a continuación:

Exponencial: presenta tasa de fallos constante, es decir, no le afecta el envejecimiento, lo que se denomina falta de memoria. Esta función es habitual durante gran parte del ciclo vital de las máquinas ya que los fallos por fatiga no empiezan a producirse hasta la parte final de su vida. Adicionalmente las piezas más desgastadas suelen ser reemplazadas antes del fallo, eliminando así la influencia del envejecimiento. Su función de supervivencia y de densidad son:

$$S(t) = 1 - F(t) = e^{-\lambda \cdot t} \tag{6.1}$$

$$f(t) = \lambda \cdot e^{-\lambda \cdot t} \tag{6.2}$$

Existen otras familias de funciones como la Weibull o la Gamma, que son generalizaciones de la exponencial. No se considerarán en este caso ya que se estima no hay datos suficientes como para ajustar todos sus parámetros.

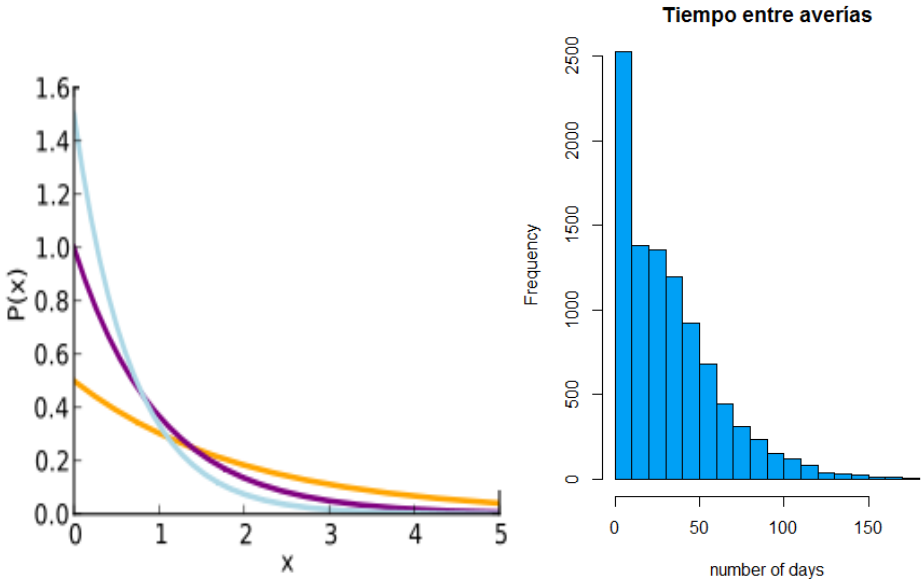


Figura- 6.14 Distribuciónes exponenciales (izquierda) e histograma de tiempo entre fallos (derecha)

Para comprobar si los datos podrían ajustarse a una exponencial se representan en forma de histograma. Con ese fin, se calcula, solo para las averías (los accidentes son, en principio, aleatorios y se comportarían como ruido en esta distribución por lo que se desprecian), el tiempo que pasa entre dos averías consecutivas para el mismo equipo (figura- 6.14).

Se puede observar que la gráfica presenta una distribución similar a la exponencial dibujada en figura- 6.14. Por tanto, parece adecuada la elección del modelo exponencial como función de supervivencia. Otro método para comprobar si esta distribución es la adecuada consiste en emplear un gráfico cuantil-cuantil (Q-Q) como el de la figura- 6.15.

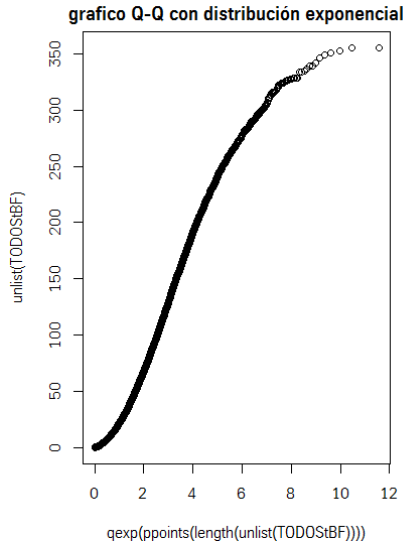


Figura- 6.15 Gráfico Q-Q respecto a dsitribución exponencial

Si todos los puntos quedan alineados los datos siguen la distribución de la función con la que se está comparando. En este caso hay colas a izquierda y derecha por lo que podría decirse que el ajuste no es perfecto, pero debería ser suficiente para hacer una aproximación inicial.

Ahora que se ha elegido una función de supervivencia, solo queda ajustar sus parámetros, en este caso λ, para lo cual se empleará, como ya se ha comentado anteriormente, el método de máxima verosimilitud.

La verosimilitud *L(λ)* es una función que indica cómo de adecuado es un parámetro λ de un modelo estadístico para explicar un conjunto de observaciones $x_1, x_2, \ldots, x_n$. Es decir, la función de verosimilitud indica cuán verosímil, o posible, es que un valor concreto del parámetro sea el que le corresponde a un conjunto conocido de medidas de una variable.

Matemáticamente, se definede la forma:

$$L(\lambda|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\lambda) \tag{6.3}$$

Es habitual trabajar con los logaritmos de la verosimilitud:

$$l(\lambda|x_1, x_2, \ldots, x_n) = Ln(L(\lambda|x_1, x_2, \ldots, x_n)) = \sum_{i=1}^{n} Ln(f(x_i|\lambda)) \tag{6.4}$$

A partir de dicha función, se puede obtener el estimador de máxima verosimilitud (EMV) derivando e igualando a cero. Este será el valor más probable del parámetro de la distribución estadística para los datos considerados, esto es, el que mejor debería ajustar las observaciones a la distribución.

Dado que se ha escogido como distribución la función exponencial, el EMV se calculde la forma:

$$L(\lambda|t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} \lambda \cdot e^{-\lambda * t_i} = \lambda^n \cdot e^{-\lambda * \sum_{i=1}^{n} t_i} \tag{6.5}$$

$$l(\lambda|t_1, t_2, \ldots, t_n) = n \cdot Ln(\lambda) - \lambda \cdot \sum_{i=1}^{n} t_i \tag{6.6}$$

$$\frac{\partial l(\lambda|t_1, t_2, \ldots, t_n)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i \tag{6.7}$$

Igualando a cero y despejando se obtiene:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\hat{t}} \tag{6.8}$$

Siendo $\hat{t}$ el tiempo medio entre fallos, también denominado MTBF como se había visto en 4.1.2.-.

La fórmula expuesta es válida siempre que se conozca el tiempo que pasa desde que se produce un fallo hasta el siguiente en todos los individuos de la muestra. Sin embargo, en este caso solo se dispone de los reportes generados desde noviembre de 2017 a noviembre de 2018 y hay muchos equipos que en ese intervalo fallan una sola vez, es decir, no se conoce cuánto tardan en fallar de nuevo.

Este fenómeno se denomina censura y es muy habitual en el análisis de supervivencia. Un claro ejemplo son los estudios médicos, donde se suministra un tratamiento a un grupo de pacientes con *n* individuos. En el periodo considerado mueren *m* de los *n* tratados, cada uno de ellos en el momento t_i desde que se le administró el tratamiento, mientras que *n-m* pacientes sobreviven tras completar el periodo de seguimiento de cada uno de ellos t_i. No se conoce cuánto más vivirán, pero se sabe que han sobrevivido al menos todo el periodo considerado. Matemáticamente, la función de verosimilitud con censuras se expresa como:

$$L(\lambda) = \prod_{i=1}^{m} f(t_i|\lambda) \cdot \prod_{i=m+1}^{n} (1 - F(x_i|\lambda)) \tag{6.9}$$

$$l(\lambda) = Ln(\lambda) = \sum_{i=1}^{m} Ln\big(f(t_i|\lambda)\big) + \sum_{i=1}^{n} Ln\big(1 - F(x_i|\lambda)\big) \tag{6.10}$$

$$L(\lambda) = \prod_{i=1}^{m} \lambda \cdot e^{-\lambda \cdot t_i} \cdot \prod_{i=m+1}^{n} (1 - F(x_i|\lambda)) = \lambda^m \cdot e^{-\lambda * (\sum_{i=1}^{m} t_i + \sum_{i=m+1}^{n} x_i)} \tag{6.11}$$

$$l(\lambda|t_1, t_2, \dots, t_n) = m \cdot Ln(\lambda) - \lambda \cdot (\sum_{i=1}^{m} t_i + \sum_{i=m+1}^{n} x_i) \tag{6.12}$$

$$\frac{\partial l(\lambda|t_1, t_2, \dots, t_n)}{\partial \lambda} = \frac{m}{\lambda} - (\sum_{i=1}^{m} t_i + \sum_{i=m+1}^{n} x_i) \tag{6.13}$$

Igualando a cero y despejando se obtiene:

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{m} t_i + \sum_{i=m+1}^{n} t_i} = \frac{m}{n * T} \tag{6.14}$$

Siendo T el tiempo medio entre fallos incluyendo los datos censurados.

Una vez que se conoce cómo calcular el parámetro solo queda decidir qué datos se usarán. La elección no es baladí ya que los datos disponibles no son los más habituales de este tipo de estudio. Hay algunos equipos para los que se conocen varios tiempos entre fallos, mientras hay otros que solo tienen un fallo y se podría considerar que están censurados por la izquierda (el tiempo de vida sería desde el inicio del estudio hasta el fallo), por la derecha (el tiempo de vida sería desde el fallo hasta el fin del estudio) o censurado por ambos lados.

Por tanto, hay que tomar dos decisiones: en primer lugar, se empleará la censura por la derecha por considerarla más representativa. Hay que tener en cuenta que si se usara la censura por la izquierda podría tomarse como tiempo inicial un instante en el que la máquina no ha sido instalada todavía, puesto que muchas fueron dadas de alta durante el desarrollo de la recogida de datos, y esto no tendría sentido.

Por otro lado, puesto que se considera como individuo de la muestra a cada una de las máquinas mantenidas, solo puede existir un dato por cada una de ellas. Los principales candidatos serían uno de todos los tiempos entre fallos de cada máquina al azar, el máximo o mínimo tiempo entre fallos o el valor medio. Se ha optado por esta última alternativa al pensar que sería más representativa de la función de supervivencia que se va a calcular. Se muestra en la figura- 6.16 la función cogiendo los datos sin y con censura y en figura- 6.17 el código necesario para calcularla.

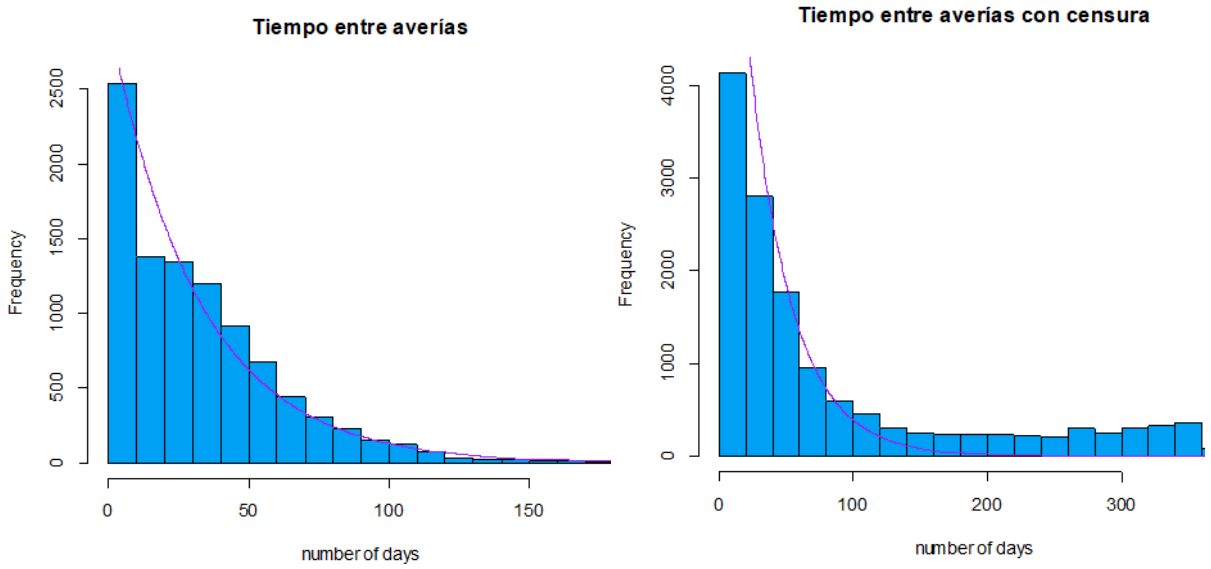


Figura- 6.16 Histograma de tiempo entre fallos sin censura (izquierda) y con ella (derecha)

```
fallosdistintos3=tapply(D$FaultareaDEF[D$accident=="averia"],
D$UnitNumber[D$accident=="averia"],function(x)length(x))
equiposconunfallo=names(fallosdistintos3)[which(fallosdistintos3==1)]
Dseveral=D[-match(unlist(equiposconunfallo),D$UnitNumber),]
TBFM=tapply(Dseveral$LoggedDate,Dseveral$UnitNumber,function(x)sum(as.numer
ic(diff(c(x))/length(x),units="days")))#EL MIN PERMITE VER OUTLIERS
TBFM=unlist(TBFM)[-which(is.na(unlist(TBFM)))]
hist(TBFM,col="blue",main="Tiempo entre averías",xlab="number of days")
Donli=D[match(unlist(equiposconunfallo),D$UnitNumber),]
TODOStBF2=(by(Donli$LoggedDate,Donli$UnitNumber,function(x)
(as.numeric(diff(c(x,MaxDate)),units="days"))))
todostbf2=unlist(TODOStBF2)[-which(is.na(unlist(TODOStBF2)))]
TUTI=c(todostbf2,TBFM)
hist(TUTI,col="blue",main="Tiempo entre averías",xlab="number of days")
lambda=length(TBFM)/sum(TBFM)
lambda2=length(TUTI)/sum(TUTI)
plot(density(TBFM))
g<- function(l) lambda*exp(-lambda*l) ## Muestra truncada
v<- seq(0, 365, by=0.01)
lines(v, g(v), col='purple')
hist(TBFM,col="blue",main="Tiempo entre averías",xlab="number of days")
g<- function(l) lambda*exp(-lambda*l)*length(TBFM)*20 ## Muestra truncada
v<- seq(0, 365, by=0.01)
lines(v, g(v), col='purple')
plot(density(TUTI))
g<- function(l) lambda2*exp(-lambda2*l) ## Muestra truncada
lines(v, g(v), col='green')
hist(TUTI,col="blue",main="Tiempo entre averías con censura",
xlab="number of days")
g<- function(l) lambda*exp(-lambda*l)*length(TUTI)*20 ## Muestra truncada
v<- seq(0, 365, by=0.01)
lines(v, g(v), col='purple')
```

Figura- 6.17 Código para estimar la distribución que sigue el tiempo entre fallos

Se puede observar que el histograma es relativamente similar a la curva exponencial, si bien el ajuste no es perfecto, lo que puede achacarse a los pocos datos disponibles o a que se requiere un modelo más complejo como el Weibull. No obstante, se acepta como válida esta aproximación inicial con censuras con λ= 0,0117 y MTBF=85,32 días. Es decir, los equipos fallan de media una vez al trimestre.

6.4.- CARACTERIZACIÓN DE LAS DELEGACIOENS

6.4.1.- Tamaño y segmento de sus escaleras

Una vez analizados los condicionantes de los propios equipos es conveniente ver cómo se distribuyen geográficamente. Las delegaciones no tienen influencia en cuanto a la construcción e instalación de las unidades, pero sí la tienen en el mantenimiento de los equipos. De ellas dependen las visitas periódicas para prevenir fallos, así como las reparaciones en caso de incidente. Además, suelen ser las delegaciones las que forman y

contratan a los operarios y en algunos casos las que deciden qué proveedor suministrará algunos repuestos, por lo que condicionan en gran medida el funcionamiento de los equipos.

- **Tamaño de las delegaciones:** en primer lugar, se presenta en la figura- 6.18 un gráfico combinando el número de equipos (eje izquierdo) y el de informes (eje derecho) para dar una noción sobre cómo de relevante es cada país. También se representa en la figura- 6.19 sobre un mapa cada delegación con una burbuja, cuyo tamaño es proporcional a su número de informes. Este segundo gráfico no aporta nueva información, pero permite tener una impresión más visual y cualitativa de la importancia de cada mercado.

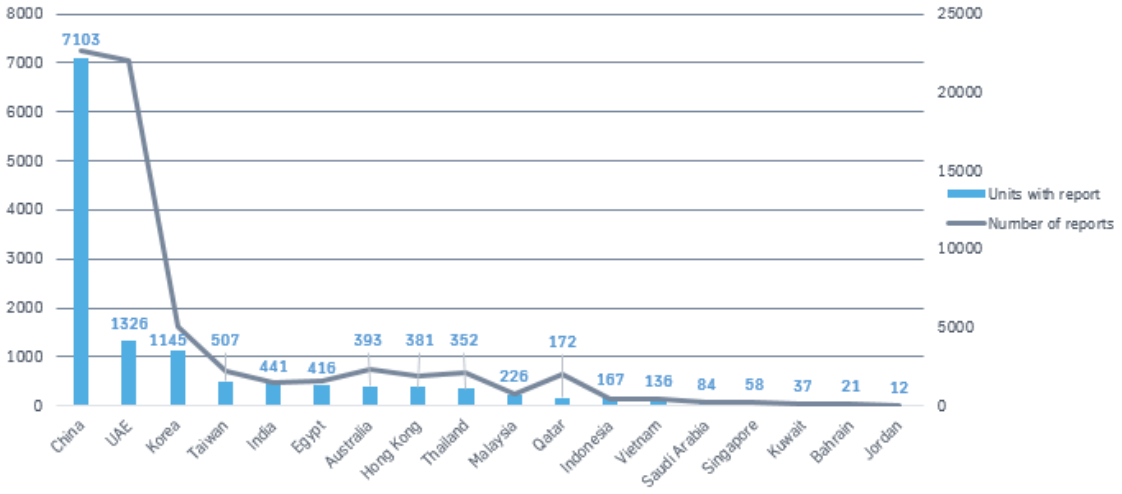


Figura- 6.18 Número de equipos y reportes de cada país

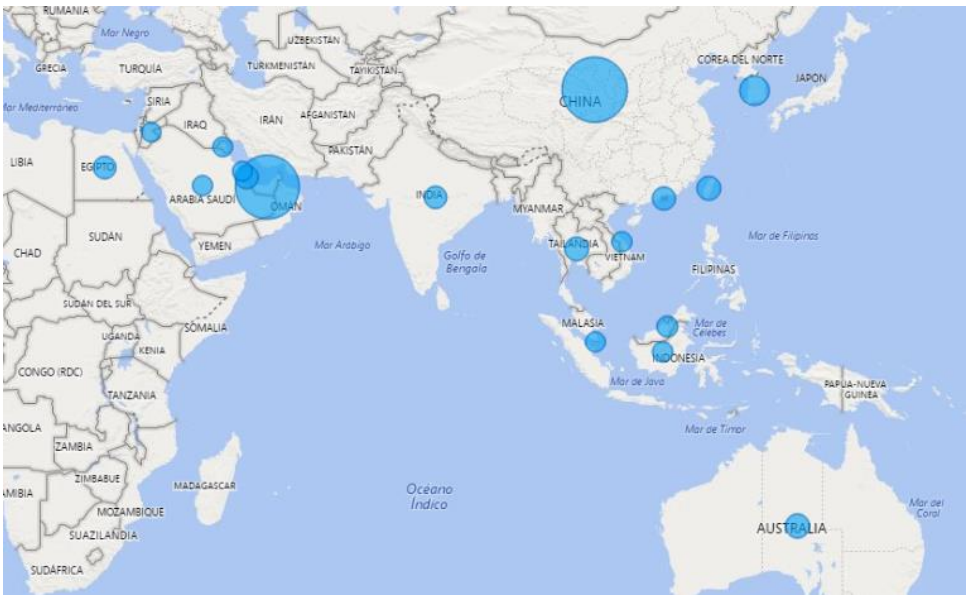


Figura- 6.19 Número de reportes de cada país en formato mapa

Se puede apreciar que China es el país con más equipos instalados, seguido muy de lejos por Emiratos Árabes Unidos (Principalmente Dubái) y Corea del Sur. Sin embargo, en cuanto a la cantidad de incidentes, el número es similar en China y UAE, lo que requerirá una investigación más exhaustiva de las causas de este fenómeno.

Para intentar capturar mejor los motivos de este elevado número de incidentes se desagregará a partir de ahora UAE en sus dos componentes: UAE, que se refiere al país en general y cuenta con 925 de los 1336 equipos, y UAEDIA con los 401 restantes, situados en el aeropuerto internacional de Dubái.

- **Segmentos en cada país:** con el fin de comprobar que la etiqueta UAEDIA se corresponde con los equipos instalados en el aeropuerto de Dubái, y para comprender mejor qué equipos mantiene cada delegación, se ofrece en la figura- 6.20 la distribución por segmento como proporción del total de equipos de cada país.

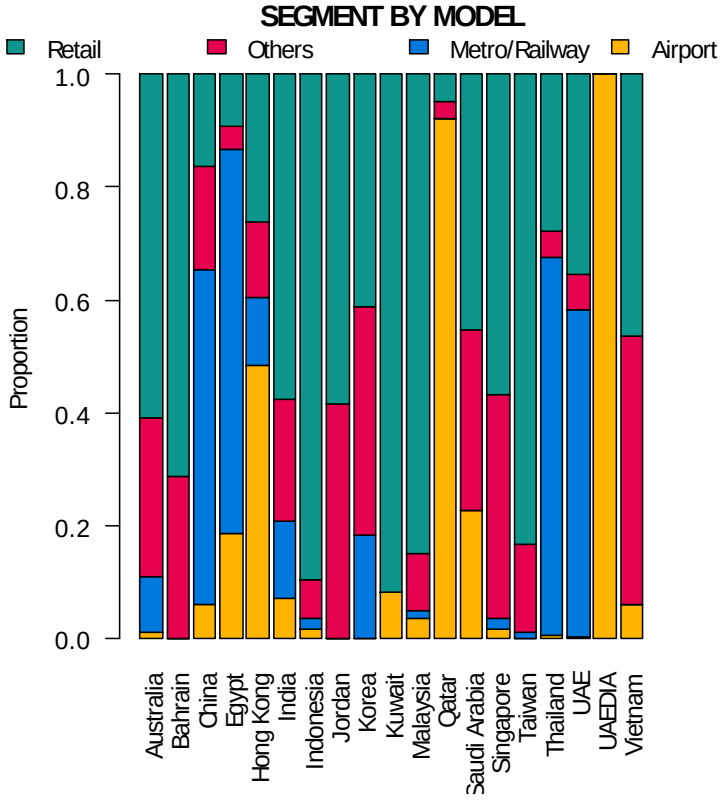


Figura- 6.20 Proporción de equipos de cada segmento en cada delegación

Se confirma que los equipos de UAEDIA pertenecen al aeropuerto. Además, se observa que existen tres tipos de países: aquellos en los que la mayoría de unidades pertenecen a aeropuertos (UAEDIA, Hong Kong y Qatar), aquellos con mayoría en el metro (China, Egipto, Tailandia y UAE) y aquellos con mayoría comercial o similar.

Esto podría condicionar diversos aspectos del mantenimiento ya que tampoco habrá una distribución homogénea de modelos entre las delegaciones.

6.4.2.- Parámetros de frecuencia del servicio de mantenimiento correctivo

Una primera aproximación para comprobar si es relevante la diferencia existente entre delegaciones en cuanto a la aplicación y modelo mayoritario, es analizar las principales variables del mantenimiento correctivo, que pueden depender del segmento y de los operarios de la delegación. Estas se detallan a continuación.

- **Accidentalidad:** se define la accidentalidad como la proporción entre el número de accidentes —aquellos con valor "accidente" en la variable *accident*— y el total de incidentes registrados. Se muestra la accidentalidad en la figura- 6.21. La línea roja indica la media global.

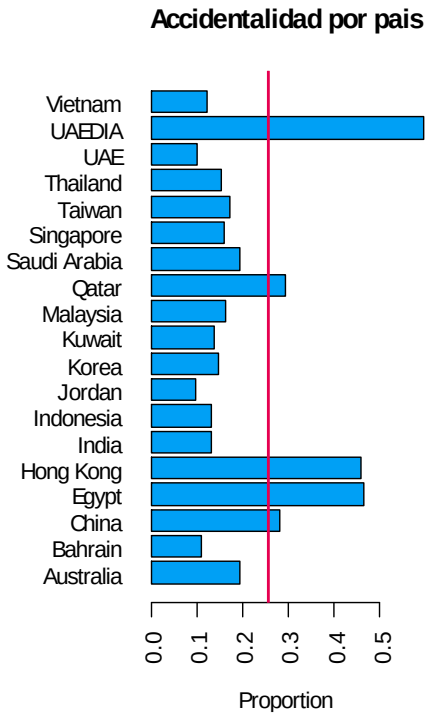


Figura- 6.21 Accidentalidad por país

Llaman la atención los valores de Hong Kong, UAEDIA, Egipto y Qatar, superiores al 30%, en contraposición al 15% de la mayoría de países. Se ha preguntado a expertos de Thyssenkrupp para encontrar una posible explicación y se ha descubierto una característica diferenciadora entre al menos dos de esas delegaciones y el resto.

En los aeropuertos de Hong Kong y Dubái trabajan cuadrillas de Thyssenkrupp dedicadas en exclusiva a la supervisión de los equipos situados en estos edificios, al contrario que en el resto de delegaciones donde los operarios suelen ser itinerantes y van allí donde se produzca el fallo, teniendo que desplazarse en furgoneta u otro medio de transporte.

En la mayoría de edificios, los incidentes menores, como algunas activaciones del pulsador de emergencia o atascos en la placa de peines, pueden ser resueltos por el propio personal del establecimiento sin necesidad de llamar a operarios de Thyssenkrupp, los cuales tardarían un tiempo en llegar y solucionarlo. Al no intervenir ningún operario de la delegación, no queda registro de esos pequeños accidentes. Por tanto, en la mayoría de edificios solo se llama a mantenimiento cuando es imprescindible, por lo que la accidentalidad será baja.

En cambio, en estos dos aeropuertos, el personal de Thyssenkrupp es el encargado de supervisar los equipos, y sus operarios son los primeros en darse cuenta del fallo, motivo por el cual son ellos quienes reparan hasta el más minimo incidente y dejan registro escrito de ello, aumentando con ello la accidentalidad.

Además, la diferencia entre UAEDIA y Hong Kong se debe a que este último incluye más edificios que el aeropuerto, por lo que mezcla ambas maneras de solucionar los incidentes menores. Si se reduce el estudio a su aeropuerto, su accidentalidad sube hasta cotas superiores al 70%, por encima incluso del de Dubái.

En Egipto y Qatar es probable que suceda algo similar, pero desde el centro de Thyssenkrupp en Gijón no conocen dicha situación de primera mano —al contrario que en los otros dos aeropuertos en los que conocen a los responsables de las cuadrillas— por lo que será necesario intentar confirmarlo mediante un análisis profundo de la base de datos.

Este hallazo reafirma la necesidad de desprenderse de los accidentes a la hora de comparar delegaciones, segmentos y modelos, ya que la accidentalidad de los aeropuertos también estará inflada frente a las demás como se ve en tabla 6.2.

Segmento	Airport	Metro/Railway	Others	Retail
Accidentalidad	49,3%	21,0%	18,1%	20,8%

Tabla 6.2 Accidentalidad por segmento

También se muestra el porcentaje global en la figura- 6.22.

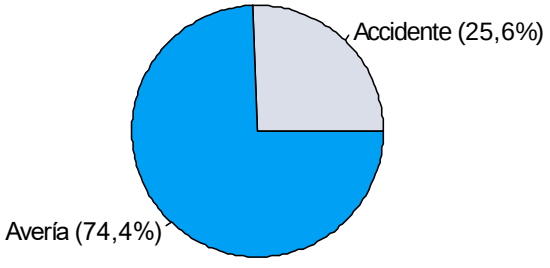


Figura- 6.22 Gráfico de pastel con los accidentes y averías

Adicionalmente, también se podría estimar, por medio de la diferencia de la accidentalidad entre los aeropuertos con delegación propia y el resto de edificios, el

número de incidentes producidos en los equipos en los cuales no se llama al personal de Thyssenkrupp. Esto permitiría hacerse una idea más certera del tiempo que no están disponibles las escaleras, al tener en cuenta incidentes menores, cuya frecuencia sería difícil de conocer por otros medios.

- **Tasa de fallos:** otra manera más completa de ver el efecto de la accidentalidad, es mediante la tasa de fallos, es decir, el número de fallos por equipo. Se muestra en la figura- 6.23 la tasa de fallos dividida entre la de accidentes y la de averías, siendo la línea roja la media de la tasa de averías.

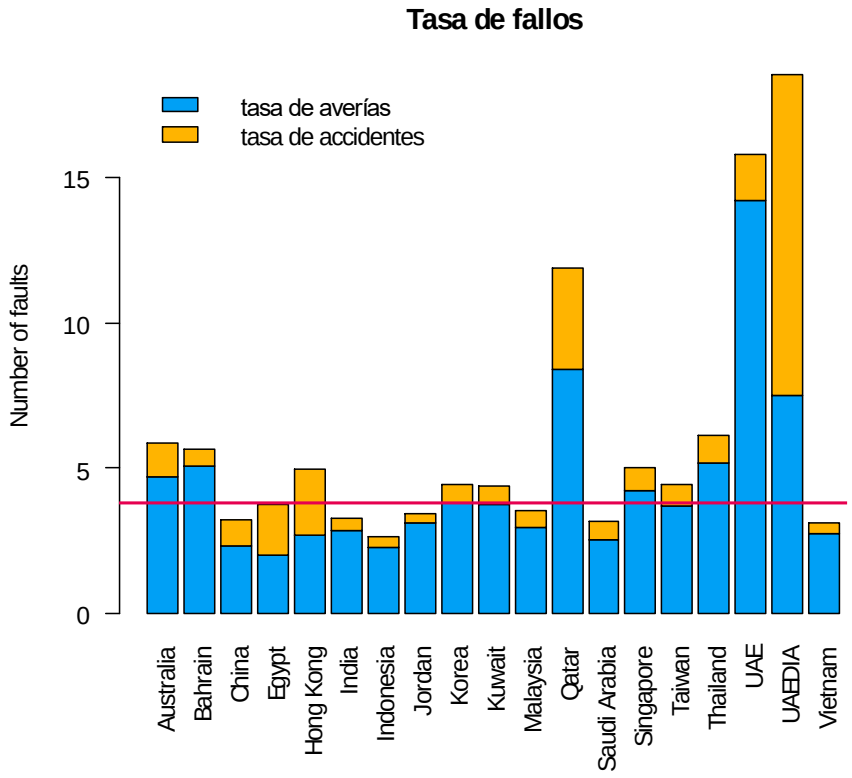


Figura- 6.23 Tasa de fallos por país

Se vuelve a apreciar la gran tasa de accidentes en Hong Kong, UAEDIA, Qatar y Egipto en proporción a su tasa de fallos total. Además, este gráfico permite ver que el gran número de incidentes que se observó originalmente en UAE parece deberse a averías reales en todo el país, siendo la tasa de averías en su aeropuerto relativamente similar a la media global de todas las delegaciones. También es destacable la gran tasa de averías en Qatar, que requerirá una investigación en profundidad. El resto de países presentan tasas muy parecidas entre sí, con una tasa de averías cercana a 3-4 y de accidentes de alrededor de 1.

6.4.3.- Parámetros de tiempo del servicio del mantenimiento correctivo

Además del número de fallos, tanto absoluto como relativo al parque de equipos, y su tipo, existe otro aspecto que define al servicio de reparación. Este es el tiempo que se tarda en

proporcionar dicho servicio. Esta característica define en gran parte la calidad percibida por el cliente y en principio depende en mayor medida del buen hacer y la organización de los operarios de cada delegación que de los equipos que haya instalados, si bien estos también tienen una influencia significativa.

Hay que destacar que existen ciertos *outliers* de duraciones que desvirtúan la media como estadístico de comparación, por ello se empleará la mediana, a la que le afectan menos dichos valores atípicos. Las principales variables temporales de la calidad del mantenimiento son:

- **Tiempo de reacción:** el tiempo que tarda el operario en llegar al lugar del incidente desde que se produce la llamada de aviso se relaciona exclusivamente con la organización de la delegación, es decir, con el número de operarios contratados, su disposición geográfica y su capacidad de respuesta. Por tanto, esta variable es la que mejor condensa el rendimiento de la delegación, al ser independiente del tipo de modelos y su uso. Se muestra el diagrama de cajas del tiempo de reacción en y el gráfico de barras con la mediana en figura- 6.24, la línea roja es la mediana global y se pintan de verde los 3 tiempos más bajos y de rojo los 3 más altos.

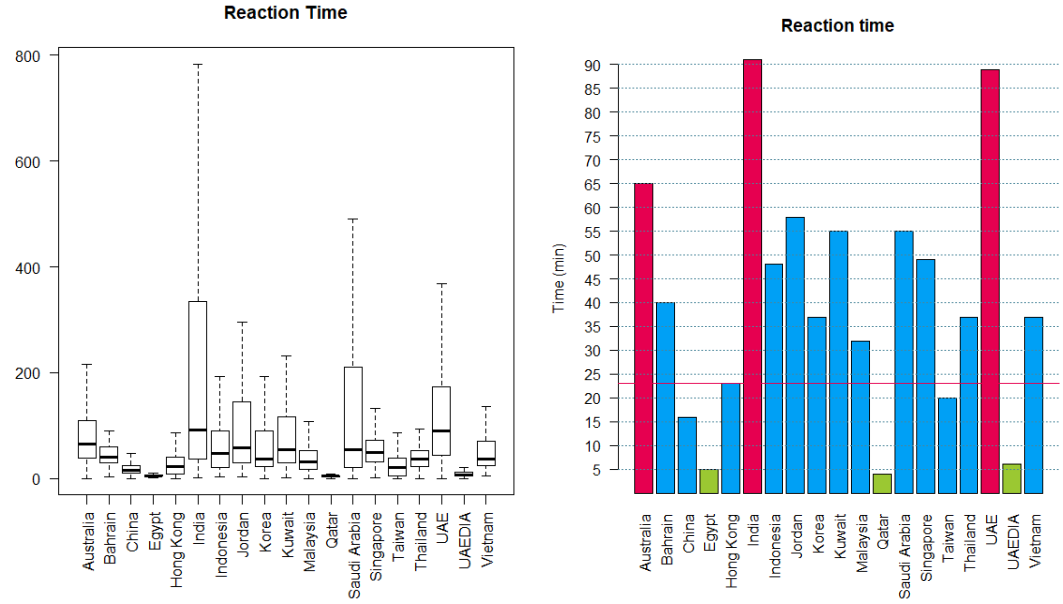


Figura- 6.24 Diagrama de cajas (izquierda) y mediana del tiempo de reacción al incidente (derecha) por país

Se puede apreciar que existe un grupo de delegaciones con un tiempo de reacción mediano inferior a 10 minutos, compuesto por las ubicadas en los aeropuertos de Dubái y Hong Kong —aunque este como país tenga una mediana de 20 minutos en su aeropuerto se reduce a 7— y por Egipto y Qatar. Estos datos de Qatar y Egipto vuelven a dar indicios de que cuentan con cuadrillas especializadas en algún edificio. De todas formas, se espera a conocer los resultados de las otras variables temporales para confirmarlo.

Por otro lado, el alto tiempo de reacción en Australia e India parece deberse a la gran amplitud geográfica de estos países y a la baja presencia de Thyssenkrupp en estos mercados, que obliga a los operarios a realizar grandes desplazamientos hasta el lugar del incidente.

Finalmente, la baja velocidad de respuesta en UAE tiene su origen en el bajo número de operarios de Thyssenkrupp en dicho mercado, puesto que una gran parte de equipos son de Mitsubishi y el contrato con Thyssenkrupp de la mayoría de los equipos entró en vigor en 2017, por lo que puede que aún no tengan contratado todo el personal necesario y les falte coordinación. Además, algunas unidades están situadas en Abu Dhabi, muy distante geográficamente de Dubái.

Duración: la otra medida temporal del mantenimiento correctivo es la duración de la reparación, es decir, el tiempo que tarda el operario desde que llega al equipo hasta que soluciona el problema. Como se puede deducir, este parámetro va a depender en parte de la pericia del técnico para reparar el equipo, y, por tanto, dará una primera medida de la calidad del servicio que proporciona cada delegación. Se muestra el diagrama de cajas del tiempo de reparación por fallo y el de barras de tiempo mediano en figura- 6.25, la línea roja es la mediana global y en verde aparecen los 3 tiempos más bajos y en rojo los 3 más altos.

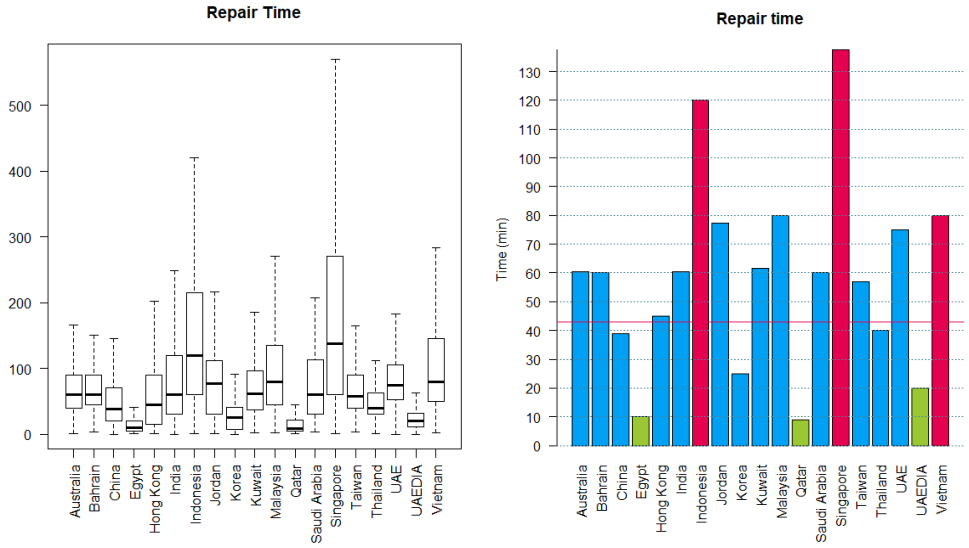


Figura- 6.25 Diagrama de cajas del tiempo de reparación por fallo para cada país

En este caso se vuelven a presentar valores excesivamente bajos del tiempo de reparación en Egipto y Qatar lo que, unido a los ya observados en el tiempo de reacción, indican que debe haber algún edificio distinto a los demás en estos países. Para confirmar si efectivamente dichos tiempos bajos se concentran en un único lugar, se analizan los registros de los principales edificios de cada país en figura- 6.26 y figura- 6.27. Se incluye encima de cada barra el número de informes de cada edificio.

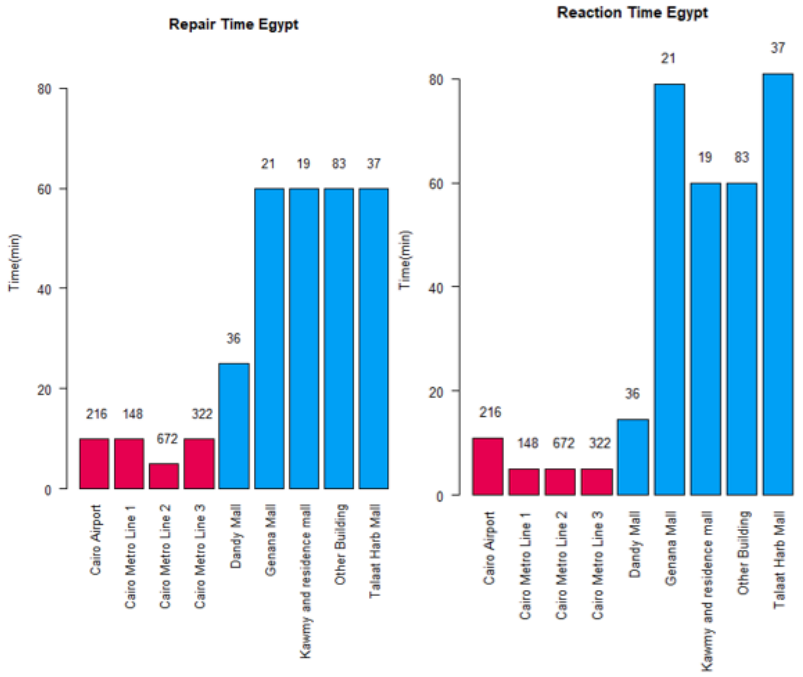


Figura- 6.26 Tiempo de reacción (izquierda) y de reparación (derecha) para cada edificio de Egipto

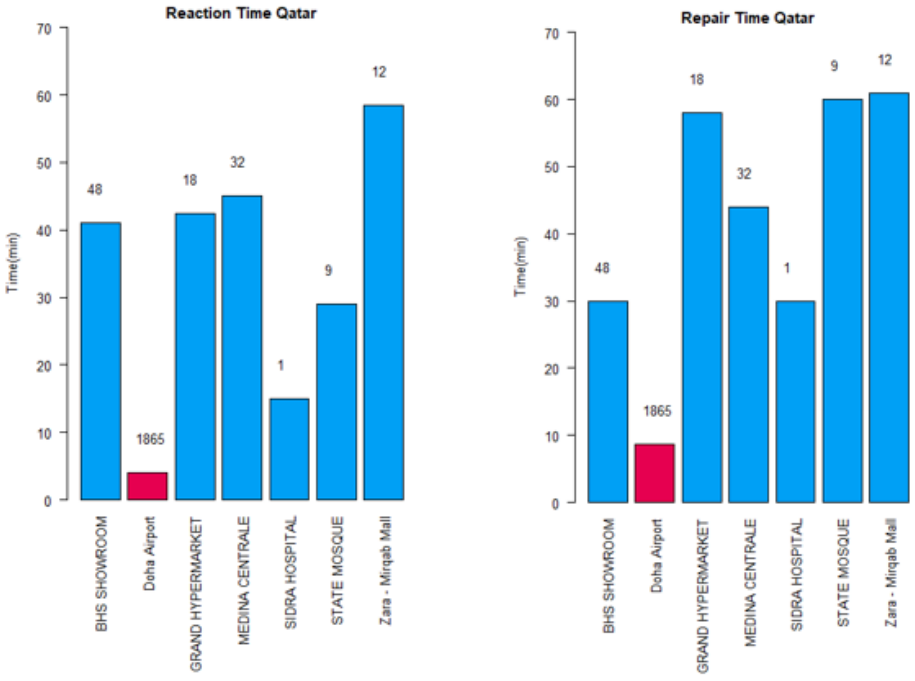


Figura- 6.27 Tiempo de reacción (izquierda) y de reparación (derecha) para cada edificio de Qatar

Se observa que en ambos países existe al menos un edificio —el metro y aeropuerto de El Cairo en Egipto y el aeropuerto de Doha en Qatar— que concentra la mayoría de los informes y tiene tiempos de actuación especialmente bajos, desvirtuando así los valores medianos del país. A falta de confirmación de Thyssenkrupp, se emplea

una última prueba para comprobar qué tipo de incidente suelen solventar en dichos edificios. Tanto en los dos aeropuertos como en el metro del Cairo el 70% de los fallos se deben a pulsaciones de interruptores de emergencia, activación de seguridad de placa de peines (relacionada con tropiezos o atascos de zapatos de pasajeros) y el *flap* del pasamanos (relacionado con golpes de los carritos de equipaje). Este es el tipo de incidente menor que en la mayoría de edificios son resueltos por el propio personal sin llamar a Thyssenkrupp, por lo que si aquí queda registro de ello es porque Thyssenkrupp trabaja como personal exclusivo en estos edificios.

Esto también permite explicar la alta tasa de averías de Qatar. Muchas de los fallos por activación de seguridades se deberán a accidentes de los pasajeros, pero en este aeropuerto no se detalla cómo sucedió el incidente, por lo que no hay elementos para valorarlo como accidente —aunque es probable que lo sea— y es asignado a avería. Por tanto, la tasa real de accidentes en Qatar se espera que sea mucho más alta, mientras la de averías debería disminuir y ser parecida a la del resto de países.

Por otro lado, al igual que en Egipto y Qatar, también es destacable el bajo tiempo de reparación en UAEDIA y en el aeropuerto de Hong Kong—10 minutos frente a los 45 del país en su conjunto—. Siguiendo el razonamiento iniciado en este apartado, esto debería indicar que los técnicos de dichas cuadrillas poseen una gran habilidad para solventar cualquier tipo de fallo en apenas minutos. Sin embargo, aunque esto pudiera ser cierto, la diferencia frente al resto de delegaciones es desmedida y no parece deberse únicamente a una mayor capacidad de los operarios. La explicación en este caso es sencilla. Tal y como se había comentado al hablar de la accidentalidad, en estos dos aeropuertos son los propios operarios de Thyssenkrupp quienes solucionan todos los incidentes menores. Dichas intervenciones suelen implicar únicamente el reinicio de la escalera y la ayuda al cliente si ha sufrido algún tipo de daño. Por tanto, su tiempo de reparación debería ser muy inferior al de una avería real y esto empuja hacia abajo a la mediana del tiempo de reparación en ambos aeropuertos. Esta hipótesis se valida al calcular la mediana del tiempo de reparación para los incidentes considerados avería y para los accidentes que se muestra en la tabla 6.3.

	Accidente	Avería
Mediana del tiempo de reparación	23 min	54 min

Tabla 6.3 Tiempo de reparación mediano para accidente y avería

Así pues, se vuelve a demostrar la necesidad de prescindir de los informes de accidentes, o, al menos, contrastar por separado el tiempo de reparación de accidente y el de avería para poder comparar las delegaciones. Si se separan ambos

tipos, tal y como se muestra en figura- 6.28, se observa que los valores de UAEDIA y Hong Kong son más cercanos a la mediana global en lo que a averías se refiere.

No obstante, el resto de las delegaciones, con un porcentaje similar entre ellas de accidentes, apenas presentan variaciones entre el tiempo de reparación de averías y el tiempo de reparación conjunto. Por ello se considera que, exceptuando en las delegaciones exclusivas, es correcto emplear el tiempo de reparación global, sin separar entre accidente y avería, para comparar la pericia y habilidad de los operarios.

Así pues, se considera que los operarios de Indonesia, Malasia, Singapur y Vietnam podrían no estar correctamente capacitados y necesitar una formación mejor.

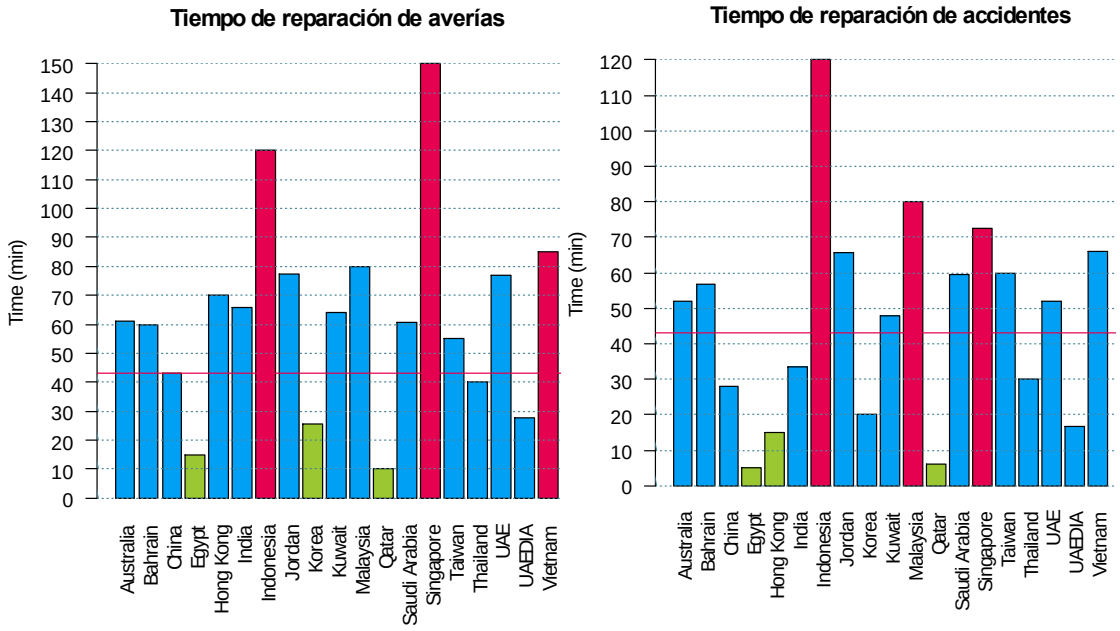


Figura- 6.28 Tiempo de reparación en avería (izquierda) y accidente (derecha)

Disponibilidad: tras analizar por separado los valores más llamativos del tiempo de reacción y reparación, conviene combinarlos para obtener el tiempo que cada fallo inutiliza la escalera. Este tiempo en el que el equipo no está disponible es el que realmente importa al cliente y definirá la calidad del servicio de cada delegación.

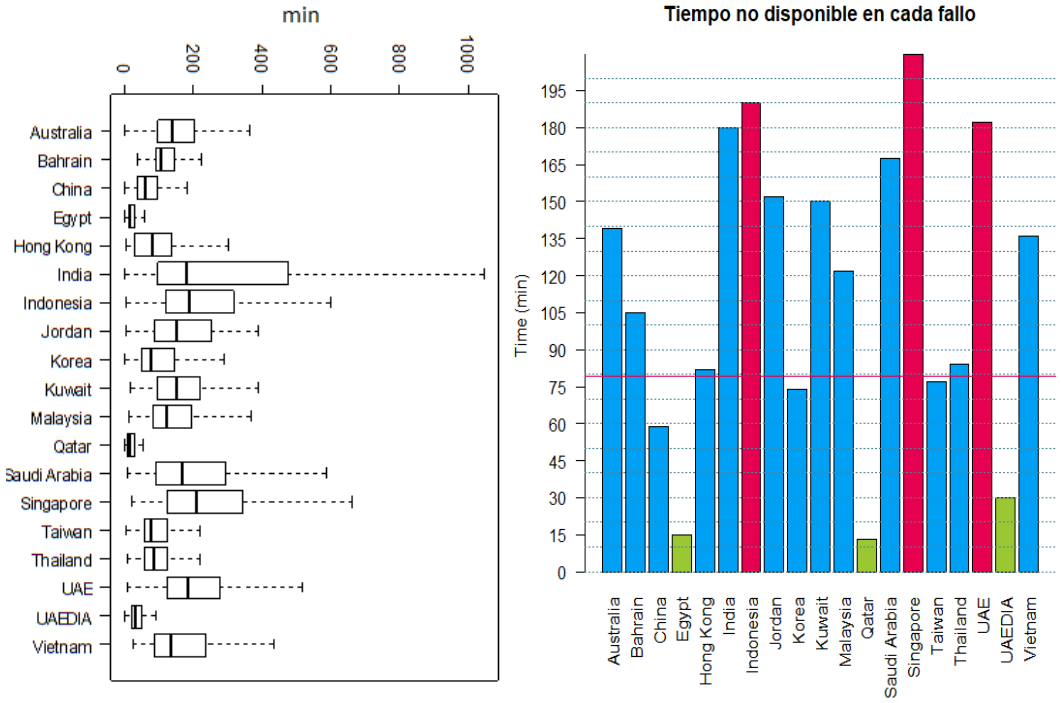


Figura- 6.29 Diagrama de cajas del tiempo no disponible por fallo para cada país

Gracias a estos gráficos, se determina que las delegaciones de Singapur, Indonesia, India y UAE son las que más tardan en solucionar los incidentes y habría que establecer algún tipo de diálogo con dichas organizaciones para encontrar las causas últimas de este peor servicio y establecer un plan de formación, aumentar la plantilla o mejorar sus medios para poder ofrecer el servicio que requiere el cliente. En contraste, China, Taiwan y Corea —Hong Kong, UAEDIA, Qatar y Egipto no se incluyen por sus particularidades— prestan un servicio de gran calidad y también sería de interés averiguar cómo se organizan para ver si se pueden trasladar sus métodos de trabajo a otras delegaciones.

6.5.- CARACTERIZACIÓN DE LOS INCIDENTES

6.5.1.- Caracterización temporal

Descubrir en qué momento del día, semana y año, se producen los incidentes es uno de los resultados más útiles de este estudio. Conocer cuándo se produce un mayor número de llamadas al servicio de mantenimiento permite organizar los turnos de los operarios y planificar correctamente la política de recursos humanos para proporcionar el nivel de servicio requerido por el cliente en cada instante. Para dar una idea inicial de esta distribución temporal se muestra el diagrama de barras por hora, día y mes, dividido entre accidente y avería, en figura- 6.30.

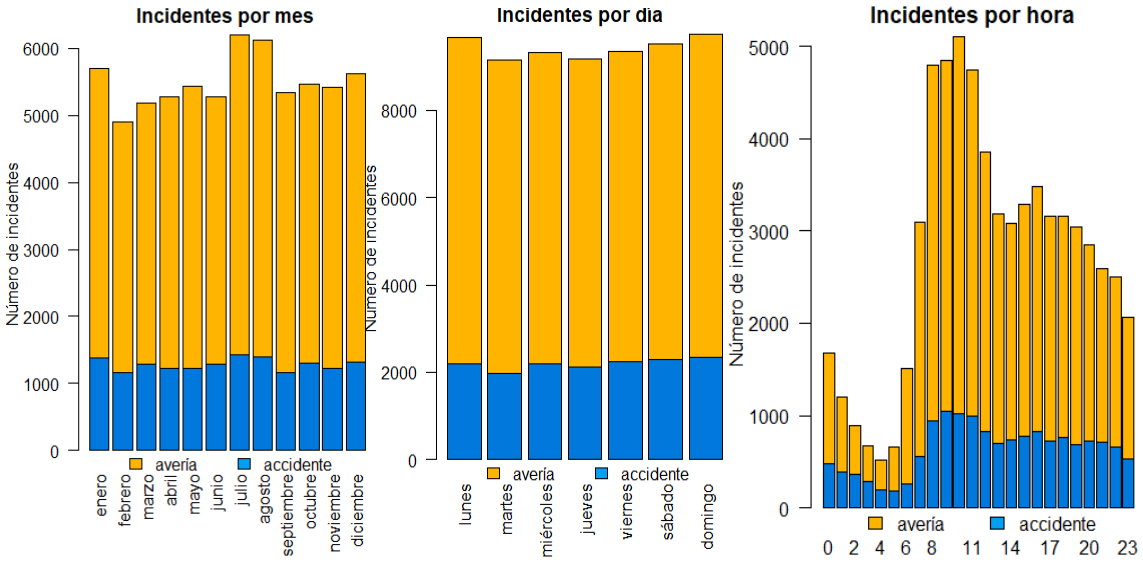


Figura- 6.30 Número de incidentes por mes (izquierda), día (centro) y hora (derecha)

En primer lugar, se hace notar que no parece haber dependencia entre el número de accidentes registrados y ninguna de las variables temporales, puesto que la proporción entre la barra azul y la amarilla permanece prácticamente constante en todos los casos. La única influencia destacable parece ser horaria, ya que se ve que durante la madrugada aumenta el porcentaje de accidentes respecto al total, si bien no parece haber una dependencia fuerte.

Así pues, se procede a analizar cada variable temporal globalmente, sin diferenciar entre averías y accidentes:

Variacion mensual: se puede apreciar una diferencia de hasta el 15% entre los meses de verano —julio y agosto— y el resto del año, con un leve repunte en diciembre y enero. Esto se puede achacar por un lado al incremento del transporte en trenes y aeropuertos durante esos meses por las vacaciones en muchos de esos países y, por otro lado, a un mayor uso de los centros comerciales a cubierto por las elevadas temperaturas exteriores en países como Kuwait, UAE o Bahrein. En cambio, el pico de diciembre y enero (figura- 6.31) solo se da en países donde es habitual celebrar la Navidad (Australia o Corea) y en aquellos cuyos equipos están situados en aeropuertos que sirven de escala a vuelos internacionales (UAEDIA, Hong Kong, Qatar); mientras en países como China, India, Vietnam o Taiwan no se da ese repunte en dichos meses. Otra tendencia que se ha observado está relacionada con los países de mayoria musulmana (UAE, Arabia Saudi, Qatar, Egipto, Bahrein, Jordania, Kuwait, Indonesia y Malasia). En ellos se celebra el Ramadán, que en 2018 empezó el 16 de mayo y acabó el 14 de junio. Si se representan los fallos por mes en estos países (figura- 6.31) se aprecia que los incidentes se reducen hasta un 20% en Ramadán frente al resto de meses. Esto se debe a que el Ramadán es un periodo de recogimiento en el que se reducen al mínimo muchas actividades, lo que explicaría ese menor número de fallos.

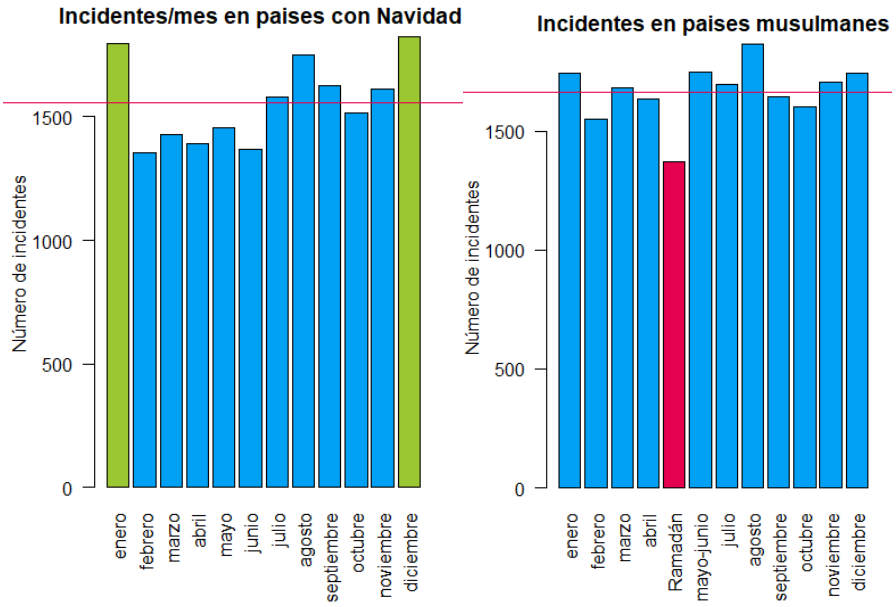


Figura- 6.31 Incidentes al mes en países con influencia de la Navidad (izquierda) y en Ramadán (derecha)

Otras fiestas, como el Año Nuevo Chino, no parecen tener influencia, probablemente por su corta longitud —de unos pocos días— que no deja huella suficiente en los incidentes registrados a lo largo de todo el mes. Se puede consultar la tabla por meses para cada país en ANEXOS III.

Variación diaria: en este caso, se observa que apenas hay diferencia entre el número de fallos registrado y el día de la semana. Esto puede llamar la atención, puesto que seria esperable que los fallos disinuyesen mucho los domingos cuando muchos establecimientos permanecen cerrados. Sin embargo, tras consultar con expertos de Thyssenkrupp, se ha visto que esta era una visión eurocentrista. En la mayoría de los países asiáticos la actividad de los centros de ocio el domingo es igual o superior a la del resto de días. Para confirmarlo se obtienen los gráficos por segmento en figura- 6.32.

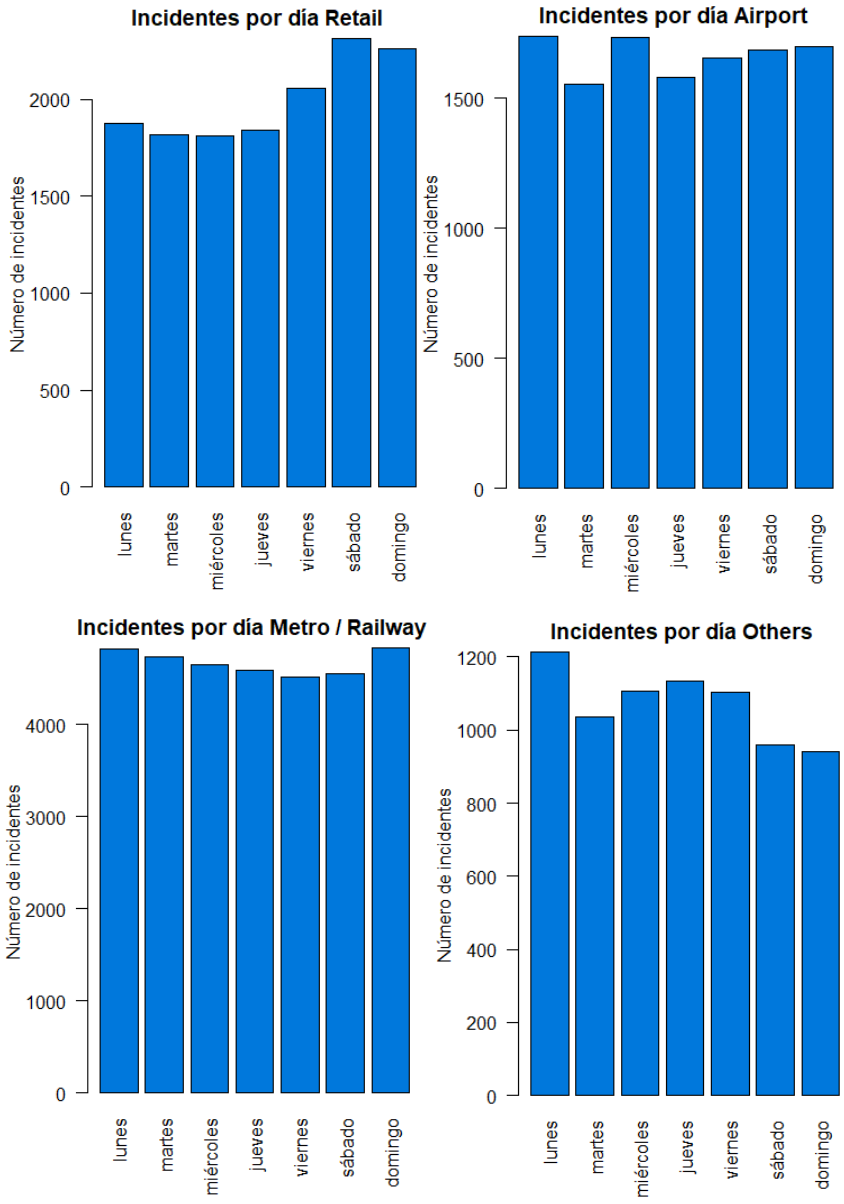


Figura- 6.32 Variación diaria en función del segmento

En los sectores de transporte (aeropuerto y metro) los incidentes tienen escasa variación con el día. En cambio, en el sector de ocio se produce un fuerte incremento de un 10-15% durante el fin de semana, mientras en el de otros (oficinas, hospitales) la actividad disminuye durante esos días, como era de esperar. Estas tendencias contrapuestas de unos y otros serían la causa de la escasa variabilidad diaria del número de incidentes globalmente.

Variación horaria: finalmente, el efecto de la hora del día es el que más importancia tiene como se mostró en la figura- 6.30. Esto es lógico, puesto que por la noche la mayoría de edificios permanecen cerrados y no registran actividad. En cambio, a primera hora de la mañana se produce la hora punta en el metro y oficinas, lo que implica una mayor carga

para los equipos y mayor posibilidad de fallos. De nuevo, parece evidente una dependencia con el segmento, por lo que se representa la curva horaria por segmento en la figura- 6.33.

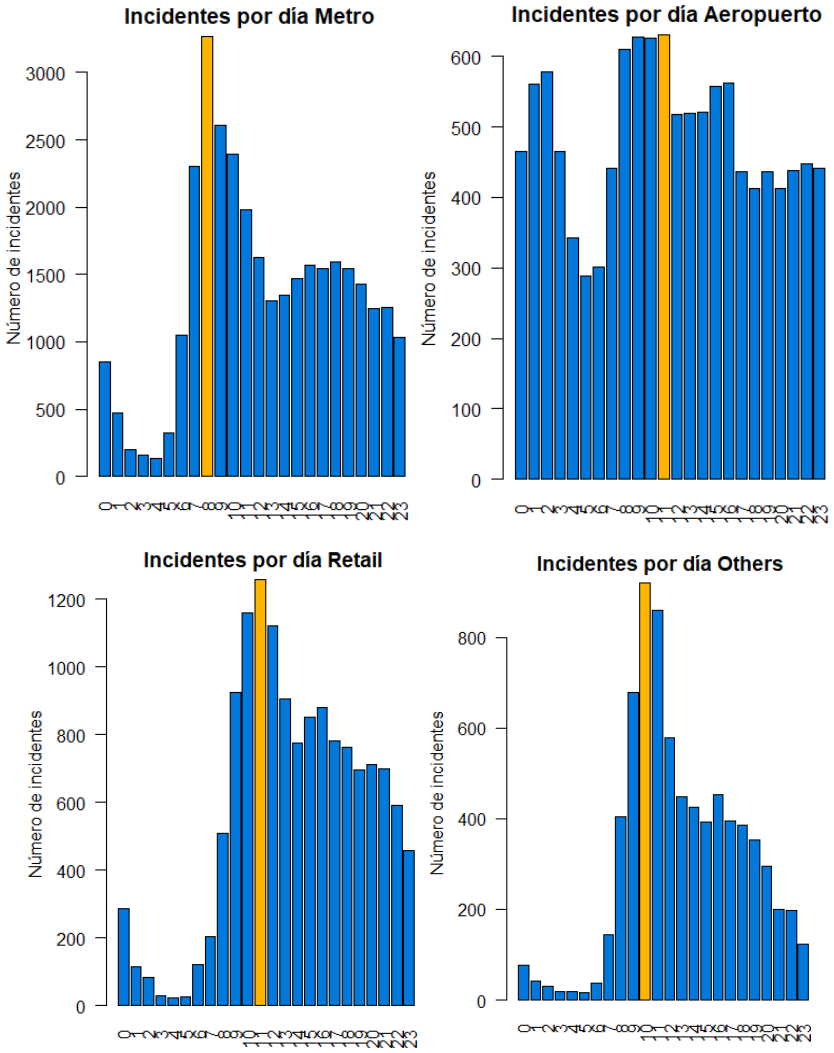


Figura- 6.33 Variación horaria en función del segmento

Como se puede ver, el metro experimenta su pico de incidentes a las 8 de la mañana, acorde con la hora de más uso. Por otro lado, en las oficinas (Others) se retrasa hasta las 10 de la mañana, dos horas más tarde que en el metro, lo que tiene sentido ya que la hora de entrada en las oficinas es posterior a la hora de mayor uso del metro. En el *retail*, este pico se retrasa aún mas y no se produce hasta las 11, lo que se explica por la menor actividad comercial a primera hora. Además, el sector del *retail* experimenta un repunte significativo a media tarde, que es cuando más ocupación suelen tener los centros comerciales. Este repunte también se da en el metro aunque es menos intenso y se reparte proporcionalmente toda la tarde.

Finalmente, sorprende la curva en el caso de los aeropuertos, con un gran número de incidentes registrados de madrugada. Esto se debe a que los aeropuertos asiáticos tienen

mucha actividad por las noches al servir de escala a vuelos intercontinentales. Por tanto, los resultados son acordes a lo esperado y demuestran la importancia de saber bien cómo se reparte por segmentos el parque de equipos de cada delegación para poder organizar mejor los turnos de trabajo.

Así pues, se observa que la variación mensual está relacionada principalmente con las costumbres de cada país. En cambio, la variación diaria y horaria se explica en mayor medida por el segmento al que pertenece cada equipo, afectando en menor medida el modo de vida de cada mercado. Para concluir, se quiere destacar que este conocimiento no es solo útil para ofrecer un mejor servicio a los clientes actuales, sino que permitirá estimar mejor cuándo se podrán producir los fallos en los equipos de nuevos clientes a partir de su segmento y localización geográfica y así saber qué contrato de mantenimiento es el más adecuado.

6.5.2.- Tipología y tiempo de reparación

La capacidad del operario, la marca, el modelo y la geometría del equipo tienen influencia en lo que a duración de la reparación se refiere. Sin embargo es evidente que este tiempo depende en mayor medida del tipo de fallo a reparar, ya que determina qué acciones hay que realizar, y, por tanto, cuánto tiempo va a llevar realizarlas. Por este motivo, sería muy útil conocer con exactitud cuánto se tarda en reparar los fallos más comunes. Entre otras ventajas, esto permitiría indicar al cliente el tiempo que no estará disponible la escalera, lo que le permitirá organizarse mejor. Además, también ayudará a estimar cuándo un operario acabará una intervención y podrá volver a contarse con él para solucionar otro incidente, optimizando así los recursos con los que cuenta la plantilla de cada delegación.

Una primera aproximación para determinar qué fallos son más probables que se produzcan y qué tiempo tienen asignados es obtener su frecuencia en los registros y su tiempo de reparación mediano. Previamente a analizar esos resultados se procede a describir someramente cada fallo que se ha podido distinguir mediante el algoritmo de clasificación de textos:

1- Balaustrada (*Balaustrade*): los principales fallos asignados tienen que ver con daños en los cristales laterales, golpes en los faldones metálicos que doblan el zócalo y pueden producir rozamiento con los escalones o tornillos sueltos que hacen que se despegue alguna de las piezas, como los cepillos de seguridad.

2-Frenos (*Brake*): los más comunes son distancia de frenado excesiva, bloqueo o desalineamiento de pastillas de freno o frenado brusco tanto del freno principal como del de emergencia o del trinquete de seguridad.

3-Cadena principal y de pasamanos (*Chains*): desgaste excesivo, movimiento a tirones o rotura de algún bulón, eslabón u otro elemento de la cadena.

4-Interruptor automático (*Circuit breaker*): salto de alguna de las seguridades eléctricas del motor trifásico o del armario eléctrico.

5-Seguridad de placa de peines (*Combplate switch*): activación del interruptor debido a una presión excesiva en los dientes del peine, habitualmente por un objeto atascado. También se incluyen fallos por tornillos sueltos en las placas de las plataformas de llegada o por rotura de algún diente del peine.

6-Interruptor de parada de emergencia (*Emergency stop*): pulsación por parte de alguna persona del botón de parada de emergencia.

7-Alarma de incendios (*Fire alarm*): activación de la alarma de incendios del edificio. No suele estar relacionada con un incendio provocado en la propia escalera.

8-Seguridad de la plataforma de llegada (*Floor plate switch*): elevación indebida de las plataformas de llegada por apertura del pozo donde va el motor y el volteo de peldaños, apareciendo un desnivel entre el suelo y la base de la escalera que puede hacer tropezar a los pasajeros o que se cuelen objetos dentro del pozo.

9-Engranajes (*Gears*): principalmente desgaste excesivo, ruido, rotura o mala transmisión tanto del reductor del motor como del piñón y rueda de las cadenas.

10-Guías (*Guides*): desalineación, vibraciones o golpes que doblan las guías tanto del pasamanos como de los peldaños.

11-Pasamanos (*Handrail*): desgaste excesivo de la cinta del pasamanos, de la polea que le transmite el movimiento, del sistema que mantiene el contacto entre polea y cinta o de los rodillos que facilitan el movimiento.

12-Flap de seguridad del pasamanos (*Handrail inlet*): activación del interruptor por impacto de algún objeto en la puerta por la que entra el pasamanos en la parte baja de la balaustrada en las plataformas de llegada.

13-Pérdida de sincronismo del pasamanos (*Handrail speed*): activación de la seguridad que indica que cada lado del pasamanos se mueve a distinta velocidad entre sí o respecto a la de referencia.

14-Interruptor de encendido (*Keyswitch*): problemas al encender o apagar el equipo, principalmente rotura del llavín o mal contacto del interruptor.

15-Luces señalizadoras (*Lighting*): golpes en los cristales que protegen las luces o en las propias lámparas y también bombillas fundidas. Principalmente se produce en las balizas de indicación de sentido en las zonas de llegada y en la balaustrada.

16-Lubricación (*Lubrication*): derrames de aceite que se filtran al exterior, fallo en el sistema de lubricación automática o fallo por depósito de aceite vacío.

17-Seguridad de desalineación de la cadena principal (*Main chain safety device*): activación de un interruptor que indica que la cadena principal está torcida lo que puede ser perjudicial para su funcionamiento.

18-Seguridad de hueco entre escalones (*Miss step sensor*): activación del interruptor que indica que hay espacio entre dos escalones consecutivos y podría introducirse un objeto extraño entre ellos.

19-Motor (*Motordrive*): vibraciones, ruidos o rotura de algún elemento del motor.

20-Pérdida de control de la velocidad del motor (*Motor over/underspeed*): embalamiento, velocidad fluctuante o insuficiente del motor eléctrico.

21-Otras seguridades (*Other safety switches*): otros interruptores de seguridad, se agrupan en una categoría general al ser un número pequeño de fallos los de cada tipo distinto de interruptor.

22-Otros fallos (*Others*): fallos no identificables por el texto o aquellos en los cuales el operario al llegar ya se encontró el problema solucionado y no necesitó hacer nada para arreglar el equipo.

23-Sistema de detección de pasajeros (*Passenger detection*): fallo en el radar o en el sensor que detecta a un pasajero y cambia la velocidad de movimiento del equipo.

24-PLC y variador de frecuencia (*PLC and inverter*): fallo electrónico del controlador o del variador de frecuencia del motor.

25-Suministro eléctrico del equipo (*Power supply*): apagón en el edificio que quita el suministro eléctrico al equipo.

26-Regriferación del variador (*Refrigeration*): colmatación del filtro, vibraciones, ruido, rotura o enfriamiento insuficiente del ventilador que refrigera el variador de frecuencia.

27-Fallo eléctrico o de relé (*Relay and wires*): rotura del aislamiento de algún cable, mala conexión entre bornes o activación errónea de algún relé.

28-Reinicio del equipo (*Reset*): el equipo vuelve a funcionar sin hacerle nada más que reiniciar la unidad. El reset no es en sí la causa del fallo, si no la acción tomada, pero no se dispone de información para asignarlo a otra causa y es mejor saber que solo fue necesario un reset que mandar estos registros directamente a otros.

29-Seguridad de hueco en el faldón de la balaustrada (*Skirt switch*): activación del interruptor que indica que hay un hueco en el lateral entre peldaños y balaustrada y podría intoducirse algún objeto extraño.

30-Peldaño (*Step*): desalineación, desgaste, adhesión de algún cuerpo u otros motivos que afectan a la apariencia o movimiento de los peldaños.

31-Conexión del peldaño a la cadena de peldaños (*Step bolt*): rotura o desalineamiento del bulón y cojinete que conectan los peldaños a la cadena de peldaños.

32-Cadena de peldaños (*Step chain*): falta de tensión, desgaste o rotura de alguno de los eslabones u otro elemento de la cadena de peldaños.

33-Inserto amarillo de límite de peldaño (*Step demarcation*): rotura o desprendimiento del inserto de plástico amarillo de algún peldaño.

34-Rodillo de peldaños (*Step roller*): rotura o desgaste del rodillo de poliuretano de los peldaños que se mueve por las guias por el arrastre de la cadena de peldaños, consiguiendo así el avance de los escalones.

35-Seguridad de hundimiento de peldaño (*Step upthrust*): activación del interruptor que indica que el peldaño se ha combado, lo que puede indicar que sufre una carga excesiva.

36-Inundación (*Water leakage*): daños por agua en algún elemento electrónico o mecánico del equipo.

Tras definir los fallos con los que se va a tratar se ofrece a continuación (figura- 6.34) la frecuencia relativa de cada uno de ellos. Los 5 más y menos comunes se colorean de rojo y verde respectivamente.

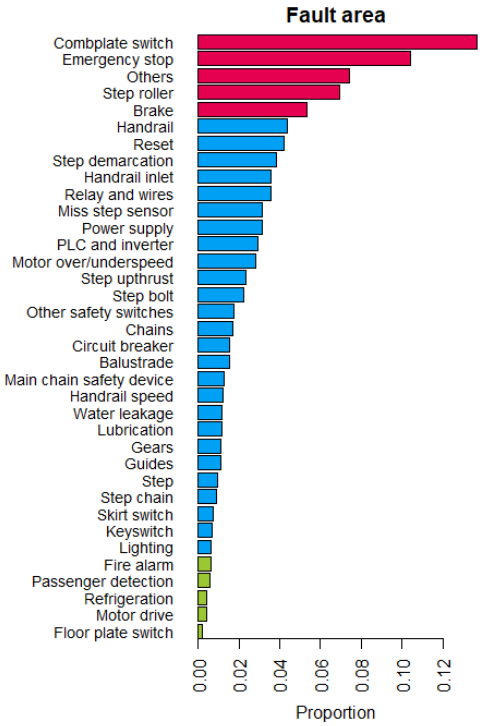


Figura- 6.34 Proporción de cada fallo

En primer lugar, se hace notar que los tres fallos más frecuentes están relacionados con factores difíciles de controlar y casi independientes del diseño y mantenimiento preventivo

que reciba el equipo. Tanto los fallos de la placa de peines, como el pulsador de emergencia como “Others” están causados principalmente por accidentes y usos inadecuados por parte de los clientes. Su prevalencia frente a los demás fallos es una buena noticia. Indica que las escaleras y pasillos rodantes son equipos muy fiables y sus incidentes están más relacionados con el uso que se hace de ellos que con una falta de calidad de sus elementos o un funcionamiento inadecuado.

De todas formas, saber que el fallo de placa de peines es tan frecuente puede llevar a Thyssenkrupp a dedicar recursos a un rediseño de este componente o a añadir alguna señalización o elemento que alerte a los pasajeros para que estén atentos. Por tanto, este dato es muy útil, ya que permitirá centrar los recursos en aquellos fallos que tienen mayor margen de mejora, y que hasta ahora se pensaba que no eran relevantes o no se les daba suficiente importancia.

Los fallos siguientes, como frenos, rodillos de peldaños y pasamanos son lógicos puesto que estos elementos son los que presentan un mayor desgaste y sufren mayor número de ciclos.

En cuanto al otro extremo del diagrama, se cumple que los fallos menos frecuentes entran mayoritariamente en una de estas dos categorías:

1. Seguridades opcionales que no vienen de serie con la escalera, por lo que gran parte del parque de equipos no contarán con ellas y no pueden reportar fallos como los del *Floor plate switch*, *Passenger detection*, *Lighting* o *Skirt switch*.
2. Elementos comerciales, fabricados externamente y que no deberían sufrir un fuerte desgaste al no estar sometidos a grandes variaciones de carga como el ventilador del variador o el mecanismo de encendido por llavín. También entraría en esta categoría el motor eléctrico que, aunque sí presentará variación de carga, será menor que en otros elementos y además los motores de corriente alterna son máquinas muy robustas y fiables que presenten pocos fallos.

El otro estadístico que define a cada fallo, además de su frecuencia, es el tiempo que conlleva su reparación. Para cuantificarlo se muestra su diagrama de cajas en la figura- 6.35. Hay que indicar que en este gráfico los extremos de la caja indican el 1º y 3º cuartil, la franja más oscura es la mediana y los bigotes (lineas de trazo discontinuo) se extienden hasta el mayor entre el máximo y 1,5 veces el rango intercuartil por la derecha y el menor entre el mínimo y 1,5 veces el rango intercuartil por la izquierda.

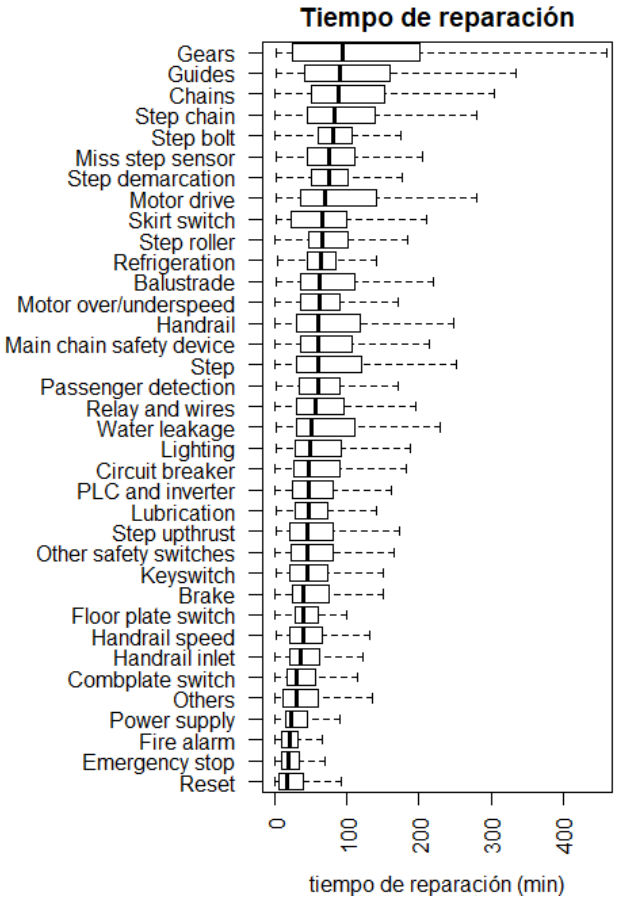


Figura- 6.35 Diagrama de cajas con el tiempo de reparación para cada fallo

Al contrario que en el gráfico de frecuencias, en este caso las primeras posiciones son copadas por fallos puramente mecánicos en los que un accidente causado por pasajeros u objetos extraños es altamente improbable. De esta manera, los fallos que más tiempo se necesita para solucionar son los engranajes, guías, cadenas y cadenas de peldaño. Esto se puede justificar por el hecho de que para acceder a esos elementos es necesario levantar peldaños o la plataforma de llegada, lo que conlleva bastante tiempo, sin contar con la propia acción correctora que implicará el cambio de piezas o corregir mecánicamente algún elemento y que consume muchos minutos.

Por otro lado, los fallos que se solucionan en menor tiempo suelen requerir como única actuación el rearme de alguna seguridad y el reinicio del equipo como *reset*, *emergency stop*, *fire alarm*, *power supply* o *Combplate switch*.

Se puede apreciar que estos fallos de menor duración coinciden en muchos casos con los más frecuentes.

Esto hace pensar que la frecuencia no es la medida adecuada para determinar en qué fallos hay que centrar los esfuerzos para reducir su incidencia. Se considera que es mucho mas

grave un fallo que, a lo largo del año, consume 100 horas de trabajo de los operarios, aunque solo se produzcan 50 incidentes de esta categoria, que un tipo de fallo que se produzca en 300 ocasiones, pero requiera únicamente 20 horas de trabajo en total. Por este motivo, se calcula el tiempo anual de reparación como:

$$tiempo\ total\ consumido\ por\ el\ fallo\ i = \sum t_i \tag{6.15}$$

Es cierto que desde un punto de vista logístico habría que incluir el tiempo de reacción en esta ponderación. No obstante, este análisis se centra en la mantenibilidad del equipo, es decir, cuánto tiempo hay que dedicar a la reparación en sí y cómo se pueden disminuir dichos tiempos mediante un rediseño, cambios en el proceso de fabricación o mejor formación. Por este motivo, no se incluye el tiempo de reacción en la ponderación empleada.

Se muestra el diagrama de frecuencia relativa ponderada en figura- 6.36. Los 5 fallos más y menos comunes se colorean de rojo y verde respectivamente.

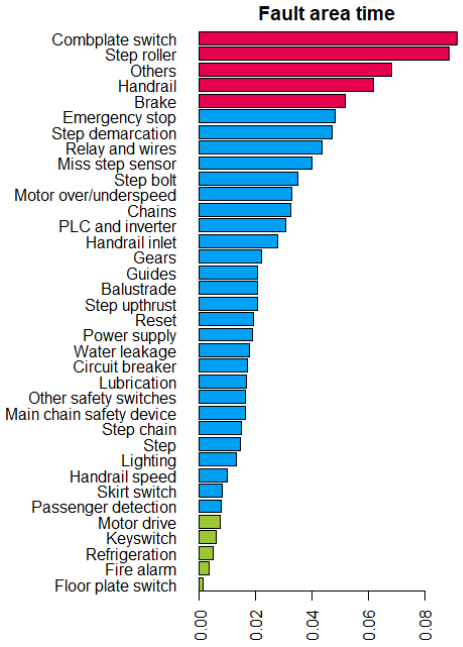


Figura- 6.36 Proporción de tiempo de reparación de cada fallo

Los cinco fallos a los que se dedica más tiempo difieren ligeramente de los más frecuentes. Pese a que la seguridad de placa de peines y la categoría “Others” siguen en la misma posición, los otros tres han variado. Suben posiciones aquellos relacionados con fallos mecánicos que exigen repuestos y retirar peldaños u otras partes de la escalera para poder solucionarlo, como el rodillo de peldaños, los frenos o el pasamanos, mientras que el botón de emergencia —al ser un fallo que implica únicamente el reinicio del equipo— baja

posiciones. En ese mismo sentido, el fallo de *Reset* baja de la posición 7 a la 18 al ser un fallo frecuente pero fácil de solucionar.

En el otro extremo las variaciones son casi nulas ya que la mayoría de los fallos eran seguridades que solo necesitan rearmarse en muchos casos, lo que consume poco tiempo.

Por tanto, como resumen de estos gráficos, habría que centrar los esfuerzos de mejora, rediseño o colocación de sensores para evitar la magnitud y frecuencia de los fallos en la placa de peines, los frenos, el pasamanos y los rodillos de los peldaños.

Fallos agrupados por subsistema: se puede apreciar que el número de fallos distintos es bastante elevado y muchos de ellos están relacionados, ya que los fallos no suelen ser totalmente independientes. Puede ser un componente concreto el elemento dañado (el rodillo de peldaños en este ejemplo) pero la causa del fallo puede deberse a rozamientos con otras partes cercanas que se hayan desviado o doblado (rozamiento con las guías) o al desgaste de otro componente con el que hace contacto (desgaste de la cadena que provoca un movimiento a tirones). Por este motivo, es recomendable hablar de subsistemas dentro de la escalera más que de fallos en elementos concretos.

Este enfoque tiene multiples ventajas: por un lado, facilita la representación e interpretación gráfica al disminuir el número de niveles a dibujar. Por otro lado, disminuye el posible error causado durante la clasificación de fallos. En la mayoría de las ocasiones en las que la asignación no haya acertado, el informe habrá sido asignado a un fallo similar, es decir, de otro elemento del mismo subsistema de la escalera. Por tanto, al agrupar dichos fallos desaparecía el error, al estar ambos en la misma categoría general. Finalmente, permitirá tener una visión de conjunto frente a la atomizada de los fallos y así se podrá ver qué parte de la escalera es la más comprometida y requiere más mantenimiento.

Por este motivo se plantea crear una variable para agrupar los fallos en función de la zona constructiva a la que pertenecen. Se detalla a continuación la agrupación realizada. Se añaden entre paréntesis el identificador de fallo de acuerdo con la lista anterior:

1. **Cadenas (Chains):** compendio de los fallos de cadena de peldaños (32), cadenas (3), desalineación de la cadena principal (17) y lubricación (16).
2. **Controlador electrónico (Controller):** incluye los fallos de PLC y variador de frecuencia (24), refrigeración del variador (26), cables y relés (27), sistema de detección de pasajeros (23), dispositivo de arranque (14) y otras seguridades (21).
3. **Sistema motriz (Driving system):** Agrupa los fallos de motor (19), velocidad del motor (20), engranajes (9), guías (10) y frenos (2).

4. **Pasamanos y balaustrada (Handrail system & Balustrade):** incluye fallos de pasamanos (11), sincronismo de pasamanos (13), flap de seguridad del pasamanos (12), balaustrada (1) y detector de hueco en los faldones (29).
5. **Plataformas de llegada (Landings):** agrupa, a petición de Thyssenkrupp, los fallos de seguridad del peine (5), de la plataforma de llegada (8) y del interruptor de emergencia (6). Se incluyen los fallos del stop de emergencia al considerar que en la mayor parte de ocasiones en que un pasajero lo activa se debe a tropiezos y caídas producidas en las plataformas de llegada por un traspiés en la placa de peines o por el cambio de velocidad entre la plataforma y los escalones.
6. **Otros (Others):** agrupa los fallos no asignados a áreas anteriores, es decir, el reset (28), suministro eléctrico (25) y diferencial (4), alarma de incendios (7), luces (15), inundaciones (36) y otros (22). Estos se suelen corresponder con fallos ajenos al propio funcionamiento del equipo.
7. **Banda de peldaños (Step band):** conjunto formado por el fallo de peldaño (30), conexión de peldaño y cadena (31), inserto del peldaño (33), rodillo de peldaño (34), seguridad de hundimiento de peldaño (35) y sensor de presencia de peldaños (18).

Una vez que se conocen los subsistemas se muestra la frecuencia de fallo, la duración mediana de reparación y la frecuencia ponderada en cada uno de ellos en figura- 6.37. Se colorean de azul los 3 subsistemas con más fallos y los 3 cuyo tiempo mediano de reparación es mayor.

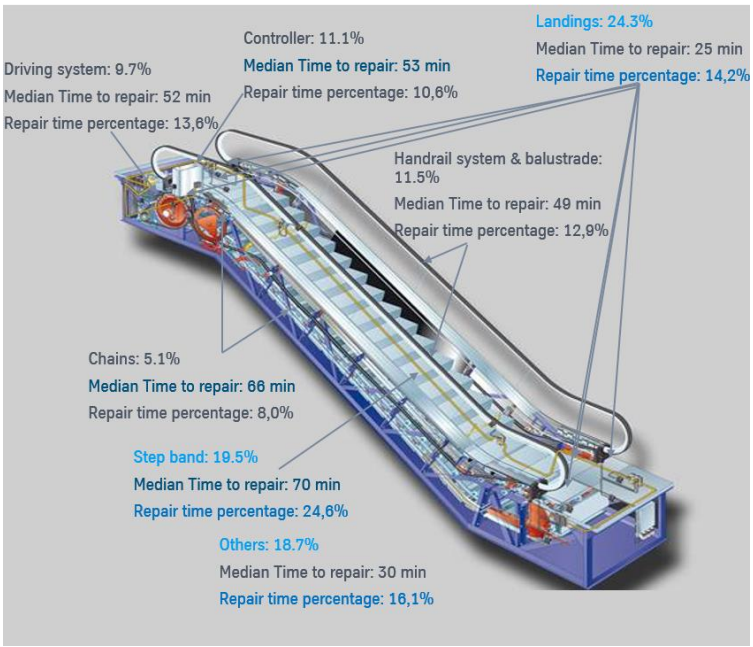


Figura- 6.37 Proporción y tiempo de cada subsistema del equipo

En esta imagen se observa que casi la mitad de los fallos se deben a la plataforma de llegada o a causas externas a la escalera. Dichos fallos suelen solucionarse en menos de media hora de mediana lo que concuerda con el análisis que se había hecho al hablar de fallos individuales. La única actuación que requieren suele ser rearmar la seguridad y reiniciar la escalera lo que consume poco tiempo al operario.

En el otro extremo, los fallos que implican mayor tiempo de reparación suelen ser poco frecuentes. Hay una excepción, los fallos de la banda de peldaños representan la segunda causa más frecuente de fallo y además son los que más tiempo necesitan para repararse.

Por tanto, después de analizar tanto los fallos por separado como por subsistemas se puede hacer una primera recomendación de en qué centrar los esfuerzos de mejora. Tanto los fallos de la placa de peines como los de la banda de peldaños copan los primeros puestos en cuanto a fallos más frecuentes y también los que más tiempo consumen al año para repararlos. Por este motivo, y dado que este trabajo se enmarca en un proyecto más general de mantenimiento predictivo, parece necesario centrar los esfuerzos de investigación en estos elementos para intentar reducir su incidencia y gravedad. Sin embargo, este es un comentario inicial en vista de los datos y corresponde analizar a Thyssenkrupp si realmente merece la pena dedicar recursos a estos dos fallos, puesto que pueden ser los más difíciles de detectar por mantenimiento predictivo —especialmente los de placa de peines al ser accidentes y tener un fuerte componente aleatorio— y puede haber otros elementos cuya ratio “recursos invertidos/beneficio obtenido” sea mayor.

Fallos agrupados por reparación: en la misma línea que la agrupación por subsistema, se ha demostrado que puede ser conveniente agrupar aquellos fallos que tengan la misma causa primaria y cuyo procedimiento de reparación y puesta de nuevo en marcha sea similar. Esta reclasificación facilitará la labor de encontrar tendencias relacionadas con características concretas de la máquina que solo deberían afectar a un tipo concreto de los fallos aquí codificados. Se detalla a continuación la agrupación realizada. Se añaden entre paréntesis el número de fallo de acuerdo con la lista inicial de fallos:

1. **Eléctrico (Electric-controller):** aquellos fallos relacionados con la parte eléctrica y electrónica del equipo. Incluye cables y relés (27), PLC y variador (24), diferencial (4), interruptor de encendido (14) y luces (15). En general, estos fallos se solucionan en el armario eléctrico de la unidad.
2. **Mecánico (Mechanical):** aquellos fallos relacionados con la parte móvil o estructural del equipo. Son la balaustrada (1), frenos (2), cadenas (3), cadena de peldaños (32), engranajes (9), guías (10), motor (19), lubricación (16), pasamanos (11), refrigeración (26), peldaño (30), conexión de peldaño y cadena (31), inserto del peldaño (33) y rodillo de peldaño (34). En la mayoría

de las ocasiones estos fallos requieren abrir el pozo de la plataforma de llegada o levantar peldaños por lo que su proceso de reparación es similar.

3. **Otros (Others):** los fallos cuya causa primaria está relacionada con un elemento externo al equipo, pero sin que se produzca la activación de ninguna seguridad, se asignan a este grupo. Son los incidentes de inundaciones (36), reset (28), suministro eléctrico (25), alarma de incendios (7), pulsador de emergencia (6) y otros (22).
4. **Seguridades (Safety):** aquellos fallos que se producen por la activación de un interruptor unido a la línea de seguridades de la máquina. Son las seguridades de placa de peines (5), plataforma de llegada (8), el flap del pasamanos (12), sincronismo del pasamanos (13), velocidad del motor (20), alineamiento de cadena (17), detección de pasajero (23), hundimiento de peldaño (35), presencia de peldaños (18), hueco en faldones (29) y otras seguridades (21). Estos fallos suelen indicar la presencia de un cuerpo extraño o el impacto de algún objeto por lo que todos suelen tener causas similares y su reparación consiste en rearmar la seguridad y comprobar que no se han producido daños en la zona afectada.

Se muestra en la figura- 6.38 el diagrama de tarta con la frecuencia de los fallos y la proporción del tiempo de reparación. El diagrama de cajas con la duración de reparación se puede ver en figura- 6.39.

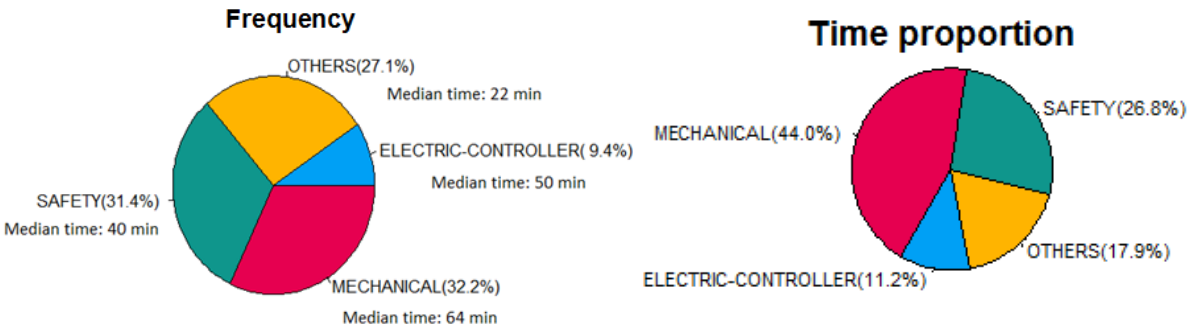


Figura- 6.38 Frecuencia y proporción de tiempo de cada tipo de fallo

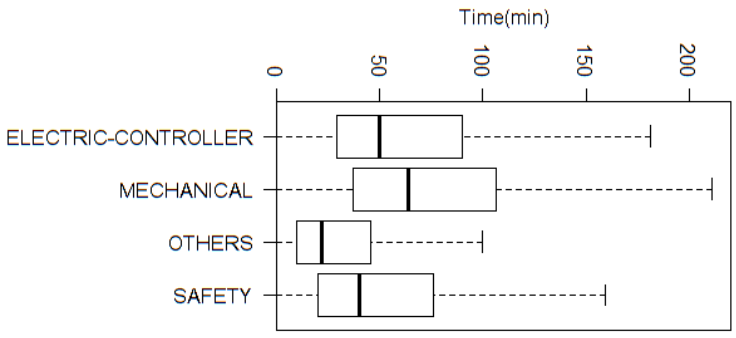


Figura- 6.39 Tiempo mediano de reparación para cada tipo de fallo

Los resultados de esta agrupación van en sintonía con todo los que se habia comentado anteriormente. Los fallos mecánicos, al incluir la mayor parte de los de peldaños, que son los más numerosos, son los más comunes y además, al requerir reemplazo de piezas y acceder a partes interiores de la escalera, consumen más tiempo. En cambio, la parte de control eléctrico del equipo no provoca un número elevado de fallos. Además, se vuelve a ver que los fallos por causas externas o para intentar proteger al usuario representan casi la mitad de los informes. Como ya se ha indicado, en algunos casos, como la placa de peines, puede intentarse un rediseño para, manteniendo el nivel de seguridad, disminuir estos fallos. Sin embargo, en la mayor parte de los casos el diseño de la seguridad viene regido por la norma y no hay mucho márgen para reducir los incidentes en este aspecto. Finalmente, se muestra el código de agrupación para ambas variables en http://bellman.ciencias.uniovi.es/~raul/Acondicionamiento.html.

6.5.3.- Caracterización del proceso de reparación

El siguiente paso, ahora que se conoce qué fallos requieren mayor atención, es determinar cómo se puede mejorar el mantenimiento que se realiza en dichos elementos. Esto es, intentar averiguar cuál es la mejor forma para disminuir tanto su frecuencia cómo el nivel de afectación en el equipo, es decir, el tiempo que el equipo no está disponible por dicho fallo. En este frente se abren tres grandes posibilidades.

1. Mejorar el componente en sí, ya sea rediseñándolo o cambiando los requerimientos de calidad en caso de que sea un elemento comprado a terceras empresas. Esto también se puede aplicar a subsistemas si se ve que hay interferencia o algún tipo de conflicto durante el funcionamiento de varios componentes contiguos.
2. Anticiparse al fallo, observando mediante sensores los primeros indicios que delatan que un elemento empieza a fallar y reparándolo o reponiéndolo antes de que la avería tenga lugar. Esto debería permitir disminuir la gravedad del fallo al repararlo en sus primeras etapas. En esos primerios estadíos el fallo no ha afectado a otras zonas y es más sencillo de reparar. Además, permitirá saber con precisión donde está el fallo, disminuyendo el tiempo que pasa el operario retirando escalones y otras partes para localizar la avería.
   También es relevante la disminución del número de fallos que se registrarían, puesto que al conocer cuándo se va a producir un fallo se puede reparar durante una visita rutinaria de mantenimiento preventivo. Por tanto, la avería no se produciría durante el funcionamiento normal de la máquina y el cliente no tendría que llamar al servicio de reparación de Thyssenkrupp.
3. Modificar la política de stocks para evitar que la reparación se posponga por no contar con la pieza adecuada para solucionar el fallo. Este punto tiene gran

relación con el anterior, ya que el principal mecanismo para garantizar que se contará con el repuesto adecuado es predecir el fallo antes de que se produzca.

Para saber cuál de estas opciones es la más adecuada en cada caso se procede a analizar los resultados de las 3 variables léxicas binarias que se codificaron mediante el procesamiento del lenguaje natural: *replaced*, *shutdown* y *max*.

Análisis de repuestos: En primer lugar, esta variable permite conocer cómo de frecuente es tener que reemplazar alguna pieza. Además, gracias a ella se podrá estimar qué penalización implica realizar un reemplazo en términos de aumento del tiempo de reparación. Para visualizar ambos parámetros se representa en un diagrama de tarta en la figura- 6.40 el porcentaje de registros en los que se indica que se sustityó una pieza ("Cambio") y en los que no fue necesario ("Ajuste"). También se indica el tiempo mediano de reparación en cada caso.

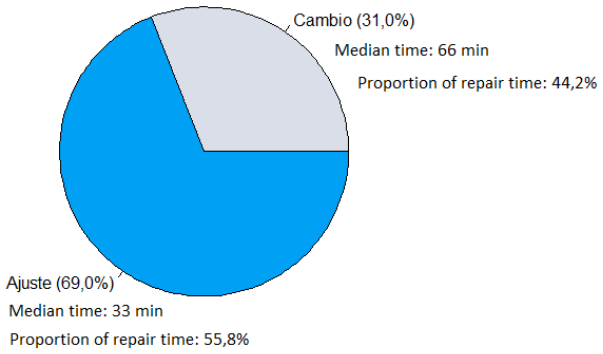


Figura- 6.40 Diagrama de tarta de la variable replaced

Se aprecia que el uso de repuestos es minoritario, debido en gran parte al elevado número de incidentes relacionados con causas externas o accidentes que no dañan ningún elemento de la máquina. Sin embargo, suelen implicar un mayor tiempo de reparación, por lo que en tiempo total su importancia es casi igual que la de los fallos que solo necesitan ajuste, rearme o reinicio del equipo.

Así pues, cualquier disminución de los fallos que implican repuestos repercutirá muy positivamente en la optimización de los recursos de las delegaciones.

En vista de la envergadura de los eventos relacionados con reemplazo de piezas, es recomendable comprobar qué fallos concretos son los que más uso hacen de ellos, puesto que permitirán saber qué elementos o subsistemas tienen una calidad o diseño deficiente que perjudica el funcionamiento habitual de la unidad. Para ello, se muestra en figura- 6.41 el porcentaje de fallos con uso de repuesto de cada subsistema y fallo individual.

	Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
Replaced	21,4%	29,6%	26,7%	29,3%	16,5%	12,3%	72,4%

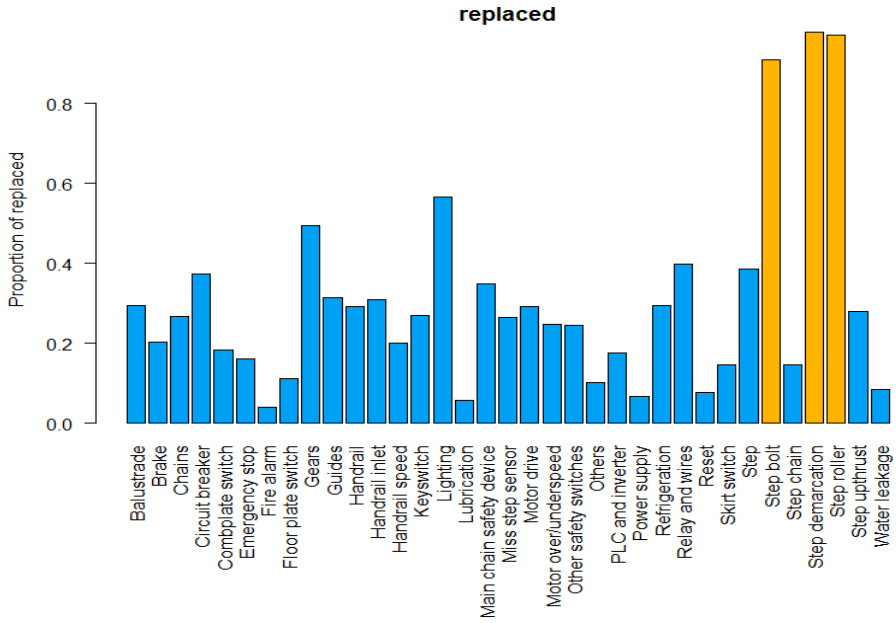


Figura- 6.41 Proporción de incidentes con uso de repuesto en cada fallo

Los fallos que, de producirse, suelen conllevar casi siempre la sustitución de la pieza afectada son los de *step bolt, step demarcation* y *step roller*. Todos pertenecen al subsistema de la *stepband,* que es el que más sufre los cambios de carga, ya que sobre él descansan los pasajeros durante su viaje, lo que puede explicar esa mayor necesidad de repuestos.

Por tanto, dado que estos tres fallos son difíciles de reparar—implican más de una hora como mínimo—, son muy frecuentes y su reparación tiene asociado un mayor coste al necesitar repuestos, sería aconsejable intentar mejorar la resistencia de dichos componentes aumentando los coeficientes de seguridad en el diseño o llegando a acuerdos con los proveedores para mejorar la calidad de esos elementos.

Análisis de fallos de larga duración: de igual manera que en el caso de los repuestos, lo primero que interesa conocer es cuántas de las incidencias no se pueden resolver con la primera visita del personal de Thyssenkrupp y es necesario dejar la escalera parada y volver más tarde para acabar la reparación. También convendría saber si esa primera visita tiene mayor duración en esos casos. Con el fin de analizarlo, se representa el diagrama de tarta con los porcentajes, la duracion mediana y el porcentaje de la duración total en figura- 6.42.

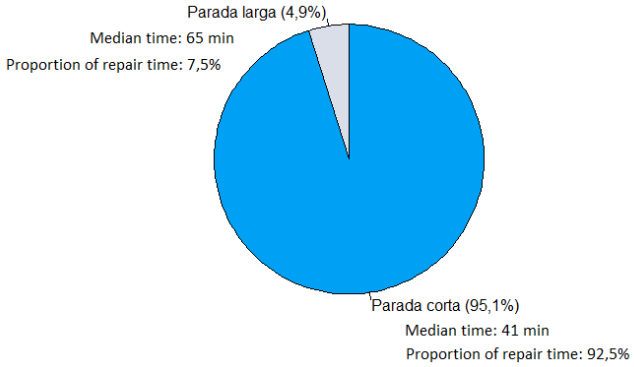


Figura- 6.42 Diagrama de tarta de la variable shutdown

El resultado de este gráfico indica que el servicio de mantenimiento parece ser de buena calidad. En el 95,1% de los incidentes el personal es capaz de solucionar el fallo y volver a poner en marcha el equipo. De todas formas, no hay que despreciar ese 4,9% de fallos que requieren una visita posterior, y más teniendo en cuenta que en ellos la primera inspección para analizar el fallo es también más larga, lo que se puede deber a que el operario intenta reparar la unidad de todas las maneras posibles antes de programar otra visita ulterior.

Pese a que el porcentaje de paradas de larga duración es bajo, está por encima de lo que los expertos de Thyssenkrupp esperaban, por lo que se analiza más en detalle este dato. Se representa la variación horaria del porcentaje de fallos que no pueden solucionarse en el momento en la figura- 6.43.

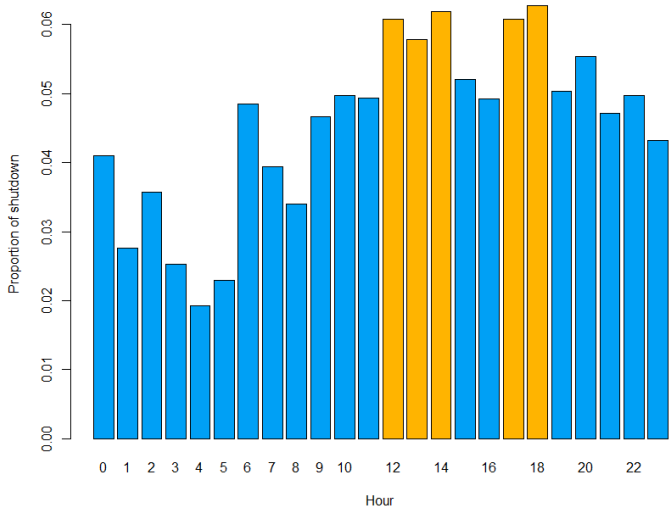


Figura- 6.43 Variación horaria del porcentaje de paradas largas

Sería de esperar que dichos fallos fueran constantes a lo largo del día, puesto que no deberían depender de una mayor actividad en el equipo, sino de contar con los medios adecuados para solucionar el fallo. Sin embargo, se observa que el porcentaje es más o menos constante de 10 de la mañana a 10 de la noche excepto por dos picos —de 12 a 14 y de 17 a 18—y experimenta una fuerte disminución por la noche. Indagando en el texto de

los registros se ha comprobado que en muchos casos la reparación se pospone no por una incapacidad técnica del operario, sino porque el cliente no quiere que se vea interrumpida la actividad diaría de su establecimiento.

El cliente llama al servicio técnico confiando en que la reparación sea sencilla y rápida. Sin embargo, cuando el operario le indica que no es un fallo de seguridad o externo si no una avería, el cliente prefiere que la reparación se realice más tarde, cuando el establecimiento esté cerrado y el ruido y otras molestias que ocasiona la reparación no afecten a los usuarios.

Además, hay que tener en cuenta el componente de seguridad. En muchos casos no se quiere cerrar con barreras el equipo —medida imprescindible en las reparaciones— para que pueda seguir sirviendo como salida de emergencia, aunque no funcione como escalera mecánica si no como escalera normal, retrasando así la reparación a una hora en la que la ocupación sea menor.

Por este motivo, el campo de parada larga no recoge adecuadamente el fenómeno de paradas que no se pueden realizar por no contar con los medios técnicos necesarios. Para adecuarla a lo que se pide, se analizarán a partir de ahora solo aquellas paradas largas que implican uso de repuestos, ya que se da por hecho que en esos casos la reparación se pospone por un criterio técnico y no por petición del cliente. Con esta nueva definición se reduce a un 1,9% sobre el total —un 6,2% de aquellos fallos que requieren repuestos— los fallos de larga duración por motivos técnicos, lo que se acerca mucho más a lo esperado por Thyssenkrupp.

Ahora que ya se conoce con precisión qué registros entran dentro de la categoría de fallo que no se ha podido reparar en una sola visita por falta de medios técnicos, se pasa a analizar qué fallos son los más comunes dentro de este grupo. Para ello se muestra el porcentaje de fallos *shutdown* en aquellos que han necesitado repuesto en cada subsistema y fallo individual en la figura- 6.44.

	Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
Shutdown	17,5%	5%	12%	10%	3%	8%	4%

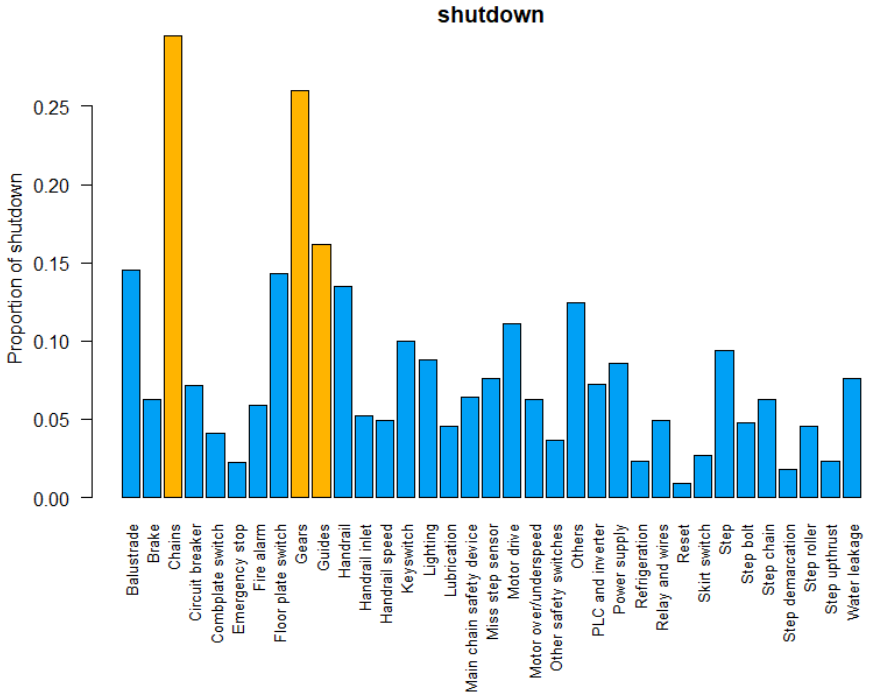


Figura- 6.44 Proporción de incidentes en los que se usó un repuesto cuya parada fue de larga duración

Se puede ver que en alrededor de un cuarto de los fallos de la cadena, engranajes y guías en los que se necesita cambiar alguna pieza, ese repuesto no está disponible en el momento. Esto es lógico, puesto que son elementos grandes y que suelen fallar poco y no merece la pena tenerlos en stock y llevarlos al lugar del incidente si no se sabe expresamente que serán necesarios.

Sería de gran interés saber cuánto se tarda en contar con dicho repuesto —equivalente a cuando se realiza la segunda visita que soluciona la avería— ya que permitiría saber el tiempo que se pasa parado el equipo esperando su reparación. Sin embargo, el operario no rellena un nuevo registro en la segunda visita ni completa el que hizo en la primera, por lo que no se tienen datos que permitan conocer esa información.

Estos fallos serían los que más se beneficiarían del mantenimiento predictivo al poder pedir el repuesto con antelación para llevarlo justo cuando se necesite. Esto evitaría tener que realizar una segunda visita, disminuyendo así en gran medida el tiempo que se pasará parado el equipo. Por tanto, es necesario analizar cómo de detectables son estos fallos.

Análisis de fallos detectables: tal y como se había hecho en los dos casos anteriores, el primer paso para saber qué fallos son los más detectables es conocer en general cómo de predecible son los fallos que se producen, es decir, en cuántos casos se notan síntomas previos —ruido, calor, vibraciones— que puedan indicar que pronto se va a producir una

avería. Se representa un diagrama de pástel con la frecuencia, el tiempo mediano y el porcentaje de tiempo total para los fallos predecibles ("Detectable") y los que no ("Suddenly") en la figura- 6.45.

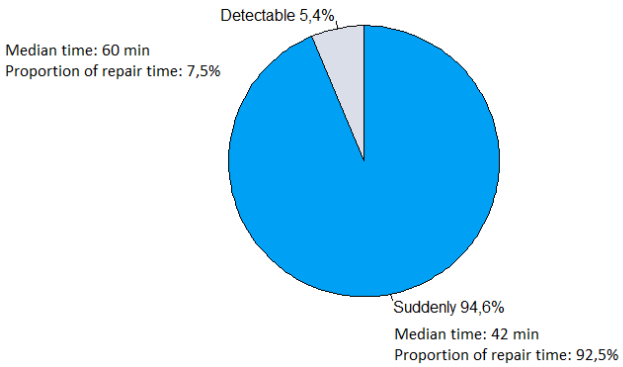


Figura- 6.45 Diagrama circular de la variable max

El porcentaje puede parecer pequeño —apenas el 5,4%— pero esto se debe principalmente a dos motivos. Hay un alto porcentaje de fallos cuya causa es externa al equipo y que son difícilmente anticipables y, por otro lado, los síntomas que se podría haber detectado han sido indicados por un ser humano. Esto implica que el ruido, nivel de vibraciones o aumento de temperatura eran muy considerables para que haya podido ser detectado por los sentidos del operario o de algún usuario. Los sensores presentan umbrales de detección mucho menores y más fiables, en cuanto a diferencias de temperatura o medida de vibraciones, por lo que sería esperable que el número de fallos realmente anticipables fuera mucho mayor al aquí indicado.

De todas formas, no hay que despreciar la capacidad informativa de esta variable y puede servir como primera aproximación para valorar un fenómeno de esta complejidad.

En esto caso la duración mediana mayor va en la misma línea de lo comentado anteriormente. Los fallos por causa externa no son detectables y estos son los que tienen menor duración, por lo que es natural que la duración de reparación de los fallos predecibles sea mucho mayor que los que suceden sin avisar.

Como se había comentado previamente, los fallos que más interesa saber si se pueden predecir son aquellos para los que se van a necesitar repuestos, ya que permitirá pedirlos solo cuando se necesiten, evitando tanto la rotura de stock que se vio que existía en las paradas largas como un excesivo almacenamiento que puede ser igual de perjudicial por su alto coste asociado. Por esta razón, se muestra a continuación en la figura- 6.46 el porcentaje de fallos predecibles de cada subsistema y fallo individual, limitado a aquellos que implican uso de repuestos.

	Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
Detectable	13,5%	3,9%	7,9%	4,1%	9,1%	4,1%	5,3%

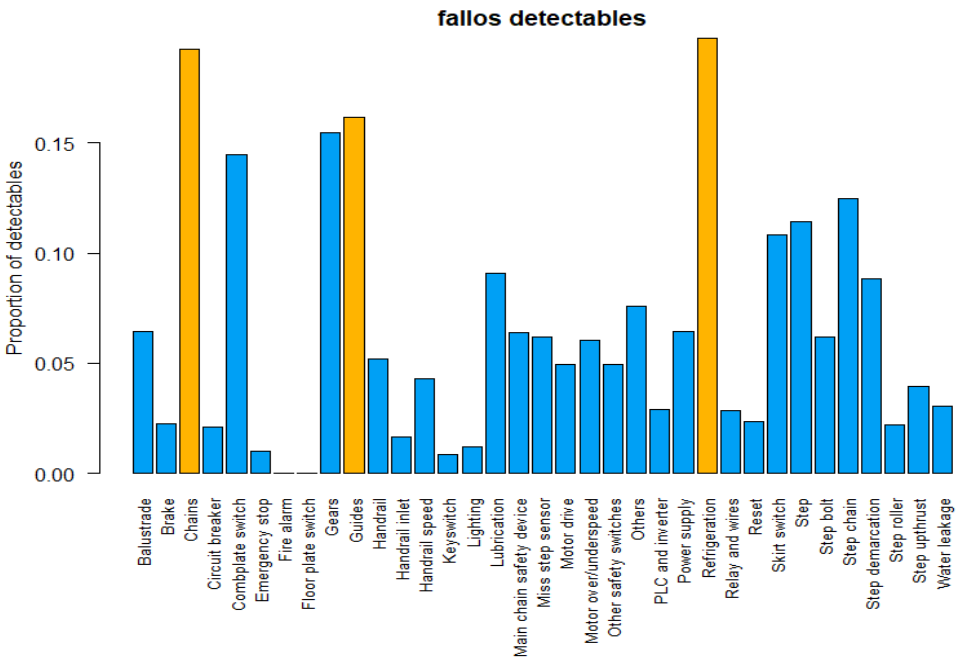


Figura- 6.46 Proporción de incidentes en los que se usó un repuesto cuya parada era detectable

En esta representación destaca que el fallo que podría predecirse en mayor número de ocasiones es el de cadenas. Esto coincide con aquel para el que no se tiene un repuesto en el momento adecuado, lo que es una muy buena noticia.

Además, el fallo de guías y el de engranajes también se puede detectar en muchas ocasiones por lo que habría que profundizar en qué sensores utilizar para intentar llevar a cabo dicha predicción, ya que se necesita mucho tiempo solucionar estos fallos y, además, su reparación implica desmontar gran parte de la escalera produciendo ruido y otras molestias. Por tanto, anticipar cuándo se van a producir estos fallos permitirá repararlos de noche o en momentos en los que se genere el menor trastorno posible al cliente, lo que redundará en una buena impresión del servicio de mantenimiento.

Como conclusión general a este apartado, el fallo de banda de peldaños es el que más margen de mejora parece tener mediante un rediseño o cambio en la calidad de los componentes. Por otra parte, los fallos de cadenas son los más proclives a no poder ser reparados en el momento, lo que debería llevar a pensar en nuevas estrategias en las delegaciones para poder detectarlo durante las visitas de mantenimiento preventivo. Finalmente, son también los fallos de cadena los que serían más fáciles de anticipar empleando sensores, por lo que la estrategia para reducir los casos de parada larga puede ir por el camino de la sensorización.

6.6.- TENDENCIAS DETECTADAS

6.6.1.- Tendencias constructivas

Se ha podido comprobar que la delegación y el segmento son las variables que más peso tienen en el número de accidentes registrados y, por tanto, también en el número de fallos totales. Además, también es destacable su influencia sobre el momento en que se producen los fallos, ya que la cantidad de incidentes varía en gran medida con la actividad a la que se vean sometidos los equipos y esta depende de los usos y horarios de cada país.

Por otro lado, también se ha constatado que el tipo de fallo y las características del proceso de reparación son lo que más condiciona el tiempo necesario para solucionarlo, sin menospreciar el efecto del operario y la organización de la delegación, especialmente en el tiempo de reacción al fallo, que depende casi exclusivamente de los medios disponibles en cada país.

Ahora se intentarán hallar tendencias más sutiles, relaciones entre los fallos y los distintos parámetros que definen el funcionamiento de cada equipo y que puedan hacerlo más o menos propenso a sufrir determinados incidentes.

Influencia del ambiente exterior: una de las primeras dudas que surgieron fue conocer cómo afectaba el ambiente exterior a las escaleras mecánicas. Cuando un fallo se produce por eventos meteorológicos debería quedar recogido en la variable *natural*, ya que en el informe habrá alguna referencia al agua que inutilizó el motor o al polvo y arena que bloqueo un rodamiento. En total un 1,3% de todos los eventos se deben a este tipo de fenómenos. Para comprobar si las escaleras exteriores los sufren en mayor medida se representa el porcentaje de fallos por el ambiente para las escaleras interiores y exteriores en figura- 6.47.

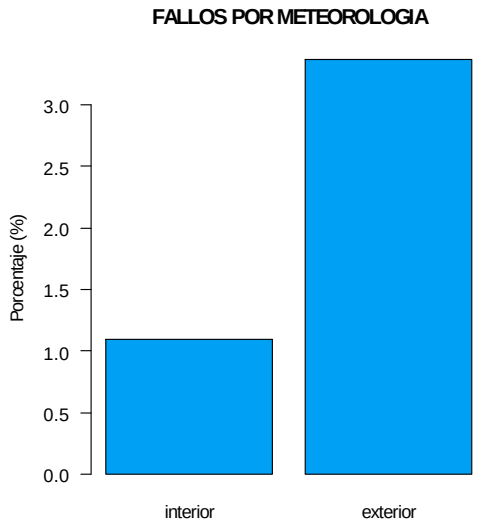


Figura- 6.47 Porcentaje de fallos por meteorología en función de la exposición de la escalera exterior

Se puede comprobar que las escaleras ubicadas a la intemperie tienen un riesgo tres veces superior de tener un fallo debido al ambiente. Esto era lo esperado, pero es importante haber cuantificado cúanto aumentaba la probabilidad de sufrir estos incidentes. Además, puede sorprender el 1% de fallos de las escaleras interiores, pero se debe principalmente a la presencia de agua de lluvia en algunas escaleras situadas en estaciones de metro semicubiertas que se consideran interiores.

Influencia de la inclinación: otro de los aspectos de seguridad que interesaba conocer era cómo afectaba la inclinación de la escalera a los incidentes relacionados con caídas de pasajeros. Esta información queda recogida en la variable *fall* que indica si en el informe se comenta algo sobre una caída de un pasajero. Se representa dicha información frente al ángulo de inclinación en figura- 6.48. Solo se emplean ángulos de escaleras ya que se ha visto que en los pasillos apenas se producen estos eventos.

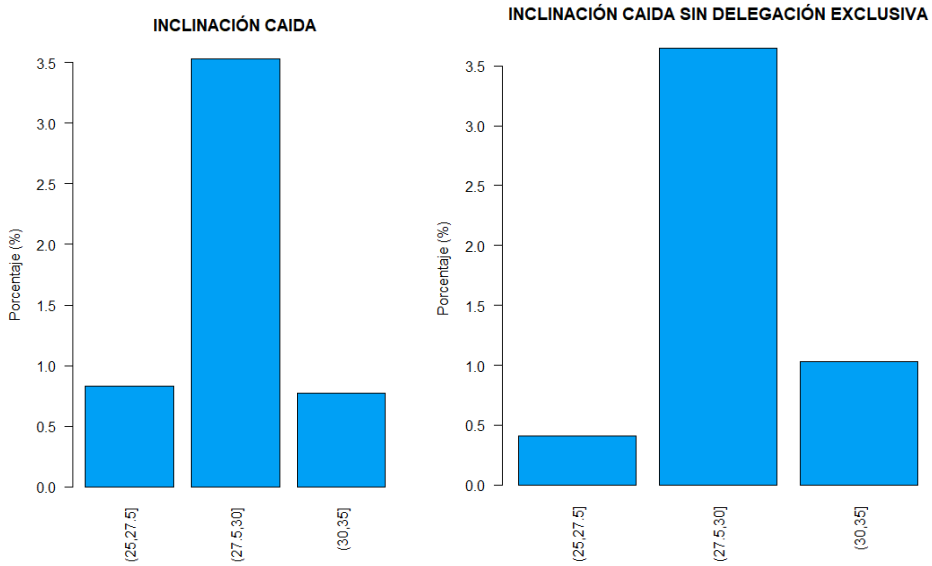


Figura- 6.48 Porcentaje de caídas en cada inclinación con (izquierda), y sin (derecha) delegaciones exclusivas

En primer lugar, se observa la diferencia que existe al excluir las cuadrillas exclusivas de un solo edificio del análisis. Dichas cuadrillas se ubican principalmente en aeropuertos, cuyas escaleras tienen mayoritariamente un ángulo de 27,5°. En ellas quedan registradas un mayor número de caídas al solucionar los operarios de Thyssenkrupp todos los incidentes, incluidos los menores, como ya se ha comentado en varias ocasiones.

En el resto de los edificios, por lo general, si no se produce el salto de ninguna seguridad que requiera rearme o no se activa el pulsador de emergencia no avisan a los operarios de mantenimiento y no queda registro de la mayoría de caidas. Esto se puede confirmar al observar en figura- 6.49 las caídas por segmento, donde se ve que son muy altas en aeropuertos en general, pero, al eliminar las delegaciones exclusivas, este segmento presenta una tasa similar al retail o al segmento de otros.

Por otro lado, la intuición inicial de que a mayor inclinación del equipo mayor número de caídas se registra no parece cumplirse. Se observa que el mayor número de incidentes por caída se da en las escaleras de 30°. Esto se puede relacionar con el segmento en el que se emplea cada inclinación de escalera como se ve en figura- 6.49.

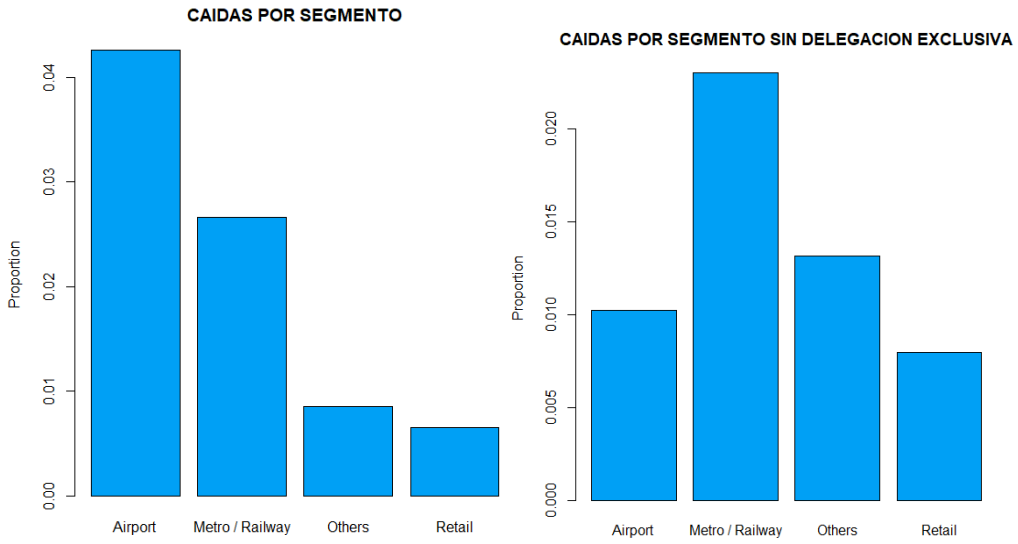


Figura- 6.49 Caídas por segmento (izquierda) y reducido a edificios sin delegación exclusiva (derecha)

Como ya se había comentado previamente en 6.2.1.-, el ángulo de 30° es mayoritario en el metro. Este se caracteriza por un uso muy intenso, en el que los pasajeros suelen ir con prisa y no suelen permanecer parados en un escalón, si no que caminan sobre la escalera, al contrario que los centros comerciales, oficinas o aeropuertos donde los usuarios suelen permanecer en un peldaño, lo que reduce enormemente el riesgo de caída. Además, en los metros existe el efecto de hora punta en el que son habituales los empujones y aglomeraciones. Esto hace que sea mucho más fácil tropezar y caer al intentar subirse o bajarse de la escalera. Estos dos fenómenos explicarían ese mayor número de incidentes relacionados con caídas.

Influencia del tamaño: siguiendo por la línea de la accidentalidad, surgió la pregunta de si el resto de las características geómetricas del equipo influían en el riesgo de sufrir, no ya una caída, si no cualquier tipo de accidente. Para comprobarlo se representa la accidentalidad frente al tiempo de viaje en figura- 6.50, como criterio alternativo se representa también los incidentes por activación de seguridades, la mayoría de las cuales se activan para disminuir los daños sobre los pasajeros en un accidente.

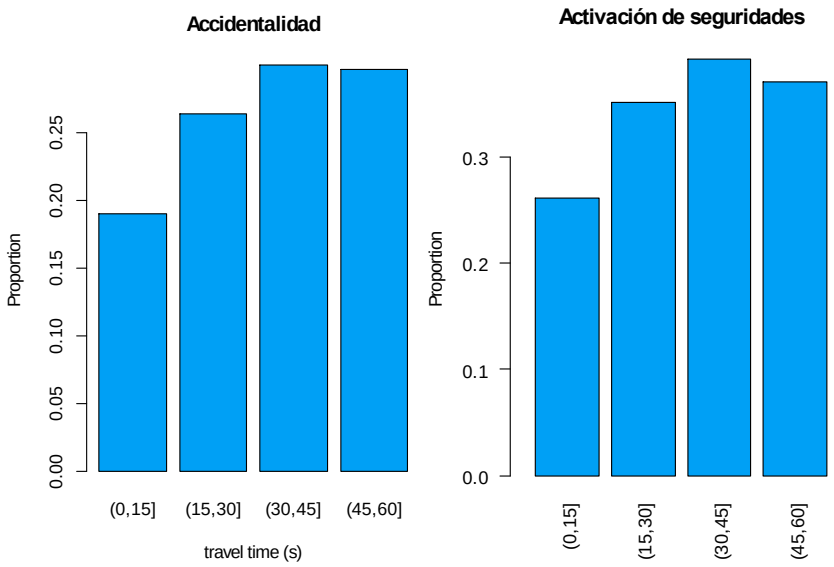


Figura- 6.50 Proporción de accidentalidad (izquierda) y activación de seguridades (derecha)

Se puede ver una tendencia creciente de accidentes a mayor tamaño del equipo que se estabiliza a partir de los 45 segundos de viaje. En este caso la dependencia con el segmento es casi despreciable, puesto que no existe un tamaño preferente para cada sector como se muestra en figura- 6.51 y, además, se esta hablando de activación de seguridades, que siempre requiere la intervención de Thyssenkrupp por lo que el efecto de las delegaciones exclusivas tampoco es significativo.

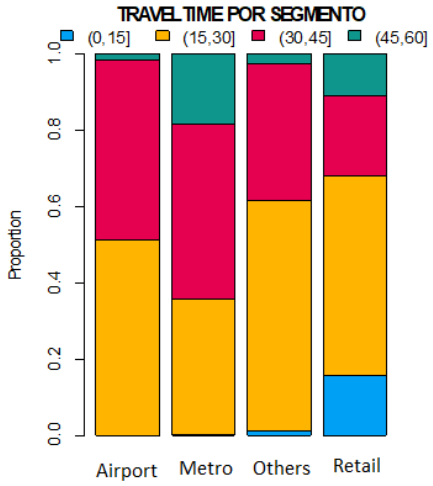


Figura- 6.51 Tiempo de viaje por segmento

Por este motivo, hay que buscar otra explicación a ese crecimiento de la accidentalidad con el tamaño.

Se cree que puede deberse, por un lado, a que el usuario tiene más tiempo para cometer alguna imprudencia al pasarse más tiempo en el equipo y, por otro lado, a que al ser un viaje más largo, es más facil que el usuario sufra distracciones, no esté atento y sin querer

introduzca algún paraguas o zapato en algún hueco o se produzca otro tipo de incidente similar.

Influencia de la tipología: otra de las hipótesis que se quería contrastar era si los pasillos sufren menos accidentes y de menor gravedad que las escaleras. Esto se explicaría por la menor carga que sufren todos los elementos al no tener que vencer a la gravedad para desplazar a los pasajeros. Para comprobar este razonamiento se representa la tasa de fallo media para ambas tipologías en la figura- 6.52:

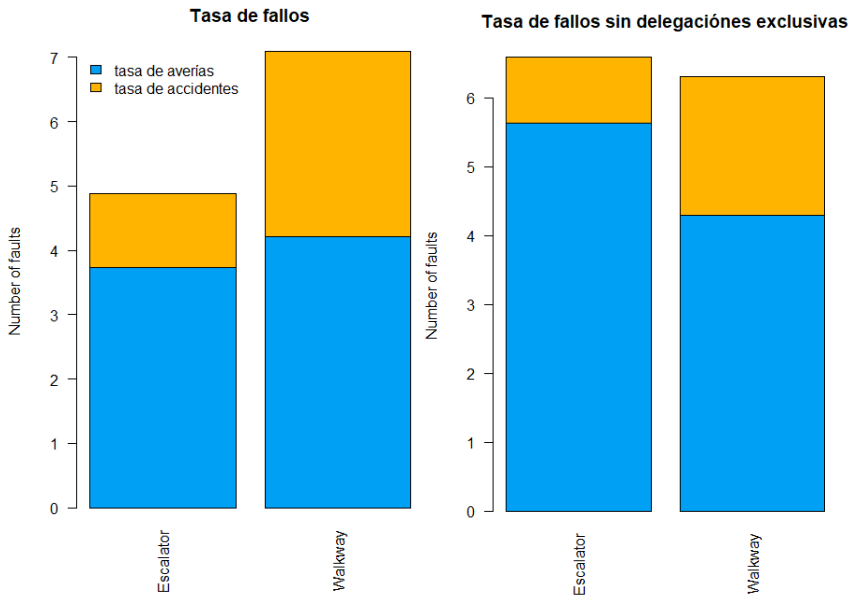


Figura- 6.52 Tasa de avería y accidente (izquierda) y restringida a edificios sin cuadrilla exclusiva (derecha)

Si no se eliminan los reportes de los aeropuertos con personal dedicado en exclusiva ni se distingue entre avería y accidente podría parecer que los pasillos tienen muchos más fallos que las escaleras. Sin embargo, ya se ha puesto de manifiesto que la verdadera medida del funcionamiento electromecánico se consigue restringiendo el análisis a las averías. Además, dado que las pasarelas son mucho más habituales en aeropuertos, es necesario excluir los equipos situados en delegaciones exclusivas para eliminar ese sesgo. Esta exclusión solo elimina 971 de los 12977 equipos por lo que el análisis no debería perder validez al conservar casi el 95% de los registros.

Al hacer ese filtrado se obtiene que, tal y como se esperaba, las escaleras sufren un mayor número de averías que los pasillos. De todas formas, aunque sufran más fallos, estos no parecen ser de más gravedad, puesto que ambos equipos tienen medianas similares en el tiempo de reparación, tal y como se ve en la tabla 6.4.

	Escalera	Pasillo rodante
Mediana del tiempo de reparación por fallo	64,7 min	65,6 min
Mediana del tiempo de reparación anual	180,2 min	170 min

Tabla 6.4 Mediana de duracion de averías tras excluir las delegaciones exclusivas

Esto parece indicar que el efecto de la gravedad no es especialmente relevante, seguramente debido a que las escaleras se diseñan con mayores coeficientes de seguridad y más robustas que los pasillos para compensar ese mayor esfuerzo por el peso.

Influencia del fabricante: uno de los aspectos en los que Thyssenkrupp tenía mayor interés era poder comparar sus propios equipos con aquellos de la competencia en los que llevaba también el mantenimiento. Se emplea como medida del desempeño el tiempo anual parado por avería —los accidentes no se incluyen ya que en principio deberían ser independientes del fabricante— al ser un buen reflejo tanto del número de fallos como de la gravedad de los mismos. También se eliminan los equipos de delegaciones exclusivas puesto que pertenecen casi en su totalidad a Thyssenkrupp y podrían sesgar los resultados. Se muestra en la figura-6.53 el diagrama de caja con los tiempos por fabricante. A partir de 3000 se cambia la escala del eje para que se pueda apreciar tanto la caja de Mitsubishi como la de los otros fabricantes. En este caso no se separa Thyssenkrupp es sus tres fábricas ya que se tratará la influencia de la fábrica más adelante en 6.6.4.-.

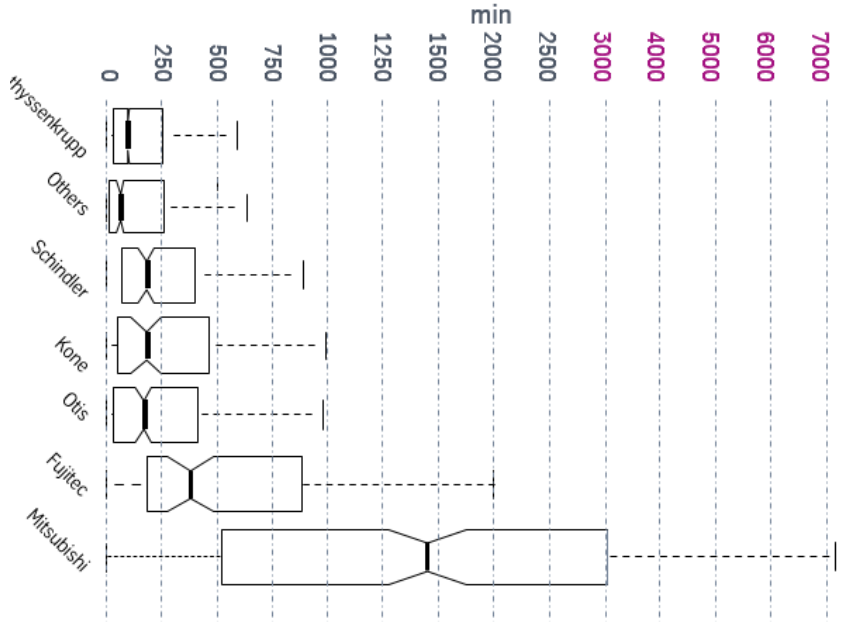


Figura- 6.53 Diagrama de cajas de tiempo de reparación de avería anual por fabricante

Se puede observar que Mitsubishi destaca por el gran tiempo parado, seguido muy de lejos por Fujitec. Estos equipos se encuentran ubicados principalmente en UAE, como se ve en la distribución por fabricante de cada país que se muestra en la figura- 6.54.

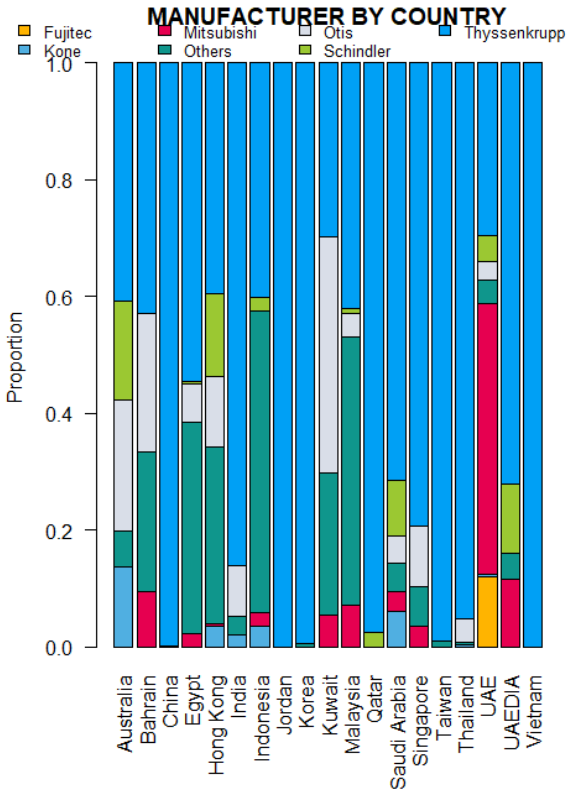


Figura- 6.54 Distribución de fabricante por país

Anteriormente ya se observó que este país era el que tenía un mayor número de fallos y esta podría ser la causa. En principio, la carga y condiciones que soportan los equipos de este país son iguales al resto, por lo que no parece que estar ubicado en UAE haga que los equipos fallen más, si no que UAE registra más fallos por tener equipos de Mitsubishi y Fujitec. No obstante, para comprobarlo se muestra la mediana del tiempo de reparación en los cuatro países con más equipos de Mitsubishi en la tabla 6.5.

EN MINUTOS	Fujitec	Kone	Mitsubishi	Other	Otis	Schindler	thyssenkrupp
Tiempo mediano anual de reparación en UAE	373	165	1786	310	206	64	196
Tiempo mediano anual de reparación en UAEDIA			320	354		220	156
Tiempo mediano anual de reparación en Bahrein			240	395	180		180
Tiempo mediano anual de reparación en Malasia			197	145	120	316	150

Tabla 6.5 Tiempo mediano anual de reparación por fabricante en países con alta cuota de Mitsubishi

Con esta comprobación se confirma que los equipos de Mitsubishi están parados más tiempo que los de Thyssenkrupp independientemente del país, sin embargo, el valor de UAE es completamente anómalo. El de Fujitec podría serlo también, pero dado que solo hay equipos de este fabricante en UAE no hay manera de contrastar sus resultados. Para intentar averiguar la causa de este tiempo atípico se examina qué tipo de fallos tienen estos fabricantes y se muestran los porcentajes de fallos eléctricos y mecánicos en la figura- 6.55.

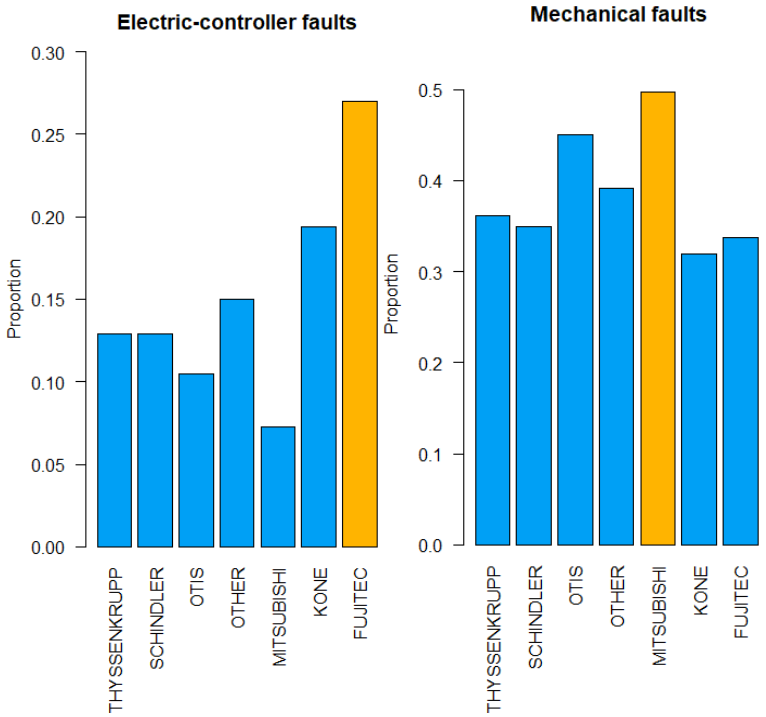


Figura- 6.55 Proporción de fallos eléctricos (izquierda) y mecánicos (derecha)

Mitsubishi es el fabricante con mayor proporción de fallos mecánicos, mientras Fujitec destaca en fallos eléctricos. Para concretar qué fallo es el más común en cada caso —lo que permitirá saber qué elemento hace que estos equipos se pasen tanto tiempo parados— se obtienen los diagramas circulares por subsistema y por fallo concreto en figura- 6.56 y figura- 6.57.

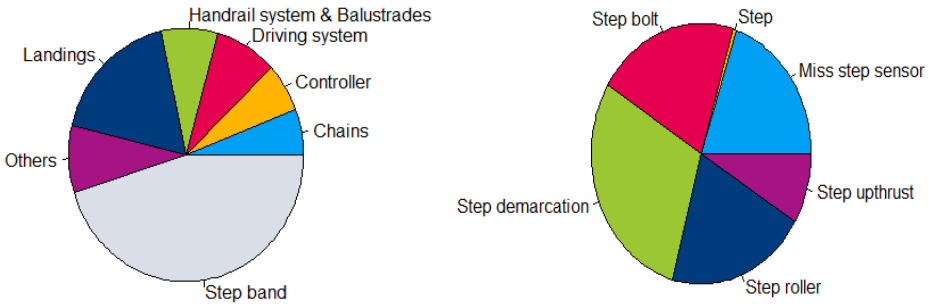


Figura- 6.56 Frecuencias de cada subsistema (izquierda) y de cada fallo de Step band (derecha) en Mitsubishi

Es destacable que en Mitsubishi los fallos mecánicos se deben especialmente a la banda de peldaños, y, dentro de este subsistema y no solo restringido a fallos mecánicos, a las demarcaciones de los peldaños, a los rodillos y su conexión con la cadena de peldaños. Es probable que los equipos de Mitsubishi instalados en UAE tengan algún defecto en este elemento o que los repuestos que se esten empleando no sean los adecuados y den problemas.

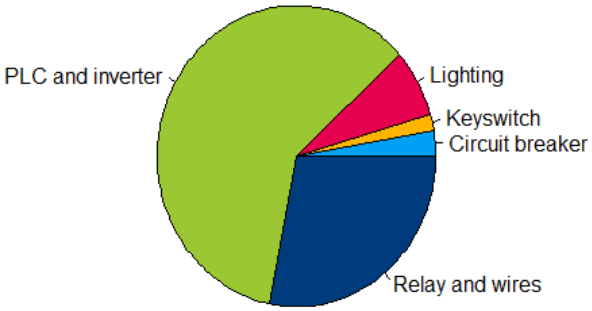


Figura- 6.57 Frecuencia de cada fallo eléctrico en Fujitec

En cuanto a Fujitec, sus fallos eléctricos se deben al equipo electrónico de control, por lo que ahora que se sabe que este elemento da problemas se puede dar una formación específica sobre el sistema de control de este fabricante para que los operarios puedan resolver mejor los fallos que surjan.

Influencia del modelo: una de las características que definen el diseño de los elementos mecánicos es el modelo del equipo. Dentro de la gama de Thyssenkrupp, el modelo Velino es el que tiene un diseño más liviano y con menores coeficientes de seguridad, seguido del Tugela y el Victoria. Por tanto, sería esperable que el tiempo que se pasen parados sea inversamente proporcional a dicha robustez. Para comprobarlo se obtiene el diagrama de cajas para cada modelo —solo para averías y sin delegaciones exclusivas— en la figura- 6.58.

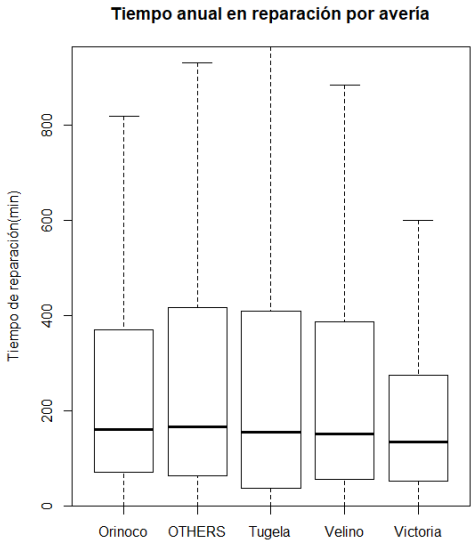


Figura- 6.58 Diagrama de cajas para cada modelo de Thyssenkrupp

El resultado muestra que apenas hay diferencia entre los distintos modelos. Esto se debería, al igual que lo que se vio entre escaleras y pasillos —se ve también aquí entre Orinoco y el resto de modelos— a que, aunque un tipo de equipo sufra más cargas que otro, —ya sea luchar contra la gravedad en escaleras frente a pasillos o el efecto de hora punta y

aglomeraciones en Tugela de metro frente a Velino de centros comerciales— el equipo se ha diseñado teniendo en cuenta ese mayor esfuerzo al que va a estar sometido. En resumen, esa mayor carga de unos modelos o segmentos no se traduce en más fallos al estar los equipos empleados para esa aplicación específicamente diseñados para ese uso.

Para confirmar la hipótesis de la correlación entre uso, resistencia del modelo y tiempo parado se va a analizar el caso que ya se había mencionado en 6.2.2.-, el grupo de escaleras Velino que se emplea en el metro en vez de en centros comerciales, que es para lo que fueron diseñadas. Si la tesis de trabajo es cierta, este grupo sufrirá un tiempo de reparación mucho mayor que lo esperado, puesto que se están utilizando en una aplicación donde la carga es superior a aquella para la que fue diseñada. Se muestra el resultado en tabla 6.6.

	Velino en comercial	Velino en otros	Velino en metro	Tugela en metro
Tiempo mediano anual de reparación	145 min	154 min	317 min	160 min
Número de equipos	2411	1822	437	2744

Tabla 6.6 Tiempo mediano de reparación y número de equipos para algunas combinaciones segmento-modelo

Tanto si se compara con la Velino en sus aplicaciones propias, como con el otro modelo (Tugela) empleado en el segmento de metro, se observa que las Velinos en el segmento de metro tienen muchos más fallos. Esto confirma que todas las escaleras funcionan bien y de manera similar en su uso esperado. Sin embargo, al emplearla en un segmento donde la carga es superior a aquella para la que fue diseñada empieza a sufrir un mayor número de averías y pasa más tiempo parada.

Por otro lado, y retomando la figura- 6.58, hay que destacar que el modelo Victoria sí que presenta menos fallos que el resto de equipos, y esto sería debido a que esta pensada para aplicaciones de carga muy elevada donde no se permiten casi fallos y se necesita una disponibilidad extremadamente alta, por lo que se diseña con coeficientes de seguridad muy altos.

Influencia del controlador: en estos últimos puntos se analizará el comportamiento de elementos constructivos concretos de los equipos para ver si las distintas opciones tienen alguna influencia en los fallos. Se empezará por ver el efecto que tiene el controlador instalado. La unidad puede ser controlada por un PLC, por un microprocesador o por lógica cableada (relés). Estas tres alternativas están disponibles para todos los modelos y tamaños por lo que la elección de uno u otro solo depende de las preferencias del cliente.

Para poder comparar el rendimiento de cada uno de los controladores se representa el tiempo parado al año por avería en la figura- 6.59 y la tasa mediana de averías en la tabla 6.7.

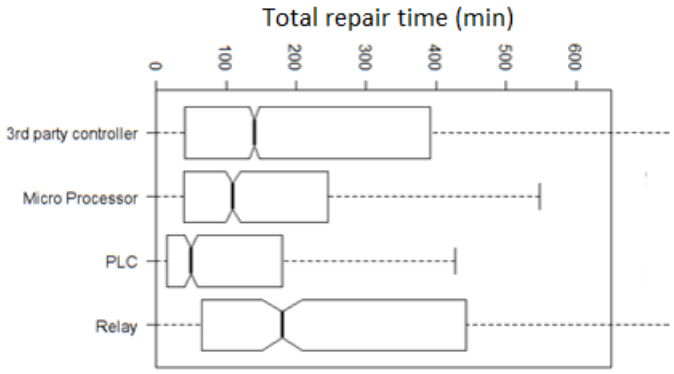


Figura- 6.59 Diagrama de cajas del tiempo de reparación anual para cada tipo de controlador

	Microprocessor	PLC	Relay	3º party controller
Tasa de averías mediana	1	2	2	2
Porcentaje de equipos con controlador conocido	10,8%	8,6%	5,1%	75,6%

Tabla 6.7 Tasa de averías por tipo de controlador

Se puede apreciar que el tiempo de reparación de los equipos con PLC es mucho menor que en el resto de los controladores. Sin embargo, el número de averías registradas es similar, lo que significa que las averías se solucionan en menos tiempo. La principal explicación que se ha encontrado es que los PLC cuentan con un sistema de diagnóstico y un display de tal manera que cuando el operario llega al equipo el fallo se indica en el display y ya sabe qué parte tiene que inspeccionar, lo que le ahorra mucho tiempo. Los equipos con microprocesador cuentan con dispositivos similares, pero a más bajo nivel, solo para ciertos fallos, y este sería el motivo por el que su tiempo de reparación es menor que el de los relés también.

Se quiere comentar que hay un 30% de equipos cuyo controlador se desconoce. Además, el resto está muy disperso y muchos se asignan a 3º party controller, por lo que los datos de PLC, relé y microprocesador que se analizan en este apartado están restringidos a solo unos 2000 equipos de los 13000 equipos.

Influencia de los componentes mecánicos: en el último punto de esta sección se intenta averiguar cómo afectan las distintas configuraciones posibles del equpo. En concreto, se verán las diferencias entre disponer de uno o dos motores, balaustrada metálica o de cristal y tener o no un perfil de metal bajo el pasamanos. Para analizar su efecto se calculará la tasa de averías para las zonas en las que se ubica cada una de las opciones, el tiempo mediano de reparación de cada una de esas averías y el número de equipos con cada configuración. Todo esto se muestra en la tabla 6.8. En esta sección no es necesario eliminar las delegaciones exclusivas al centrarse en averías y no emplear datos de las zonas con mayor accidentalidad como las plataformas de llegada.

MOTOR	Single	Dual
Tasa averías cadenas o sistema motriz	0,54	0,46
Mediana repair time cadenas o sistema mot	58 min	55 min
Nº de equipos	5564	1262
BALAUSTRADA	Cristal	Metal
Tasa averías balaustrada	0,07	0,03
Mediana *repair time* balaustrada	58 min	42 min
Nº de equipos	6956	1672
PASAMANOS	Con perfil	Sin perfil
Tasa averías pasamanos	0,29	0,34
Mediana *repair time* pasamanos	63 min	40 min
Nº de equipos	5370	3258

Tabla 6.8 Influencia de distintas configuraciones mecánicas

- **Motor single/dual:** podría pensarse que tener dos motores, uno para cada lado, conseguiría que las cargas fueran más equilibradas y hubiera menos fallos. Sin embargo, parece que la influencia de esta variable es casi nula y tanto los equipos *Single* como los *Dual* tienen fallos similares del sistema motriz y cadenas y se tarda lo mismo en repararlos.
- **Balaustrada de cristal/metal:** en este caso sí que parece haber una diferencia. Los equipos con balaustrada de cristal tienen el doble de fallos en este elemento y además se tarda más en repararlos que los que tienen balaustrada metálica. Esto podría deberse a que es más fácil que se rompa el cristal. Además, los faldones de las balaustradas de cristal pueden ser más endebles ya que los equipos con laterales metálicos son de uso intensivo y tienen un zócalo más resistente.
- **Pasamanos con/sin perfil:** esta última variable tampoco parece tener una influencia clara. Presentan una tasa de averías similar en ambas opciones, aunque los tiempos de reparación son muy distintos. Esto se debe a que el pasamanos, al contrario que en los dos casos anteriores, presenta fallos con tiempos de reparación muy dispares. Por eso, es necesario ver la tasa de avería para cada uno de ellos en tabla 6.9.

Tasa de avería	Sincronismo	Flap	Pasamanos	Guías
Con perfil	0,03	0,06	0,15	0,05
Sin perfil	0,09	0,1	0,12	0,03

Tabla 6.9 Tasa de avería en cada fallo del pasamanos

Se observa que los fallos de sincronismo y flap de seguridad —cuyo tiempo de reparación es muy corto al ser seguridades— son mayores en los equipos sin perfil, mientras los fallos de las guías y el pasamanos son mayores cuando sí hay perfil. Estos fallos son más difíciles de reparar al tener que abrir el equipo para arreglarlo, lo

que explicaría ese mayor tiempo de reparación de cada fallo en los equipos con perfil bajo pasamanos.

Así pues, aunque hay diferencia entre las tasas de cada fallo individual, no parece que el efecto de tener o no perfil bajo pasamanos sea muy significativo al tener igual tasa global y poder explicar la diferencia de tiempos por el tipo de fallo de cada una y no por una mayor gravedad de los fallos en una frente a otra. De todas formas, un estudio más exhaustivo en laboratorio podría comprobar si esto es cierto.

6.6.2.- Tendencias del mantenimiento preventivo

CONFIDENCIAL

6.6.3.- Tendencias con el año de fabricación

Otra relación que se espera encontrar es entre la antigüedad del equipo y el tiempo anual que se pasa en reparación. La hipótesis de partida es que cuanto más viejo sea el equipo más averías presentará y más tiempo estará parado al año. Para comprobarlo se obtiene, al igual que en el caso anterior, los diagramas de cajas y la tasa de averías en función del lustro en el que fue fabricada la unidad. Se representa en la figura- 6.60.

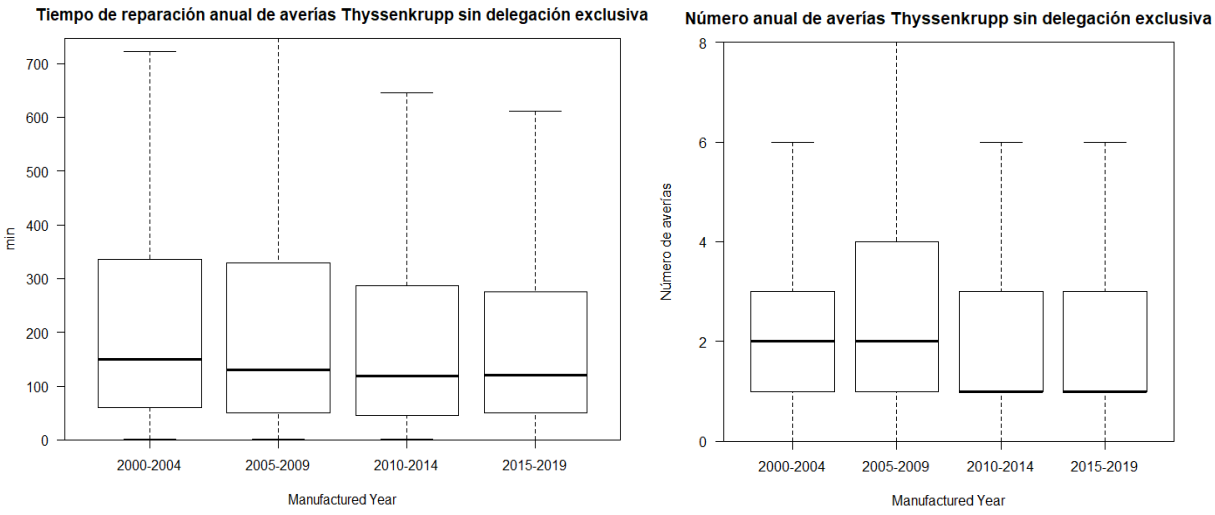


Figura- 6.60 Tiempo anual de reparación de averías (izquierda) y número de averías (derecha) en función del lustro de fabricación

En este caso, aunque la tendencia es muy débil, si se observa una cierta disminución del tiempo de reparación y el número de averías cuanto más moderno es el equipo. De todas formas, la variación es mínima por lo que parece que su influencia no es muy significativa. Esto se podría explicar porque muchos de los elementos del equipo se van reponiendo con los años notándose así menos la antigüedad de la unidad.

Otra de los datos que más interesan es saber cómo va evolucionando en función del año de fabricación el fallo más habitual. Esto se muestra en la figura- 6.61.

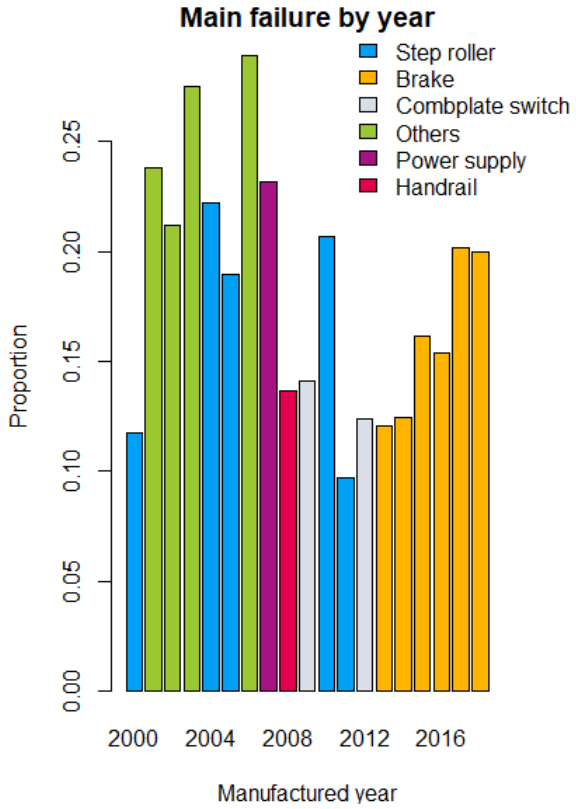


Figura- 6.61 Fallo más habitual en función del año de fabricación

El resultado muestra que durante la primera década de los 2000 se fueron alternados los fallos externos con los de rodillos de peldaño. En los últimos años de esa década y principios de la siguiente no hay una tendencia y se alternan distintos fallos como el pasamanos o la seguridad de placa de peines. Finalmente, en los últimos años se aprecia un incremento paulatino de los fallos relacionados con los frenos.

Esto podría delatar algún fallo en el diseño, montaje o fabricación de este elemento, ya que no es un dato puntual y además parece que se va agravando con los años. Para intentar dar un poco de luz a la causa de este fallo de frenos y también al de rodillo de peldaños —son los dos más graves puesto que el resto son puntuales o relacionados con causas externas al equipo— se analizarán los fallos por fábrica, ya que se cree que el fallo puede venir de ahí.

6.6.4.- Tendencias con la fábrica (*step roller* y *brake*)

CONFIDENCIAL

6.7.- OTROS MODELOS ESTADÍSTICOS

6.7.1.- Predicción de accidente o avería

Una vez analizadas las principales tendencias que se han encontrado en la base de datos, es momento de plantear cómo mejorar el mantenimiento predictivo de los equipos. Para ello, se desarrollarán una serie de modelos de aprendizaje automático que anticipen información sobre el incidente. Esto permitirá a la delegación priorizar unos fallos sobre otros y dotar de los medios que considere más adecuados para cada caso al operario antes de que vaya a solucionar el incidente.

En primer lugar, sería conveniente conocer si un incidente es una avería o un accidente, ya que la manera de solucionarlos es muy distinta. Los accidentes requieren una intervención inmediata para garantizar que ningún pasajero se ha herido y que ningún objeto personal ha sufrido daños e intentar minimizar la gravedad del suceso. En cambio, las averías, si bien deben ser también reparadas con presteza, no afectan a nada externo al equipo por lo que no requieren una intervención tan urgente.

Actualmente, gran parte de los incidentes llegan a través de llamadas al *call center,* por lo que el interlocutor de la llamada podría indicar si es un accidente —y hay que acudir cuanto antes— o es una avería y no es necesaria tanta celeridad. Sin embargo, ya hoy en día algunas paradas son comunicadas por los propios controladores de los equipos, los cuales tienen conexión a Internet y pueden enviar avisos en caso de que se detengan. En estos casos no habría manera de conocer la naturaleza del incidente, por lo que sería muy útil contar con una herramienta que pudiera, a partir de variables del equipo (tamaño, inclinación, fabricante, localización, modelo, historial de fallos previos, antigüedad y datos constructivos) y variables del fallo conocidas antes de llegar al lugar del incidente (hora, día y mes), predecir si es accidente o avería y así asignar prioridad a unos fallos frente a otros.

Podría también pensarse que estos equipos, al igual que mandan señal de parada, podrían indicar cuál es la causa de dicha parada, pero esto es solo posible para algunas seguridades y, aunque se conociera la seguridad activada, esto no va necesariamente ligado a que el incidente sea un accidente o una avería por lo que este modelo de predicción sigue siendo muy útil.

Así pues, se entrenarán dos modelos distintos —un árbol de clasificación y una red neuronal— con las mismas variables predictoras para ver cuál consigue mejores resultados.

Variables predictoras: muchas de estas variables ya formaban parte de la base de datos. Se pueden dividir entre perteneciente al equipo y pertenecientes al fallo y se pueden ver en la tabla 6.10.

Variables del equipo	UnitType	Angle	TravelTime	exclusive	newsegment	Manufacturer_en	ModeloBase
	EscalatorOutdoorUnit	Balaustrade	DriveType_en	EscalatorDrive	HandrailPerfil	Controller	newyear
Variables del incidente	Hour	Day	Month	Dayyear			

Tabla 6.10 Variables predictoras consideradas en los modelos

La variable *exclusive* es binaria y su valor 1 indica que el equipo está mantenido por una delegación con dedicación exclusiva. Como se ha visto en apartados anteriores, esto influye mucho en el número de accidentes y averías que registra un equipo. Por otro lado, la variable *newsegment* es la que recoge la información sobre el segmento —metro, aeropuerto, comercial y otros— en el que está la unidad.

Tras realizar varias pruebas preliminares se ha observado que el resultado no era del todo satisfactorio, por lo que se ha pensado cómo se podría aumentar la información disponible y así facilitar la labor de predicción. Se ha decidido crear una nueva serie de variables, indicativas del histórico de fallos, las cuales recogen cuáles fueron los dos fallos previos que tuvo cada equipo. En total son 12 variables, en 4 grupos de 3. La primera de cada grupo indica el fallo inmediatamente anterior, la segunda el fallo previo al inmediatamente anterior y la tercera es la combinación de las dos anteriores. Por otra parte, el primer grupo (ac) indica si dichos fallos previos fueron accidentes o averías, el segundo (fp) el tipo de fallo —mecánico, eléctrico, de seguridad u otro— el tercero el subsistema al que afectó (ar) y el cuarto grupo el fallo concreto (ty). Se quiere destacar que se ha empleado el número de fallo, subsistema y tipo empleado en 6.5.2.- en vez de su denominación completa para que sea más compacto. Se dispone el código para la creación de estas variables en la figura- 6.62:

```
FaultAreaName_en2=tapply(D$FaultAreaName_en2,D$UnitNumber,c)
FaultAreaName_en2 = FaultAreaName_en2[-which(sapply(FaultAreaName_en2,
is.null))]
faulty=as.data.frame.array(FaultAreaName_en2)
faulty$FaultAreaName_en2=as.character(faulty$FaultAreaName_en2)
faulty$FaultAreaName_en2 = gsub("[c(),]","",
as.character(faulty$FaultAreaName_en2))
faulty$FaultAreaName_en2 = gsub("[:]"," ",
as.character(faulty$FaultAreaName_en2))
faulty=tidyr::separate(faulty,FaultAreaName_en2,
into=as.character(c(1:176)),sep=" ",fill="right")
DD$ty = NA # fallo previo
DD$ty2 = NA # fallo anteprevio
```

```
for (e in rownames(faulty)) {
DD[which(DD$UnitNumber==e)[-1],"ty"] =
as.character(faulty[e,1:(sum(!is.na(faulty[e,]))-1)])
numfallos = sum(!is.na(faulty[e,]))
if (numfallos>2) DD[which(DD$UnitNumber==e)[-(1:2)],"ty2"] = as.character(f
aulty[e,1:(numfallos-2)])}
D$ty=factor(DD$ty)
D$ty2=factor(DD$ty2)
D$ty3=as.integer(D$ty)+as.integer(D$ty2)/100
D$ty3=factor(D$ty3)
```

Figura- 6.62 Código para creación de variables del histórico de fallos

Esta mejora del modelo tiene cierta similaridad con los procesos markovianos en tiempo discreto [28], los cuales intentan predecir a partir de un suceso presente cuál será el acontecimiento siguiente que tendrá lugar. Esto se suele visualizar mediante las matrices de transición. Estas indican por filas la probabilidad condicionada de que suceda A sabiendo que ha sucedido B. Se pueden construir a partir del primer fallo anterior o de los dos previos. Se pone como ejemplo la matriz de transición para el tipo de fallo anterior en tabla 6.11.

	MECÁNICO	ELÉCTRICO	SEGURIDAD	OTRO
MECÁNICO	51%	7%	24%	17%
ELÉCTRICO	26%	21%	28%	25%
SEGURIDAD	25%	8%	44%	23%
OTRO	20%	9%	26%	46%

Tabla 6.11 Matriz de transición por filas del tipo de fallo

Se observa que en todos los casos lo más probable después de un tipo A es que se produzca otro fallo A, salvo en el caso de los eléctricos, que es más probable que se produzcan después de una seguridad, aunque casi no hay diferencia con el resto de las probabilidades condicionadas. Como se puede intuir, esta información es muy relevante para predecir cómo será el próximo fallo que tendrá un equipo, por lo que se incluirán estas 12 variables en los modelos realizados.

También destacar que no todas las variables presentes serán utilizadas por el modelo, ya que será el propio algoritmo el que escoja aquellas que realmente sirven para predecir el resultado que se quiere conseguir. De hecho, hay muchas como *Day* o *EscalatorDrive* que ya se ha visto en los gráficos que apenas tienen influencia y no serán una gran ayuda para el modelo. Por este motivo, despues de las pruebas iniciales, se ha reducido el número de variables para disminuir el tiempo de computación y emplear solo aquellas que se han revelado importantes.

Conjunto de datos: se emplea para el entrenamiento de ambos modelos una muestra aleatoria equilibrada, es decir, con igual número de accidentes que de fallos, para que no haya sesgo hacia uno de los dos. Se escoge como tamaño de cada grupo el 80% del menor

de ambos, en este caso los accidentes que representan un 25% de los fallos registrados. Así pues, el *TrainingData* son 27084 registros mientras el *TestData* serán los restantes.

Modelos: se detallan a continuación las particularidades y el código de cada modelo.

- **Árbol de clasificación:** se ha escogido un cp=0,005 despues de varias pruebas. Se puede ver en http://bellman.ciencias.uniovi.es/~raul/complete.html el árbol de clasificación, así como el conjunto de reglas que define. Este modelo solo emplea como variables ac3, ty3, fabricante, modelo y segmento. Esto parece indicar que el momento del incidente no es relevante, siendo el modelo de equipo y su historial lo que mejor define su comportamiento futuro.
- **Red neuronal:** se ha empleado una red con una sola capa oculta con 5 neuronas, 1431 entradas y 1 salida, lo que implica 7166 pesos distintos a ajustar. Se pueden ver los valores de los pesos finales en ANEXOS IV.

Se muestra el código para ambos modelos en la figura- 6.63:

```
#%%%%%%%%%%%%%%%%%%%%%% ARBOL CLASIFICACION DE ACCIDENTES %%
input_one=D[which(D$accident==1),]
input_zero=D[which(D$accident==0),]
set.seed(100)
input_ones_training_rows <- sample(1:nrow(input_one), 0.8*nrow(input_one))
input_zeros_training_rows <- sample(1:nrow(input_zero),0.8*nrow(input_one))
training_ones <- input_one[input_ones_training_rows, ]
training_zeros <- input_zero[input_zeros_training_rows, ]
trainingData <- rbind(training_ones, training_zeros)
test_ones <- input_one[-input_ones_training_rows, ]
test_zeros <- input_zero[-input_zeros_training_rows, ]
testData <- rbind(test_ones, test_zeros)
nnet(accident~EscalatorOutdoorUnit+Balaustrade+DriveType_en+EscalatorDrive+
Handrailperfil+Hour+UnitType_en+Angle+newsegment+ModeloBase+Manufacturer_en
+exclusive+TravelTime+ac+ac2+ac3+fp+fp2+fp3+ar+ar2+ar3+ty+ty2+ty3,
trainingData,size=5,MaxNWts=200000,rang=3e-5,maxit=150)->neurnet2#RED
rpart (accident~Hour+UnitType_en+Angle+newsegment+ModeloBase+Manufacturer_e
n+exclusive+TravelTime+ac+ac2+ac3+fp+fp2+fp3+ar+ar2+ar3+ty+ty2+ty3,
trainingData,control=rpart.control(cp=0.005)) -> arbol.af#ARBOL
```

Figura- 6.63 Código para crear modelos de árbol de clasificación y red neuronal para predecir accidentes

Finalmente, hay que comparar el acierto de cada modelo al predecir los resultados del *TestData*. Se muestra dicho porcentaje junto a la matriz de confusión de cada caso en tabla 6.12.

	CON EL TESTDATA							
		REALES				REALES		
		accidente	averia			accidente	averia	
PREDICCIÓN	accidente	71,82%	37,36%	PREDICCIÓN	accidente	75,78%	29,71%	
	averia	28,18%	62,64%		averia	24,22%	70,29%	
	ÁRBOL DE CLASIFICACIÓN				RED NEURONAL			
	Tasa de acierto		71,03%		Tasa de acierto		71,34%	
	CON TODA LA POBLACIÓN							
		REALES				REALES		
		accidente	averia			accidente	averia	
PREDICCIÓN	accidente	72,58%	35,02%	PREDICCIÓN	accidente	80,80%	23,65%	
	averia	27,42%	64,98%		averia	19,20%	76,35%	
	ÁRBOL DE CLASIFICACIÓN				RED NEURONAL			
	Tasa de acierto		70,63%		Tasa de acierto		78,17%	

Tabla 6.12 Matriz de confusión y tasa de acierto de los modelos creados para predecir si es accidente

Se puede ver que ambos modelos tienen resultados similares, aunque la red neuronal es ligeramente mejor. La única ventaja de este árbol de clasificación es que permite saber de forma sencilla cómo se clasifican los registros, lo que puede ayudar a entender los patrones que encuentra.

Hay que llamar la atención sobre el hecho de que no se consigue un acierto superior al 80%. Esta cifra puede parecer muy baja, pero hay que tener en cuenta que el fenómeno del accidente tiene un componente puramente aleatorio que hace que sea muy difícil de predecir su ocurrencia, por lo que se puede considerar que el resultado obtenido es aceptable.

De todas formas, hay que contrastar estos modelos con la alternativa 0, esto es, clasificar todos los incidentes como averías —con lo que se acertaría en un 74% de las ocasiones— y no dar prioridad a ningún fallo frente a otro.

Para comparar ambos modelos hay que tener en cuenta que indicar que un fallo es un accidente cuando realmente es una avería tiene un coste asociado α, ya que obliga a actuar más rápido de lo que sería necesario, por lo que se hay que movilizar antes al operario y disponer de medios suficientes para llegar al establecimiento. Por otro lado, indicar que un fallo es una avería cuando realmente es un accidente tiene un coste β, ligado en este caso a la peor imagen que recibirá el cliente al no haber actuado suficientemente rápido. Estos costes son difíciles de cuantificar, pero permiten comparar la bondad de ambos modelos:

$$Coste\ modelo\ 0 = 0 \cdot 0{,}74 \cdot \alpha + 1 \cdot 0{,}26 \cdot \beta = 0{,}26 \cdot \beta \quad (6.1)$$

$$Coste\ modelo\ ML\ = 0.24 \cdot 0{,}74 \cdot \alpha + 0.19 \cdot 0{,}26 \cdot \beta = 0.17 \cdot \alpha + 0.05 \cdot \beta \quad (6.2)$$

El modelo de *Machine Learning* será más beneficioso si:

$$Coste\ modelo\ 0 > Coste\ modelo\ machine\ learning \tag{6.3}$$

$$0{,}26 \cdot \beta > 0.05 \cdot \beta + 0.17 \cdot \alpha \tag{6.4}$$

$$\beta > 0.81 \cdot \alpha \tag{6.5}$$

Por tanto, si el coste de equivocarse con los accidentes β es similar o mayor al de equivocarse con las averías α el modelo de *Machine Learning* es mejor que la alternativa 0. Esto en principio debería cumplirse, puesto que el daño reputacional de no acudir rápido a un accidente debería implicar un mayor coste a la larga, por lo que se recomienda emplear este modelo en el mantenimiento real; si bien sería mejor que pase previamente por unas fases de testeo y mejora.

6.7.2.- Predicción del subsistema de fallo

De manera similar al modelo anterior, seria conveniente anticipar a qué subsistema afecta un incidente, ya que esto permitiría llevar repuestos específicos de dichos elementos y reduciría el tiempo que tiene que dedicar el operario a explorar el equipo y encontrar dónde se produjo el fallo. Es decir, intentaría llevar a cabo el papel del display en los equipos con PLC con diagnóstico, pero para cualquier unidad y no solo restringido a ciertos fallos.

Las variables predictoras iniciales y los modelos usados son análogos a los del subapartado anterior. La única diferencia estriba en el *TrainingData*, ahora compuesto por una muestra equilibrada cuyo tamaño para cada subsistema es en torno al 80% del subsistema con menos fallos. Esto se traduce en unos 3000 fallos por cada uno, sumando el *TrainingData* 21000 informes y siendo el resto *TestData*. Al igual que en el caso anterior, se entrena con un menor número de registros respecto a la comprobación, pero es la única manera de asegurar que la muestra esté equilibrada y no haya sesgo hacia ningún subsistema por ser mayoritario. Se explican las particularidades de cada modelo a continuación:

- **Árbol de clasificación:** se ha escogido un cp=0,002 tras varias pruebas. No se emplean las variables ar3 ni las de ty al observar que el resultado era mejor cuando se excluían del modelo. Se puede ver en http://bellman.ciencias.uniovi.es/~raul/complete.html el árbol de clasificación, así como el conjunto de reglas que define. Este modelo emplea como variables ar, fabricante, modelo, país y año de fabricación, segmento y presencia de delegación exclusiva. Esto indica que el historial de fallos o el momento del incidente no es tan importante como la antigüedad y modelo para predecir el próximo fallo.

- **Red neuronal:** se ha empleado una red con una sola capa oculta con 15 neuronas, 1437 entradas y 7 salidas. Esto se traduce en un total de 21682 pesos a determinar. Se pueden ver sus valores finales en ANEXOS V. Hay que destacar que el número de neuronas y el número de capas se han escogido tras varias pruebas intentando que el modelo no sea muy complejo y se consigue un resultado aceptable.

Se puede ver el cógido para ambos modelos en figura- 6.64 y se muestran los resultados de ambos modelos en tabla 6.13.

```
#%%%%%%%%%%%%%%%%% ARBOL CLASIFICACION DE SUBSISTEMA DE FALLO    %%%%%%M=do.
call(c, by (D, D$Faultareapublic, function(x) sample(x$id,3000)))
trainingData=D[match(M,D$id),]
testData=D[-match(M,D$id),]

rpart (Faultareapublic~Controller_en+Balaustrade+factorycountry+newyear+Hou
r+UnitType_en+Angle+newsegment+ModeloBase+Manufacturer_en+exclusive+TravelT
ime+factorycountry+max+ar+ar2+ac+ac2+ac3+fp+fp2+fp3, trainingData,control=r
part.control(cp=0.002)) -> arbol.af

nnet (Faultareapublic~EscalatorOutdoorUnit+factorycountry+newyear+Balaustra
de+DriveType_en+EscalatorDrive+Handrailperfil+Hour+UnitType_en+Angle+newseg
ment+ModeloBase+Manufacturer_en+exclusive+TravelTime+factorycountry+max+ar+
ar2+ar3+ac+ac2+ac3+fp+fp2+fp3+ty+ty2+ty3, trainingData,size=15,maxit=10,Max
NWts=45000,rang=1e-6) -> neur3
```

Figura- 6.64 Código para crear el árbol de clasificación y red neuronal para predecir subsistema de fallo

		CON EL TESTDATA						
		REALES						
		Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
PREDICCIÓN	Cadenas	29,86%	10,83%	15,03%	12,33%	7,31%	10,39%	6,54%
	Controlador	22,19%	34,59%	24,16%	28,72%	18,19%	29,99%	18,48%
	Sistema motriz	7,12%	8,65%	23,43%	8,04%	4,85%	8,35%	5,08%
	Pasamanos y balaustrada	12,33%	9,62%	9,13%	18,51%	8,26%	8,92%	7,13%
	Plataforma de llegada	9,32%	15,89%	9,98%	16,61%	46,94%	15,91%	7,70%
	Otros	3,84%	8,80%	6,65%	4,65%	3,36%	18,11%	1,97%
	Banda de peldaños	15,34%	11,62%	11,63%	11,15%	11,09%	8,33%	53,10%
	ÁRBOL DE CLASIFICACIÓN							
	Tasa de acierto							36,06%
		CON TODA LA POBLACIÓN						
		REALES						
		Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
PREDICCIÓN	Cadenas	26,63%	11,14%	15,18%	12,34%	7,17%	10,47%	6,59%
	Controlador	22,47%	34,18%	24,69%	28,12%	18,12%	29,80%	18,33%
	Sistema motriz	7,61%	8,71%	22,97%	8,14%	4,76%	8,14%	5,11%
	Pasamanos y balaustrada	11,92%	9,99%	9,52%	18,91%	8,39%	9,03%	7,09%
	Plataforma de llegada	10,73%	15,94%	9,59%	16,96%	47,28%	15,79%	7,91%
	Otros	4,73%	8,38%	6,73%	4,46%	3,32%	18,39%	1,96%
	Banda de peldaños	15,93%	11,66%	11,31%	11,06%	10,96%	8,39%	53,01%
	ÁRBOL DE CLASIFICACIÓN							
	Tasa de acierto							34,70%

Tabla 6.13 Matriz de confusión y tasa de acierto para cada subsistema del árbol de clasificación

		CON EL TESTDATA						
		REALES						
		Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
PREDICCIÓN	Cadenas	20,00%	6,06%	11,29%	6,96%	8,42%	9,78%	3,21%
	Controlador	0,00%	12,12%	9,68%	6,09%	5,05%	4,89%	1,92%
	Sistema motriz	0,00%	36,36%	27,42%	16,52%	5,39%	21,74%	6,41%
	Pasamanos y balaustrada	80,00%	9,09%	11,29%	30,43%	12,46%	24,46%	7,05%
	Plataforma de llegada	0,00%	15,15%	12,90%	13,91%	47,81%	11,41%	38,46%
	Otros	0,00%	21,21%	24,19%	22,61%	4,71%	23,37%	2,56%
	Banda de peldaños	0,00%	0,00%	3,23%	3,48%	16,16%	4,35%	40,38%
	RED NEURONAL							
	Tasa de acierto							35,80%
		CON TODA LA POBLACIÓN						
		REALES						
		Cadenas	Controlador	Sistema motriz	Pasamanos y balaustrada	Plataforma de llegada	Otros	Banda de peldaños
PREDICCIÓN	Cadenas	50,00%	6,78%	7,27%	4,12%	6,91%	7,00%	4,04%
	Controlador	10,53%	37,29%	8,18%	4,12%	3,99%	4,67%	1,52%
	Sistema motriz	21,05%	25,42%	47,27%	13,92%	4,52%	15,95%	6,06%
	Pasamanos y balaustrada	13,16%	6,78%	10,00%	52,58%	11,17%	19,46%	6,57%
	Plataforma de llegada	5,26%	8,47%	7,27%	9,79%	55,85%	8,95%	34,85%
	Otros	0,00%	13,56%	14,55%	13,40%	3,99%	40,86%	2,02%
	Banda de peldaños	0,00%	1,69%	5,45%	2,06%	13,56%	3,11%	44,95%
	RED NEURONAL							
	Tasa de acierto							48,62%

Tabla 6.14 Matriz de confusión y tasa de acierto para cada subsistema de la red neuronal

El porcentaje de acierto vuelve a ser similar para ambos modelos en el *TestData*, del entorno del 35%. Con toda la población, la tasa de acierto es superior en la red neuronal, aunque sigue siendo muy baja, de nuevo debido en parte a la aleatoriedad del fenómeno y en parte al aumento del número de categorías posibles a las que se puede asignar un fallo. De hecho, si se asignaran al azar se acertaría 1 de cada 7 veces, un 14,29%, por lo que el azar funciona mejor que la red neuronal en el *TestData* en el controlador y se comporta casi igual en "Otros". En el árbol, el azar casi es mejor que el modelo en "Otros" y en el pasamanos y balaustrada.

Así pues, a la vista de los resultados este modelo puede considerarse una orientación inicial, pero será necesario construir modelos distintos, con un mayor número de datos y nuevas variables para conseguir resultados significativos que se puedan aplicar con fiabilidad al mantenimiento real.

6.7.3.- Análisis ANOVA de la influencia del operario en el tiempo de reparación

Finalmente, desde Thyssenkrupp han indicado que sería de gran utilidad conocer cómo de influyente es el operario en el tiempo de reparación, es decir, si la habilidad del técnico condiciona mucho cuánto se tarda en arreglar un fallo. Esto les permitiría saber si es muy rentable invertir en formación y cursos para los trabajadores u ofrecer mejores condiciones de trabajo a aquellos con más experiencia. Si por el contrario se obtiene que no influye la pericia del trabajador a la hora de arreglar el equipo, la política de recursos humanos podría ser muy distinta.

Para conocer dicha información se va a emplear un análisis ANOVA, también llamado análisis de la varianza, que mide si la media de una variable continua es igual para distintas poblaciones definidas en función de variables factor.

Esta prueba estadística requiere que se cumplan ciertas hipótesis, como la normalidad de la distribución de la duración y que las distintas varianzas sean similares. Sin embargo, en este caso solo se emplea el ANOVA solo como una medida descriptiva de cuánto influyen los distintos parámetros, sin importar el valor concreto que se obtiene. Por este motivo, no se comprobarán dichas hipótesis.

La base de datos cuenta con un gran número de trabajadores registrados —más de 3000— por lo que es necesario reducir el número de operarios considerado para que se pueda llevar a cabo el análisis de manera correcta. Se ha decidido restringir el análisis solo a UAEDIA. Esto se debe a que tiene una delegación exclusiva de Thyssenkrupp, por lo que los equipos tienen todos condiciones equivalentes de uso al estar en el mismo edificio y además todo su mantenimiento corre a cargo de un grupo conocido de técnicos. Además, esta delegación es de las que cuenta con mayor número de fallos por operarios, con lo que la población considerada es mayor. Se pueden ver los datos de fallos y operarios en tabla 6.15.

	Nº de operarios	Nº de equipos	Nº de fallos
UAE	69	401	7426

Tabla 6.15 Tamaño de la población de fallos en UAEDIA

En primer lugar, hay dos variables cuya influencia ya se ha apreciado al analizar los gráficos, como son si el fallo es accidente o avería y el tipo de fallo concreto, ya que la duración es muy distinta si es una activación de pulsador de emergencia o un problema en la caja de engranajes. Por tanto, estas dos variables deben ser tenidas en cuenta como efectos fijos en el análisis ANOVA, ya que se conoce de antemano que cada uno de sus subgrupos tendrá un efecto relevante en la media.

Así pues, se obtendrá un modelo ANOVA mixto, empleando como efectos fijos la variable accidente y el fallo y como efectos aleatorios el identicador del operario. Se muestra el resultado en la figura- 6.65.

```
ANOVA-Type Estimation of Mixed Model:
-------------------------------------

        [Fixed Effects]

                                int                       accidentaccidente                       accidentaveria
                          35.592059                               -3.953592                             0.000000
        FaultAreaName_en2Balustrade                FaultAreaName_en2Brake                FaultAreaName_en2Chains
                          14.985288                               11.850095                             7.421891
     FaultAreaName_en2Circuit breaker      FaultAreaName_en2Combplate switch      FaultAreaName_en2Emergency stop
                           3.081368                               -5.635438                           -12.907187
          FaultAreaName_en2Fire alarm     FaultAreaName_en2Floor plate switch                FaultAreaName_en2Gears
                          -9.029364                                0.988464                             8.848271
              FaultAreaName_en2Guides               FaultAreaName_en2Handrail          FaultAreaName_en2Handrail inlet
                         -10.398017                                5.485922                            -5.528163
     FaultAreaName_en2Handrail speed               FaultAreaName_en2Keyswitch                FaultAreaName_en2Lighting
                          -5.027944                               -7.680410                            10.523654
        FaultAreaName_en2Lubrication FaultAreaName_en2Main chain safety device       FaultAreaName_en2Miss step sensor
                           4.928366                               10.906251                             9.774586
         FaultAreaName_en2Motor drive    FaultAreaName_en2Motor over/underspeed  FaultAreaName_en2Other safety switches
                          -3.290346                                5.986598                            -6.166366
              FaultAreaName_en2Others        FaultAreaName_en2PLC and inverter          FaultAreaName_en2Power supply
                         -13.275302                               -3.581347                            -7.475998
       FaultAreaName_en2Refrigeration         FaultAreaName_en2Relay and wires                FaultAreaName_en2Reset
                          25.011745                                5.481023                            -7.462625
        FaultAreaName_en2Skirt switch                FaultAreaName_en2Step                 FaultAreaName_en2Step bolt
                          -4.823989                                9.884239                             7.225885
          FaultAreaName_en2Step chain      FaultAreaName_en2Step demarcation          FaultAreaName_en2Step roller
                          14.621847                               18.931062                            31.446946
       FaultAreaName_en2Step upthrust           FaultAreaName_en2Water leakage
                          -4.570925                                0.000000


        [Variance Components]

  Name                 DF         SS              MS          VC          %Total    SD        CV[%]
1 total                1031.051495                            243.496758  100       15.604383 66.944898
2 PrimaryEmployeeNumber 65        391372.61353    6021.117131 56.661974   23.270115 7.527415  32.293621
3 error                6990       1305975.137055 186.834784  186.834784  76.729885 13.668752 58.640782
```

Figura- 6.65 Modelo ANOVA mixto para ver la influencia de la pericia del operario

Se puede apreciar que se toma como dato de comparación de los efectos fijos el caso de avería y fallo por agua. Se observa que la diferencia entre accidente y avería no es muy elevada en el caso de UAEDIA —cerca de 5 minutos— mientras que sí hay diferencias sustanciales en función del fallo, pudiendo llegar a más de media hora si se compara el fallo de activación del pulsador de emergencia frente al de rodillo de peldaños, demarcación de peldaños o cadena de peldaños. Así pues, parece confirmarse que la variable que más condiciona la duración de la reparación es el elemento donde se produce el fallo.

Por otro lado, si se analiza la parte inferior de figura- 6.65 se puede ver como el operario que soluciona el incidente explica alrededor de un tercio de la variabilidad de la duración de la reparación en cada subgrupo, siendo los dos tercios restantes debidos a efectos aleatorios no tenidos en cuenta en este modelo. Así pues, la pericia del operario parece bastante importante y condiciona en parte cuánto se tarda en resolver un fallo. Por este motivo, sería recomendable garantizar a los técnicos una formación adecuada con un reciclaje continuo y darles unas buenas condiciones laborales. Se muestra el código para llegar a este resultado en la figura- 6.66.

```
DD=D
DD=DD[DD$X.U.FEFF.CountryName_en=="UAEDIA",]
modelo=(anovaMM(Duration2~accident+FaultAreaName_en2+
(PrimaryEmployeeNumber),DD[DD$Duration2<100,]))
```

Figura- 6.66 Código para análisis ANOVA

7.- Conclusiones

Se ha llevado a cabo el análisis de los informes de mantenimiento correctivo de las escaleras mecánicas y pasillos rodantes mantenidos por Thyssenkrupp en el área de Asia-Pacífico en el periodo que va de noviembre de 2017 a noviembre de 2018. Se han aplicado distintas técnicas estadísticas con una doble finalidad: por un lado, dar una forma adecuada a la base de datos y por otro hallar las relaciones y dependencias existentes entre las distintas variables que la componen. De dicho trabajo se desprenden las siguientes conclusiones:

- Se ha elaborado un procedimiento de acondicionamiento de los datos extensible a informes procedentes de otros países o periodo temporal.
- Se ha desarrollado, mediante el uso de técnicas de procesamiento del lenguaje natural, un algoritmo de clasificación del área de fallo, a partir de la descripción del incidente y su reparación previamente traducida por el procedimiento de acondicionamiento, con una tasa estimada de acierto superior al 90%. Además, se ha creado un criterio único para definir distintas características del incidente como si es un accidente o avería.
- Se han construido modelos de redes neuronales y árboles de clasificación con un acierto cercano al 75% para distinguir entre avería y accidente a partir de características del equipo y del incidente que se pueden conocer antes de mandar a un operario a repararlo. En la predicción del subsistema del fallo, el acierto se estima en un 40%. Modelos más precisos que partan de esta base e incluyan datos de sensores instalados en el equipo podrían ayudar a establecer una prioridad entre incidentes o ayudar al operario a decidir qué repuestos llevar o qué parte inspeccionar primero de un equipo, lo que permitirá disminuir el tiempo de reparación y mejorar el servicio que se presta al cliente.

Respecto a las tendencias detectadas en la población de equipos estudiados se pueden extraer las siguientes conclusiones:

- Los equipos de Thyssenkrupp siguen una función de supervivencia cercana a la exponencial, con un parámetro λ=0,0117, lo que equivale a una avería cada 85,3 días. Además, cada equipo se pasa de mediana cerca de 4 horas al año parado por incidentes.
- Los fallos más comunes están relacionados con causas externas al equipo o con un uso inadecuado del mismo, siendo los más habituales la seguridad de placa de peines y la activación del pulsador de emergencia.

- Se ha detectado un aumento en el número de fallos relacionados con los frenos en los equipos fabricados desde el año 2013, produciéndose en este elemento un 20% de las averías en los equipos de 2017. **CONFIDENCIAL.**
- **CONFIDENCIAL.**
- Se observa un alto número de averías relacionadas con la banda de peldaños en los equipos de Mitsubishi en el metro de Dubái. Esto no es extensible al resto de equipos Mitsubishi por lo que se achaca a defectos particulares de esas unidades. También es destacable el número de averías electrónicas de las escaleras de Fujitec en el mismo metro.
- El equipo más habitual mantenido por Thyssenkrupp en Asía-Pacífico se corresponde con una escalera mecánica Thyssenkrupp modelo Tugela, con un desnivel de 4 a 6 metros, un ancho de peldaño de 1000 mm, una velocidad de 0,5 m/s y una inclinación de 30°, ubicada en una estación de metro o tren China.
- El país en el que estén ubicadas las escaleras tiene una influencia capital en el tiempo de reacción al incidente y en la variación mensual del número de fallos. En cambio, el segmento en el que se emplee el equipo afecta tanto a la variación horaria y semanal como al perfil de uso y al modelo de escalera recomendado.
- Se han detectado al menos 4 edificios (Aeropuertos de El Cairo, Doha y Dubái y Metro de El Cairo) cuyo mantenimiento corre a cargo de delegaciones exclusivas, es decir, operarios siempre presentes en dicho emplazamiento para solucionar cualquier incidente. Estos edificios se caracterizan por una alta tasa de accidentes, los cuales se solucionan en poco tiempo en comparación con las averías, y un tiempo de reacción mínimo.
- Los países —sin tener en cuenta aquellos con delegaciones exclusivas— que menos tardan en responder y solucionar los incidentes son China, Taiwan y Corea, mientras Singapur, Indonesia, India y UAE presentan tiempos de respuesta y reparación mucho mayores.
- El tiempo medio necesario para solucionar un incidente depende principalmente del elemento afectado y el subsistema al que pertenecen. La pericia del operario representa un tercio de la variabilidad de la duración de la reparación.
- Los fallos a los que se dedica anualmente más tiempo son los de seguridad de placa de peines y los de rodillo de peldaño.
- Los fallos más graves —refiriendose con esto a que requieren un mayor tiempo de reparación— se circunscriben al sistema motriz del equipo, en concreto a las guías, engranajes y cadenas, que requieren más de hora y media de mediana para ser reparados.
- Estos tres fallos —engranajes, cadenas y guías— son también los que más fácilmente se pueden detectar antes de que sucedan y, al menos en el caso de cadenas, es

también uno de los fallos en los que se requieren varias visitas para completar la reparación al no contar con el repuesto adecuado en el momento.

- Los equipos situados a la intemperie tienen el triple de riesgo de sufrir algún fallo debido a fenómenos meteorológicos que aquellos situados en espacios cubiertos.
- Los equipos en los que se registran un mayor número de incidentes relacionados con caídas son los de 30° de inclinación, se cree que debido a su uso en metros donde la hora punta y las prisas pueden llevar a un mayor número de esto tipo de incidentes.
- A mayor tiempo de viaje en el equipo mayor accidentalidad se registra, probablemente por despistes de los pasajeros durante su uso.
- No hay diferencia en la tasa de averías y la gravedad de los incidentes entre pasillos rodantes y escaleras mecánicas, ni entre los distintos modelos de Thyssenkrupp, siempre que se use cada uno para la función para la que fue diseñado. En cambio, se ha comprobado que emplear modelos comerciales para aplicaciones de gran carga se relaciona con un mayor número de averías y de mayor gravedad.
- Los equipos cuyo controlador es un PLC necesitan menos tiempo para ser reparados que aquellos con un microprocesador o control por relés, probablemente por las capacidades de diagnostico del PLC que facilitan la tarea de reparación al operario.
- El resto de variaciones mecánicas, como el número de motores o la presencia de perfil bajo el pasamanos, no parecen tener influencia en los incidentes registrados. La excepción es el material de la balaustrada, registrando mayor número de fallos las de cristal frente a las metálicas.
- **CONFIDENCIAL.**
- la antigüedad del equipo no parecen tener ninguna influencia en el número y gravedad de las averías registradas.

8.- Planificación

En este apartado se detalla la programación temporal de las tareas del proyecto, con la finalidad de describir cómo se ha llevado a cabo la realización del mismo. Para ello, el trabajo se divide en las tareas que se indican en la tabla 8.1.

Tarea	Comienzo planificado	Fin planificado	Comienzo real	Fin real	Variación de duración
Diseño de etapas de proyecto	lun 28/01/19	mar 29/01/19	lun 28/01/19	mar 29/01/19	0 días
Estudio preeliminar de base de datos	lun 28/01/19	mié 30/01/19	lun 28/01/19	mié 30/01/19	0 días
Recopilación y estudio de documentación	mar 29/01/19	lun 18/02/19	mar 29/01/19	lun 25/02/19	5 días
Aprendizaje de Power BI	mar 05/02/19	lun 11/02/19	mar 05/02/19	jue 07/02/19	-2 días
Acondicionamiento de base de datos	mar 05/02/19	lun 25/02/19	mar 05/02/19	vie 01/03/19	4 días
Traducción	lun 25/02/19	mié 27/02/19	lun 04/03/19	lun 18/03/19	8 días
Aprendizaje de R	lun 25/02/19	mar 05/03/19	vie 08/02/19	lun 25/02/19	5 días
Creación de modelos de clasificación	vie 01/03/19	lun 11/03/19	lun 04/03/19	vie 22/03/19	8 días
Estudio completo de base de datos	lun 11/03/19	vie 29/03/19	lun 11/03/19	vie 29/03/19	0 días
Creación de modelos de mantenimiento predictivo	vie 29/03/19	jue 11/04/19	vie 29/03/19	jue 11/04/19	0 días
Documentación	lun 28/01/19	vie 07/06/19	lun 28/01/19	vie 07/06/19	0 días
Comprobación y revisión	vie 07/06/19	mié 12/06/19	vie 07/06/19	mié 12/06/19	0 días

Tabla 8.1 Tareas del proyecto

Por tanto, este trabajo se desarrollará a lo largo de 12 tareas desde el 28/1/19 al 12/6/19 con una duración aproximada de 100 días a lo largo de 6 meses.

Se puede apreciar como no se ha cumplido estrictamente la planificación inicial. Esto se debe principalmente a dos hechos:

En primer lugar el trabajo estaba configurado para manejar la base de datos en sus etapas iniciales con Power BI, pero tras un par de días de manejo de esta herramienta se vio que no se adaptaba adecuadamente a este proyecto. Esto llevo a un menor tiempo de aprendizaje de este programa y un adelanto y mayor tiempo de dedicación para el aprendizaje de R.

Por otro lado, como ya se habia comentado previamente, la traducción se realizó empleando la versión gratiuta de *Google Translate*, lo que conllevó demoras y tiempos de espera que alargaron mucho más de lo esperado esta tarea, con el consiguiente retraso de los trabajos que dependían de esta, como la creación de modelos de clasificación y el acondicionamiento de la base de datos.

De todas formas, estas pequeñas variaciones en la planificación inicial no afectaron finalmente a los tiempos de entrega, pudiéndose cumplir los plazos establecidos.

Finalmente, para la representación temporal de las tareas del proyecto se empleará un diagrama de Gantt, ya que es una técnica muy extendida por su sencillez y facilidad de control, se puede observar en la figura- 8.1. En él, la línea azul es la planificación inicial y la negra la real y las tareas siguen el mismo orden vertical que en la tabla 8.1.

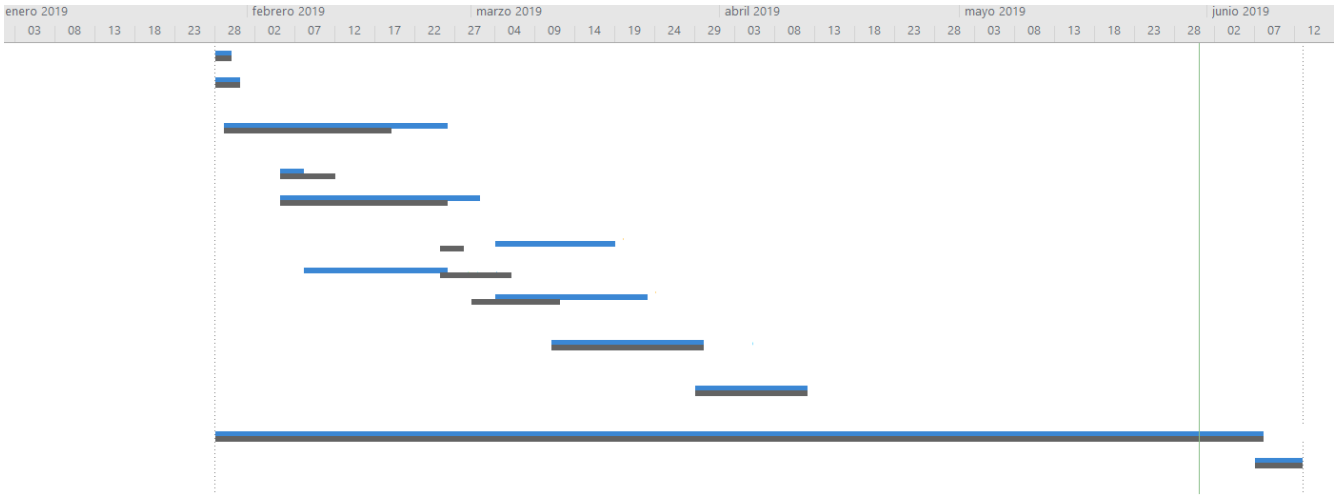


Figura- 8.1 Diagrama de Gantt del proyecto

9.- Presupuesto

9.1.- COSTES PARCIALES DEL PROYECTO

Para facilitar la elaboración del presupuesto, los costes han sido divididos en tres apartados que se detallan a continuación: equipos informáticos y software, mano de obra y otros conceptos.

9.1.1.- Coste de equipos informáticos y software

Este apartado incluye tanto los costes de adquisición de los dispositivos informáticos empleados como las licencias de los distintos programas informáticos usados. Dado que estos equipos y programas se seguirán empleando en futuros trabajos, solo se computará el coste proporcional a la duración del proyecto. Esta se estima en unos 6 meses. Se empleará amortización constante de los equipos a lo largo de su vida útil.

CONCEPTO	COSTE TOTAL (€)	TIEMPO DE USO (años)	TIEMPO DE AMORTIZACIÓN (años)	IMPORTE TOTAL (€)
Ordenador portatil	900	0.5	3	150
Raspberry PI	60	0.5	3	10
Windows 10 Professional	70	0.5	3	12
Microsoft Office 2016 Professional	255	0.5	3	43
COSTE TOTAL (€)				215

Tabla 9.1 Coste de equipos informáticos y software

9.1.2.- Coste de mano de obra

En este apartado se incluye las horas de estudio y elaboración de la documentación, así como las horas de programación.

CONCEPTO	HORAS DE TRABAJO(h)	COSTE HORARIO (€/h)	COSTE TOTAL (€)
Recopilación de información	80	30	2.400
Programación	250	30	7.500
Redacción de la documentación	200	30	6.000
COSTE TOTAL (€)			15.900

Tabla 9.2 Coste de mano de obra

9.1.3.- Otros conceptos

En este apartado se incluyen todos aquellos costes no considerados en apartados anteriores. No se incluye el precio de la consulta de bibliografía y otro material técnico debido a que se han usado las bases de datos de las que dispone la Universidad de Oviedo (AENORmás y ScienceDirect).

CONCEPTO	IMPORTE (€)
Papelería y material de oficina	20
Impresión y encuadernación	30
Transporte	400
COSTE TOTAL	450

Tabla 9.3 Coste de otros conceptos

9.2.- COSTE TOTAL DEL PROYECTO

Se incluye en el presupuesto final la suma de los apartados anterior aumentado en:

- **Gastos generales:** Debidos a consumos eléctricos, gastos de personal administrativo y otros costes imprevistos que no se incluyen en la planificación. Se valoran en un 10% sobre el coste total bruto.
- **Beneficio industrial:** 12% sobre el coste total bruto
- **I.V.A.:** 21% sobre el coste total bruto.

CONCEPTO	COSTE (€)
Coste equipos informáticos y software	215
Coste de mano de obra	15.900
Costes de otros conceptos	450
Coste total bruto	16.565
Gastos generales (10%)	1.656,50
Beneficio industrial (12%)	1.987,80
Coste total sin impuestos	20.209,30
I.V.A (21%)	4.243,95
COSTE TOTAL DEL PROYECTO	24.453,25

Tabla 9.4 Presupuesto del proyecto de investigación

El presupuesto del proyecto asciende a un total de **VEINTICUATRO MIL CUATROCIENTOS CINCUENTA Y TRES EUROS CON VEINTICINCO CÉNTIMOS.**

RAÚL CABEZAS RODRÍGUEZ

GIJÓN, a 12/06/2019

10.- Trabajos futuros

Este TFM abre una nueva línea de trabajo dentro de Thyssenkrupp para aprovechar todo el potencial de los datos generados por la actividad propia de esta compañía. Entre los posibles desarrollos futuros para aumentar la cantidad de datos disponibles y conseguir un mejor uso de estos destacan:

1. Obtención de más datos y mejor organizados:

- Simplificar el formulario de recogida de datos de los informes de mantenimiento correctivo. Se ha comprobado que muchas variables no son realmente necesarias y otras se podrían sustituir gracias a las técnicas aquí desarrolladas.
- Emplear una estructura de base de datos con tablas relacionales. En principio sería necesaria construir una tabla con las variables de clientes, otra con las del operario, una tercera con las características del equipo y la última con los campos del incidente. Esto permitiría limitar la información que se le pide al operario a la de la tabla de incidente y los identificadores de cliente, operario y equipo. Esto debería reducir el número de campos incompletos por registro, ya que solo se exigirá al operario aquella información que conoce, obteniendo el resto por medio de otras fuentes como sistemas ERP u otros sistemas que se utilicen en cada delegación.
- Aumentar la base de datos con resultados de años previos o de otros países, lo que permitiría mejorar los modelos basados en *Machine Learning* al poder realizar un aprendizaje más completo.
- Ampliar la base de datos con los informes de mantenimiento preventivo, que podrían complementar la información de los registros de mantenimiento correctivo.
- Combinar esta base de datos con los datos de sensorización del proyecto MAX de Thyssenkrupp para tener más información disponible.

2. Mejora de modelos actuales:

- Utilizar la computación en la nube para extender el aprendizaje no supervisado a toda la base de datos en vez de a pequeñas muestras, garantizando así que se obtienen resultados similares al algoritmo de clasificación propio.
- Intentar hallar alguna correlación en el histórico de fallos de cada equipo, es decir, averiguar si existe alguna relación entre qué fallo ha sufrido un equipo y cuál es el próximo fallo que va a sufrir. Actualmente es difícil de desarrollar al tener solo datos de un año, pero con varios años de datos sería una vía a explorar con mucho potencial.
- Desarrollar modelos de predicción centrados en predecir cuándo se va a producir el próximo incidente, en vez de los modelos actuales que intentan predecir qué tipo de

incidente se producirá y cómo de grave es. Al igual que en el punto anterior, para poder desarrollar estos modelos se necesita contar con un histórico de datos con más años.

- Mejorar el algoritmo de clasificación automática de área de fallo, ampliando el número de lexemas considerado y obteniendo etiquetas que permitan llevar a cabo técnicas de aprendizaje supervisado.

3. Uso de la información en campo:

- Desarrollo de una aplicación en Rshiny o PowerBI para facilitar la consulta de resultados a las delegaciones territoriales, pudiendo personalizar los gráficos y tablas a cada país. Esto permitiría que cada país comparase su rendimiento cada cierto tiempo —trimestral o semestralmente— con sus resultados históricos y los de otros países y así ver si está cumpliendo los estándares de calidad necesarios. También se podrían crear indicadores KPI que permitieran ver de forma intuitiva qué aspectos son mejorables en una delegación en relación al resto.
- Emplear alguna herramienta de *Machine Learning* en la nube como Microsoft Azure para desarrollar modelos más complejos como redes neuronales con más capas y así ver si se consigue un mejor mantenimiento predictivo. Algunos de estos modelos podrían intentar predecir la duración de la reparación a partir de datos como el operario que la reparará, el edificio y las características del equipo o estimar cuál es el fallo más probable en un equipo en función de sus características y su patrón de uso. Además, al ser en la nube podrían funcionar en tiempo real y mostrar información al operario antes de que inicie la reparación o a la delegación sobre cuánto tiempo debería tardar el operario en reparar el fallo mejorando la organización y la toma de decisiones.

11.- Bibliografía

[1] «https://www.nvdo.nl,» [En línea]. Available: https://www.nvdo.nl/efnms/ . [Último acceso: 13 4 2019].

[2] «Macro Economics report M4C for the MRO market in Europe,» KPMG Advisory CVBA/SCRL, 2015.

[3] «https://www.maintworld.com,» [En línea]. Available: https://www.maintworld.com/EFNMS/Maintenance-A-Necessary-and-Important-Function-in-the-Future. [Último acceso: 13 4 2019].

[4] I. Alsyouf, «The role of maintenance in improving companies' productivity and profitability.,» *International Journal of Production Economics,* vol. 105, nº 10.1016/j.ijpe.2004.06.057. , pp. 70-78, 2007.

[5] L. Al-Sharif, «Asset Management of Public Service Escalators,» *Elevator World,* nº June, p. 98, 1999.

[6] «https://medium.com,» [En línea]. Available: https://medium.com/datos-y-ciencia/qu%C3%A9-diablos-es-ciencia-de-datos-f1c8c7add107 . [Último acceso: 14 4 2019].

[7] M. g. institute, «Big data: The next frontier for innovation, competition and productivity,» 2011.

[8] M. g. institute, «THE AGE OF ANALYTICS: COMPETING IN A DATA-DRIVEN WORLD,» 2016.

[9] E. &. s. affairs, «Wo r l d U r b a n i z at i o n Prospec t s : Th e 2 0 1 8 R e v i s i o n,» United Nations, 2018.

[10] E. &. S. affairs, «World population ageing,» ONU, 2017.

[11] «https://www.prnewswire.com,» [En línea]. Available: https://www.prnewswire.com/news-releases/elevator-industry-2016-2020-sales-volume-maintenance-forecast-research-reports-589306791.html. [Último acceso: 14 4

2019].

[12] Descripción técnica Velino Base, Catalogo de Ventas, Thyssenkrupp, 2005.

[13] M. B. M. Abella, «Mantenimiento Industrial,» Leganés, 2003.

[14] P. Peeling, «Big Data and machine learning for predictive maintenance,» 2017.

[15] D. Madigan, *Introduction to survival analysis,* University of Columbia, 2004.

[16] M. H. Andreas Kaplan, «Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence,» *Business Horizons,* vol. 62, nº January-February, pp. 15-25, 2019.

[17] H. J. E. Eduardo Morales, *Aprendizaje por refuerzo,* INAOE, 2019.

[18] Amazon AWS, «https://docs.aws.amazon.com,» [En línea]. Available: https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/when-to-use-machine-learning.html. [Último acceso: 25 4 2019].

[19] I. I. A. M. Pedro Larranaga, *Árboles de clasificación,* Departamento de Ciencias de la Computaci´on e Inteligencia Artificial, Universidad del Pa´ıs Vasco.

[20] A. T. Quevedo, Sistema de reconocimiento de caracteres manuscritos usando Redes Neuronales Convolucionales implementado en Python, Sevilla: Universidad de Sevilla, 2017.

[21] H. J. E. Eduardo Morales, *Clustering,* INAOE.

[22] D. Peña, Análisis de datos multivariantes, 2002, Mcgraw Hill.

[23] S. d. l. F. Fernández, *Análisis de correspondencia simple y múltiple,* Madrid: Universidad autónoma de Madrid, 2011.

[24] EliteDataScience, «https://elitedatascience.com,» [En línea]. Available: https://elitedatascience.com/data-cleaning. [Último acceso: 24 4 2019].

[25] P. Grabinski, «https://www.kdnuggets.com,» [En línea]. Available: https://www.kdnuggets.com/2018/12/feature-engineering-explained.html. [Último

acceso: 4 5 2019].

[26] T. Risueño, «https://blog.bitext.com,» [En línea]. Available: https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/. [Último acceso: 5 05 2019].

[27] J. Brownlee, «https://machinelearningmastery.com,» [En línea]. Available: https://machinelearningmastery.com/gentle-introduction-bag-words-model/. [Último acceso: 05 05 2019].

[28] A. Jimenez, *Cadenas de Markov en tiempo discreto,* Madrid: UPM.