Trabajo Fin de Máster

Técnicas estadísticas y herramientas de visualización aplicadas al Business Intelligence

María Díaz Cao

Máster en Técnicas Estadísticas

Curso 2022-2023

Propuesta de Trabajo Fin de Máster

Título en galego: Técnicas estatísticas e ferramentas de visualización aplicadas á intelixencia de negocio
Título en español: Técnicas estadísticas y herramientas de visualización aplicadas al Business Intelligence
English title: Statistical techniques and visualization tools applied to Business Intelligence
Modalidad: Modalidad B
Autor/a: María Díaz Cao, Universidade da Coruña
Director/a: Javier Tarrío Saavedra, Universidade da Coruña; Salvador Naya Fernández, Universidade da Coruña
Tutor/a: Manuel Domínguez Basteiro, Cofrico
Breve resumen del trabajo: Desarrollo de soluciones al problema planteado por la empresa de selección de fechas para un modelo de machine learning, utilizando el análisis de datos, y estudio y aplicación de gráficos de control no paramétricos.
Recomendaciones:
Otras observaciones:

Don/doña Javier Tarrío Saavedra, Titular de universidad de la Universidade da Coruña, don/doña Salvador Naya Fernández, Catedrático de universidad de la Universidade da Coruña, don/doña Manuel Domínguez Basteiro, Desarrollador I+D+i de Cofrico, informan que el Trabajo Fin de Máster titulado

Técnicas estadísticas y herramientas de visualización aplicadas al Business Intelligence

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

En A Coruña, a 5 de junio de 2023.

TARRIO SAAVEDRA JAVIER - DNI 53164914T

Digitally signed by TARRIO SAAVEDRA JAVIER - DNI 53164914T
DN: c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=CERTIFICADO ELECTRONICO DE EMPLEADO PUBLICO, ou=MATEMATICAS, ou=18373, serialNumber=IDCES-53164914T, sn=TARRIO SAAVEDRA, givenName=JAVIER, cn=TARRIO SAAVEDRA JAVIER - DNI 53164914T
Date: 2023.06.05 18:39:44 +02'00'

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

76353731X SALVADOR NAYA (R: Q6550005J)

Firmado digitalmente por 76353731X SALVADOR NAYA (R: Q6550005J)
Fecha: 2023.06.05 18:10:48 +02'00'

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

El/la tutor/a:
Don/doña Manuel Domínguez Basteiro

El/la autor/a:
Don/doña María Díaz Cao

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

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

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

Índice general

Resumen

Resumen

En la empresa Cofrico se lleva a cabo una detección y análisis de anomalías de las máquinas que intervienen en el proceso de refrigeración. Uno de los métodos que se utilizan para este fin combina técnicas de machine learning con los gráficos de control del Control Estadístico de la Calidad.

En relación con el programa de machine learning, a la empresa le interesaba buscar un método automático que seleccionara las fechas de entrenamiento y calibrado. En este trabajo se recogen las soluciones planteadas, junto con los resultados que aportaron.

Por otra parte, se consideraron también distintas opciones alternativas a los gráfcios de control actuales, estando entre ellas algunos gráficos de control no paramétricos. Se describe en el trabajo el estudio realizado sobre gráficos de control y la aplicación a datos reales de la empresa, analizando los resultados.

Abstract

The company Cofrico carries out a detection and analysis of anomalies in the machines involved in the refrigeration process. One of the methods used for this purpose combines machine learning techniques with Statistical Quality Control control charts.

Regarding the machine learning program, the company was interested in finding an automatic method that would pick the training and calibration dates. In this work the proposed solutions are collected, along with the results they provided.

On the other hand, different alternative options to the current control charts were also considered, including some non-parametric control charts. The study carried out on control charts and the application to real data of the company is described in the work, analyzing the results.

Capítulo 1

Introducción

Comenzamos el trabajo con una presentación de Cofrico, la empresa en la que fueron realizadas las prácticas. Durante las prácticas me incorporé al equipo de desarrollo de la empresa en Bergondo, formado por físicos, ingenieros, matemáticos y desarrolladores web, entre otros. Como se trata de una empresa de refrigeración, a continuación, describimos brevemente una instalación frigorífica, con los elementos que la conforman. Por úñtimo, exponemos las motivaciones del trabajo y los ámbitos en los que se centrará.

1.1. Sobre la empresa Cofrico

Cofrico es una empresa de refrigeración y climatización que nació en Burela en 1985. Sus líneas de negocio son la refrigeración industrial y comercial, la eficiencia energética y la climatización industrial y comercial, centrándose siempre en el ahorro y la eficiencia energética. Se encarga del diseño, proyecto, ejecución y mantenimiento de las instalaciones. Destaca desde su inicio la apuesta por la calidad y la innovación tecnológica, con lo que se convirtió en una de las principales empresas del país en este sector, con 10 sedes en España además de la central y más de 200 trabajadores, destacando dos patentes registradas en ahorro de energía.

Figura 1.1: Instalaciones de la sede de Cofrico en Bergondo

La diferenciación de la empresa se basa en aplicar la tecnología en la industria, mediante la refrigeración ecológica con CO_2 y el mantenimiento 4.0: mantenimiento preventivo y predictivo. Esto es un reflejo de sus valores: innovación, calidad del servicio, eficiencia energética y compromiso con el medio ambiente.

El compromiso con sus valores llevó a la empresa a invertir en I+D+i, configurando un departamento propio. De aquí surgió un sistema de monitorización, mantenimiento predictivo y automatización de las instalaciones, la plataforma Gradhoc. Se trata de un software que recoge todos los datos medidos en las instalaciones, mostrándolos para información de los clientes, y utilizándolos para optimizar las máquinas y para realizar un mantenimiento, preventivo y predictivo, de ellas. Con esto se consigue el ahorro en energía y la mejora en la eficiencia de las instalaciones, llegando a reducir consumos y costes hasta un 50 %.En ella se combinan tecnologías como IoT, gemelos digitales e inteligencia artificial. Es un software especializado en refrigeración comercial e industrial que gestiona las instalaciones.

1.1.1. Instalaciones frigoríficas

Como la empresa está centrada principalmente en instalaciones frigoríficas, comentamos brevemente la estructura de estas. Esto nos permitirá también conocer mejor los datos utilizados en el trabajo, que se corresponden con las variables de una de las máquinas.

El proceso de frío consiste principalmente en un ciclo de compresión que recoge el calor del ambiente a refrigerar para después expulsarlo a una zona que se encuentra a mayor temperatura gracias a un fluido térmico denominado refrigerante. En este proceso participan evaporadores, compresores, condensadores y válvulas de expansión.

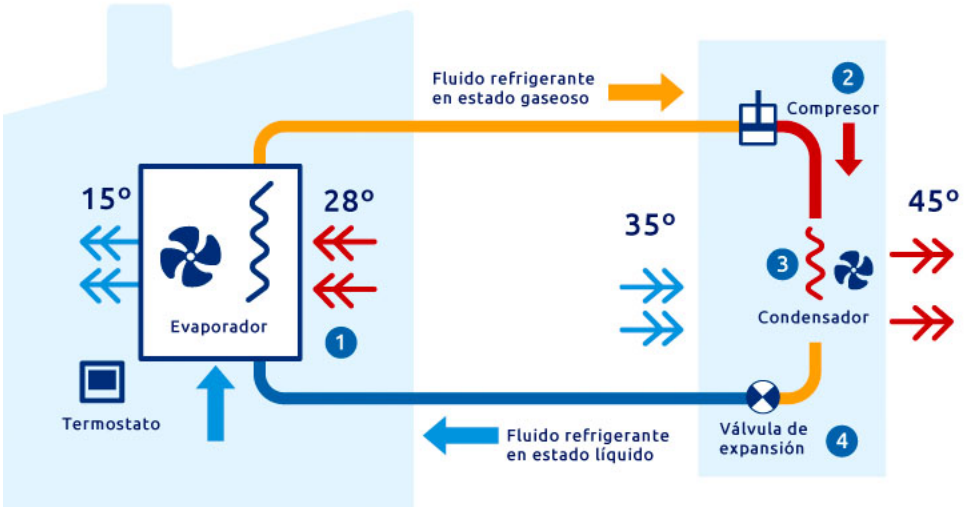


Figura 1.2: Ciclo de frío

El compresor es la parte que se encarga de generar la diferencia de presiones que se realizan durante el ciclo. Este elemento aspira el refrigerante en condiciones de baja presión y baja temperatura y lo comprime hasta las condiciones adecuadas para expulsar el calor en el condensador.

El evaporador es donde se realiza el intercambio de calor. Este se encarga de extraer calor de la estancia que se quiere refrigerar con lo que el aire ambiente se enfría mientras que el fluido térmico aumenta su temperatura.

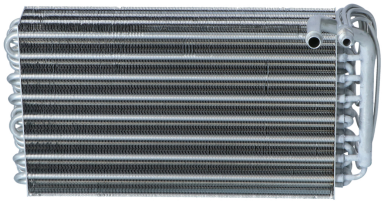
El condensador es el elemento encargado de condensar el vapor y evacuar el calor de condensación al exterior mediante un fluido de intercambio.

La válvula de expansión se encarga de acondicionar el refrigerante que sale del condensador, bajando su presión y su temperatura y adecuándolo para volver a extraer calor de la estancia que queremos refrigerar.

Todas estas máquinas están monitorizadas, y se registran valores de las temperaturas importantes en cada parte del ciclo, presiones, porcentaje de apertura de las válvulas y demás. Estas variables son

las que se analizan permitiendo realizar el mantenimiento preventivo y predictivo.

(a) Compresor



(b) Evaporador

(c) Condensador

(d) Válvula de expansión

Figura 1.3: Componentes de una instalación frigorífica

1.2. Planteamiento del trabajo.

Como comentamos en la sección anterior, la empresa dispone de un software para controlar las instalaciones.

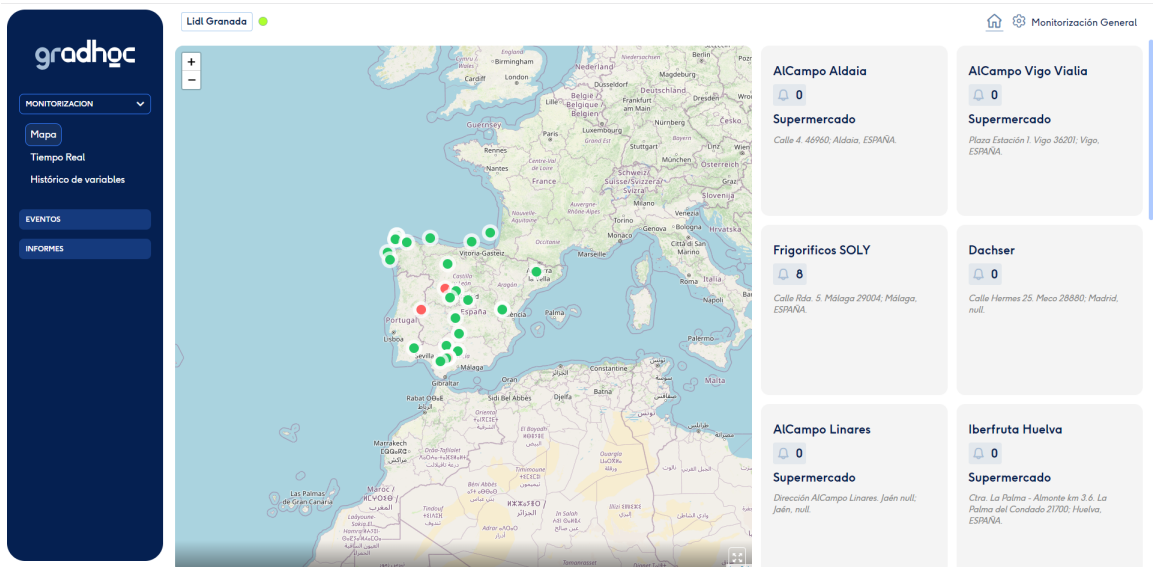


Figura 1.4: Aplicación Gradhoc

Dentro este se dispone de un programa de detección de anomalías. En él se combina un modelo de machine learning con el control estadístico de la calidad para detectar fallos. Esta idea surgió por parte del jefe del equipo hace unos años, y fue desarrollada durante las mismas prácticas del máster en Técnicas Estadísticas del curso pasado. Como resultado surgieron ideas a futuro de interés para la empresa, siendo la principal el entrenamiento automático del modelo.

Durante el inicio de las prácticas se trabajó en esta línea, planteando y desarrollando distintas soluciones. En el capítulo 2 se describe más detalladamente el tema, con las complicaciones asociadas, y luego las distintas soluciones y sus resultados.

Por otra parte, más en relación con el control estadístico de la calidad, está la idea planteada por los tutores. En su momento, además del programa de machine learning se creó un método de detección de fallos. En este se utilizan técnicas de control estadístico de la calidad, los gráficos de control, para estudiar las desviaciones de los datos. La propuesta fue probar otros gráficos de control no paramétricos, para lo cual se utilizó como referencia el trabajo de Chakraborti y Graham (2019) [5].

En el capítulo 3 introduciremos el control estadístico de la calidad y los gráficos de control paramétricos que se utilizaban y en el capítulo 4 describiremos los nuevos gráficos de control no paramétricos.

Por último, en el último capítulo, se aplican los gráficos de control descritos a datos reales de la empresa, medidos en sus máquinas de frío. También se comparará los resultados conseguidos con los gráficos anteriores frente a los nuevos.

1.3. Control estadístico de la calidad

Introducimos brevemente la rama del control estadístico de la calidad, que es la que se centra este trabajo. Damos una idea general de ella y de su importancia, y en el capítulo 3 se describe con mayor profundidad la metodología utilizada.

El control estadístico de la calidad es una rama que ganó gran interés en los últimos años. Una influencia importante para este su desarrollo fue su relación con la industria.

La cantidad de procesos industriales que se realizan aumentó en gran manera en las últimas décadas, pero esta forma de producción conlleva una mayor posibilidad de fallos. En un proceso intervienen máquinas, personal, materiales y más partes. Todo esto es susceptible de deteriorarse o desviarse de

su funcionamiento correcto, dando lugar a productos peores o defectuosos o a disminuir la producción. Estas son situaciones que conviene evitar, ya que empeorarían los resultados y supondrían un aumento de los costes.

En los últimos años las empresas realizaron un proceso de digitalización. Los fabricantes están incorporando nuevas tecnologías como el internet de las cosas (IoT), análisis y cloud computing, IA y machine learning en las instalaciones de producción y en todas sus operaciones. Esto llevó a la situación actual de la industria, la Industria 4.0 [6], que está cambiando y mejorando la forma en que las empresas producen sus productos. Como parte de este proceso de mejora se encuentran técnicas estadísticas que permiten controlar el comportamiento de las partes involucradas de un proceso industrial, de forma que ayudan a detectar las desviaciones frente al comportamiento habitual y prevenir los fallos. Esta detección de anomalías está siendo de gran interés.

En este trabajo nos centraremos en una de las herramientas más importantes del control estadístico de la calidad, que son los gráficos de control. Estos permiten detectar cuándo un proceso se está desviando de su comportamiento habitual. Primero se describen de forma teórica y a continuación los aplicaremos a los datos de la empresa.

Capítulo 2

Caso de estudio en la empresa Cofrico

2.1. Problemas planteados por la empresa

Como se mencionó en la introducción, la empresa utiliza un modelo de machine learning para ajustar y predecir los valores de variables medidas en evaporadores para luego controlar el funcionamiento de sus máquinas.

Cuando se dispone de un número grande de observaciones de las variables, el procedimiento habitual en machine learning consiste en dividir la muestra en muestra de entrenamiento y muestra de calibrado o test. Con el conjunto de datos de entrenamiento se construyen los modelos y con el conjunto de datos de calibrado se evalúa el rendimiento del modelo, que sirve para aproximar los errores que se cometerían con datos nuevos.

Tal y como están planteados los modelos, que se utilizan para luego usar gráficos de control con objeto de monitorizar el funcionamiento de los evaporadores, es importante que las muestras de entrenamiento y test se tomen cuando las máquinas están funcionando correctamente, no cuando hay alguna anomalía.

Esta selección de las muestras de entrenamiento y calibrado se realizaba manualmente, visualizando los datos y tomando las fechas más estables. En una instalación puede haber desde unos pocos evaporadores hasta más de doscientos, lo que hace que este procedimiento no sea el más adecuado, por el tiempo que requiere.

Surgió así el interés de buscar una forma de que la selección se realice de forma automática, lo que permitiría a la vez automatizar la puesta en funcionamiento del programa de machine learning en instalaciones nuevas, evitando la intervención humana.

Un problema presente en esta automatización es el de la idoneidad de la muestra. Como se comentó, los datos deben corresponderse con momentos de funcionamiento normal pero dada la naturaleza de los datos, puede resultar complicado. Como se describirá con más detalle en el capítulo 5, las variables presentan picos en los valores que se repiten varias veces al día. Es importante que esto no se detecte como desviación sino como el comportamiento adecuado.

Con esto el problema sería diseñar un método que permita, dado un histórico de datos, seleccionar rangos de fechas de buen funcionamiento. Describimos las soluciones que se pensaron, cómo se aplicaron y sus resultados.

2.2. Soluciones planteadas.

Durante el trabajo de un evaporador se distinguen dos etapas diferenciadas: los desescarches y el comportamiento estable. Los desescarches se realizan de forma periódica, varias veces al día y durante

estos las variables medidas sufren subidas elevadas en sus valores, como mostramos en la gráfica.

Esto dificulta seleccionar las fechas porque estos picos forman parte del comportamiento habitual, pero se detectarían como desviaciones, y es complicado distinguir si son picos correspondientes a desescarches adecuados o a un cambio en el comportamiento.

Ante esto se probaron distintos procedimientos. En primer lugar, se pensó en utilizar reglas físicas.

2.2.1. Reglas físicas.

Durante el proceso de frío, hay ciertas relaciones entre las variables que se deben cumplir y que indican que el proceso de frío se realiza correctamente. Las reglas consideradas son:

- La temperatura de impulsión debe ser menor que la temperatura de retorno, y menor también que la de evaporación.
- La temperatura de evaporación no debe ser mayor que la de retorno.
- La temperatura de fin de desescarche, si supera los cero grados en algún pico, lo debe superar en todos.

Utilizando datos de momentos de comportamiento adecuado de los evaporadores estas reglas físicas se cumplían muy poco. Se intentó descartar los desescarches, que permite eliminar las zonas de mayor variación de los datos, con lo que quizá sería más sencillo que se cumplieran, pero no se mejoró. Por último, se probó también a usar las medias por horas para ver si estas lo cumplían, de nuevo los resultados no fueron satisfactorios. Esta solución está muy condicionada por la colocación de las sondas que monitorizan las variables. Si estas están mal colocadas, puede que las temperaturas no se registren adecuadamente y no cumplan las reglas.

2.2.2. Comparación por grupos

Dentro del nuevo paradigma introducido por la industria 4.0, de la que hablamos en la introducción, es cada vez más frecuente el uso del gemelo digital como un modelo que reproduce la realidad de forma digital. La empresa dispone de otro software de gemelos digitales que permite agrupar los evaporadores teniendo en cuenta sus características. Esto se podría utilizar para entrenar todos los evaporadores clasificados en el mismo grupo comparando con uno que se entrenaría a mano. Cuando se quiere entrenar una nueva instalación, se obtienen gráficas de un representante para cada grupo. Se seleccionan las fechas de entrenamiento y calibrado para estos, y durante el proceso de entrenamiento y calibrado de la instalación, el programa entrena el resto de evaporadores de cada grupo.

Los pasos a seguir de este procedimiento para seleccionar las franjas de fechas son los siguientes:

1. - Se selecciona un representante para cada grupo, que se entrena y calibra.
2. - Con el modelo entrenado del representante se predice utilizando los datos de los otros evaporadores del grupo y se calculan los residuos.
3. - Con los residuos se buscan las fechas en las que son menores de forma continua, escogiendo cinco días para el entrenamiento y dos para el calibrado.
4. - Se entrena el modelo para cada evaporador del grupo con las fechas seleccionadas.

De esta forma se entrenarían todos los evaporadores del grupo, y realizándolo en todos los grupos, todos los de la instalación.

2.2.3. Comparación por características

Los evaporadores tienen un tipo de servicio diferente según estén en una central de positiva, donde los productos solo se refrigeran, o de negativa, donde se mantienen congelados. Además, tienen asociado un tipo de agrupación. Estas dos características permiten clasificar los evaporadores y agruparlos.

Con estos grupos, se plantea una situación similar a la anterior. Los pasos a seguir en este procedimiento serían los mismos que en el caso anterior, lo único que cambia son los grupos y que en este caso no habría que entrenar y calibrar. Un representante del grupo podría servir para seleccionar la fecha de los demás, pero se enfocó de forma un poco distinta. En un archivo json, se guardaron para cada combinación de tipo de servicio con tipo de agrupación, un representante, que está entrenado y calibrado. Cuando queremos entrenar en una nueva instalación, esto nos permite consultar los modelos entrenados y se evita la primera parte del método anterior. En el resto se procede igual. Se recupera el modelo entrenado de la base de datos, se predice y se analizan los residuos con los datos del nuevo evaporador, seleccionando las franjas donde son menores.

Capítulo 3

Gráficos de control paramétricos

Como propuesta por parte de los tutores, en relación con el control estadístico de la calidad, se planteó probar alguna alternativa a los gráficos de control utilizados para detectar los fallos o desviaciones en el comportamiento de las máquinas. Comenzamos en este capítulo introduciendo el control estadístico de la calidad, revisando la asignatura Control Estadístico de la calidad, Naya y Tarrío (2022) [1] y describiendo los gráficos utilizados actualmente. A continuación, en otro capítulo, describiremos las nuevas técnicas.

3.1. Introducción al control estadístico de la calidad

La calidad puede ser definida y entendida de diversas formas. Una forma de medir la calidad sería ver si el producto cumple ciertas características deseadas o ciertos requisitos, que pueden compararse con unos estándares. Esto sería, que es adecuado para su uso. Otra forma es definir la calidad fue dada por Montgomery (2009) [17] como la inversa de la variabilidad.

El control estadístico de la calidad consiste en la aplicación de diferentes técnicas estadísticas a distintos procesos, industriales, administrativos, comerciales... Con estas técnicas lo que se pretende es comprobar que el proceso de interés cumple los requisitos de calidad necesarios o exigidos y mejorar la calidad de los productos.

Este es un ámbito de gran interés actualmente, gracias al desarrollo de técnicas para registrar, guardar y procesar grandes volúmenes de datos. Esto permite registrar el comportamiento y las características del proceso, con lo que se tiene un conocimiento muy preciso sobre el funcionamiento de las máquinas o de las variables de interés de los procesos.

En todo proceso se producen variaciones entre los bienes producidos, resultado de la variabilidad presente en ellos. Lo que interesa es distinguir si las variaciones que se producen son las esperables en base a la variabilidad propia del proceso o si, por el contrario, se deben a un mal funcionamiento. El control estadístico busca identificar las causas de variación para luego minimizarlas.

Con esto, la mejora de la calidad consiste en la reducción de la variabilidad en los procesos y en los productos. Para asegurar la calidad de los productos, un proceso de producción debe ser estable y todo lo relativo a él, operadores, ingenieros, máquinas, personal de control y gestión, debe buscar una mejora en el rendimiento, reduciendo la variabilidad. En esta línea, el control estadístico de procesos es una herramienta para conseguirlo.

En esta idea de reducir la variabilidad de un proceso destaca la metodología seis sigma. Esta busca mejorar los procesos centrándose en reducir la variabilidad apoyándose en herramientas y métodos estadísticos. Para aplicarlo se llevan a cabo cinco etapas, que constituyen el ciclo DMAIC. Las herramientas principales del control estadístico de procesos son los gráficos de control y el análisis de capacidad de procesos.

Un concepto importante del control estadístico de la calidad son las causas asignables y no asig-

nables. En el proceso de producción intervienen una serie de variables, que se distinguen en dos tipos: las entradas controlables y las entradas incontrolables.

Todas son fuentes de variabilidad, la diferencia es que unas tienen causas asignables de variabilidad y por tanto producen efectos predecibles, mientras que las otras tendrían causas no asignables que aparecen con efectos combinados, no predecibles de antemano e inherentes a la incertidumbre del proceso productivo. Las entradas controlables serían por ejemplo la maquinara, los materiales o las condiciones de los trabajadores. Estas sí tendrían causas asignables, sobre las que podemos actuar para mejorarlas. Entradas no controlables podrían ser las condiciones meteorológicas en la construcción o problemas de la red eléctrica, sobre las que no tenemos capacidad de acción.

Esto nos lleva a uno de los conceptos más importantes, el de proceso bajo control. Un proceso se dice que está bajo control cuando no tiene causas asignables de variabilidad, únicamente intervendría la variabilidad intrínseca del proceso.

3.2. Gráficos de control.

Los gráficos de control son uno de los procedimientos principales y más simples del control estadístico de la calidad. Estos sirven para monitorizar un proceso, de forma que permite detectar desviaciones de su comportamiento habitual, distinguiendo si esta desviación se debe a una causa asignable conocida o si, por el contrario, se trata de una causa especial o extraordinaria.

Considerando una característica de control es necesario controlar tanto su media como su variabilidad para asegurar que se encuentra bajo control. Los gráficos de control representan la herramienta más importante en el análisis de las variaciones de los procesos de producción o servicios.

Existen distintos tipos de gráficos, pero todos tienen una estructura similar. En ellos se representa la variable que estamos monitorizando a lo largo del tiempo, uniendo los valores que toma mediante líneas. Esto permite identificar patrones que indican cambios en el rendimiento del proceso. El valor medio de la variable graficada se muestra con una línea horizontal, y se representan también otras dos rectas, el límite inferior y superior. Estos serían los límites de referencia para determinar que el proceso se está desviando y se corresponderían con las regiones de rechazo del contraste de hipótesis.

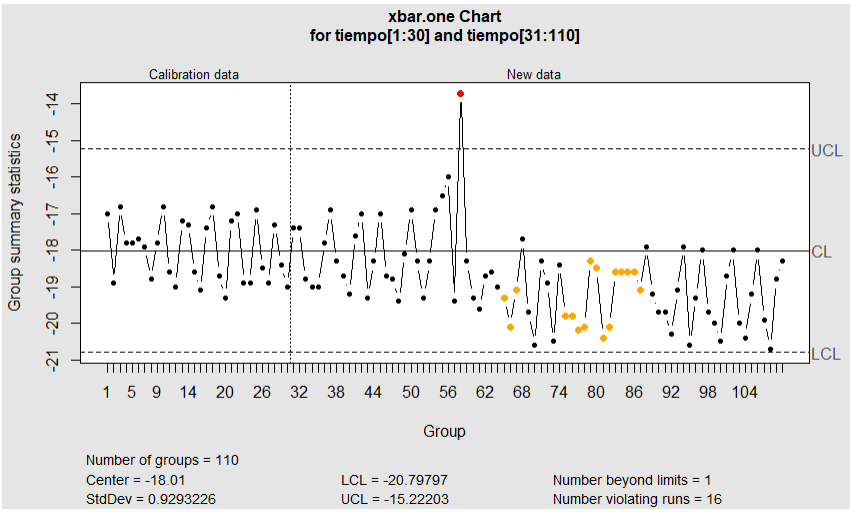


Figura 3.1: Gráfico de control.

Los gráficos de control se basan en un contraste de hipótesis, en el que se considera como hipótesis nula, H_0, que el proceso está bajo control, mientras que la hipótesis alternativa, H_1, es que el proceso se desvía del control. Entonces, para controlar que el proceso se mantiene bajo control, debemos realizar el contraste para cada muestra nueva que se vaya tomando de las variables.

Cuando se lleva a cabo un contraste se tienen en cuenta los dos errores que se pueden cometer, que son, el error tipo I, rechazar H_0 cuando es cierta, y el error de tipo II, no rechazar H_0 cuando es falsa. En el control estadístico de la calidad el error de tipo I se conoce como riesgo del vendedor y el error tipo II riesgo del comprador.

Los gráficos de control se conciben y diseñan diferente según el conocimiento que se tenga sobre la distribución del proceso bajo control. En función de esto, se distinguen dos fases en su construcción.

Cuando se dispone de un conocimiento completo de la distribución del proceso y de sus parámetros, los datos se tratan en la fase II, comenzando a monitorizarlos cuanto antes para detectar desviaciones.

Por otro lado, cuando no conocemos toda esta información, debe realizarse un análisis en la fase I para caracterizar la variación del proceso en condiciones normales y así estimar los límites de control adecuados para el monitorizado en la fase II.

3.3. Medidas para los resultados de los modelos

Cuando monitorizamos un proceso utilizando gráficos de control nos interesa estudiar los resultados aportados, para saber si está realizando bien el control. Para eso describimos a continuación dos medidas relativas a la detección de un fallo.

3.3.1. Longitud media de racha.

El average run length (ARL) o longitud media de racha es el número medio de muestras necesarias para detectar la primera muestra fuera de control.

$$\text{ARL} = \frac{1}{1-\beta} \tag{3.1}$$

Donde β es el riesgo del comprador. La probabilidad de detectar una desviación en la primera muestra considerada es $1-\beta$m en la segunda $\beta(1-\beta)$ y en la k-ésima muestra $\beta^{k-1}(1-\beta)$. Entonces, el número de muestras que sería necesario analizar antes de detectar el cambio sigue una distribución geométrica, cuya suma coincide con la expresión anterior;

$$\sum_{k=1}^{\infty} k\beta^{k-1}(1-\beta) = \frac{1}{1-\beta} \tag{3.2}$$

Este valor puede calcularse de forma analítica o mediante simulación. Es una medida importante a la hora de diseñar los gráficos de control porque, cuando el proceso está bajo control, el ARL debe ser alto, para evitar falsas alarmas, pero cuando el proceso está fuera de control debe ser bajo, para detectar cuanto antes la desviación y el mal funcionamiento.

3.3.2. Tiempo medio hasta la señal.

Otro concepto relacionado con los gráficos es el tiempo medio hasta la señal o ATS (Average Time to Signal). Se trata del tiempo promedio que le llevaría al gráfico señalar que el proceso está fuera de control. Su cálculo es sencillo, solo se necesita conocer el tiempo entre muestras, h, y el ARL. Resulta sencillo obtener una fórmula aproximada para el valor de ATS si conocemos el tiempo entre muestras, h:

$$ATS = ARL \cdot h \tag{3.3}$$

3.4. Control multivariante.

En un proceso de producción es común que haya varias características importantes para monitorizar y generar una alarma si estas se desplazan fuera de los límites operativos normales predefinidos. Podría

realizarse un control univariante de cada una de ellas, pero estaríamos desaprovechando parte de la información de los datos porque las variables de un proceso suelen estar bastante relacionadas entre sí. La rama del control de la calidad que se encarga del estudio de conjuntos de datos multidimensionales es el control multivariante.

En un análisis multivariante se estudian todas las características de forma simultánea. De esta forma, a parte de la información individual de cada variable, se tiene en cuenta también el efecto de las interacciones entre las variables, con lo que se utiliza también la información aportada a través de las covarianzas.

A parte de ignorar información contenida en los datos, cuando se realizan solo los controles univariantes se aumenta de forma considerable el error. Considerando dos variables, con sus correspondientes límites de control, la probabilidad conjunta de que las variables estén fuera de control se correspondería con la multiplicación de sus probabilidades individuales. Como se trata de número pequeños, se obtiene una probabilidad mucho menor de que la variable bidimensional esté fuera de control, aumentando el error de tipo I. De igual modo, disminuye la probabilidad conjunta de que la variable esté bajo control respecto a las individuales, lo que llevaría a una mala calidad del producto.

Con todo esto, los gráficos individuales supondrían un aumento en la detección de falsas alarmas, lo que sería un problema grave y haría necesarias herramientas para evitarlo. Una de las soluciones multivariantes más sencillas consiste en realizar una corrección en el nivel de significación. Esto supondría seguir usando gráficos univariantes con límites de control tipo Bonferroni, que corrigen la probabilidad de error de tipo I teniendo en cuenta el número de variables independientes. Se dividide la probabilidad de error de tipo I, α, entre el número de variables, con lo que se contrarresta el problema de las comparaciones múltiples.

Esto aportó una solución al problema, pero más interesantes son los métodos que utilizan la estructura de correlación de los datos y las técnicas de la estadística multivariante, llevando al control multivariante de procesos.

El desarrollo del control multivariante está muy ligado al de la estadística multivariante, apoyándose en técnicas como la profundidad de datos, el análisis de componentes principales, el análisis discriminante o el clúster. Veremos y describiremos métodos de control multivariante basados en algunas de ellas. El gran impulso para el desarrollo de estos métodos fueron los intereses de la industria, junto con las mejoras computacionales de los ordenadores.

Actualmente la investigación sobre técnicas de control multivariante está experimentando cierto cambio al dirigirse cada vez más a procedimientos gráficos, que permitan resumir, de forma visual, grandes cantidades de datos, así como hacia el estudio de modelos multivariantes no paramétricos o semiparamétricos, en los que muchas de las hipótesis sobre las distribuciones, impuestas por los modelos clásicos del control multivariante, se pueden relajar notablemente.

3.4.1. Datos multivariantes

Introducimos a continuación los conceptos básicos del contexto multidimensional. En este caso, se están controlando p características, teniendo así una variable aleatoria p-dimensional, $\vec{X} = (X_1, X_2, ..., X_p)$ compuesta por p variables aleatorias unidimensionales $X_1, X_2, ..., X_p$, que se corresponden con cada una de las características monitorizadas, que pueden ser continuas o discretas. Las observaciones pueden ser observaciones individuales o una combinación de ellas, como la media muestral.

Para esta variable aleatoria p-dimensional se define su vector de medias y la matriz de varianzas-covarianzas. La esperanza de la variable multidimensional se define como el vector compuesto por la media individual de cada variable y la matriz de varianzas covarianzas recoge las covarianzas de todos los pares de variables,

$$\mu = E(\vec{X}) = \begin{pmatrix} E(X_1) \\ \dots \\ E(X_p) \end{pmatrix} \tag{3.4}$$

$$\Sigma = Cov(\vec{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12}... & \sigma_{1p} \\ \sigma_{21} & \sigma_{22}... & \sigma_{2p} \\ \dots & ..... & \\ \sigma_{p1} & \sigma_{p2}... & \sigma_{pp} \end{pmatrix} \tag{3.5}$$

La matriz de varianzas-covarianzas es simétrica y semidefinida positiva.

Los gráficos de control multivariante para el vector de medias se diseñan para detectar cambios a lo largo del tiempo con respecto al vector de medias calculado con datos del proceso bajo control.

3.4.2. Estimación μ y Σ.

Mostramos la estimación de estos valores en el caso multidimensional general. Suponiendo que disponemos de m muestras, con n observaciones cada una de las p variables, siendo x_{ijk} la observación i-ésima de la variable j en la muestra k. En primer lugar, se calcula la media y la varianza muestral para cada muestra:

$$\begin{aligned} \bar{x}_{jk} &= \frac{1}{n}\sum_{i=1}^{n} x_{ijk}, \quad j = 1, ..., p; k = 1, 2, ..., m \\ s^2_{jk} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_{ijk} - \bar{x}_{jk})^2 \quad j = 1, ..., p; k = 1, 2, ..., m \end{aligned} \tag{3.6}$$

La covarianza entre dos variables j_1, j_2 de la submuestra k sería:

$$s^2_{j_1 j_2 k} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij_1 k} - \bar{x}_{jk})(x_{ij_2 k} - \bar{x}_{jk}) \quad j_1 \neq j_2; k = 1, 2, ..., m \tag{3.7}$$

Una vez calculados para cada subgrupo, se promedian en los m subgrupos obteniendo la estimación global para cada variable,

$$\begin{aligned} \bar{\bar{x}}_j &= \frac{1}{m}\sum_{k=1}^{m} \bar{x}_{jk}, \quad j = 1, ..., p, \\ \bar{s}^2_j &= \frac{1}{m}\sum_{k=1}^{m} s^2_{jk}, \quad j = 1, ..., p, \\ \bar{s}_{j_1 j_2} &= \frac{1}{m}\sum_{k=1}^{m} s^2_{j_1 j_2 k}, j_1 \neq j_2 \end{aligned} \tag{3.8}$$

Con esto las estimaciones obtenidas son:

$$\bar{\bar{x}} = \begin{pmatrix} \bar{\bar{x}}_1 \\ \dots \\ \bar{\bar{x}}_p \end{pmatrix} \tag{3.9}$$

$$S = \begin{pmatrix} \bar{s}^2_1 & \bar{s}_{12}... & \bar{s}_{1p} \\ \bar{s}_{21} & \bar{s}^2_2... & \bar{s}_{2p} \\ \dots & ..... & \\ \bar{s}_j & \bar{s}_j... & \bar{s}^2_p \end{pmatrix} \tag{3.10}$$

Cuando se trata con observaciones individuales se presenta un problema a la hora de estimar la matriz de covarianzas Σ. Se propusieron diversos estimadores, entre ellos el estimador habitual de esta matriz, que se obtiene agrupando todas las observaciones. Otra opción es utilizar las diferencias entre pares de observaciones sucesivos.

3.5. Gráficos de control EWMA y MEWMA

En esta sección comentaremos brevemente los gráficos de control que se están utilizando actualmente para la detección de anomalías en la empresa. Los describimos teóricamente y mostramos sus principales características.

3.5.1. EWMA

El gráfico de medias móviles con ponderación exponencial o gráfico EWMA es un gráfico de medidas individuales que fue propuesto por Roberts (1959) [19]. Su implementación es sencilla al igual que su interpretación.

Se basa en el estadístico:

$$Z_t = \lambda x_1 + (1-\lambda)Z_{t-1}, \quad 0 < \lambda < 1, \tag{3.11}$$

que es la media ponderada de la observación actual y todas las observaciones anteriores, dándole mayor peso a la última. Se establecen también los límites de control inferior y superior, en base a una muestra de calibrado, para monitorizar el proceso. El valor inicial Z_0, suele ser el valor objetivo. Las observaciones pueden ser valores individuales del proceso o medias muestrales calculadas de un diseño de muestreo. Se considera que el proceso está fuera de control cuando Z_i cae fuera de los límites de control, teniendo en cuenta una serie de criterios.

Cuando las observaciones x_i son independientes e idénticamente distribuidas (i.i.d.) con la misma varianza, σ_Y^2, la varianza del estadístico de control se corresponde con:

$$\sigma^2(Z_t) = [(1-(1-\lambda)^{2t})\lambda/(2-\lambda)]\sigma_Y^2 \tag{3.12}$$

A excepción de los casos en los que λ sea pequeño, el efecto del valor inicial desaparece pronto, y la varianza converge a su valor asintótico,

$$\sigma_Z^2 = \frac{\lambda}{2-\lambda}\sigma_Y^2. \tag{3.13}$$

Considerando $\lambda = 2/(t+1)$, la varianza de Z_t se aproxima por $Var(Z_t) \approx \sigma^2/t$, y los límites de control serían $\hat{\mu} \pm 3\sqrt{\frac{\hat{\sigma}^2}{t}}$.

Estos gráficos son muy efectivos para detectar pequeños cambios en la media, pero su comportamiento no es tan bueno cuando se trata de cambios de mayor tamaño.

3.5.2. MEWMA

Los gráficos MEWMA, gráficos de medias móviles ponderadas exponencialmente para el caso multivariante es una extensión del gráfico EWMA al caso multivariante. Estos fueron desarrollados por Lowry et al (1992) [15]. Para estudiar el comportamiento conjunto se asume que las X_i, $i = 1, 2, 3...$, son vectores aleatorios independientes que siguen una distribución normal multivariante con vector de medias μ_i, $i = 1, 2, ...$ respectivamente. La extensión natural al caso multivariante sería definir vectores de EWMA:

$$Z_i = \Lambda X_i + (I - \Lambda)Z_{i-1}, \tag{3.14}$$

siendo $Z_0 = 0$ y $\Lambda = diag(\lambda_1, ..., \lambda_p)$, $0 < \lambda_j \leq 1$, $j = 1, 2, ..., p$.

Para analizar la información aportada por estos Z_i y detectar con ello las desviaciones, se utiliza el estadístico:

$$T_i^2 = Z_i^t \Sigma_{Z_i}^{-1} Z_i. \tag{3.15}$$

Siendo $\Sigma_{Z_i}^{-1}$ la inversa de la matriz de varianzas-covarianzas de los Z_i. A partir de la matriz de varianzas-covarianzas, Σ de partida, se calcularía como:

$$\Sigma_{Z_i} = \frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]\Sigma. \tag{3.16}$$

En este caso, se tiene solo un límite de control, el superior, sobre el cual se proponen diferentes aproximaciones para calcularlo.

Cuando las hipótesis de independencia no se cumplen, tanto en el caso univariante como en el multivariante, las propiedades de la longitud media de racha (Average Run Length, ARL) de los gráficos se pueden ver afectadas en gran medida y las señales de proceso fuera de control podrían dejar de ser significantes, perdiendo toda la utilidad del gráfico. Ante esto, se desenvolvieron métodos basados en las series temporales múltiples para el caso de observaciones multivariantes autocorreladas como el de Valipour et al (2013) [18] o el de Reynols y Lu (1997) [22]. También se desarrollaron gráficos de control no paramétricos para cuando no se puede asumir una distribución para los datos. En el siguiente capítulo nos centramos en estos gráficos no paramétricos, introduciendo las ideas en las que se basan y comentando algunos interesantes y de utilidad para nuestros datos.

3.6. T^2 de Hotelling

Describimos en esta sección otro gráfico de control paramétrico, que se basa en el estadístico T^2, desarrollados ambos por Hotelling (1947) [11] y que permiten el estudio de variables multidimensionales en las que hay relación entre las variables.

Dada una distribución p-dimensional, con vector de medias μ y matriz de covarianzas Σ para este gráfico se utiliza el estadístico

$$\chi_0^2 = n(\bar{x}-\mu)^t\Sigma^{-1}(\bar{x}-\mu), \tag{3.17}$$

que se corresponde con la distancia de Mahalanobis entre el vector que de la media muestral y el vector de la media poblacional y este tendría como límite superior de control $UCL = \chi^2_{\alpha,p}$

Como el vector de medias y la matriz de covarianzas suelen ser desconocidos, se sustituyen por las estimaciones S y $\bar{\bar{x}}$, con lo que la estimación del estadístico sería:

$$T^2 = n(\bar{x}-\bar{\bar{x}})^t S^{-1}(\bar{x}-\bar{\bar{x}}), \tag{3.18}$$

el gráfico que se obtiene graficando este estadístico es el T^2 de Hotelling. Se trata de un gráfico de control que no depende de la dirección del fallo, la detección solo depende de la magnitud de la desviación. El estadístico sigue una distribución F de Snedecor con p y $(mn-m-p+1)$ grados de libertad.

Como límites de control se utilizan:

$$\begin{aligned} UCL =& \frac{p(m-1)(n-1)}{mn-m-p+1}F_{\alpha,p,mn-m-p+1} \\ LCL =& 0 \end{aligned} \tag{3.19}$$

en caso de utilizarlo en la fase I, y cuando se pasa al monitorizado de la fase II el límite de control adecuado sería:

$$\begin{aligned} UCL =& \frac{p(m+1)(m-1)}{m^2-mp}F_{\alpha,p,m-p} \\ LCL =& 0 \end{aligned} \tag{3.20}$$

3.7. Evaluación de los gráficos

Cuando se aplican los gráficos de control es interesante evaluar los resultados que aportan, para saber si estos son fiables. Entre los métodos para analizar la evaluación de gráficos de control están las versiones muestrales del ARL y ATS, descritos antes. Por otra parte, tratando los gráficos como si fueran un método de clasificación, se puede evaluar su comportamiento mediante la matriz de confusión y medidas como la especificidad, sensibilidad y precisión.

3.7.1. Métodos de clasificación

Para evaluar los gráficos se estudiarán como si fuese un método de clasificación, para lo cual nos apoyamos en los apuntes de la asignatura Aprendizaje Estadístico, del MTE [8]. Cuando se trabaja con un modelo de clasificación, lo habitual es calcular la matriz de confusión de los datos de test. En general se consideran dos categorías, positivo y negativo, y se genera una tabla de contingencia de los valores reales frente a lo estimado por el modelo. En nuestro caso, estas dos categorías serían, como positivo que haya desviación, y como negativo que esté bajo control. Se denotan por verdaderos positivos (TP) las observaciones que se detectaron como fuera de control, y en efecto la máquina se había desviado; falsos positivos (FP) se detecta desviación, pero no la había. Serían verdaderos negativos (TN) observaciones bajo control en las que no se detectó desviación, y falsos negativos (FN), los que se estiman bajo control habiendo desviación. Con esto se construye la siguiente tabla, que es la denominada matriz de confusión,

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	Verdaderos positivos	Falsos negativos
Bajo control	Falsos positivos	Verdaderos negativos

Cuadro 3.1: Matriz de confusión de un método de clasificación.

Con los valores recogidos en la tabla se pueden calcular otras medidas generales de la calidad de la estimación de las nuevas observaciones.

Sensibilidad

Miden la proporción de positivos que fueron bien detectados respectivamente.

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

Especificidad.

Análogo a la anterior, pero en este caso para los negativos.

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

Precisión global.

Miden la proporción de positivos y negativos que fueron bien detectados respectivamente.

$$ACC = \frac{TP + TN}{TP + FN + FP + FN}$$

Precisión balanceada La precisión global es una medida que puede presentar problemas cuando las clases no están balanceadas. Cuando nos encontramos en este caso existen otras medidas de la

precisión global que buscan evitar este problema, como la precisión balanceada (balanced accuracy, BA):

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

3.7.2. ARL y ATS muestral.

Antes describimos el ARL desde un punto de vista teórico, pero se utilizará otro enfoque basándose en los datos, que ayudará también a evaluar la bondad de los gráficos.

En nuestro caso, como conocemos los datos, sabemos cuándo comenzó a fallar el evaporador. Esto nos permite calcular un ARL muestral, que consiste en contar el número de observaciones que fueron necesarias, desde que se empezó la desviación, para que el gráfico comenzara a detectarla.

Cuanto antes se detecten las desviaciones, antes se permite estudiar, tratando de solucionarlo de la forma más rápida. Esto puede hacer que fallos que supondrían problema grave si no se corrigen, puedan no ir a más.

De igual modo, utilizando el ARL muestral calculado, se calculará el ATS, teniendo en cuenta que los datos están tomados cada 5 minutos. ($h = 5$).

Capítulo 4

Gráficos de control no paramétricos.

Como se comentó en el anterior capítulo, los gráficos paramétricos pueden resultar muy restrictivos por las hipótesis que deben asumir sobre la distribución de los datos, y el incumplimiento de estas hipótesis da lugar a mal comportamiento de los gráficos y a la pérdida de su utilidad.

Los gráficos de control expuestos hasta ahora solo son válidos cuando la variable sigue una distribución normal y no tienen en cuenta la correlación que puede haber entre variables ni que la aproximación de Bonferroni sobreestima la probabilidad de error de tipo I.

La rama de los gráficos de control no paramétricos ha crecido notablemente en los últimos años. Estas técnicas se basan en métodos no paramétricos, o de distribución libre, lo que supone una alternativa más robusta cuando trabajamos con datos de los que desconocemos su distribución. Cuando asumimos hipótesis como la normalidad, las propiedades bajo control solo se cumplirán bajo las hipótesis de distribución aceptadas. Si nos desviamos de estas, puede afectar empeorando los resultados de los gráficos. En el caso de los gráficos no paramétricos las hipótesis que se piden son más sencillas, como simetría o continuidad de los datos.

Los gráficos de control no paramétricos son de especial interés en la fase I. Por lo general, en ella nuestro conocimiento sobre los datos y su distribución es escaso. Estos gráficos tienen la ventaja de que no se asume ninguna distribución para los datos. En la fase II podríamos, una vez lista la fase 1, utilizarla para buscar una distribución paramétrica apropiada. Aún así, presenta inconvenientes, por lo que parece más recomendable continuar también en ella con los no paramétricos.

4.1. Gráficos de análisis para la fase I

Los gráficos de control de la fase I han aumentado su interés en los últimos años ya que, una selección errónea de los límites de control empeora de forma considerable el comportamiento de los gráficos. Con ellos se busca comprobar si las observaciones de una o varias características de la calidad provienen todas de una distribución bajo control o, por el contrario, de una distribución en la que los parámetros han cambiado.

Los métodos de distribución libre en esta fase han ganado importancia debido a los inconvenientes asociados a no cumplirse las hipótesis de distribución asumidas. Esto empeora el comportamiento y la sensibilidad, llevando a un aumento de la probabilidad de considerar un proceso como inestable estando estable.

Teniendo en cuenta la importancia de esta fase I, se han desarrollado otros métodos que realizan un análisis previo. Con ellos lo que se quiere es buscar zonas de cambio en los datos, tanto en localización como en escala.

4.1.1. Gráfico de segmentaciones recursivas y permutaciones (RSP)

Se trata de un procedimiento de distribución libre para observaciones univariantes propuesto por Capizzi, Masarotto (2013) [3]. El método utiliza segmentaciones recursivas y permutaciones para detectar cambios, tanto únicos como múltiples, en la media o en la escala de la variable.

Para detectar los cambios en la media, se realiza el contraste de hipótesis de que el proceso está bajo control frente a que la media del proceso sufrió cierto número de cambios aislados o de paso. El número de cambios de paso que se desea detectar es un parámetro elegido por el usuario, pero se recomienda utilizar

$$K = \text{máx}\left(3, \text{mín}\left(50, \left[\frac{m}{15}\right]\right)\right),$$

El estadístico de control para detectar los cambios aislados sería:

$$T_0 = \max_{i=1,\dots,m} |\bar{x}_i - \bar{\bar{x}}|,$$

siendo $\bar{x}_i$ la media del i-ésimo subgrupo y $\bar{\bar{x}}$ la media global de todas las observaciones.

Cuando utilizamos métodos no paramétricos, detectar cambios aislados solo es posible cuando tenemos los datos en subgrupos. Un valor atípico aislado en una secuencia de observaciones individuales no se puede detectar sin información adicional sobre la distribución. En nuestro caso, esta parte se omitiría.

Nos centramos por tanto en la detección de variaciones de paso en la media. Para esto el contraste de hipótesis asociado sería:

$$H_0 : \text{el proceso está bajo control}$$

$$H_{1,k} : E(x_{ij}) = \begin{cases} \mu_0 & 0 < i \leq t_1, \\ \mu_0 & t_1 < i \leq t_2, \\ \dots & \\ \mu_0 & t_k < i \leq m, \end{cases} \tag{4.1}$$

donde los valores temporales $t_1, ..., t_K$ de los cambios y el valor desviado de la media en cada caso $\mu_1, ..., \mu_K$ son desconocidos.

Para el cálculo de los estadísticos y de los puntos de cambio se utiliza un enfoque de segmentación recursiva hacia delante.

En este algoritmo, el intervalo se divide en k subintervalos que, en cada etapa se irán dividiendo, buscando un potencial punto de cambio, que se seleccionaría maximizando la expresión:

$$\sum_{i=1}^{k+1} (\hat{t}_i - \hat{t}_{i-1})(\bar{x}(\hat{t}_{i-1}, \hat{t}_i) - \bar{\bar{x}})^2, \tag{4.2}$$

condicionado en cada etapa por los resultados de la anterior. En la expresión, $\hat{t}_i$, representarían los puntos de cambio nuevos calculados, que dan lugar a otra partición, y $\bar{x}(t_1, t_2)$ es la media de cada intervalo, $\bar{x}(t_1, t_2) = \frac{1}{t_2 - t_1} \sum_{i=t_1+1}^{t_2} \bar{x}_i$.

El estadístico de control T_k coincide con el valor máximo calculado de la expresión anterior.

Por último, el estadístico global de control sería:

$$W = \max_{k=0,\dots,K} \frac{T_k - E_0(T_k)}{var_0(T_k)}. \tag{4.3}$$

Este no se puede calcular porque depende de la distribución de los datos, pero se utiliza una aproximación empleando permutaciones.

4.1.2. Mphase1

Los mismos autores desarrollaron también un procedimiento para datos multivariantes, Capizzi y Masarotto (2017) [4]. Se aplica tanto a individuales como con subgrupos y en ellos la necesidad de asumir alguna distribución para los datos se elimina utilizando un procedimiento de permutaciones. Estos gráficos de control se basan en los rangos con signo multivariantes, combinando los signos y los rangos de la profundidad de Mahalanobis.

Se consideran m subgrupos ordenados en el tiempo de tamaño n de p variables, y denotamos $\mathbf{X}_{ij}$ al vector g-dimensional con la j-ésima observación del i-ésimo grupo. En el caso de que se trate de observaciones individuales y no en subgrupos sería $n = 1$, y no se podrían detectar cambios aislados, como se comentó en la sección anterior.

Etapa 1.- En esta etapa se realiza la estandarización de los datos y la asignación de los correspondientes rangos signados multivariantes, $\mathbf{r}_{ij}$. Esta transformación mejora los resultados de los gráficos cuando la distribución del proceso tiene colas pesadas o está sesgada.

Considerando $\mathbf{z}_{i,j}$ las observaciones estandarizadas, los rangos con signos multivariantes se calculan:

$$\mathbf{u}_{i,j} = \begin{cases} 0, & \mathbf{z}_{i,j} = 0 \\ \dfrac{\sqrt{F^{-1}_{\chi^2_g}\left(\frac{r_{i,j}}{1+mn}\right)}}{||\mathbf{z}_{i,j}||}\mathbf{z}_{i,j}, & \mathbf{z}_{i,j} \neq 0 \end{cases} \tag{4.4}$$

donde $||x||$ denota la norma euclídea del vector, $r_{i,j}$, el rango de $||z_{i,j}||$ dentro de $||z_{1,1}||, ..., ||z_{m,n}||$ y $F_{\chi^2_g}(\cdot)$ es la distribución de una variable aleatoria χ^2 con g grados de libertad.

Las siguientes etapas se centran en ajustar el modelo de regresión lineal múltiple

$$\mathbf{u}_{ij} = \vec{\beta}_{comun} + \sum_{t=2}^{m-1} \beta_{paso,t} I(i \geq t) + \sum_{t=1}^{m} \beta_{aislado,t} I(i = t) + \varepsilon_{i,j}, \tag{4.5}$$

donde los $\beta' s$ son los vectores de parámetros desconocidos g-dimensionales e I es la función indicadora. El $\vec{\beta}_{comun}$ representaría el nivel estable de los rangos signados, $\beta_{paso,t}$, sería un cambio de nivel que empieza en el tiempo t y afecta a todo lo demás y $\beta_{aislado,t}$ es un cambio de nivel que afecta solo a la observación en tiempo t.

Con esto, comprobar la estabilidad del modelo sería lo mismo que realizar el contraste de hipótesis:

$$\begin{aligned} H_0: \quad & \beta_{paso,t} = 0, t = 2, ..., m-1, \\ & \beta_{aislado,t} = 0, t = 1, ..., m \\ H_1: \quad & \exists \beta_{paso,t} \text{ o } \beta_{aislado,t} \neq 0 \end{aligned} \tag{4.6}$$

Etapa 2.- En la segunda etapa se utiliza un algoritmo de búsqueda hacia delante para seleccionar, entre los $2m - 2$ vectores de parámetros β, las que podrían ser verdaderas desviaciones según los datos. Se selecciona un parámetro K, $K < m$, que es el número máximo de variaciones que queremos seleccionar, siendo recomendado $K = \text{mín}(50, [\sqrt{m}])$.

Con este algoritmo se ajusta un modelo con un número creciente de parámetros. Como criterio para el ajuste del modelo se minimiza la suma de residuos al cuadrado condicionada por las desviaciones detectadas en las etapas anteriores del algoritmo

$$\sum_{i=1}^{m}\sum_{j=1}^{n} ||\mathbf{u}_{ij} - \hat{\mathbf{u}}_i^{(k)}||^2, \tag{4.7}$$

siendo $\hat{\mathbf{u}}_i^{(k)}$ los valores ajustados en la etapa k.

Tras cada paso se calcula la varianza explicada:

$$T_k = n \sum_{i=1}^{m} ||\hat{\mathbf{u}}_i^{(k)}||^2 - mn||\bar{\bar{\mathbf{u}}}||^2, \tag{4.8}$$

siendo $\bar{\bar{\mathbf{u}}} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{r}_{i,j}/mn$

Etapa 3.- Se trata de la etapa de test. En la anterior se calcularon K estadísticos de test T_k para detectar cambios y ahora estos se combinan en un estadístico general y se calcula un único p- valor. El estadístico teórico sería

$$W_{obs} = \max_{k=1,\ldots,K} \frac{T_k - E_0(T_k)}{var_0(T_k)}$$

Su cálculo no es posible porque depende de la distribución de los datos bajo control, pero se utiliza una aproximación mediante permutaciones.

Etapa4.- En la última etapa, cuando se rechaza que el proceso sea estable, se refina el modelo utilizando el algoritmo LASSO adaptativo y un criterio de información. Con esto se eliminan algunos cambios innecesarios introducidos por el algoritmo de la etapa 2 y se identifican también el subconjunto de variables involucradas en cada variación.

4.2. Gráficos de control no paramétricos

En esta sección introducimos los gráficos de control no paramétricos de forma teórica. Se describe primero la profundidad de datos, que es en lo que se basan estos gráficos, mostrando dos de las profundidades utilizadas. A continuación, se describen los gráficos de control no paramétricos basados en la profundidad de datos, (Regina Liu 1995) [14], y cómo se construyen los gráficos.

4.2.1. Gráficos de control no paramétricos basados en la profundidad de datos

Estos gráficos de control se basan en la idea de reducir cada dato multivariante a un único índice univariante. Este índice se corresponde con la clasificación, centro-exterior, inducida por la profundidad de datos. En todo este procedimiento solo se utilizan los datos disponibles, por lo que es un enfoque no paramétrico. Como no se definen en base a ninguna suposición a cerca de un modelo paramétrico para los datos, permite aplicarlos en un número mucho más amplio de estudios que otros gráficos como los MEWMA. Además, permiten controlar de manera simultánea los cambios en localización y en escala de un proceso, lo que resulta de gran interés.

Profundidad de datos y estadístico.

La profundidad de datos es una función que indica, en algún sentido, cómo de "profundo" está situado un punto con respecto a una nube de puntos dada en un d-espacio, o con respecto a una distribución de probabilidad. La profundidad define un centro de la nube, que es el conjunto de puntos más profundos, y mide cómo de lejos se encuentra otro punto respecto a estos.

En base a esto, se puede definir un rango en el espacio euclídeo multidimensional mediante en cálculo de la profundidad de las observaciones.

Profundidad de Tukey.

Como función de profundidad para los gráficos de control utilizaremos la profundidad de Tukey (1975) [20] o profundidad del semiespacio. La profundidad de Tukey de un vector x respecto a la matriz de datos bajo control, denotémosla D, se calcula como:

$$TD(x) = \#\{y \in D \mid d(x,y) \leq d(x, y_x[k])\} - \#\{y \in D \mid d(x,y) < d(x, y_x[k])\},$$

$d(x, y)$ es la distancia entre x e y y $y_x[k]$ es el k-vecino más cercano de x, seleccionando k como la mediana del rango, redondeando hacia abajo al entero más cercano.

Sea $x \in R^p$, la profundidad de x con respecto a una medida de probabilidad P en R se define como la mínima probabilidad que queda determinada por cualquier semiespacio cerrado que contenga x, es decir,

$$HD(x, P) = inf\{P(H) : H \text{ es un semiespacio cerrado que contenga a } x\},\ x \in R^p.$$

Profundidad de Mahalanobis.

Otra profundidad fue propuesta por Mahalanobis (1936) [16]. Dada una distribución G, p -dimensional con media μ y matriz de covarianzas S, la distancia de un punto x respecto a G se define

$$DM(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}. \tag{4.9}$$

A partir de la profundidad de datos se definen los estadísticos que se utilizarán para los gráficos de control. Consideremos G una distribución p-dimensional e $Y_1, ..., Y_m$, m observaciones aleatorias de G, en condiciones bajo control. Sean ahora $X_1, X_2, ...$ nuevas observaciones del proceso, suponiendo que los X_i siguen una distribución F si el proceso ya no se encuentra bajo control. Denotando por $D_G(\cdot)$ la profundidad utilizada, podemos utilizarlo para ordenar los $Y_1, ..., Y_m$ obteniendo los estadísticos de orden $Y[1], ..., Y[m]$. $Y[m]$ sería el punto de más central y, cuanto menor sea la profundidad, más alejado estará de la distribución G. Con esto se define el estadístico de rango

$$r_G(y) = P\{D_G(Y) \leq D_G(y) \mid Y \sim G\}.$$

En caso de que la distribución sea desconocida, se sustituye por la distribución empírica

$$r_{G_m}(y) = \frac{\#\{D_{G_m}(Y_j) \leq D_{G_m}(y), j = 1, ..., m\}}{m}$$

Entonces, con una muestra nueva $\{X_1, ..., X_m\} \sim F$, interesará contrastar si las distribuciones G y F son iguales, lo que indicaría que el proceso sigue bajo control. Cuando las distribuciones son desconocidas, se sustituyen por su versión empírica.

Los contrastes de hipótesis que se realizan para crear los gráficos de control son contrastes de igualdad de los vectores de medias de cada distribución, y de igualdad de matrices de covarianzas.

Veremos en este capítulo varios gráficos de control basados en este concepto de profundidad de datos.

4.2.2. *r*-charts.

Se utiliza como estadístico el rango de las observaciones utilizando la profundidad escogida, que definimos en la sección anterior.

El contraste asociado sería:

$$\begin{cases} H_0 : F = G, \\ H_1 : F \text{ y } G \text{ difieren en posición o escala.} \end{cases} \tag{4.10}$$

Con esto, para una nueva muestra $\{X_1, ..., X_n\}$ se calculan los rangos $\{r_G(X_1), ..., r_G(X_n)\}$ o $\{r_{G_m}(X_1), ..., r_{G_m}(X_m)\}$ caso de no conocer la distribución de los datos pero sí disponer de una muestra de referencia $\{Y_1, ..., Y_m\}$. Lo que se muestra en el gráfico son los estadísticos de rangos a lo largo del tiempo. La línea central del gráfico de control se corresponde con $CL = 0{,}5$ y como límite inferior se establece la tasa de falsa alarma, α. El límite superior en este caso carece de sentido ya que, cuanto

mayor es el rango, más central es el dato en la nube de puntos, y esto ya estaría cubierto por la línea central.

Los valores pequeños son los que indican que hay pocos puntos de la nube considerada como referencia más lejanos. Un rango pequeño, pone de manifiesto una posible desviación entre G y F. Esto puede deberse tanto a un cambio en localización o en escala de las variables.

Sin embargo, resulta importante destacar el caso en el que el estadístico está cercano al 1. Si los valores de r_{G_m} se acercan a 1, esto significa una disminución de la dispersión de los datos y, por lo tanto, una ganancia en la precisión y mejora en el proceso.

4.2.3. *Q*-charts.

Estos gráficos serían la versión no paramétrica de los gráficos $\bar{x}$. Se busca no señalar como fuera de control cuando el proceso está realmente bajo control, a pesar de que algunos valores individuales de la muestra estén fuera de control. Para ello, igual que en la versión paramétrica, se grafican las medias de los subgrupos. Se utiliza como estadístico:

$$Q(G, F_q^j) = \frac{1}{q}\sum_{i=1}^{q} r_G(X_i) \tag{4.11}$$

o de nuevo, en caso de que la distribución no se conozca,

$$Q(G_m, F_q^j) = \frac{1}{q}\sum_{i=1}^{q} r_{G_m}(X_i), \tag{4.12}$$

donde F_q^j es la distribución empírica de las observaciones de X en el subgrupo j.

Para el gráfico se consideran por tanto subconjuntos consecutivos de tamaño q y se representan los valores de $Q_j(G_m, F_q^j)$, que son los promedios de los rangos de las observaciones de ese subgrupo, $\{r_{G_m}(X_1), ..., r_{G_m}(X_n)\}$.

Los límites de control se determinan en función del tamaño de las submuestras. Si $q \geq 5$, $CL = 0{,}5$ y $LCL = 0{,}5 - Z_\alpha(12q)^{1/2}$ o $LCL = 0{,}5 - Z_\alpha\sqrt{\frac{1}{12}\left(\frac{1}{m} + \frac{1}{q}\right)}$, el primero cuando conocemos la distribución G y el segundo cuando se utiliza la empírica. En el caso de que el tamaño de los subgrupos sea más pequeño, $q < 5$, $CL = 0{,}5$ igual y $LCL = \frac{(q!\alpha)^{1/q}}{q}$

4.2.4. *S*-chart.

En este caso el gráfico sería la versión no paramétrica del CUSUM. Para tomar la decisión de si el proceso está bajo control se analiza toda la muestra o la mayor parte de esta. Es más efectivo que los anteriores para detectar pequeños cambios en el proceso. Se basa en el estadístico:

$$S_n(G) = \sum_{i=1}^{n}(r_G(X_i) - \frac{1}{2})$$

que tendría límites de control $CL = 0$ y $LCL = -Z_\alpha\left(\frac{n}{12}\right)^{\frac{1}{2}}$, y en caso de no conocer la distribución,

$$S_n(G_m) = \sum_{i=1}^{n}(r_{G_m}(X_i) - \frac{1}{2})$$

con límites de control $CL = 0$ y $LCL = -Z_\alpha\sqrt{n^2\frac{\left(\frac{1}{m}+\frac{1}{n}\right)}{12}}$. El estadístico calcula la desviación del rango de cada observación respecto del valor "objetivo".

En caso de que el tamaño muestral sea grande, es recomendable estandarizar el estadístico. De esta forma, los límites inferiores, que en ambos casos son curvas, se transforman en rectas. Esto sería:

$$S_n^*(G) = \frac{S_n(G)}{\sqrt{\frac{n}{12}}}$$

$$S_n^*(G_m) = \frac{S_n(G_m)}{\sqrt{n^2 \frac{\left(\frac{1}{m}+\frac{1}{n}\right)}{12}}}.$$

De esta forma, los límites de control serían, $CL = 0$ y $LCL = -Z_\alpha$

Capítulo 5

Aplicación de los gráficos y análisis de los resultados.

Una vez descritos de forma teórica los gráficos de control no paramétricos, en este capítulo se aplicarán a los datos tomados de la empresa y analizaremos los resultados que aportan.

5.1. Datos reales.

Como se describió en la introducción, en una instalación frigorífica hay varias máquinas que intervienen en el proceso del frío.

Las máquinas que se están controlando con el programa de machine learning y detección de fallos son los evaporadores. El evaporador se encarga, en el proceso de frío, del intercambio de calor entre los fluidos refrigerantes. En él tiene lugar el paso de la energía térmica desde un medio al otro: mientras uno de ellos se enfría, el otro se calienta y se evapora.

Los evaporadores cuentan con un sistema de desescarche que, mediante resistencias, entra en funcionamiento cada cierto tiempo para evitar formaciones de hielo no deseadas. Los sistemas de desescarche se basan en aportar calor al evaporador de forma que se produzca la fusión del hielo y escarcha acumulados. Como consecuencia del desescarche, el evaporador pasa de ser el punto más frío de la cámara a ser el punto con mayor temperatura. Esto se ve reflejado en las variables, que presentan subidas considerables de sus valores.

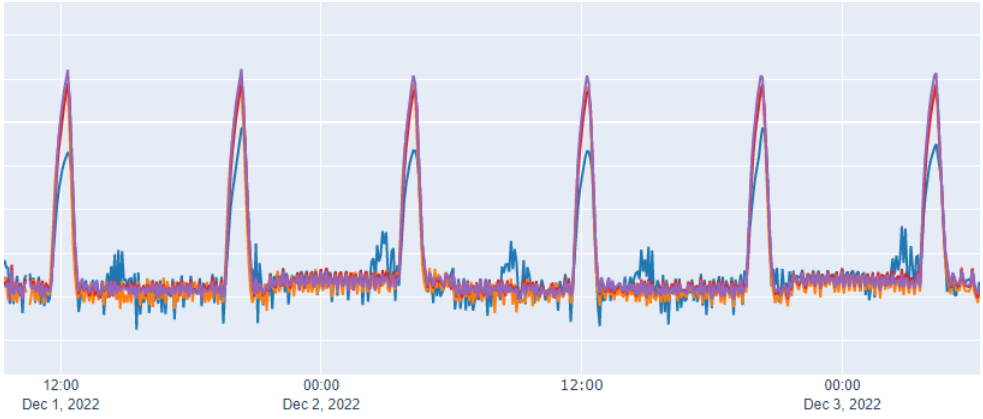


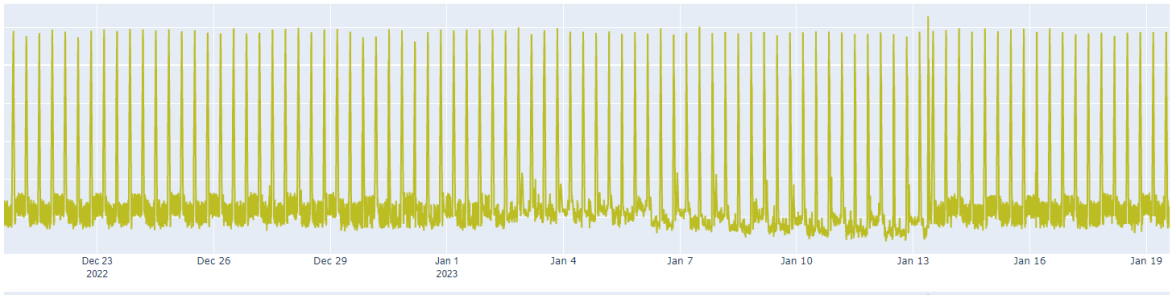
Figura 5.1: Variables de un evaporador.

En la figura 5.1 los picos se corresponden con los desescarches. Estos dificultan el control de la calidad, ya que se detectarían como puntos fuera de control, cuando realmente forman parte del comportamiento usual de la máquina. Para solucionar esto, se entrenaron modelos de machine learning para los evaporadores, y el control se realiza a los residuos del modelo.

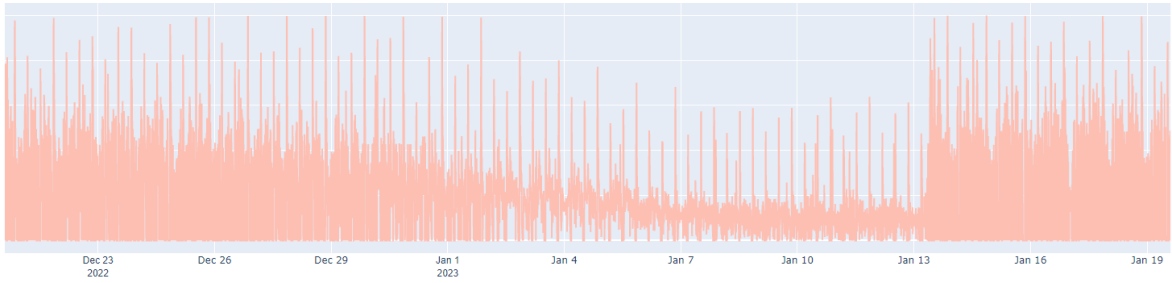
Con esto, se guardaron tres tipos de datos: los reales de las variables del evaporador, los residuos del modelo de machine learning y también las predicciones del modelo.

En este caso, las variables que se seleccionaron para el análisis de los evaporadores en el trabajo anterior [7] son la temperatura de evaporación, la temperatura de impulsión, la temperatura de fin de desescarche, el porcentaje de apertura de la válvula AKV y la temperatura de retorno. Esta selección se realizó teniendo en cuenta el conocimiento del ciclo de frío, tomando las variables que se consideraron más importantes para el correcto funcionamiento de las máquinas, y cuya desviación supondría un mayor problema.

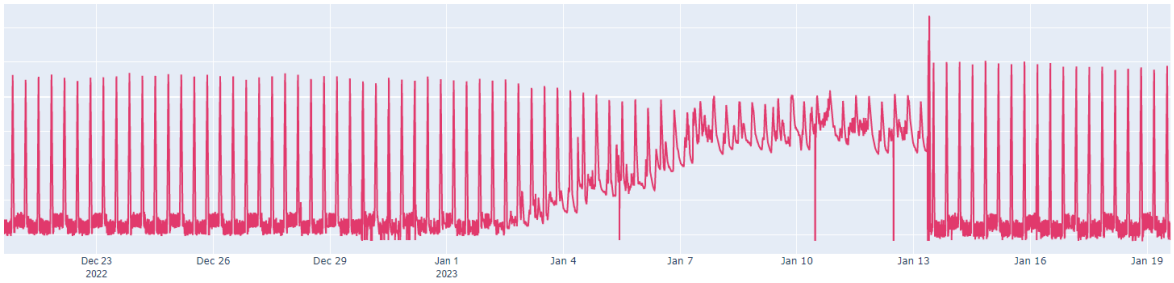
Se tomó un histórico de datos durante el cual se produce un fallo del evaporador. Se cogieron varios días anteriores en los que el comportamiento era el usual, y unos días después del fallo cuando estaba de nuevo en el comportamiento correcto.



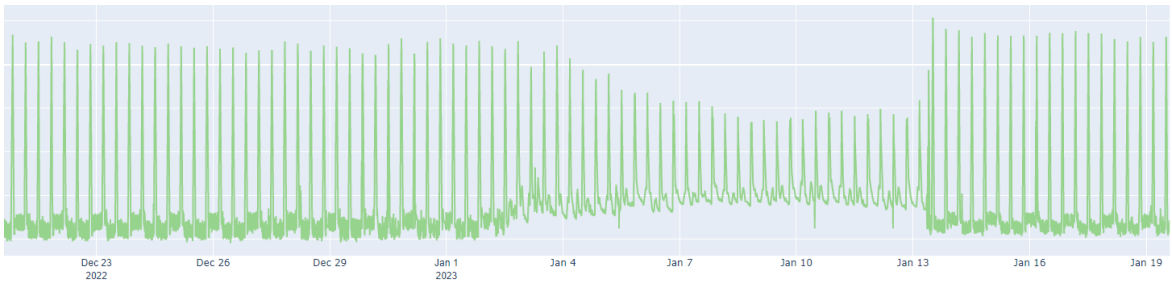
(a) Termperatura de fin de desescarche.



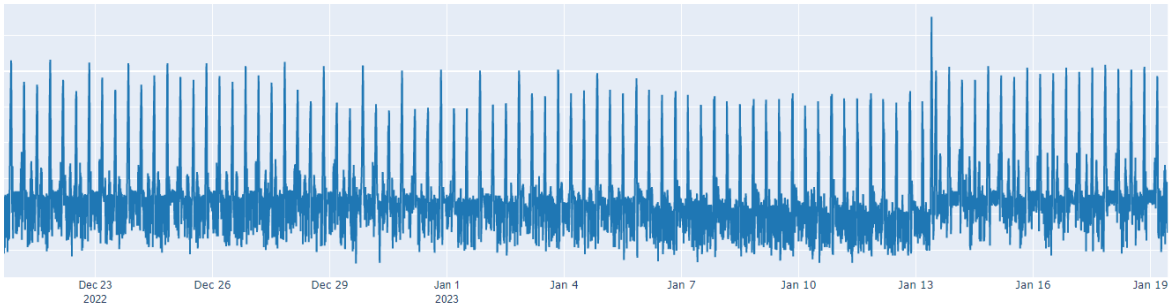
(b) Porcentaje de apertura de la AKV.



(c) Temperatura de impulsión.



(d) Temperatura de retorno.



(e) Temperatura de evaporación.

Figura 5.2: Valores reales de las variables.

Mostramos en la figura 5.2 las variables reales que se utilizan para la detección de fallos en este evaporador. Vemos como en los datos se refleja lo comentado. En ellas hay una primera zona en la que las variables tienen un comportamiento normal. Luego, empieza a verse una ligera desviación de los datos, que va aumentando de manera considerable. Finalmente, la máquina vuelve al comportamiento habitual del principio.

Con esto, nos interesa aplicar los distintos gráficos de control descritos en las secciones anteriores a estos datos concretos, para ver los resultados que aportan. Como se comentó, el control de procesos se realiza con los residuos del modelo de machine learning. Hemos graficado las variables reales para ver el cambio real en ellas, pero trabajaremos con los residuos. Mostramos la diferencia en la variabilidad presente en los valores reales y en los residuos cuando el evaporador está trabajando en la forma habitual.

```
> var(na.omit(valores_reales)[1:2300, ])
          96        50         89        97        98
96 23.782546  27.94587  -7.478795 25.273200 24.653935
50 27.945868  36.97401 -23.951983 32.986379 32.087046
89 -7.478795 -23.95198 529.521596 -8.535002 -8.930314
97 25.273200  32.98638  -8.535002 30.193895 29.262808
98 24.653935  32.08705  -8.930314 29.262808 28.440659
> var(na.omit(residuos)[1:2300, ])
           96         50         89          97          98
96  1.7990953 -0.3828403   1.499793 -0.10783465  0.47440032
50 -0.3828403  1.4159666  -5.551733 -0.15471939 -1.14773591
89  1.4997925 -5.5517329 354.388388  1.21839134  4.28694192
97 -0.1078346 -0.1547194   1.218391  0.25642821  0.09453355
98  0.4744003 -1.1477359   4.286942  0.09453355  1.22402705
```

Figura 5.3: Varianzas de las variables reales y de los residuos analizados.

Vemos como hay diferencias importantes, teniendo los residuos varianzas mucho menores que facilitan su estudio.

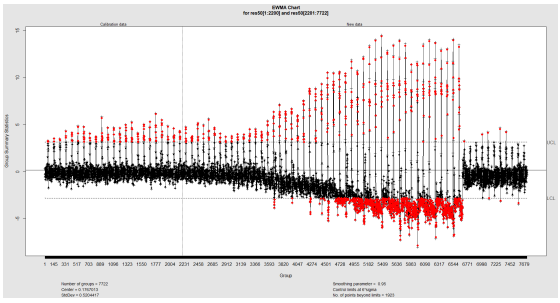
A continuación, se aplican los gráficos descritos en el capítulo anterior utilizando los paquetes de R proporcionados por los autores.

5.2. EWMA

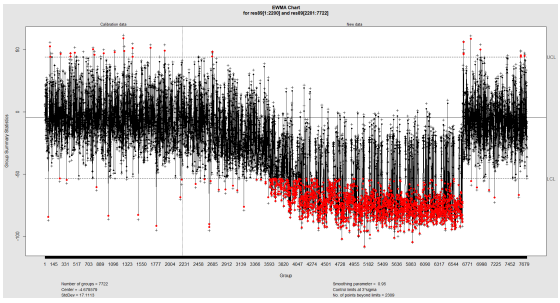
Como se comentó antes, en la empresa se llevó a cabo el desarrollo de estos gráficos EWMA, pero actualmente no se están utilizando para la detección de desviaciones. Se planteó implementarlos, dado que el estudio individual nos permitiría saber cuál es la variable que se está alejando del comportamiento habitual, aportando una información más precisa sobre el fallo que puede estar sucediendo.

Ante esto, estudiamos los resultados aportados por los gráficos EWMA aplicados a las variables individuales, viendo si realmente son interesantes, y determinando qué variables aportan mejores resultados, y serían por tanto las más relevantes para el monitorizado.

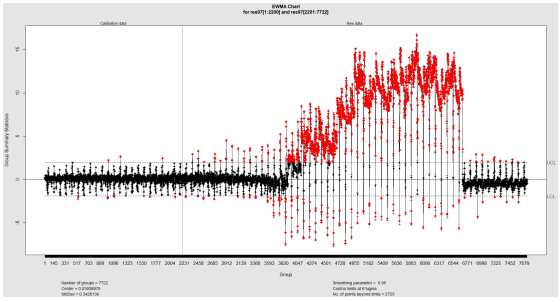
Se seleccionó como parámetro de suavizado $\lambda = 0{,}95$, con lo que el método tendría poca memoria, y sería mejor para detectar fallos grandes. Para los límites de control, teniendo en cuenta la variabilidad de los datos, se seleccionaron 6σ, con lo que detectaríamos cambios de gran tamaño. La única variable en la que no se hizo así es con el porcentaje de apertura de la AKV, porque era más razonable considerar límites 3σ.



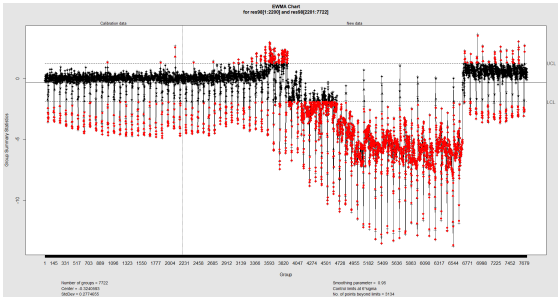
(a) Gráfico EWMA aplicado a la temperatura de fin de desescarche.



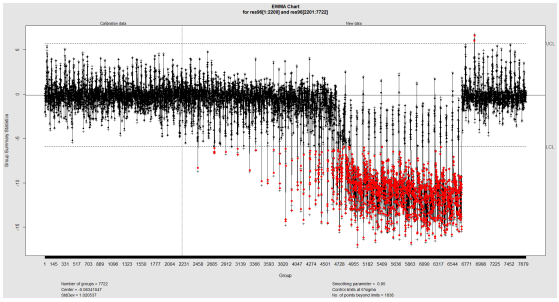
(b) Gráfico EWMA aplicado al porcentaje de apertura de la AKV.



(c) Gráfico EWMA aplicado a la temperatura de impulsión.



(d) Gráfico EWMA aplicado a la temperatura de retorno.



(e) Gráfico EWMA aplicado a la temperatura de evaporación.

Figura 5.4: Gráficos EWMA para las variables univariantes.

En la figura 5.4 se observa que en general todos detectan bien el fallo sucedido. Para analizar cómo han clasificado los datos mostramos la matriz de confusión de cada uno y las medidas asociadas.

Real \ Detec	OOC	IC
OOC	1707	1687
IC	80	2048

(a) Matriz de confusión del gráfico EWMA para la temperatura de fin de desescarche.

Real \ Detec	OOC	IC
OOC	2263	1131
IC	21	2107

(b) Matriz de confusión del gráfico EWMA para el porcentaje de apertura de la AKV.

Real \ Detec	OOC	IC
OOC	2661	733
IC	53	2075

(c) Matriz de confusión del gráfico EWMA para la temperatura de impulsión.

Real \ Detec	OOC	IC
OOC	2680	714
IC	254	1874

(d) Matriz de confusión del gráfico EWMA para la temperatura de retorno.

Real \ Detec	OOC	IC
OOC	1826	1568
IC	12	2116

(e) Matriz de confusión del gráfico EWMA para la temperatura de evaporación.

Figura 5.5: Matrices de confusión de los EWMA para las variables univariantes.

De las matrices de confusión vemos que en las variables que mejor se detectan las desviaciones son la temperatura de retorno y la temperatura de impulsión, y luego el porcentaje de apertura de la válvula. En el caso de la temperatura de fin de desescarche y la de evaporación la detección de desviaciones es muy mala, dejando sin señalar casi la mitad de los datos, por lo que no serían una opción recomendable. En estas variables, como consecuencia, sí se analiza bien el proceso bajo control, pero los resultados son igual de buenos que los de la temperatura de impulsión y el porcentaje de la válvula, por lo que estas serían las variables más fiables para monitorizar. En el caso de la temperatura de retorno en este sentido es algo peor, pero no parece suficiente para descartarla. En cuanto a falsas alarmas, la peor es la temperatura de retorno, lo que puede restarle fiabilidad, y la temperatura de fin de desescarche, que ya establecimos que no era una buena opción.

En otra tabla recogemos la especificidad, sensibilidad y precisión de cada gráfico.

	TPR	TNR	ACC	BA
Temperatura de retorno	0.79	0.881	0.825	0.836
Temperatura de evaporación	0.54	0.994	0.714	0.766
Temperatura de fin de desescarche	0.5	0.96	0.68	0.7325
Temperatura de impulsión	0.78	0.98	0.85	0.8795
Porcentaje apertura AKV	0.67	0.99	0.79	0.829

Cuadro 5.1: Medidas de las variables.

Observamos que en general estos gráficos detectan peor las observaciones fuera de control. Los mejores resultados son como vimos para la temperatura de retorno, de impulsión y el porcentaje de la AKV, siendo malos para las otras dos. Cuando se trata de determinar si el proceso está bajo control sí son fiables, aportando buenos resultados con las 5 características. En la precisión global y balanceada los resultados no son muy malos, pero no los deseados en el caso de la temperatura de evaporación y de fin de desescarche. En resumen, de las cinco variables los mejores resultados se obtienen para la temperatura de impulsión y para el porcentaje de la AKV, por lo que serían las primeras a tener en cuenta para realizar un control univariante. Luego la temperatura de retorno sería también una opción aceptable. Además del resultado obtenido con el análisis, como un evaporador es un intercambiador de calor, la selección de estas dos temperaturas es coherente con el proceso físico que se realiza.

5.3. MEWMA

Utilizaremos ahora el gráfico MEWMA, con el que actualmente se detectan las desviaciones y se generan los eventos. Nos interesa analizar los resultados de este gráfico, estudiando su fiabilidad ante un fallo grande, y compararlos con otras alternativas multivariantes, paramétricas y no paramétricas. Este tipo de gráficos tienen buen comportamiento cuando se trata de desviaciones pequeñas, pero vemos que pueden ocurrir fallos grandes, por lo que veamos si también serían adecuados. Mostramos en la siguiente figura la detección realizada.

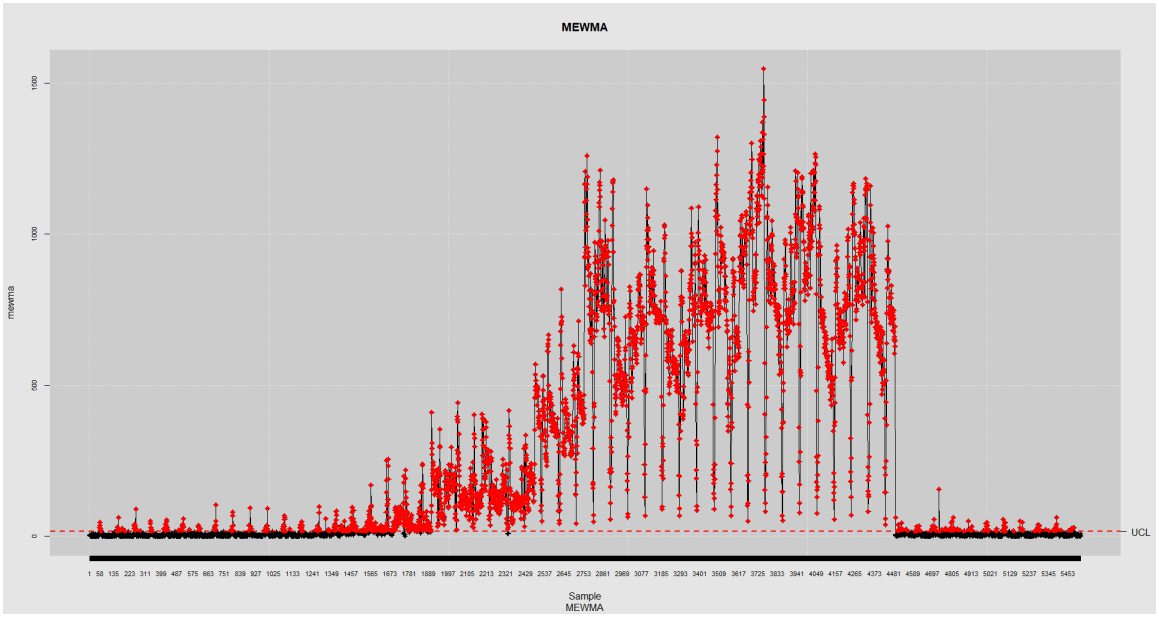


Figura 5.6: Gráfico MEWMA de los datos multivariantes.

De nuevo la figura 5.6 nos aporta una primera idea. En él parece que la desviación grande fue bien detectada, pero que hubo antes varias observaciones detectadas como desviación que no deberían serlo. Calculamos la matriz de confusión y las medidas de clasificación para tener un análisis fiable.

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	3003	391
Bajo control	244	1884

Cuadro 5.2: Matriz de confusión del gráfico MEWMA.

Vemos que en este caso el gráfico tiene un comportamiento similar cuando el proceso está bajo control y cuando se desvía de este. Los resultados obtenidos son bastante buenos, aunque quizá 244 falsas alarmas es un número que convendría que fuera menor, pero esto podría realizarse con un análisis y reglas posteriores. En comparación con los gráficos individuales, detecta más verdaderas desviaciones, pero también da más falsas alarmas.

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} = \frac{3003}{3003+391} = 0{,}885, \\ TNR &= \frac{TN}{TN+FP} = \frac{1884}{1884+244} = 0{,}885, \\ ACC &= \frac{TP+TN}{TP+FN+TN+FP} = \frac{3003+1884}{3003+391+1884+244} = 0{,}885, \\ BA &= \frac{TP+TN}{2} = \frac{0{,}885+0{,}885}{2} = 0{,}885. \end{aligned} \tag{5.1}$$

La sensibilidad y la especificidad muestran cómo efectivamente el método se comporta igual en las dos clases. La sensibilidad es mejor que la del caso univariante, pero no la especificidad, aunque el

resultado global sí es mejor. Con esto, podría ser interesante utilizar ambos gráficos, univariantes y multivariantes, aprovechando así las ventajas de cada uno.

5.4. T^2 Hotelling.

El otro gráfico multivariante paramétrico descrito fue el de Hotelling. Este es más adecuado cuando los datos están correlados. Comprobamos los resultados que aporta a los datos, en primer lugar, mostrando el gráfico.

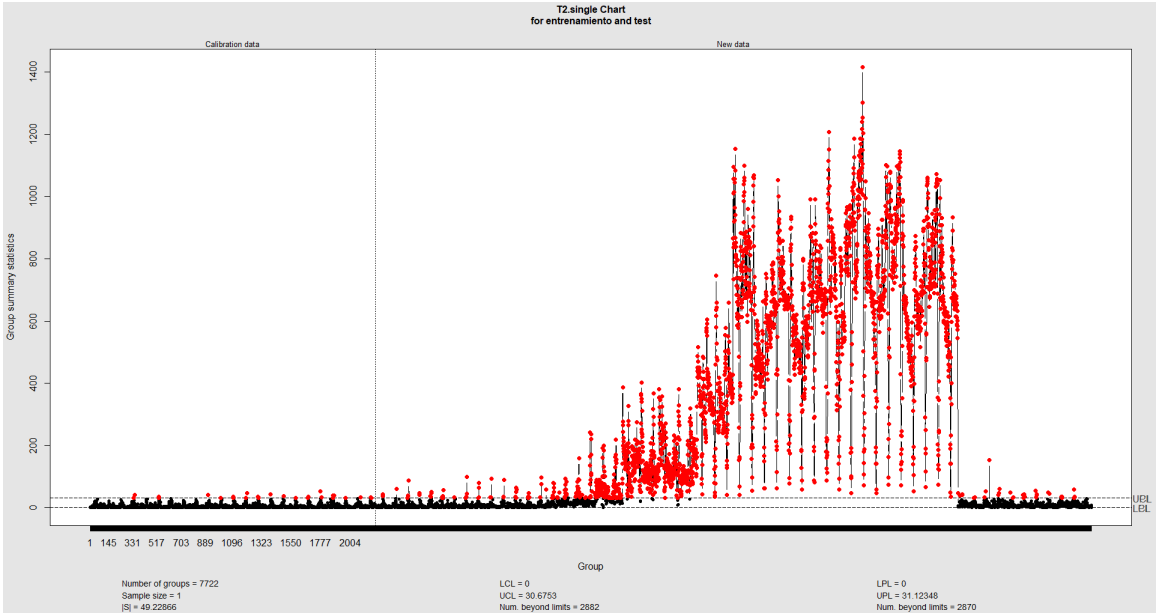


Figura 5.7: Gráfico T^2 de Hotelling de los datos multivariantes.

El resultado es muy similar al anterior, pero parece que en este caso se detectan menos falsas alarmas que con el MEWMA, manteniendo la buena detección de las observaciones relativas al fallo. Para realizar el análisis mostramos en la siguiente tabla la matriz de confusión.

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	2780	614
Bajo control	64	2064

Cuadro 5.3: Matriz de confusión del gráfico T^2 de Hotelling.

En este caso, como se vio en la figura 5.7, el número de falsas alarmas es menor, siendo ahora 64, lo que supondrá un aumento de la especificidad. Baja el número de observaciones fuera de control detectadas, pero aumenta la precisión a la hora de determinar que están bajo control. Las medidas de clasificación en este caso son:

$$TPR = \frac{TP}{TP+FN} = \frac{2780}{2780+614} = 0{,}819,$$
$$TNR = \frac{TN}{TN+FP} = \frac{2067}{2067+61} = 0{,}97,$$
$$ACC = \frac{TP+TN}{TP+FN+TN+FP} = \frac{2780+2064}{2780+614+2067+61} = 0{,}877.$$
$$BA = \frac{TP+TN}{2} = \frac{0{,}819+0{,}97}{2} = 0{,}895,$$

Respecto al gráfico MEWMA, la especificidad aumenta, permitiendo un mejor control del proceso bajo control, que se realiza muy bien, pero disminuye la sensibilidad, lo que supone una peor detección de fallos. Aun siendo peor, el valor de TPR es bueno, y teniendo en cuenta que se disminuye el número de falsas alarmas este gráfico aporta mejores resultados que el MEWMA, por lo que podría ser más interesante.

5.5. *r*-chart

Pasamos ahora a los gráficos no paramétricos multivariantes, comenzando por los gráficos r. Estos son los más adecuados para detectar fallos grandes y utilizaremos las distintas profundidades descritas en el capítulo anterior.

5.5.1. Profundidad de Tukey.

Comenzamos utilizando el gráfico r, utilizando la profundidad de Tukey. Mostramos en la figura 5.8 el gráfico r. En este se observa la línea central y el límite de control inferior. En este caso como límite inferior se estableció $\alpha = 0{,}005$ porque en base a las variables, era lo más adecuado para no aumentar demasiado las falsas alarmas.

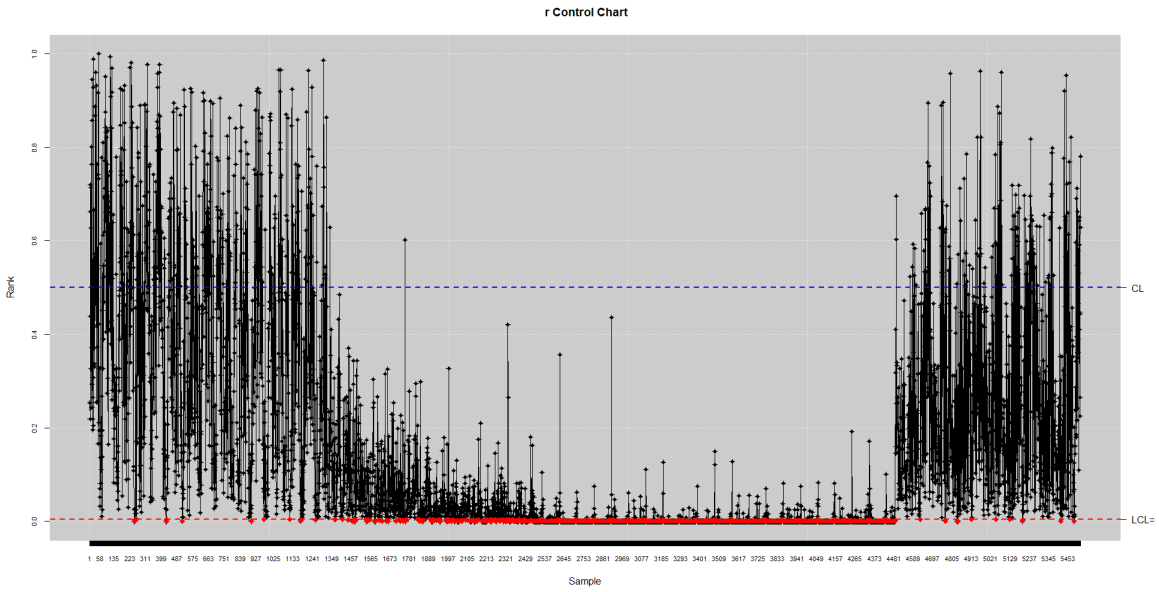


Figura 5.8: Gráfico r aplicado a los datos multivariantes.

En este caso, cuando el proceso está bajo control observamos en la figura 5.8 que se señalan muy

pocas desviaciones, por lo que debería haber pocas falsas alarmas, pero cuando está fuera de control hay bastantes observaciones que no son detectadas como tal.

Mostramos la matriz de confusión de este gráfico.

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	2050	1394
Bajo control	34	2044

Cuadro 5.4: Matriz de confusión del gráfico r.

Este gráfico es el que menos verdaderos positivos tiene, por lo que es el peor detectando cuando el proceso está fuera de control. En cuanto a verdaderos negativos es algo mejor que el MEWMA, pero peor que el gráfico de Hotelling, con lo que tampoco es mejor al detectar el comportamiento adecuado. Lo único que mejora son las falsas alarmas, teniendo en este caso solo 34.

Calculamos también la precisión, especificidad y sensibilidad.

$$TPR = 0{,}595,$$
$$TNR = 0{,}984,$$
$$ACC = 0{,}744.$$
$$BA = 0{,}7895$$

La tasa de verdaderos positivos muestra lo que comentamos, siendo la menor de las tres con diferencia. La especificidad es mayor, pero la diferencia con el de Hotelling es pequeña, y es el que tiene peor precisión, con lo que sería la peor opción de las tres hasta ahora, a pesar de ser no paramétrico.

5.5.2. Profundidad Mahalanobis.

Probamos el mismo gráfico r pero cambiando la distancia que se utiliza para calcular la profundidad de datos, siendo en este caso la de Mahalanobis.

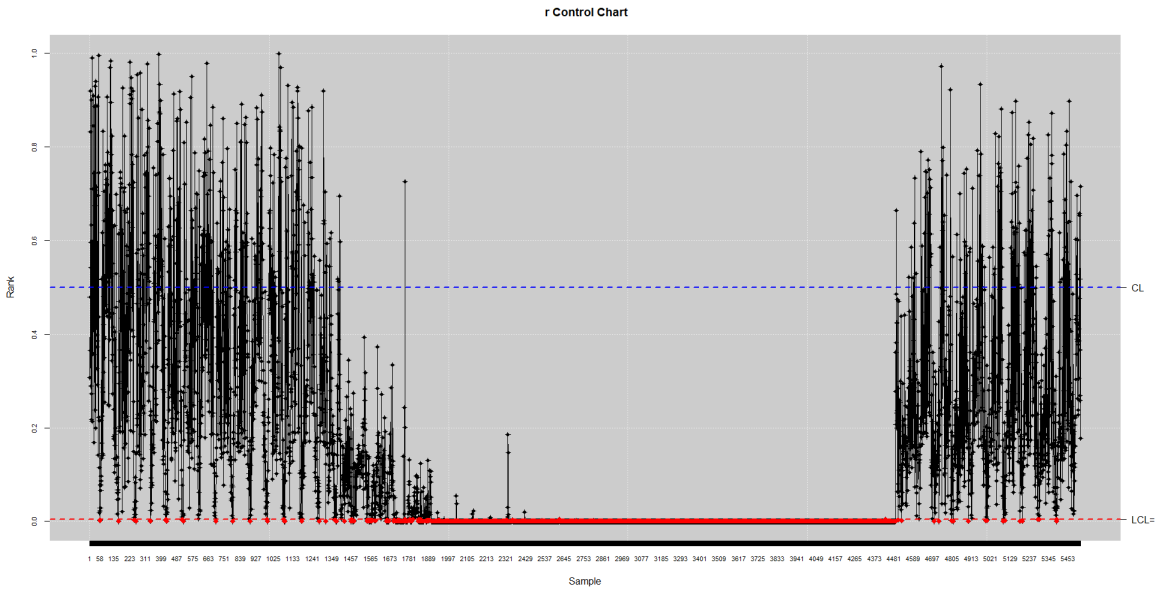


Figura 5.9: Gráfico r aplicado a los datos multivariantes utilizando la profundidad de Mahalanobis.

El cambio de la profundidad utilizada parece que mejora considerablemente los resultados. Vemos que en este caso la detección de las desviaciones es mucho mejor, sin aumentar, parece, las falsas alarmas.

Para confirmarlo se muestra también la matriz de confusión.

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	2743	701
Bajo control	45	2033

Cuadro 5.5: Matriz de confusión del gráfico r utilizando la profundidad de Mahalanobis.

Nos fijamos que los valores obtenidos son muy similares a los de la T^2 de hotelling, siendo sus matrices de confusión casi iguales. Con respecto al otro gráfico r mejora la detección de las desviaciones de forma considerable, sin empeorar apenas los resultados cuando el proceso está bajo control. Se tiene un aumento en las falsas alarmas, pero este no es muy significante.

Las medidas de capacidad de clasificación asociadas son:

$$TPR = 0{,}796,$$
$$TNR = 0{,}978,$$
$$ACC = 0{,}865$$
$$BA = 0{,}887$$

La sensibilidad tiene un valor un poco bajo, menor que el del gráfico de Hotelling, por lo que detectaría algo peor las desviaciones. La especificidad es mejor, pero la diferencia es pequeña, y la precisión global y balanceada es también algo más pequeña. Con todo esto, de los gráficos utilizados hasta ahora, el de Hotelling seguiría siendo el más adecuado, aunque los resultados de este serían

casi iguales, y también muy adecuados. Teniendo en cuenta que estamos utilizando la profundidad de Mahalanobis y que el gráfico de Hotelling se basa en la distancia de Mahalanobis, parece razonable que los resultados aportados sean similares.

5.5.3. Profundidad de máxima verosimilitud

Como última profundidad considerada para este gráfico utilizamos la de máxima verosimilitud, (Fraiman et al. 1997) [10].

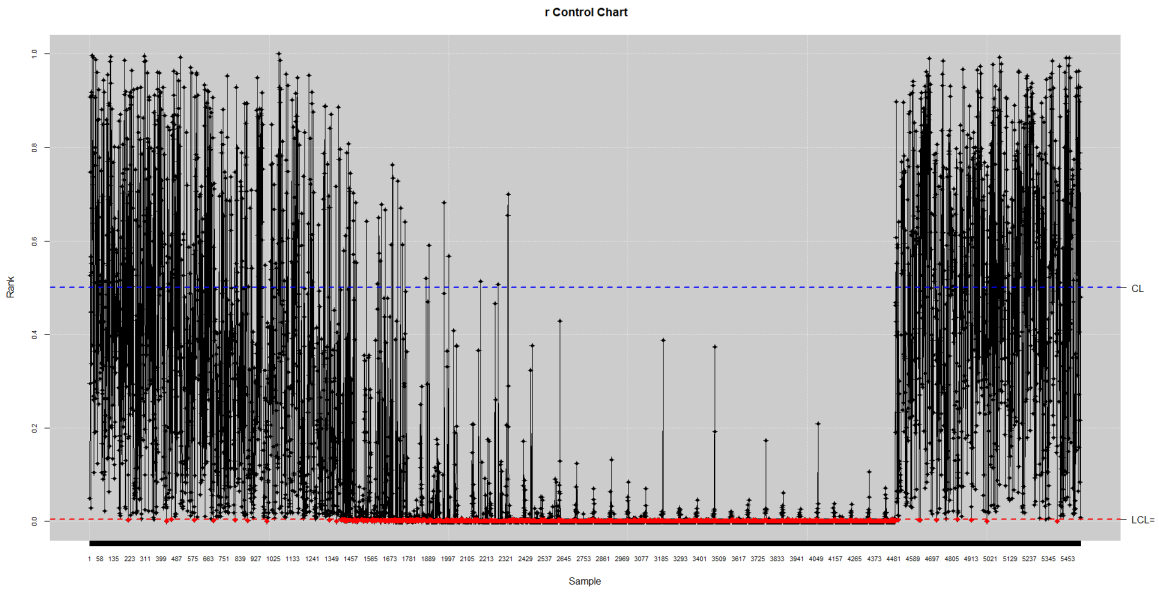


Figura 5.10: Gráfico r aplicado a los datos multivariantes con la profundidad de Mahalanobis.

En este caso parece que el gráfico de la figura 5.10 deja muchas observaciones desviadas sin detectar, pero aporta buenos resultados para el proceso bajo control, sin muchos falsos positivos.

Para completar los resultados se calculó la matriz de confusión del gráfico,

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	2485	959
Bajo control	14	2064

Cuadro 5.6: Matriz de confusión del gráfico r.

El gráfico r con esta profundidad es el que produce el menor número de falsas alarmas, dando muy buenos resultados cuando el proceso está bajo control. Quizá lo peor serían las desviaciones que no

detecta, pero con todo no sería tan mala opción.

$$TPR = 0{,}722,$$
$$TNR = 0{,}993,$$
$$ACC = 0{,}824$$
$$BA = 0{,}858$$

La tasa de verdaderos positivos tiene un valor menor de lo deseado, lo que indica que hay valores desviados que no se están captando. Por otro lado, la tasa de verdaderos negativos es muy buena, con lo que se comporta muy bien cuando el proceso está bajo control. Por último, su precisión y la precisión balanceada no tienen malos resultados, pero en general todos son peores que los del gráfico de Hotelling.

5.6. MCUSUM

Antes de aplicar el otro gráfico no paramétrico S-chart, primero utilizamos su versión paramétrica, el MCUSUM [12]. Con esto queremos ver, teniendo en cuenta que los datos no son normales, si los resultados de los métodos no paramétricos son mejores.

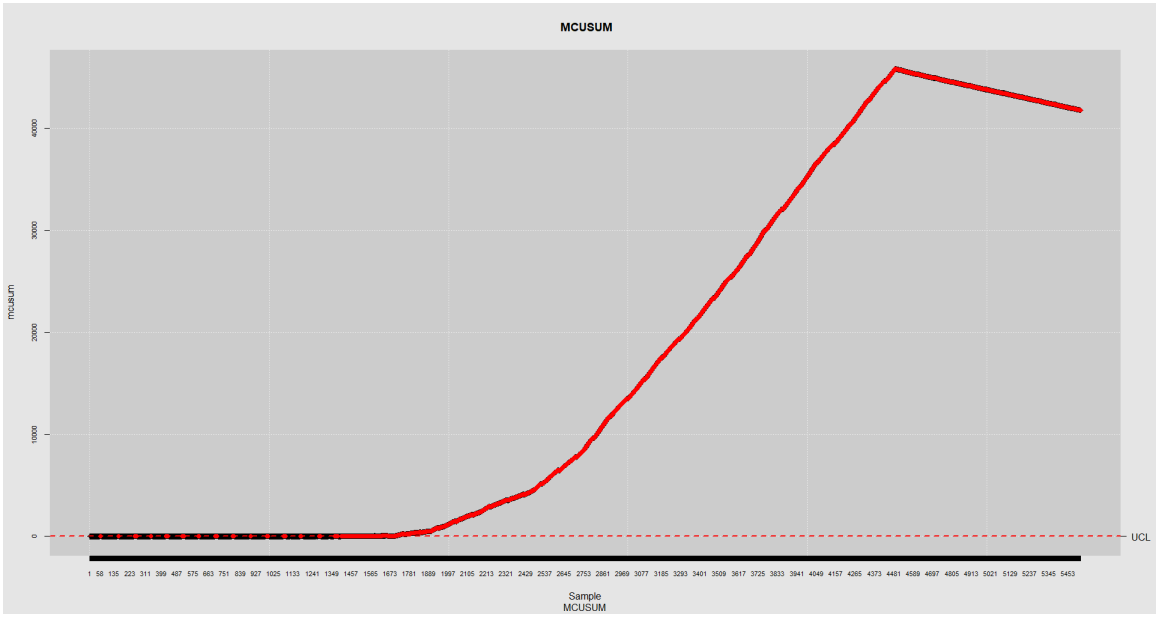


Figura 5.11: Gráfico MCUSUM de los datos multivariantes.

Este gráfico, al ir acumulando las desviaciones, vemos en la figura 5.11 como se desvía mucho del límite de control, pero cuando vuelven los datos al comportamiento habitual, la desviación es tan grande, que baja un poco pero no es capaz de volver al límite de control. Como este gráfico está diseñado para detectar desviaciones pequeñas, en nuestro caso, con un fallo tan grande no es el adecuado.

La matriz de confusión es:

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	3272	122
Bajo control	962	1166

Cuadro 5.7: Matriz de confusión del gráfico MCUSUM.

observamos como cuando el proceso está bajo control los resultados son muy malos. Detecta bien la desviación, pero seguiría marcando como fuera de control durante mucho más tiempo del deseado, produciendo muchas falsas alarmas. Calculamos las medidas de capacidad de clasificación,

$$TPR = 0{,}964,$$
$$TNR = 0{,}548,$$
$$ACC = 0{,}804.$$
$$BA = 0{,}756$$

Muestran lo mismo que se comentó. Las precisiones no son tan malas, pero si tuviéramos más datos seguramente empeorarían, porque los seguiría detectando desviación.

5.7. *S*-chart

Se trata del otro gráfico basado en la profundidad de datos aplicable a medidas individuales. En este caso el paquete no disponía de una función para nuestro caso, por lo que se calculó utilizando las fórmulas descritas. El resultado obtenido se muestra en la gráfica.

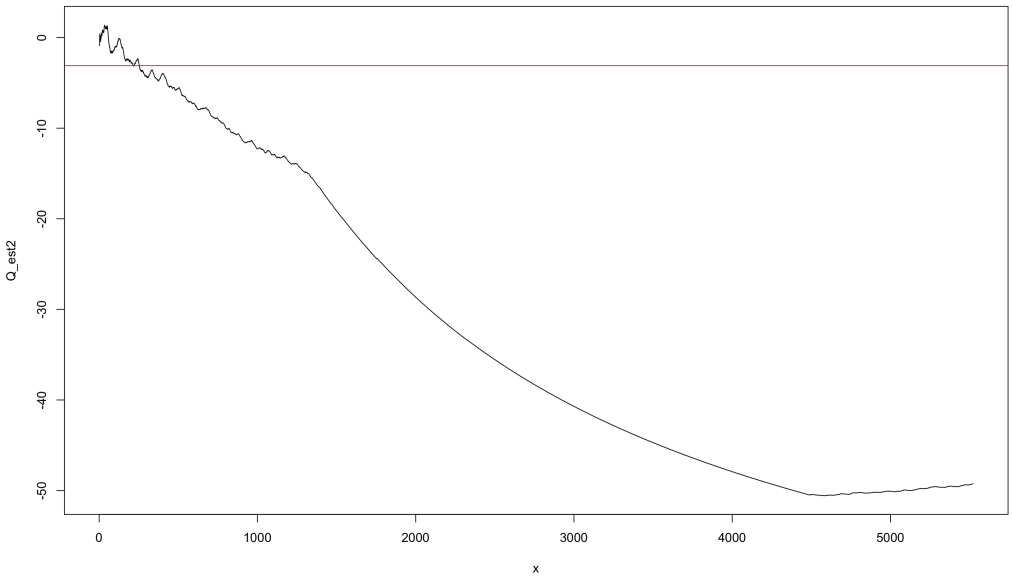


Figura 5.12: Gráfico S aplicado a los datos multivariantes.

En este caso se observa en la figura 5.12 como la media acumulada de los rangos va disminuyendo de manera suave, teniendo algún pico creciente, pero luego el descenso es mucho más pronunciado. Al final, vemos que empieza a subir de nuevo. Como comentamos sobre este gráfico, cuando se analiza una nueva observación se utiliza toda la información anterior disponible, lo que da lugar a seguir detectando como fuera de control las observaciones una vez está el proceso en su funcionamiento normal.

Se muestra la matriz de confusión asociada.

Valor real \ Detección	Fuera de control	Bajo control
Fuera de control	3444	0
Bajo control	1833	245

Cuadro 5.8: Matriz de confusión del gráfico *S*.

En este caso todas las desviaciones fueron detectadas pero el número de falsas alarmas es enorme.

$$TPR = 1,$$
$$TNR = 0{,}118,$$
$$ACC = 0{,}668.$$
$$BA = 0{,}559$$

Con la versión no paramétrica los resultados son los mismos. Las desviaciones se van acumulando, lo que hace que no vuelva a la zona de control cuando lo hace el proceso. La precisión es mala, obteniendo resultados peores que en el caso paramétrico. La comparación de los resultados no parece muy adecuada teniendo en cuenta que no son los gráficos adecuados para el tipo de fallo que ocurrió.

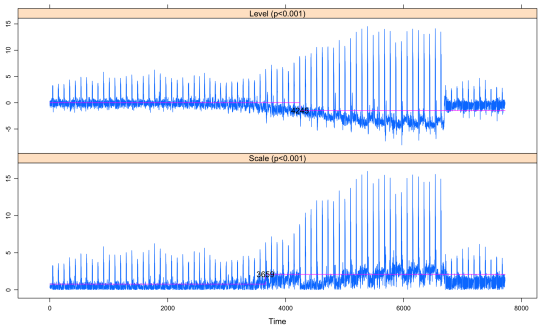
5.8. RSP

Nos centramos a partir de aquí en realizar el análisis de la fase I. Comenzamos aplicando primero el método RSP para cada variable de manera individual.

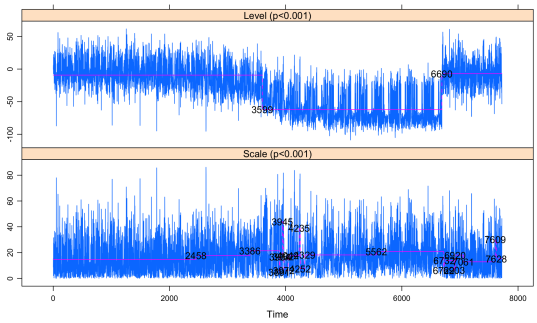
Este gráfico busca reconocer los puntos de cambio de manera retrospectiva, para luego estimar los límites de control en la fase I y pasar a monitorizar la variable en la fase II, por eso no aparecen límites de control. En la figura vemos que las variaciones de nivel se capturan perfectamente en todas las variables.

Los cambios en la escala también se capturan bien en general. Quizá para la temperatura de fin de desescarche y para la temperatura de impulsión, le faltaría al final otro punto de cambio. En la de fin de desescarche es menos evidente, pero la de impulsión parece suficiente variación como para ser detectada. Por otra parte, en el porcentaje de apertura de válvula detecta más cambios. Esta es una variable complicada de analizar porque ya tiene mucha variación, entonces puede ser lo que influya en que detecte quizá más variaciones de las de interés.

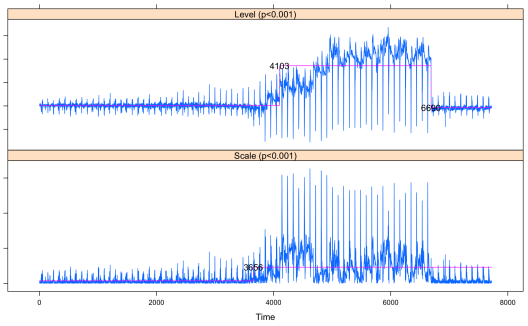
El resultado del gráfico es bueno, ya que permite detectar bien las desviaciones. Con esto, los límites de control estimados serían adecuados, ya que se podrían descartar las zonas de cambio, y permitiría conseguir una fase II de monitorizado adecuada.



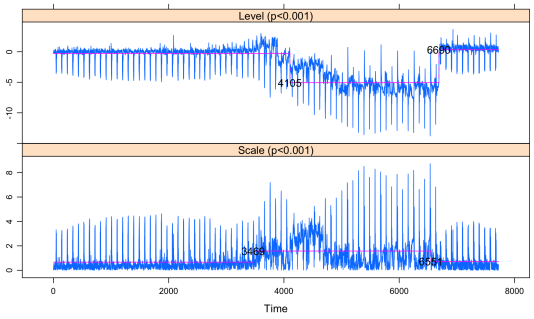
(a) Gráfico RSP aplicado a la termperatura de fin de desescarche.



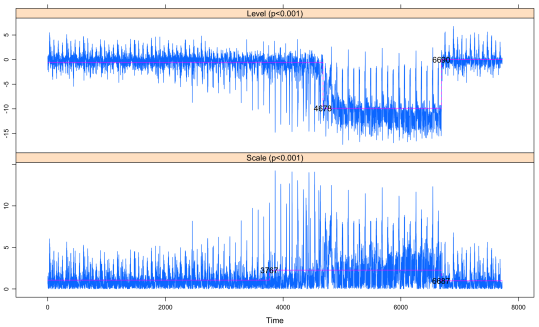
(b) Gráfico RSP aplicado al porcentaje de apertura de la AKV.



(c) Gráfico RSP aplicado a la temperatura de impulsión.



(d) Gráfico RSP aplicado a la temperatura de retorno.



(e) Gráfico RSP aplicado a la temperatura de evaporación.

Figura 5.13: Gráficos de segmentaciones recursivas y permutaciones (RSP).

En los gráficos de la figura 5.13 se observa que estos gráficos detectan muy bien las zonas de cambio de los datos.

Como en este caso lo que se busca es únicamente detectar los puntos de cambio, la matriz de confusión no tiene sentido. En su lugar calculamos el ARL muestral, teniendo en cuenta que sabemos cuándo se comenzó a producir el fallo. En el caso de la temperatura de fin de desescarche, el primer punto de cambio se detecta en la observación 3659, por lo que el $ARL_m = 203$, como se toman observaciones cada cinco minutos, un $ATS_m \approx 17$ horas para detectar la desviación. Para el porcentaje de apertura de válvula tenemos un primer punto de cambio detectado, en la observación 2458. Se trata

de una característica que presenta mucha variabilidad, por lo que esta detección en su escala puede ser fruto de su comportamiento, pero por otra parte también influye mucho en el comportamiento del evaporador, por lo que podría ser un primer indicativo, antes de que el fallo sea detectable en todas las variables. No lo tendremos en cuenta por no ser fiable. Los siguientes puntos de cambio detectados son el 3386 y el 3599, en escala y localización respectivamente. Ambos antes de detectar la desviación de la máquina, por lo que el análisis de esta variable se adelantaría al de las demás en la detección. Para la temperatura de impulsión se detecta en el 3656, en este caso $ARL_m = 200$, muy similar a la primera, y su ATS_m serían también cerca de 17 horas. Para la temperatura de retorno $ARL_m = 19$ y el ATS_m sería 1 hora y 35 minutos y para la temperatura de evaporación $ARL_m = 311$, lo que supone casi 26 horas para detectar la desviación.

Vemos que en todos los casos es el cambio en la escala el que primero señala la desviación, seguido un poco después por el cambio en la localización.

5.9. dfphase1

En este caso el gráfico utilizado será el de análisis de la fase I multivariante, que se muestra en la figura 5.14. Se trata también de análisis para la fase I, que de forma retrospectiva busca analizar si hay variaciones en la distribución. De nuevo no tenemos límites de control.

Como parámetro K se selecciona el valor recomendado, que en este caso es $K = 50$, que será el número de desviaciones detectado.

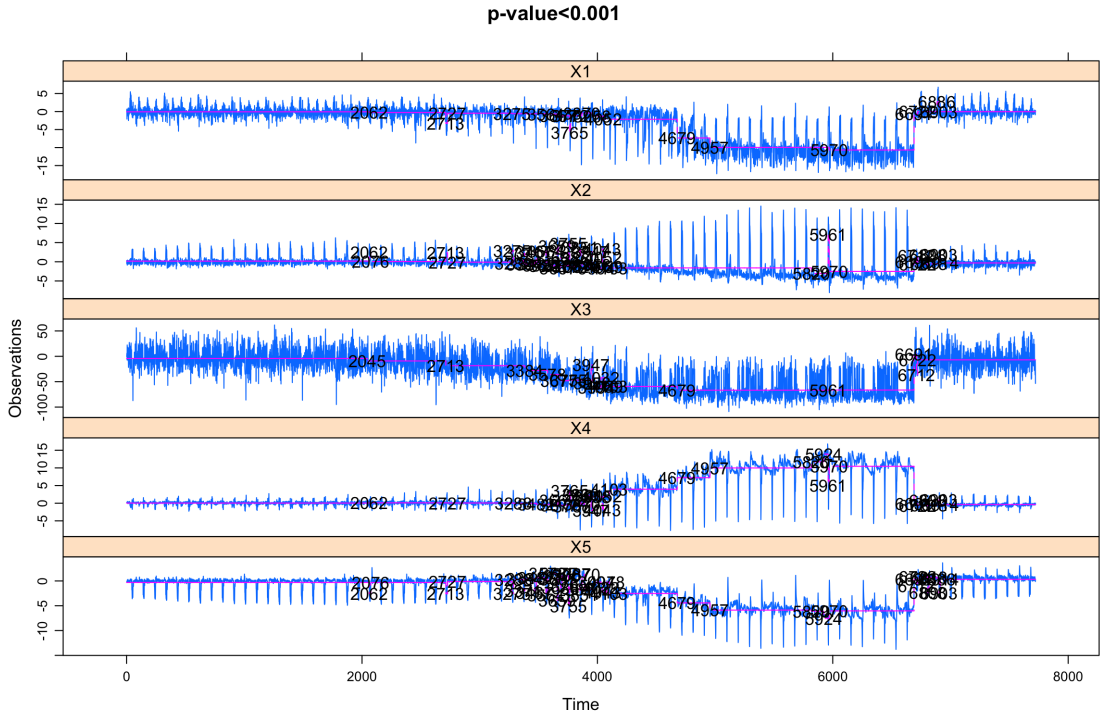


Figura 5.14: Gráfico de control multivariante de la fase I para las cinco variables.

Observamos que en general detecta bastante bien las zonas en las que la distribución empieza a variar en todas las variables. Como ya se explicó, al tratarse de medidas individuales, solo se tratan desviaciones de paso. En comparación con los gráficos individuales anteriores, vemos que aquí, antes de cada variación, se señalan un número mayor de observaciones como fuera de control. Al tratarse de

un gráfico multivariante, todas las variables se analizan de manera conjunta, puede ser lo que permita detectar mayores desviaciones en cada variable y empezar a detectarlas antes.

De los 50 puntos de cambio detectados, 9 son anteriores a empezar a detectar la desviación. De estos el 2713 sí sería relevante porque marca una desviación visible en una de las variables. El 3371 y el 3384 también porque empiezan a marcar el cambio. Los siguientes 41 puntos sí aportan información interesante para un análisis de la fase I, quizá sobrarían alguno de los últimos. Con esto, de los 50 puntos detectados por el gráfico, la proporción de utilidad es 0.82.

Capítulo 6

Conclusiones y líneas futuras.

En este trabajo comenzamos describiendo el problema planteado por la empresa y las distintas soluciones que se propusieron.

De las soluciones que se probaron para conseguir un método automático para la selección de fechas de entrenamiento y calibrado, la que utiliza reglas físicas fue la peor. Con esta no se consiguieron resultados y fue descartada.

El método de comparación por grupos si aportó mejores resultados. No evita la intervención humana en el proceso, pero reduce su tiempo. Se redujeron en un 70 % el número de gráficas que había que supervisar para seleccionar las fechas.

Por último, el método de comparación por características fue el de mayor interés para la empresa, ya que a futuro permite prescindir de la intervención humana. De momento aún sigue siendo necesaria en algunos casos, ya que el json aún no está completo de todo. Hay que tener en cuenta, a parte de las dos características mencionadas, el número de variables con las que se entrenó el modelo de xgboost. No siempre se dispone de las mismas variables en todas las instalaciones. Aun así, con este método se disminuye casi totalmente el factor humano.

Por otra parte, se han estudiado varios métodos de control estadístico de la calidad, viendo su aplicación para la empresa Cofrico.

Con el estudio realizado sobre los gráficos EWMA se vio que estos aportan información interesante para el control de las variables, y complementan los que se utilizan actualmente, que son los MEWMA. El monitorizado de una variable permitiría conocer mejor el tipo de fallo que se produce, y los resultados del gráfico univariante son buenos. Se seleccionaron además de las variables que se utilizan para el control multivariante, las que tuvieron mejores resultados univariantes, y que serían por tanto las características críticas para la calidad.

Entre otros métodos utilizados están distintos gráficos de control multivariantes, diferentes a los actuales. Con esto se quería ver si otras alternativas ofrecen mejores resultados, sobre todo los procedimientos no paramétricos, teniendo en cuenta que los datos no cumplen las hipótesis de normalidad. En esta línea se ha visto que los mejores resultados se obtienen con el gráfico de Hotelling, que es paramétrico, y también con el gráfico r con la profundidad de Mahalanobis.

Por último, dentro del control estadístico de la calidad tendríamos los procedimientos de análisis para la fase I. Como estos métodos detectan zonas de cambio, para la empresa podrían servir como otra solución para el problema planteado de selección de las muestras de entrenamiento y calibrado. Se vio que estos métodos detectaron bien las zonas de cambio en las que se producía el fallo, lo que nos permitiría descartarlas al seleccionar las muestras. El inconveniente que presenta esto es que el paquete solo está disponible en R, pero en la empresa se trabaja con Python. Como línea futura se propone implementar también en Python los paquetes disponibles ya en R.

Durante el desarrollo de este trabajo surgieron otras ideas que quedaron pendientes de desarrollo. Entre ellas está realizar un estudio de la fiabilidad y del tiempo entre fallos; también el estudio de técnicas de inteligencia artificial para la detección de anomalías, como los propuestos por Huang y Wu

(2022) [13] y Tritscher et al (2023) [21].

Bibliografía

[1] Salvador Naya y Javier Tarrío Saavedra (2020). Apuntes Control Estadístico de la Calidad. Máster en Técnicas Estadísticas. Universidade da Coruña.

[2] Capizzi G, Masarotto G (2018). Phase I Distribution-Free Analysis with the R Package dfphase1. in Frontiers in Statistical Quality Control 12, eds: Sven Knoth and Wolfgang Schmid. doi:10.1007/978-3-319-75295-2_1

[3] Capizzi G. , Masarotto G. (2013), Phase I Distribution-Free Analysis of Univariate Data. Journal of Quality Technology, 45, pp. 273-284, doi: 10.1080/00224065.2013.11917938.

[4] Capizzi G. and Masarotto G. (2017), Phase I Distribution-Free Analysis of Multivariate Data, Technometrics, 59, pp. 484–495, doi: 10.1080/00401706.2016.1272494.

[5] Chakraborti S. & Graham M A (2019): Nonparametric (distribution-free) control charts: An updated overview and some results, Quality Engineering

[6] Colaboradores de Wikipedia. Cuarta Revolución Industrial. Wikipedia, La enciclopedia libre. https://es.wikipedia.org/w/index.php?title=Cuarta_Revolución_Industrial#. Accedido 4 de junio de 2023.

[7] Domínguez Basteiro Manuel (2022). Detector de fallos en maquinaria mediante Técnicas Estadísticas y Machine Learning. Trabajo de Fin de Máster. Máster en Técnicas Estadísticas.

[8] Fernández Casal R, Costa Bouzas J, Oviedo de la Fuente O (2021). Aprendizaje Estadístico. https://rubenfcasal.github.io/aprendizaje_estadistico/.

[9] Flores M, Fernandez Casal R, Naya S and Tarrio-Saavedra J (2022). qcr: Quality Control Review. R package version 1.4. https://CRAN.R-project.org/package=qcr

[10] Fraiman R, Liu R Y, and Meloche J. Multivariate density estimation by probing depth. Lecture Notes-Monograph Series, pages 415–430, 1997.

[11] Hotelling, H. (1947). "Multivariate Quality Control," Techniques of Statistical Analysis, Eisenhart, Hastay, and Wallis (eds.), McGraw-Hill, New York.

[12] Healy JD. A note on multivariate CUSUM procedures. Technometrics. 1987; 29: 409-412.

[13] Huang Z and Wu Y, .ª Survey on Explainable Anomaly Detection for Industrial Internet of Things,"2022 IEEE Conference on Dependable and Secure Computing (DSC), Edinburgh, United Kingdom, 2022, pp. 1-9, doi: 10.1109/DSC54232.2022.9888874.

[14] Liu. R Y (1995) Control Charts for Multivariate Processes, Journal of the American Statistical Association, 90:432, 1380-1387.

[15] Lowry, C. A., W. H. Woodall, C. W. Champ, and S. E. Rigdon (1992). "A Multivariate Exponentially Weighted Moving Average Control Chart," Technometrics, Vol. 34(1), pp. 46–53.

[16] Mahalanobis P C, On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India, pp. 49–55, 1936.

[17] Montgomery D. C. (2005/09) Introduction to Statistical Quality Control. Wiley.

[18] Reynolds M R, Lu Chao-Wen (1997), Control charts for monitoring processes with autocorrelated data. Nonlinear Analysis: Theory, Methods & Applications, Volume30, pp 4059-4067.

[19] Roberts, S. W. (1959). "Control Chart Tests Based on Geometric Moving Averages," Technometrics, Vol. 42(1), pp. 97–102.

[20] Tukey J W, Mathematics and the picturing of data. In Proceedings of the international congress of mathematicians, volume 2, pages 523–531, 1975.

[21] Tritscher J, Krause A and Hotho A (2023) Feature relevance XAI in anomaly detection: Reviewing approaches and challenges. Front. Artif. Intell. 6:1099521. doi: 10.3389/frai.2023.1099521

[22] Valipour M, Banihabib M E, Reza Behbahani S M (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. Journal of Hydrology 476, pp 433-441.

Código R utilizado

```
library(qcc)
#EWMA
grupos<-1:7722
grupos<-as.vector(grupos)
x89<-as.vector(t(nonandatos['89']))
res89 <- qcc.groups(x89, grupos)
ewma89 <- ewma(res89[1:2200], lambda=0.95, nsigmas=33,
newdata=res89[2201:7722],restore.par=FALSE)
desviaciones<-ewma98$violations

#MEWMA
library(qcr)
data.mqcd <- mqcd(nonandatos[1:2200,])
res.mqcs <- mqcs.mewma(data.mqcd, lambda=0.95)
plot(res.mqcs, title =" MEWMA")
data.mqcd.new <- mqcd(nonandatos[2201:7722,])
Xmv=matrix(res.mqcs$mean,byrow=TRUE)
S = data.frame(matrix(res.mqcs$S,byrow=FALSE,nc=5))
res.mqcs.new <- mqcs.mewma(data.mqcd.new,Xmv=Xmv,S=S, lambda=0.95)
plot(res.mqcs.new, title ="MEWMA")
desviaciones<-res.mqcs.new$violations

#Hotelling
entrenamiento<-nonanresiduos[1:2200,]
test<-nonanresiduos[2201:7722,]
qq = mqcc(entrenamiento, type = "T2.single",
      pred.limits = TRUE,newdata=test, confidence.level=0.99999)
desviaciones<-qq$violations$beyond.pred.limits

#r-chart
matriz<-as.matrix(nonanresiduos)
E<-matriz[1:2200,]
t<-matriz[2201:7722,]
data <- npqcd(t, E)
# -- Tukey
res.npqcs <- npqcs.r(data, method = "Tukey", alpha = 0.005)
plot(res.npqcs, title = " r Control Chart")
res.npqcs <- npqcs.r(data, method = ""Mahalanobis", alpha = 0.005)
res.npqcs <- npqcs.r(data, method = ""LD", alpha = 0.005)

#MCUSUM
```

```
data.mqcd <- mqcd(nonandatos[1:2200,])
res.mqcs <- mqcs.mcusum(data.mqcd, k=3, h=8)
plot(res.mqcs, title =" MCUSUM")
data.mqcd.new <- mqcd(nonandatos[2201:7722,])
Xmv=matrix(res.mqcs$mean,byrow=TRUE)
S = data.frame(matrix(res.mqcs$S,byrow=FALSE,nc=5))
res.mqcs.new <- mqcs.mcusum(data.mqcd.new,Xmv=Xmv,S=S, k=3, h=8)
plot(res.mqcs.new, title =" MCUSUM")
desviaciones<-res.mqcs.new$violations

#analisis tp, tn, fp, fn
fn=0
fp=0
tp=0
tn=0
for (i in 1:length(desviaciones)){
  posicion<-desviaciones[i]
  if (posicion<=2200){
    next
  }
  if (posicion<3456){
    fp=fp+1
  } else{
    if(posicion<6850){
      tp=tp+1
    } else{fp=fp+1}
  }
}

x<-1:7722
no_desviaciones<-c()
for(i in 1:7722) {
  a=0
  equis<-x[i]
  for(j in 1:length(desviaciones)) {
    if (equis==desviaciones[j]){
      a=1
    }
  }
  if(a!=1){
    no_desviaciones[(length(no_desviaciones) + 1)] <- equis
  }
}

for (i in 1:length(no_desviaciones)){
  posicion<-no_desviaciones[i]
  if (posicion<=2200){
    next
  }
  if (posicion<3456){
    tn=tn+1
  } else{
```

```
    if(posicion<6850){
      fn=fn+1
    } else{tn=tn+1}
  }
}

library(dfphase1)
nonan96<-nonandatos['X96']
rsp(t(nonan96), plot = TRUE, L = 1000, seed = 11642257, alpha = 0.05,
    maxsteps= min(50, round(NROW(nonan96)/15)), lmin = min(10, round(NROW(nonan96)/10)))

data<-as.matrix(nonandatos)
l<-dim(data)[1]
x<-as.array(t(data),c(5,l))
mod<-mphase1(x, plot = TRUE, post.signal = TRUE, isolated = FALSE, step = TRUE,
             alpha = 0.001, gamma = 0.8,K=50,
             lmin = 5, L = 100, seed = 11642257)

entrenamiento<-residuos[1:2304,]
entrenamiento96<-residuos[1:2304,"96"]

# S-chart
library(DepthProc)
matriz<-as.matrix(residuos_sinnan)
Entrenamiento<-matriz[1:2300,]
Test<-matriz[2301:7722,]

depth_entrenamiento<-c()
for (j in 1:2300){
  Yi=Entrenamiento[j,]
  d_yi=depthTukey(Yi,Entrenamiento)[1]
  depth_entrenamiento[(length(depth_entrenamiento) + 1)] <- d_yi
  print(j)
}
depth_test<-c()
m=dim(Entrenamiento)[1]
rangos<- c()
for (i in 1:5422){
  rango=0
  y=Test[i,]
  d_y=depthTukey(y,Entrenamiento)[1]
  depth_test[(length(depth_test)+1)]<-d_y
  for (j in 1:2300){
    d_yi=depth_entrenamiento[j]
    if (d_y>=d_yi){
      rango=rango+1
    }
  }
  rango=rango/m
  rangos[(length(rangos) + 1)] <- rango
  print(i)
}
```

```
#Q-chart
Q_est2<-c()
for (i in 1:5422){
  suma=0
  r_i=rangos[1:i]
  for (j in 1:length(r_i)){
    suma=suma+(r_i[j]-0.5)
  }
  suma<-suma/(sqrt(i^2*(1/2300+1/i)/12))
  Q_est2[(length(Q_est2) + 1)] <- suma
  print(i)
}
x<-1:5422
plot(Q_est2~x, type="l")
plot(Q_est[288:576]~x[288:576], type="l")
abline(h=-3.090232, col="red") #alpha=0.001
plot(Q_est~x, type="l")
abline(h=-3.090232, col="red") #alpha=0.001
```