Universidad de Valladolid

ESCUELA DE INGENIERÍAS INDUSTRIALES

UNIVERSIDAD DE VALLADOLID

ESCUELA DE INGENIERIAS INDUSTRIALES

Grado en Ingeniería Mecánica

Diagnóstico de anomalías basadas en técnicas de manifold learning y control estadístico de procesos para mejora de la calidad.

Autor:

Caminero Marcos, Alicia

Tutor(es):

De la Fuente Aparicio,
María Jesús.
Departamento de Ingeniería de Sistemas y Automática

Valladolid, Noviembre 2022.

ÍNDICE DE CONTENIDOS

RESUMEN

Este trabajo tiene como objetivo mejorar la calidad de un proceso en una planta industrial mediante distintas técnicas de detección y diagnóstico de fallos (FDD) basadas en datos. Esto es debido a que, con la cuarta revolución industrial, también llamada Industria 4.0, se ha dado paso a la automatización de los procesos industriales utilizando las últimas tecnologías. Como consecuencia, surge la recogida masiva de datos a lo largo de dichos procesos, utilizando sensores que nos permitan la lectura las variables o factores deseados.

En este trabajo se pretende estudiar diferentes métodos que permiten el tratamiento de los datos recogidos en un proceso industrial, y poder detectar los fallos que potencialmente puedan suceder. Son técnicas basadas en reducir la dimensión del espacio de datos inicial de manera lineal, lo que nos permitiría trabajar con menor número de variables, las cuales emplearemos en analizar el comportamiento de la planta industrial.

Por un lado, se utilizarán métodos de análisis estadístico multivariante como el Análisis de Componentes Principales (PCA), el Análisis Canónico de Correlación (CCA) y el Análisis Canónico de Correlación Concurrente (CCCA). Y por otro lado se aplicará un algoritmo de aprendizaje no supervisado llamado Locally Linear Embedding o LLE que reduce la dimensión de manera lineal conservando la vecindad de los datos iniciales.

Los datos que se van a utilizar para la aplicación de los métodos mencionados son obtenidos de la planta química Tennessee Eastman, donde su proceso es muy utilizado por la comunidad científica como banco de pruebas.

Finalmente se procederá a realizar una comparación entre los diferentes métodos, tanto desde el punto de vista de resultados como de metodología, además se realiza un breve estudio de trabajo futuro donde se plantean mejoras u otras líneas de acción asociadas a este estudio.

ABSTRACT

The fourth industrial revolution, also called Industry 4.0, has given way to the automation of industrial processes using the latest technologies. Consequently, massive data collection arises throughout these processes, using sensors that allow us to read the desired variables or factors. Due to this, the need to maintain control over the process has been created to guarantee its quality, and the service offered or the final product.

In this academic work we intend to study the different methods that allow the treatment of the data collected from the process to be able to detect possible failures. They are techniques based on reducing the dimension of the initial data space linearly. This would allow us to work with fewer variables, what will help us to analyze the behavior of the industrial plant.

On the one hand, Multivariate Statistical Analysis Methods (MSCP) such as Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA) and Canonical Concurrent Correlation Analysis (CCCA) will be used. And on the other hand, an unsupervised learning algorithm called Locally Linear Embedding or LLE will be used, which reduces the dimension linearly while preserving the neighborhood of the initial data.

The data that will be used for the application of the mentioned methods are obtained from the Tennessee Eastman chemical plant, where its process is widely used by the scientific community as a testing benchmark.

Finally, a comparison will be made between the different methods, analyzing the results and the methodology, as well as a brief future work study text where improvements or other lines of action associated with this study are proposed.

CAPÍTULO I: INTRODUCCIÓN

1.1. INTRODUCCIÓN

La globalización y la alta demanda hacen que los procesos industriales sean cada vez más complejos y los tiempos de fabricación sean menores, sin renunciar a la calidad del producto o servicio final, ya que se deben satisfacer las necesidades actuales de los clientes. Por ello el control de la calidad en un proceso productivo es esencial en la industria actual, o como hoy en día se denomina, Industria 4.0.

A través del control de calidad se pretende mantener factores imprescindibles del proceso en un rango de valores, lo cual nos permite garantizar que el proceso es seguro y no suponga ningún riesgo, además de que el producto final cumpla con las especificaciones y exigencias del consumidor. De igual manera, la implementación de un sistema de calidad repercute en la economía de la empresa, aumentando su eficiencia y disminuyendo los productos de desecho y evitando tiempos de parada innecesarios.

Un buen control de la calidad conlleva que el proceso este cada vez más automatizado y se realicen las tareas de supervisión correspondientes. Como consecuencia de ello se emplean métodos de control estadístico multivariante, que buscan detectar el fallo de manera temprana, y de esa forma nos permita tener una pronta respuesta sobre el proceso y toma de decisiones para devolver el proceso a una situación de normalidad. Alguno de los métodos que se emplean son el Análisis de Componentes Principales (PCA), el Análisis Canónico de Correlación (CCA) y el Análisis Canónico de Correlación Concurrente (CCCA)

Estos métodos trabajan sobre las variables del proceso, obtenidas mediante multitud de sensores a lo largo de este, y monitoreadas ya sea cada cierto intervalo de tiempo o de manera continuada. Como la cantidad de variables en un proceso puede ser una cifra bastante alta, estos métodos se encargan de reducir la dimensión de los datos iniciales, de manera lineal, sin perder información, y de esta manera simplificar el tratamiento de estos datos. Complementariamente, se emplean dos estadísticas, una para monitorizar la variabilidad del espacio reducido por el método correspondiente llamada estadística de Hotelling o T^2, y la otra llamada Q o SPE que analiza el residuo formado por reducir la dimensión del espacio inicial.

También existen otras técnicas, en especial para datos de alta dimensión, basadas en el aprendizaje múltiple o “manifold learning”, como por ejemplo el método LLE (Locally Linear Embedding), un algoritmo de aprendizaje automático que busca la proyección de los datos no lineales a una dimensión menor mediante la búsqueda de vecinos cercanos.

1.2. OBJETIVOS

Con la realización de este trabajo se pretende analizar un proceso industrial con la finalidad de detectar los posibles fallos que pueden suceder y poder tratarles de forma temprana, para que el daño producido por estos sea lo más leve posible económica y materialmente, y como consecuencia, mejorar la calidad del proceso.

Para ello emplearemos diferentes métodos de análisis estadístico multivariante como el Análisis de Componentes Principales (PCA), también en su versión dinámica (DPCA), el Análisis Canónico de Correlación (CCA), añadiendo su versión regularizada y también dinámica (DCCA), y finalmente el método de Análisis Canónico de Correlación Concurrente (CCCA). Además, se incluirá una herramienta llamada Locally Linear Embedding (LLE) complementaria a cada uno de los métodos y que potencialmente puede mejorar el resultado de estos.

Con los resultados obtenidos, se realizará una comparativa entre los diferentes métodos y finalmente se obtendrá la conclusión de cual o cuales son las técnicas que mejor se adaptan al proceso.

Los datos que utilizaremos en este estudio son obtenidos de la planta química Tennessee Eastman, un proceso utilizado ampliamente por la comunidad científica como banco de pruebas.

1.3. ORGANIZACIÓN DE LA MEMORIA

La organización que se ha adoptado para presentar este trabajo consta de 6 capítulos:

El primer capítulo consta de una pequeña introducción sobre el tema a tratar posteriormente, aportándole un marco contextual y explicando los diferentes objetivos que va a conllevar la realización de este trabajo.

El segundo capítulo, explica de forma teórica la metodología de los diferentes métodos de análisis estadístico multivariante que se van a desarrollar en este estudio, además de enmarcar históricamente su origen y su posterior desarrollo hasta lo que conocemos hoy en día. Además, se hablará sobre una herramienta que puede ser añadida al método con la capacidad de mejorar los resultados de este.

En el tercer capítulo se explica el proceso Tennessee Eastman, su historia y porque se ha elegido como banco de pruebas para este trabajo. También se presentan las diferentes variables que forman parte de dicho proceso y los fallos que posteriormente serán analizados.

En el cuarto capitulo, se presenta la aplicación práctica de cada uno de los diferentes métodos de análisis estadístico multivariante, su metodología paso a paso para finalmente realizar la detección de los diferentes fallos. También se muestran los resultados obtenidos de las diferentes estadísticas al aplicar cada uno de los métodos.

En el quinto capitulo, se manifiestan las conclusiones en base a los resultados obtenidos de los diferentes métodos, aplicados en este trabajo, y herramientas añadidas que potencialmente pueden mejorar dichos métodos. De igual modo, se dan unas sugerencias que pueden llevarse a cabo en estudios futuros relacionados con este tema.

Por último, se muestra la bibliografía que se ha consultado para la realización de este trabajo.

CAPÍTULO II: MARCO TEÓRICO

2.1. CONTROL DE CALIDAD

La Industria que hoy en día conocemos, cuenta con departamentos y personal cualificado en cada área de su proceso productivo, para controlar la calidad del proceso, del producto o del servicio que ofrezcan.

Para obtener un buen producto final se deben cumplir tanto las especificaciones del producto o servicio como aquellas acordadas con el cliente, así como las normativas de seguridad tanto en los puestos de trabajo como a lo largo del proceso hasta la salida de la fábrica, esto garantiza que el proceso sea seguro. Además, es beneficioso que tanto la planta como el producto se mantengan en óptimas condiciones y que sean económicamente viables. [1].

El control de calidad nos ayudará, en última instancia, en la toma de decisiones sobre el proceso ante un posible estado anómalo, pero para ello, primero se deben detectar los fallos, lo que nos lleva a la monitorización del proceso, y segundo, el análisis de estos. [1]. Para obtener un buen control de la calidad, es primordial que los instrumentos de medida sean apropiados a cada variable, además de contar con una excelente red de comunicación industrial, lo que nos facilitará la temprana detección del fallo y su correspondiente diagnóstico y así poder intervenir lo antes posible.

2.1.1. ENFOQUE HISTÓRICO

El control de la calidad ha existido desde el inicio de la historia, cuando el hombre seleccionaba aquellos productos que podía consumir y los que no, pasando por los gremios de la Edad Media donde ofrecer productos de calidad era la base del mercado y para ello los aprendices pasaban largos periodos de su vida aprendiendo el oficio a cargo de los artesanos.

A finales del siglo XIX, con la llegada de la Revolución Industrial surgió la producción en serie, dando lugar a los procesos industriales, y la especialización del trabajo. Todo esto, trajo consigo la necesidad de implementar sistemas de control de calidad para el producto final y con ello, mejorar la calidad del proceso. La gran mayoría de productos de aquella época eran simples, por lo que el control de la calidad la llevaba a cabo el propio operario. Más tarde llegó la Segunda Guerra Mundial y con ello, la estandarización de los procesos de producción, debido a la imposibilidad de satisfacer individualmente a cada cliente ya que la demanda cada vez era más alta y los productos más complejos. Es aquí cuando surge la figura de empleado dedicado al control de calidad, y los operarios dejan de realizar dicha función. Cabe destacar, que en esta época el control de la calidad se basaba en la inspección del producto final.

El matemático Walter Shewhart, diseña en 1924, una gráfica estadística que permite controlar todas las variables de un producto, lo cual permitió controlar la calidad de la producción en serie a menor costo. El objetivo de este método consistía en elevar la productividad y disminuir los errores, aplicando la estadística de manera eficiente y obteniendo como resultado una mejora del costo-beneficio en las líneas de producción.

Mas tarde, estalló la Segunda Guerra Mundial, donde el control de la calidad pasa a ser fundamental, sobre todo en la industria bélica. Se realizaron estudios sobre cómo elevar la calidad aplicando el control estadístico del proceso, lo que llevo a los norteamericanos a crear un sistema de certificación de la calidad, es decir, un aseguramiento de la calidad. Actualmente esta certificación es conocida como ISO 9000. También se establecieron las primeras normas de calidad en el mundo, lo que permitió elevar los estándares de calidad drásticamente, disminuyendo el número de pérdidas de vidas humanas [2].

Hoy en día, en plena Cuarta Revolución Industrial, o Industria 4.0, la calidad ha evolucionado en la Gestión de Calidad Total. Esta nueva fase incluye lo ya existente hasta ahora; inspección, control de calidad y aseguramiento de la calidad e introduce la participación conjunta del consumidor y del proveedor. Se basa en la mejora continua de resultados en cada área de actividad de la empresa utilizando los recursos disponibles y a menor costo [3].

2.2. CONTROL ESTADÍSTICO DE PROCESOS (SPC)

El Control Estadístico de Procesos, más conocido como SPC (“Stadistical Process Control”), es una herramienta de Control de Calidad basado en técnicas y métodos estadísticos, que nos permiten controlar el proceso, mediante la monitorización de este y empleando gráficos de control, donde aparecen representadas las variaciones del proceso. La finalidad de estos métodos es determinar si el resultado de un proceso está acorde con el diseño del producto o servicio correspondiente [4]. El pionero en emplear gráficos de control, como hemos mencionado con anterioridad, fue Walter Shewart, en 1924.

Las variaciones anómalas son fácilmente localizables en los gráficos de control, lo que nos facilita una temprana respuesta al fallo, disminuyendo el posible daño causado tanto a la calidad del proceso como al producto final y la prevención de problemas graves. La correcta implementación de un SPC nos puede reducir el coste y el tiempo de producción [5].

2.2.1. VARIABILIDAD Y CAPACIDAD DEL PROCESO PRODUCTIVO

Todo proceso productivo sufre variaciones, estas pueden ser originadas por diferentes fuentes. Uno de los métodos más populares para identificar las causas de la variabilidad en un proceso es el diagrama causa-efecto de Kaoru Ishikawa, también conocido como método de las 6M o espina de pescado debido a su particular representación, que puede verse en la figura 1 [6].

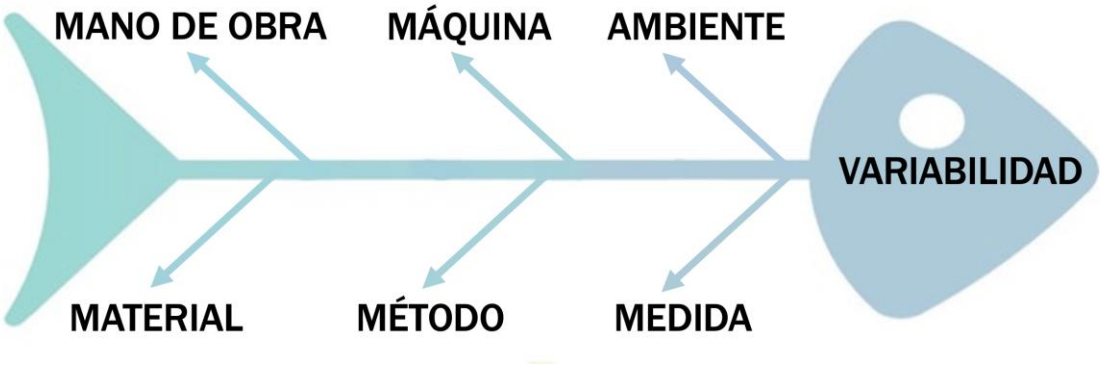

Figura 1. Diagrama de las 6 M's

En este diagrama de las 6M, el problema a solucionar sería la variabilidad del proceso, dividido en 6 elementos clave que serán analizados individualmente para finalmente, revelar cuales de ellos son los causantes del problema.

Una clasificación más sintetizada del origen de estas variaciones, y con la que vamos a trabajar en este estudio, es [1]:

- Causas comunes: son causas consustanciales al proceso, originan variaciones naturales, permanentes en el tiempo y no son evitables. Si las variaciones del proceso tienen únicamente esta fuente, el proceso es considerado bajo control estadístico, ya que dan lugar a una distribución estable y replicable en el tiempo.

- Causas especiales o asignables: son causas ajenas al proceso productivo debidas a una perturbación concreta de éste, provocando variaciones impredecibles de manera intermitente. Estas causas son las que generan los fallos que habrá que detectar. Si las variaciones del proceso además de tener carácter natural también se producen por causas especiales, el proceso se declara fuera de control.

En el caso de que las variables tengan origen natural, los datos obtenidos de las mediciones de estas, suelen agruparse en una distribución predecible y estable en el tiempo. Mientras que, si es una causa especial, las muestras

obtenidas tenderán a organizarse en distribuciones inesperadas e inestables como puede verse en la figura 2.

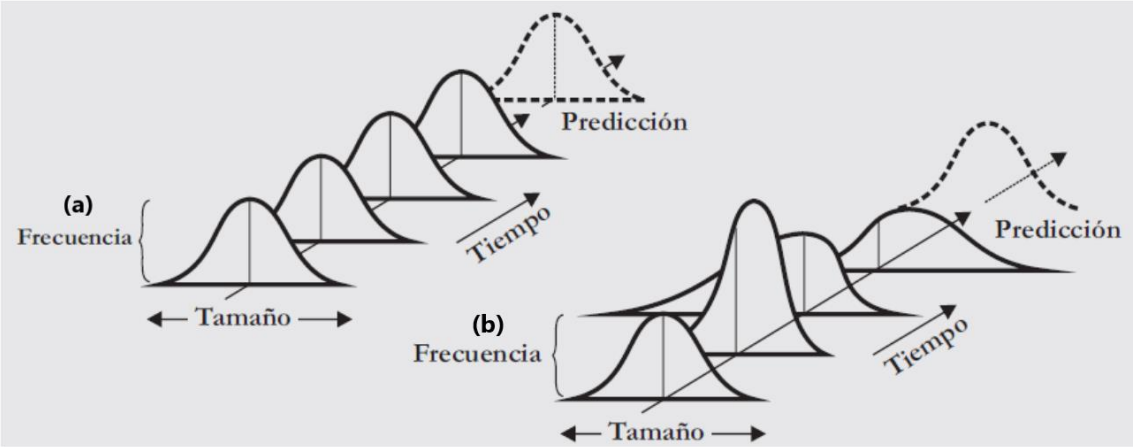

Figura 2. a) distribución estable y predecible debido a que las causas de variación son únicamente naturales. b) distribución inestable y no predecible ya que se presentan causas de variación asignables [4].

Para completar el estudio de la variabilidad de un proceso o producto, debemos determinar si este se encuentra dentro de las especificaciones establecidas. Para ello de definen unos límites que pueden ser naturales del proceso o establecidos voluntariamente por una serie de especificaciones (Figura 3).

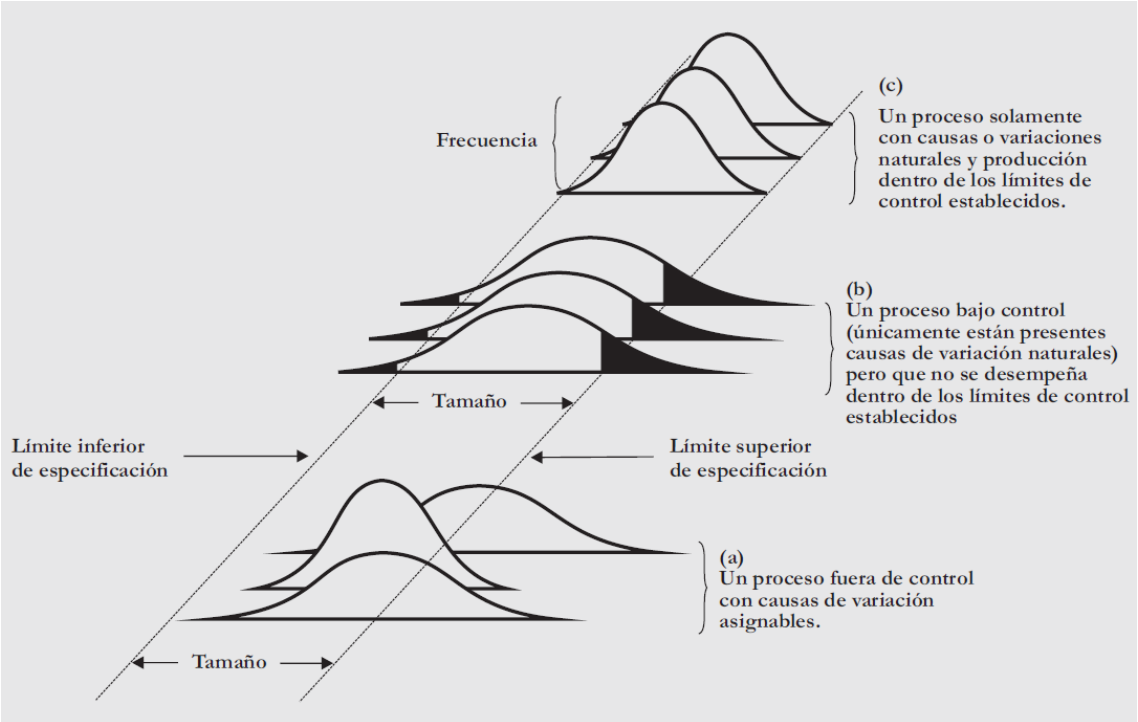

Figura 3. a), b) y c) tipos de salida del proceso [4].

En la Figura 3 se muestran tres tipos de salidas del proceso. La Figura 3 (a) muestra un proceso fuera de control debido a causas especiales de variación, que como se puede ver no cumple especificaciones de calidad. La Figura 3 (b) muestra un proceso bajo control ya que solo ocurren causas comunes de variación pero que se encuentra fuera de los límites de especificación, y por tanto el proceso se dice que no es capaz de producir con la calidad adecuada. La Figura 3 (c) muestra un proceso bajo control con causas comunes de variación y que se encuentra dentro de los límites de calidad establecidos, y por tanto se dice que es un proceso capaz de producir con la calidad adecuada.

2.2.2. GRÁFICOS DE CONTROL

Los gráficos de control son una herramienta que nos permite analizar estadísticamente el proceso productivo y detectar la presencia de factores especiales. Como se ha mencionado anteriormente, fue el Dr. Shewhart quien los desarrolló con la finalidad de estudiar si un proceso se encuentra o no bajo control estadístico.

El proceso se considera que está bajo control estadístico cuando sobre él solo actúan causas naturales, para definirlo así, se emplean los gráficos de control ya que nos permite visualizar cuando se debe intervenir en el proceso para modificar una tendencia no deseada [7].

Los gráficos de control Shewhart, muestran de manera ordenada en el tiempo los datos obtenidos de medir un parámetro de calidad. Para ello emplea una línea central donde está representada la media de las observaciones de la variable y dos límites de control, uno superior y otro inferior, generalmente a una distancia de 3 desviaciones estándar de la media, teniendo un 99,73% de probabilidad de que las muestras tomadas estén entre los límites de control. En la figura 4, podemos ver representado uno de estos gráficos.

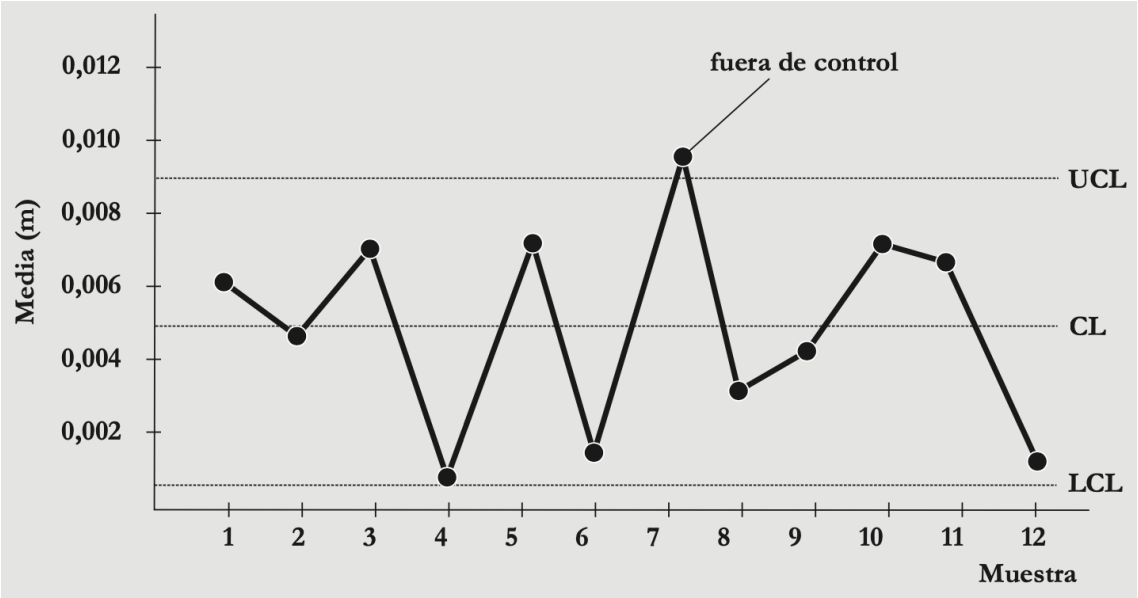

Figura 4. Gráfico de control Shewhart. Se representa la línea central indicando la media de los datos(CL) y los límites de control tanto superior (UCL) como inferior (LCL) [4].

Además, estos gráficos suelen tener otros elementos opcionales a mayores, como, por ejemplo, los limites superiores e inferiores de advertencia, situados normalmente a dos desviaciones estándar de la línea central, de esta manera se alerta de que el proceso se está alejando de la normalidad permitiendo tomar medidas preventivas. En la figura 5 que se muestra a continuación, se muestran varios tipos de límites y su distancia a la línea central.

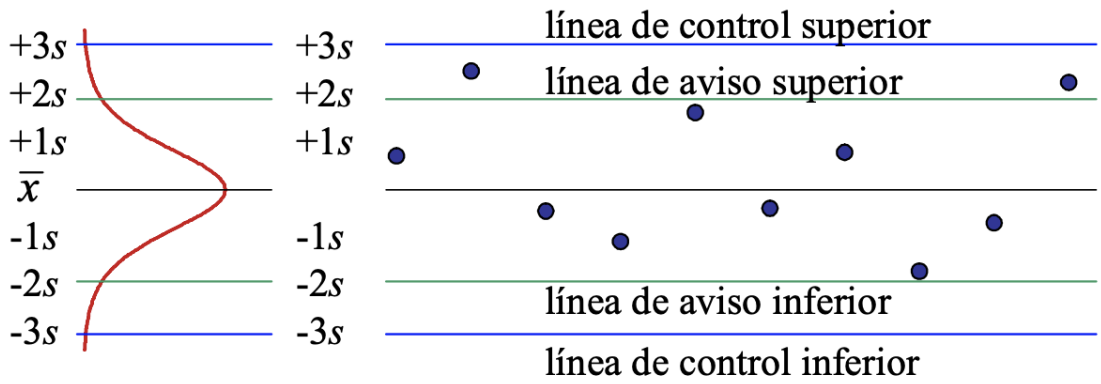

Figura 5. Límites de aviso y de control en un gráfico de control [8].

El objetivo principal de los gráficos de control es detectar las muestras que están fuera de los límites de control, cuando esto sucede, se debe investigar la causa especial de la alarma. Además, estos diagramas, permiten detectar sucesiones de puntos que podrían crear problemas en la calidad del proceso, por lo que es común diseñar una serie de normas a observar y su

correspondiente respuesta en el proceso. Por ejemplo, nueve puntos consecutivos por debajo o por encima de la línea media, se deben localizar los productos potencialmente afectados, realizar estudios sobre ellos con el fin de decidir si siguen adelante en el proceso o no y localizar la causa de esta tendencia para ponerla solución [9].

Existen otros gráficos como los CUSUM y los EWMA, que emplean toda la información aportada por las observaciones de las variables y aplican la acumulación de datos. Son una alternativa a los gráficos de Shewhart, donde solo se utiliza la información que aporta la última observación [7].

2.2.3. CONTROL ESTADÍSTICO MULTIVARIANTE (MSPC)

Hasta ahora, las técnicas mencionadas para el análisis estadístico de un proceso han sido de carácter univariable, es decir, un único factor es analizado con independencia del resto. Sin embargo, a lo largo de un proceso productivo existen multitud de variables, que no se pueden tratan de manera individual, ya que muchas de ellas están relacionadas entre sí.

El Control Estadístico Multivariante o MSPC, es un conjunto de métodos que permite analizar y estudiar estadísticamente un conjunto de variables de un proceso en bloque. Es posible que, al observar gran cantidad de variables, parte de la información obtenida de estas es redundante, por lo que se suelen aplicar métodos que reduzcan la dimensión de estas.

2.3. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El Análisis de Componentes Principales o PCA (“Principal Component Analysis”), es un método multivariante de análisis estadístico, que se incluye dentro de los métodos de reducción de la dimensión. Se aplica cuando existe un número elevado de variables en el proceso y mediante proyecciones ortogonales, persigue reducir la dimensión de estas a un conjunto de variables más pequeña linealmente no correlacionadas llamadas componentes principales. Las componentes principales son combinación de las variables originales del proceso y mantiene la estructura de correlación de las mismas [10].

Se puede explicar la reducción de las m dimensiones del espacio original a las a dimensiones finales de manera geométrica, como se muestra en la figura 6. Para ello ajustaremos los datos originales sobre la figura de un elipsoide, donde cada eje es un vector que define una componente principal. Los ejes de menor tamaño (línea verde) presentan una baja varianza y por tanto la información que aportan es mínima, mientras que los ejes de mayor tamaño (línea roja) aportan gran cantidad de información del proceso.

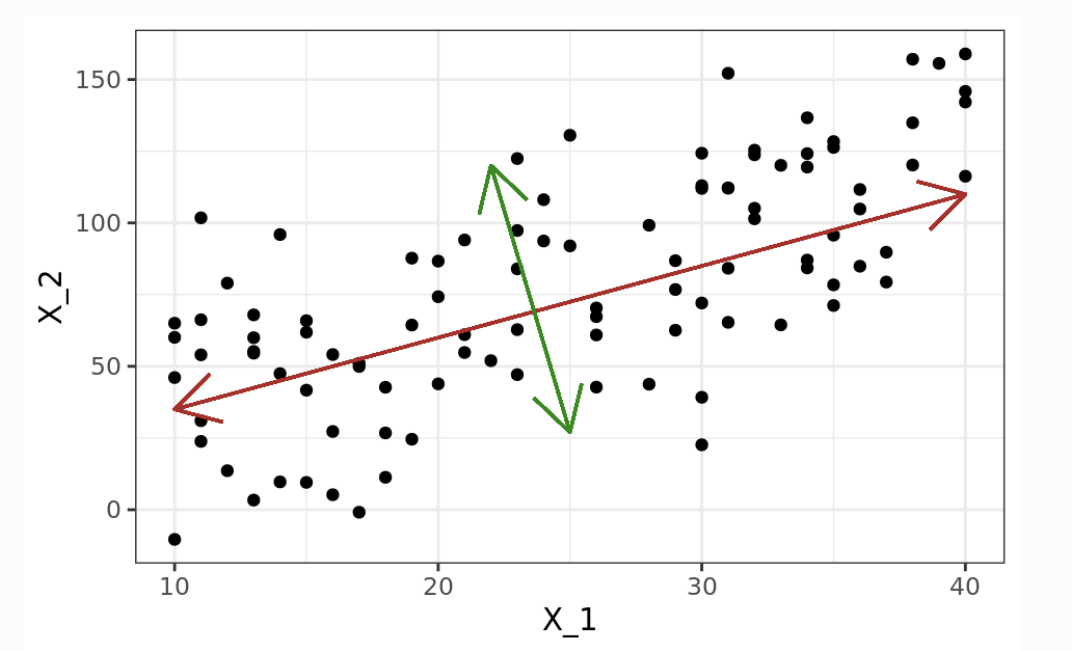

Figura 6. Ejes de una elipse perteneciente a un conjunto de datos con dos variables [11].

Finalmente, se escogen los componentes principales que explican la mayor parte de la variabilidad de los datos originales, obteniendo así un espacio de trabajo de a dimensiones menor que el original [12].

Antes de aplicar este método, se necesita un conjunto de datos de comportamiento normal, es decir, datos del proceso sin variables con causas especiales. Además, habría que realizar un pretratamiento de estos datos, eliminando aquellos que provengan de una mala medida y que no sean representativos. Las variables deben normalizarse a media cero y varianza uno, de esta manera todas las variables estarán en la misma escala para un correcto análisis. Para normalizar, se resta a cada variable su media y se divide entre su desviación estándar [5].

Los datos que se van a analizar forman una matriz X ϵ R^{nxm}, siendo n el número de observaciones de cada una de las m variables del proceso.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & & x_{2m} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad (1)$$

A partir de esta matriz, se calcula la matriz de correlación R, que tendrá valores uno en la diagonal y una dimensión de mxm:

$$R = \frac{1}{n-1} \cdot X^t \cdot X \tag{2}$$

Sobre R aplicaremos la descomposición de valores singulares y definiremos dos matrices; la primera será una matriz diagonal $\Lambda \in R^{mxm}$ que contiene los valores propios reales positivos y en orden decreciente ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$), siendo $\lambda_i = \sigma_i^2$. La segunda será una matriz ortogonal $V \in R^{nxm}$ que albergará los vectores propios de R, llamados 'scores' en PCA [13].

Una vez obtenidas estas matrices con los datos iniciales, pasamos a reducir la dimensión del espacio de trabajo, para ello utilizaremos los ''a'' valores propios más grandes, hasta que la varianza acumulada sea superior o igual a un cierto porcentaje deseado, como por ejemplo un 90%.

$$\frac{\lambda_1}{traza(\Lambda)} \cdot 100 + \frac{\lambda_2}{traza(\Lambda)} \cdot 100 + \frac{\lambda_3}{traza(\Lambda)} \cdot 100 + \cdots \geq 90 \tag{3}$$

Estos ''a'' valores propios empleados en la suma acumulada de la varianza, se almacenarán de manera decreciente en la matriz $S_a \in R^{axa}$ y de la misma manera se guardarán en la matriz $P \in R^{mxa}$ los vectores propios pertenecientes a los ''a'' valores propios. Con esta última matriz, calcularemos la matriz de componentes principales T, que no es más que los datos originales proyectados en este espacio de dimensión reducida.

$$T = X \cdot P \tag{4}$$

Con las matrices T y P podemos volver al espacio dimensional original.

$$\hat{X} = T \cdot P^t \tag{5}$$

Y definimos la matriz de residuos E como la diferencia entre X y $\hat{X}$.

$$E = X - \hat{X} \tag{6}$$

Finalmente, podemos deducir que, el conjunto de datos originales se compone por los componentes principales y por un residuo o ruido.

$$X = \hat{X} + E \tag{7}$$

Otro de los métodos de análisis multivariante que aplicaremos en este trabajo es DPCA ("Dynamic Principal Component Analysis"), se trata de aplicar PCA a

un conjunto de datos dinámicos que formarán una nueva matriz X a la cual también se normalizará a media cero y varianza uno.

Los datos dinámicos los obtenemos a partir de los datos originales, que son las variables medidas en el instante actual, y de los datos desplazados correspondientes a instantes de tiempo pasados o retardos. Construiremos la matriz X concatenando la matriz de datos originales con las matrices de datos con retardos:

$$X = [X_t\ , X_{t-1}\ , \dots , X_{t-i}] \tag{8}$$

Siendo t el instante actual e i el número de instantes de retardo.

2.3.1. ESTADÍSTICAS EMPLEADAS PARA LA MONITORIZACIÓN DEL PROCESO MEDIANTE PCA

Para monitorizar el proceso a partir del Análisis de Componentes Principales o PCA, se emplean distintas estadísticas, que establecen el umbral de fallo y permiten ver el estado del proceso de forma univariable mediante una gráfica.

Primeramente, aplicaremos la estadística de Hotelling (T^2):

$$T^2 = x \cdot P \cdot \Lambda_a^{-1} \cdot P^t \cdot x^T \tag{9}$$

Donde x ϵ R^{1xm}, es decir, cada una de las filas de X y Λ_a contiene las a primeras filas y columnas de Λ, siendo a el número de componentes principales.

Para saber si se ha producido un fallo, esta estadística tiene que superar un umbral:

$$T_\alpha^2 = \frac{(n^2-1)\cdot a}{n\cdot(n-a)} \cdot F_\alpha(a, n-a) \tag{10}$$

Donde n es el número de observaciones, a el número de componentes principales y $F_\alpha(a, n-a)$ es el valor correspondiente a la distribución de Fisher-Snedecor con n-a grados de libertad y siendo α el nivel de significancia, es decir, el valor de la probabilidad de obtener falsas alarmas, que en este trabajo será $\alpha = 0{,}01$ [13].

A continuación, aplicaremos la estadística Q o también conocida como SPE, la cual monitoriza las variables restantes a las componentes principales, es decir las variables residuo.

$$Q = r^t \cdot r \tag{11}$$

Siendo r el vector de residuos:

$$r = (I - P \cdot P^t) \cdot x^t \tag{12}$$

Al igual que en la estadística anterior, se establece un umbral para determinar si el proceso está o no fuera de control.

$$Q_\alpha = \frac{\sigma_Q^2}{2 \cdot \mu_Q} \cdot Chi^2 \left(\alpha , 2 \cdot \frac{\mu_Q^2}{\sigma_Q^2} \right) \tag{13}$$

Donde μ_Q y σ_Q es la media y la desviación típica de Q respectivamente, y α el nivel de significancia.

Estas mismas estadísticas son aplicables al método DPCA.

2.4. ANÁLISIS CANÓNICO DE CORRELACIÓN (CCA)

El Análisis Canónico de Correlación o CCA ("Canonical Correlation Analysis") es otro método de análisis multivariante en el que se extrae la estructura de correlación multidimensional entre dos grupos de variables.

CCA se focaliza en llevar al máximo la correlación entre la calidad y los datos del proceso, optimizando las dimensiones existentes, pero dejando a un lado la magnitud de las variaciones de los datos o varianza. Además, sufre problemas de colinealidad, muy típica de los datos de un proceso. Para solventar este defecto se emplea el método CCA con regularización, del cual hablaremos más adelante.

Tanto CCA como PCA son métodos para el estudio estadístico multivariante de procesos industriales basado en variables de proceso y variables de calidad. A diferencia de los métodos basados en CCA, PCA solo es efectivo para analizar las variaciones en las variables del proceso (X), pero no es capaz de extraer información de las variables de calidad (Y), lo que puede generar alarmas producidas por perturbaciones que no afectan en las variables de calidad [14].

Al igual que en el método anterior, utilizaremos primeramente un conjunto de datos de comportamiento normal de n observaciones por cada m variables del proceso. De este espacio, obtendremos dos matrices, la primera sería la matriz de proceso $X \in R^{nxr}$ donde r sería el número de variables medidas del proceso, y la segunda sería la matriz de calidad $Y \in R^{nxp}$ siendo p el número de variables de calidad. Además, antes de aplicar CCA, hay que realizar un

pretratamiento de estas dos matrices de manera independiente normalizando los datos a media cero y varianza uno.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1r} \\ x_{21} & x_{22} & & x_{2r} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nr} \end{pmatrix} \tag{14}$$

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & & y_{2p} \\ \vdots & & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} \tag{15}$$

Para obtener los valores propios tanto de X como de Y, es necesario definir dos nuevas matrices cuadradas A y B. A estará formada a partir de X, por lo que $A \in R^{rxr}$ y $B \in R^{pxp}$ se obtendrá a partir de Y.

$$A = X^t \cdot X \tag{16}$$

$$B = Y^t \cdot Y \tag{17}$$

Definimos D_X como la matriz diagonal que alberga los autovalores de A y V_X la matriz que contiene los autovectores de A. De la misma manera, asignaremos D_Y para los valores propios de B y V_Y para los vectores propios de B [15]. Con estas matrices calculamos los siguientes factores:

$$\Sigma_{xx}^{-1/2} = V_X \cdot D_X^{-1/2} \cdot V_X^t \tag{18}$$

$$\Sigma_{yy}^{-1/2} = V_Y \cdot D_Y^{-1/2} \cdot V_Y^t \tag{19}$$

Para continuar construyendo el algoritmo de CCA, utilizamos los factores calculados para construir la matriz $Z \in R^{rxp}$, de la cual obtendremos los valores singulares.

$$Z = \Sigma_{xx}^{-1/2} \cdot X^t \cdot Y \cdot \Sigma_{yy}^{-1/2} \tag{20}$$

De la factorización en valores singulares obtenemos tres matrices, la primera es $U_Z \in R^{rxp}$ con columnas ortogonales, la segunda es la matriz diagonal $D_Z \in R^{pxp}$, y la última es la matriz ortogonal $V_Z \in R^{pxp}$.

El siguiente paso es calcular dos matrices R y C, muy similares en concepto a la matriz P en PCA. Disminuiremos la dimensión inicial de U_Z y V_Z, para

obtener la matriz R ϵ R^{rxl} y la matriz C ϵ R^{pxl}, donde l es el número determinado de componentes principales, que en este caso se denominan variables latentes. En este trabajo, el número de variables latentes será 2.

$$R = \Sigma_{xx}^{-1/2} \cdot U_Z \tag{21}$$

$$C = \Sigma_{yy}^{-1/2} \cdot V_Z \tag{22}$$

Esto, nos va a permitir obtener las matrices de componentes principales, T para el conjunto de variables canónicas de X y U para el grupo de variables canónicas de Y.

$$T = X \cdot R \tag{23}$$

$$U = Y \cdot C \tag{24}$$

Una vez que hemos obtenido las matrices de componentes principales, podemos construir el modelo CCA, para ello definiremos dos nuevas matrices, P y Q, también llamadas matrices de carga.

$$P = X^t \cdot T \tag{25}$$

$$Q = Y^t \cdot T \tag{26}$$

A partir de este punto se puede deducir que:

$$X = T \cdot P^t + E_X = X_{proyectada} + E_X \tag{27}$$

$$Y = T \cdot Q^t + E_Y = Y_{proyectada} + E_Y \tag{28}$$

Siendo $X_{proyectada}$ e $Y_{proyectada}$ las matrices X e Y respectivamente proyectadas en el espacio de variables canónicas de X, mientras que E_X es la matriz residuo de X y E_Y la matriz residuo de Y.

En el caso de que existiera una fuerte colinealidad de las variables tanto de X como de Y, se emplearía un método de análisis multivariante llamado CCA con regularización. Como su nombre indica, este método, está basado en el método tradicional de CCA, pero permite solventar los problemas de colinealidad comunes de un proceso industrial, que suelen provocar un mal acondicionamiento de los datos a la hora de invertir matrices [14].

El modelo de CCA con regularización es idéntico al modelo de CCA, la única diferencia reside en las ecuaciones 18 y 19 que sufren una liguera modificación.

$$\Sigma_{xx}^{-1/2} = V_X \cdot (D_X + K_1 \cdot I)^{-1/2} \cdot V_X^t \quad (29)$$

$$\Sigma_{yy}^{-1/2} = V_Y \cdot (D_Y + K_2 \cdot I)^{-1/2} \cdot V_Y^t \quad (30)$$

Donde I es la matriz identidad, K_1 y K_2 hacen referencia a dos valores calculados en [14] para la misma planta y que en este trabajo corresponden a $1e^{-3}$ y $68e^{-3}$. Se añade la suma de estos términos de regularización con el objetivo de obtener valores propios distintos de cero, lo que implica unos valores adecuados para poder realizar la inversa de la matriz.

Teniendo como base CCA Regularizado podemos construir otros métodos de análisis multivariante, como el Análisis Canónico de Correlación Dinámico (DCCA).

Este nuevo procedimiento trata de modificar las matrices de datos iniciales y sobre ellas aplicar CCA con regularización. Se aplica de manera similar a DPCA, concatenando la matriz de datos original con las matrices de datos con retardo, pero teniendo en cuenta que hay dos matrices de datos, X e Y.

$$X = [X_t\ , X_{t-1}\ , \dots , X_{t-i}] \quad (31)$$

$$Y = [Y_t\ , Y_{t-1}\ , \dots , Y_{t-i}] \quad (32)$$

Siendo i el número de instantes de retardo y t el instante actual.

2.4.1. ESTADÍSTICAS EMPLEADAS PARA LA MONITORIZACIÓN DEL PROCESO MEDIANTE CCA

Para monitorizar el proceso a partir del Análisis Canónico de Correlación o CCA, se emplearán las mismas estadísticas que con PCA, salvando algunas diferencias que se mostrarán a continuación.

En primer lugar, se aplicará la estadística de Hotelling (T^2):

$$T^2 = T \cdot \Delta^{-1} \cdot T^t \quad (33)$$

Donde $\Delta = {}^{1}/_{(n-1)} \cdot T^t \cdot T$, siendo n el número de observaciones.

El umbral de esta estadística para detectar si el fallo se ha producido, se calcula de la misma manera que en PCA (ecuación 10).

En segundo lugar, aplicaremos la estadística Q o SPE que monitoriza las variables residuo tanto de X como de Y. En este caso, se aplicará esta estadística de manera individual, es decir, habrá una estadística Q para X, que llamaremos Q_X, y también una estadística Q para Y, denominada Q_Y.

$$Q_X = E_X \cdot E_X^t = (X - T \cdot P^t) \cdot (X - T \cdot P^t)^t \quad (34)$$

$$Q_Y = E_Y \cdot E_Y^t = (Y - T \cdot Q^t) \cdot (Y - T \cdot Q^t)^t \quad (35)$$

Cada una de estas dos estadísticas Q tendrá su propio umbral, que se obtiene de manera idéntica que en PCA.

$$Q_{X\,\alpha} = \frac{\sigma_{Q_X}^2}{2 \cdot \mu_{Q_X}} \cdot Chi^2\left(\alpha\,, 2 \cdot \frac{\mu_{Q_X}^2}{\sigma_{Q_X}^2}\right) \quad (36)$$

$$Q_{Y\,\alpha} = \frac{\sigma_{Q_Y}^2}{2 \cdot \mu_{Q_Y}} \cdot Chi^2\left(\alpha\,, 2 \cdot \frac{\mu_{Q_Y}^2}{\sigma_{Q_Y}^2}\right) \quad (37)$$

Donde μ_{Q_X} y μ_{Q_Y} es la media de Q_X y Q_Y respectivamente, σ_{Q_X} es la desviación típica de Q_X, σ_{Q_Y} la desviación típica de Q_Y y α el nivel de significancia.

Estas tres estadísticas se aplican de manera idéntica en los métodos de CCA con regularización y DCCA.

2.5. ANÁLISIS CANÓNICO DE CORRELACIÓN CONCURRRENTE (CCCA)

Otro de los métodos de análisis estadístico multivariante ampliamente utilizado en los procesos industriales es el Análisis Canónico de Correlación Concurrente o CCCA ("Concurrent Canonical Correlation Analysis"), que como su propio nombre indica está basado en el método CCA.

El método CCA, como ya hemos comentado antes, tiene dos deficiencias principalmente. La primera eran los problemas de colinealidad existentes entre dos grupos de variables, muy común en los procesos industriales y que se solucionaba empleando CCA con regularización. Y la segunda, trataba de la varianza o variabilidad de los datos que eran ignorados por el método.

El Análisis Canónico de Correlación Concurrente se emplea para detectar los fallos relevantes que afectan a la calidad, analizando la información aportada

por la variabilidad de los datos y las correlaciones entre las variables del proceso (X) y las variables de calidad (Y). Para ello, descompone el espacio de datos original en cinco subespacios [16].

1- Subespacio de correlación, formado por las variables canónicas obtenidas mediante CCA regularizado.

2- La parte restante de las variables de proceso se dividen en dos subespacios, el primero sería el subespacio principal de proceso y el segundo, el subespacio residual de proceso.

3- Lo mismo ocurre con la parte restante de las variables de calidad, donde habrá un subespacio principal de calidad y otro subespacio residual de calidad.

Para construir el algoritmo de CCCA, como ya se ha mencionado, primero aplicaremos CCA con regularización de la misma manera a como se describe en el apartado anterior. Obteniendo finalmente las matrices $R \in R^{rxl}$ (ecuación 21), $T \in R^{nxl}$ (ecuación 23), $P \in R^{rxl}$ (ecuación 25) y $Q \in R^{pxl}$ (ecuación 26).

A continuación, debemos calcular la parte despreciada de la matriz de variables de proceso, que denominaremos X_C.

$$X_C = X - T \cdot R^{\dagger} \tag{38}$$

Donde, $R^{\dagger} = (R^t \cdot R)^{-1} \cdot R^t$.

Aplicaremos el método PCA a X_C, y definiremos una nueva matriz a la que llamaremos A_C, muy similar en concepto a la matriz R en PCA.

$$A_C = \frac{1}{n-1} \cdot X_C^t \cdot X_C \tag{39}$$

Extraeremos los vectores propios de esta matriz, quedándonos con los l_X primeros autovectores, que en este estudio tendrá un valor de 20, y almacenándolos en la matriz $P_X \in R^{rxlx}$.

Se calcula la matriz de componentes principales de X_C, que llamaremos T_X, y con ella obtendremos la matriz residuo de X_C, que denominaremos X_{CC}.

$$T_X = (X - T \cdot R^{\dagger}) \cdot P_X \tag{40}$$

$$X_{CC} = X_C - T_X \cdot P_X \tag{41}$$

Aplicaremos los mismos pasos, pero con la matriz de variables de calidad (Y), calculando primeramente la parte despreciada de esta matriz.

$$Y_C = Y - T \cdot Q^T \tag{42}$$

Realizamos nuevamente el método PCA a una nueva matriz que definiremos como B_C y reuniremos los l_Y primeros autovectores, que en este trabajo tendrá un valor de 2, en la matriz $P_Y \in R^{rxly}$.

$$B_C = \frac{1}{n-1} \cdot Y_C^t \cdot Y_C \tag{43}$$

También calcularemos la matriz de componentes principales T_Y y la matriz residual de Y_C.

$$T_Y = (Y - T \cdot Q) \cdot P_Y \tag{44}$$

$$Y_{CC} = Y_C - T_Y \cdot P_Y \tag{45}$$

Una vez obtenidos todos estos subespacios aplicando el algoritmo de CCCA, es hora de aplicar las estadísticas para monitorizar el proceso.

2.5.1. ESTADÍSTICAS EMPLEADAS PARA LA MONITORIZACIÓN DEL PROCESO MEDIANTE CCCA

Se monitorizará el proceso con las mismas estadísticas que hemos aplicado hasta el momento. Lo particular con CCCA, como se ha mencionado antes, es que se obtienen 5 subespacios finales, dos de ellos son residuales por lo que se les aplicará la estadística Q y al resto la estadística T^2.

Comenzamos con el subespacio obtenido de emplear el método CCA sobre la matriz de datos original, en el que utilizaremos la estadística de Hotelling o T^2 (ecuación 33), y su umbral será el mismo que el explicado en el método PCA (ecuación 10).

Se empleará también la estadística de Hotelling en monitorizar los subespacios de X_C e Y_C.

$$T_X^2 = T_X \cdot \Delta_X^{-1} \cdot T_X^t \tag{46}$$

$$T_Y^2 = T_Y \cdot \Delta_Y^{-1} \cdot T_Y^t \tag{47}$$

Siendo $\Delta_X = {}^{1}/_{(n-1)} \cdot T_X^t \cdot T_X$, con n número de observaciones, y por similitud $\Delta_Y = {}^{1}/_{(n-1)} \cdot T_Y^t \cdot T_Y$.

Los umbrales correspondientes a cada estadística anterior será:

$$T_{X_\alpha}^2 = \frac{(n^2-1)\cdot l_X}{n\cdot(n-l_X)} \cdot F_\alpha(l_X, n-l_X) \tag{48}$$

$$T_{Y_\alpha}^2 = \frac{(n^2-1)\cdot l_Y}{n\cdot(n-l_Y)} \cdot F_\alpha(l_Y, n-l_Y) \tag{49}$$

Donde n es el número de observaciones, l_X y l_Y el número de variables latentes asociadas a X o Y y F_α el valor correspondiente a la distribución de Fisher-Snedecor siendo α el nivel de significancia con valor 0,01.

A continuación, aplicaremos la estadística Q o SPE sobre X_{CC} (ecuación 41) e Y_{CC} (ecuación 45) de manera individual, es decir, sobre el espacio residual de X_C e Y_C.

$$Q_X = X_{CC} \cdot X_{CC}^t \tag{50}$$

$$Q_Y = Y_{CC} \cdot Y_{CC}^t \tag{51}$$

Y los umbrales correspondientes serán los mismos que los explicados en el método CCA (ecuaciones 36 y 37 respectivamente).

2.6. LOCALLY LINEAR EMBEDDING (LLE)

LLE ("Locally Linear Embedding") es una técnica empleada para reducir la alta dimensionalidad de los datos iniciales. Este tipo de herramientas, tienen como finalidad obtener representaciones más pequeñas y fáciles de manejar, manteniendo la mayor cantidad posible de información del conjunto de datos principal.

Otro de los muchos métodos que se incluye dentro de la reducción dimensional, y que ya hemos mencionado con anterioridad, es el método PCA, pero este no es apto para análisis de datos con alta dimensionalidad [17].

El método LLE, es un algoritmo de aprendizaje no supervisado, es decir, no se necesita una variable de destino para reducir la dimensionalidad de manera lineal, que calcula puntos de interés con baja dimensionalidad manteniendo la vecindad de los datos de entrada, o lo que es lo mismo, busca el conjunto de vecinos más cercano a cada punto proyectándolo en una dimensión

inferior. Una representación del espacio inicial con el punto de interés y sus vecinos más cercanos, se muestran en la figura 7.

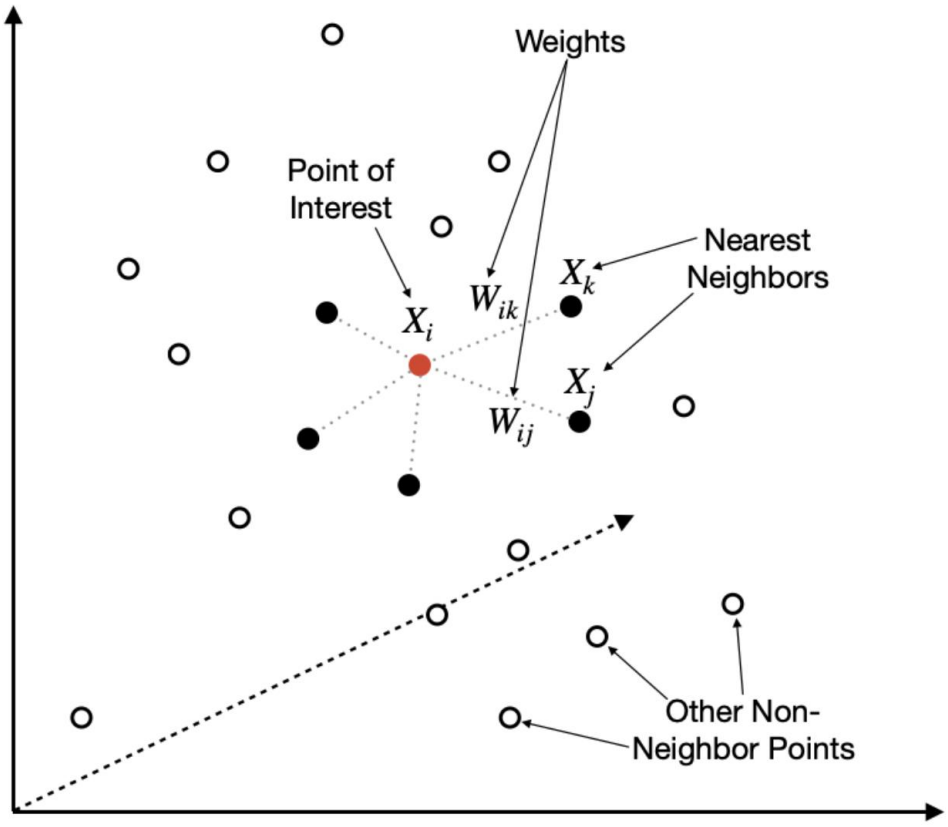

Figura 7. Punto de interés y sus vecinos más cercanos pertenecientes al espacio inicial de alta dimensionalidad [18].

Este algoritmo obtiene la posición del punto de interés x_i basándose en sus vecinos x_j. Para ello, necesita un conjunto de pesos para cada x_i que mejor lo describa como combinación lineal de sus vecinos. Los pesos w_{ij} son valores que designan cuanto contribuye x_j al reconstruir el punto x_i y se almacenan en la matriz de pesos W. Esto lleva a que se cumplan dos hipótesis, la primera es que si x_j no es vecino de x_i su peso w_{ij} será nulo y la segunda es que la suma de cada fila de la matriz de pesos es igual a 1, es decir, $\sum_j w_{ij} = 1$ [19].

LLE realiza el mapeo de vecindarios minimizando el siguiente error cuadrático:

$$\Theta(\mathrm{W}) = \sum_i \left\| x_i - \sum_j w_{ij} x_i \right\|^2 \tag{52}$$

Hasta ahora se ha trabajado sobre el espacio dimensional inicial R^{nxm}, pero el objetivo de este algoritmo es reducir la dimensionalidad a una mucho menor R^{nxd}, tal que *d* << *m*. Para ello, utilizando la misma matriz de pesos W obtenida anteriormente, se proyecta el punto x_i en el espacio reducido R^{nxd}, es decir, cada punto x_i se asigna a un punto ϕ_i en el espacio reducido, de manera que se conserve el vecindario, y eso se consigue minimizando la siguiente ecuación [20]:

$$\Psi(\phi) = \sum_i \left\| \phi_i - \sum_j w_{ij}\phi_i \right\|^2 \tag{53}$$

En esta ecuación 53 a diferencia de la ecuación 52 los pesos w_{ij} se mantienen fijos y la minimización se aplica sobre los puntos ϕ_i optimizando sus coordenadas. En la figura 8 se muestra el espacio dimensional reducido, el punto proyectado ϕ_i y los vecinos de este manteniendo la matriz de pesos.

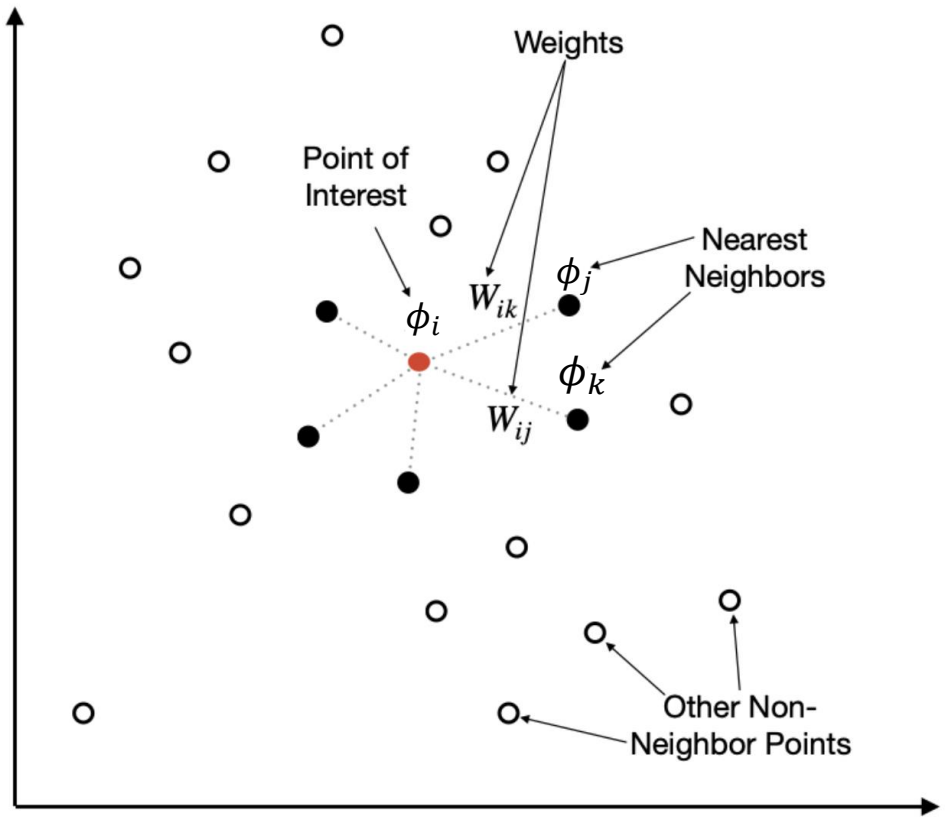

Figura 8. Punto de interés proyectado y sus vecinos más cercanos pertenecientes al espacio dimensional reducido [18].

Este algoritmo será aplicado de manera complementaria dentro de los métodos de análisis estadístico multivariante como PCA, DPCA, CCA, DCCA y CCCA. En estos tres últimos, se aplicará LLE sobre la matriz X y de la misma manera, sobre Y.

CAPÍTULO III: PLANTA TENNESSEE EASTMAN

3.1. CASO DE ESTUDIO Y DESCRIPCIÓN DE LA PLANTA TENNESSEE EASTMAN

El Proceso Tennessee Eastman (TEP), publicado por J.J. Downs y E.F. Vogel, surge a finales del siglo XX como proyecto de Eastman Chemical Company en colaboración con la Universidad de Tennessee, creando un modelo real que sirva como propósito de desarrollar, estudiar y evaluar los métodos emergentes del control de procesos multivariable [21].

Actualmente, este proceso es extensamente utilizado como banco de pruebas debido a su estructura altamente no lineal y su gran volumen de variables, que permiten múltiples posibilidades de estudio en la ingeniería de procesos, como por ejemplo la detección de anomalías en el proceso mediante control estadístico.

La planta TEP, tiene como resultado dos productos (G , H), que se obtienen a partir de cuatro reactivos (A, C, D, E), un inerte (B), y un subproducto (F), haciendo un total de ocho componentes. Las reacciones que tienen lugar en el proceso son las siguientes: [22]

Producto 1	$A_{(gas)} + C_{(gas)} + D_{(gas)} \rightarrow G_{(liq)}$
Producto 2	$A_{(gas)} + C_{(gas)} + E_{(gas)} \rightarrow H_{(liq)}$
Subproducto	$A_{(gas)} + E_{(gas)} \rightarrow F_{(liq)}$
Subproducto	$3D_{(gas)} \rightarrow 2F_{(liq)}$

Todas las reacciones que aparecen en el proceso son irreversibles y exotérmicas, además, la velocidad de reacción depende de la temperatura. La reacción que tiene como producto final G, tiene una sensibilidad a la temperatura mayor que el resto, debido a que su energía de activación es más alta.

Como puede verse en la figura 9, el proceso tiene 5 elementos clave donde se realizan las operaciones pertinentes: reactor, condensador, separador de líquido y gas, compresor y destilador. Otros elementos que encontramos son bombas y elementos de control como válvulas, indicadores, etc.

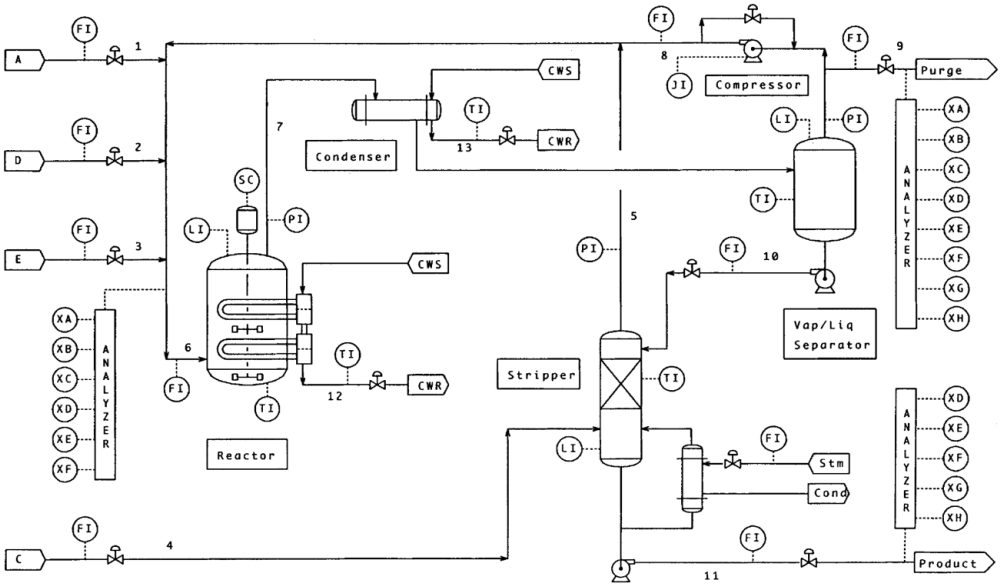

Figura 9. Proceso Tennessee Eastman [23].

Los reactivos gaseosos A, D y E entran en el reactor, donde las reacciones en fase gaseosa son catalizadas mediante un catalizador en fase liquida no volátil disuelto, que permanece de manera continua en el reactor. En este primer elemento, también se realiza un proceso de enfriamiento para eliminar el calor de la reacción.

A la salida del reactor se obtienen productos en estado vapor y productos no reaccionados, ambos son conducidos a un condensador, donde sufrirán una disminución de temperatura, y más tarde a un separador liquido-vapor. Al final de este, los productos no condensables son recirculados mediante un compresor centrifugo al reactor, volviendo junto con los reactivos A, D y E.

Los productos condensables pasan a una columna de destilación o stripper donde se separan los restos de reactivos que potencialmente podrían haber quedado, con la acción del reactivo C, y que son recirculados al reactor. Del destilador se obtienen los productos G y H que serán disociados aguas abajo del proceso, lo cual no es objeto de estudio. También se producen un inerte B y un subproducto F, ambos son purgados en la corriente de vapor del separador líquido-vapor.

3.2. DATOS DEL PROCESO

El Proceso Tennessee Eastman cuenta con un total de 12 variables manipuladas, lo que ofrece 12 grados de libertad, y 41 variables medidas. De las 12 variables manipuladas (tabla 1) o grados de libertad, 9 son válvulas de flujo, 2 son válvulas de control de temperatura y la restante pertenece a la velocidad de agitación. Las 41 variables medidas (tabla 2) se dividen en dos

tipos; las primeras 22 que se monitorizan de manera continua (XMEAS(1-22)) y el resto, que suman un total de 19 variables, se obtienen de los analizadores que se encuentran distribuidos en la planta y que dan valores únicamente cada cierto intervalo de tiempo (XMEAS(23-42)) [5].

	VARIABLE	UDS.
XMV (1)	Flujo de alimentación D (corriente 2)	kg/h
XMV (2)	Flujo de alimentación E (corriente 3)	kg/h
XMV (3)	Flujo de alimentación A (corriente 1)	kscmh
XMV (4)	Flujo de alimentación A y C (corriente 4)	kscmh
XMV (5)	Válvula de recirculación del compresor	%
XMV (6)	Válvula de purga (corriente 9)	%
XMV (7)	Flujo líquido del separador LV (corriente 10)	m^3/h
XMV (8)	Flujo líquido producto de la columna de stripping (corriente 11)	m^3/h
XMV (9)	Válvula de vapor de la columna de stripping	%
XMV (10)	Flujo de agua de refrigeración del reactor	m^3/h
XMV (11)	Flujo de agua de refrigeración del condensador	m^3/h
XMV (12)	Velocidad de agitación del reactor	rpm

Tabla 1. Variables manipuladas del proceso [22]

	VARIABLE	UDS.
XMEAS (1)	Flujo de alimentación de A (corriente 1)	kscmh
XMEAS (2)	Flujo de alimentación de D (corriente 2)	kg/h
XMEAS (3)	Flujo de alimentación de E (corriente 3)	kg/h
XMEAS (4)	Flujo de alimentación de A y C (corriente 4)	kscmh
XMEAS (5)	Flujo de recirculación (corriente 8)	kscmh
XMEAS (6)	Flujo de alimentación del reactor (corriente 6)	kscmh

XMEAS (7)	Presión en el reactor	kPa
XMEAS (8)	Nivel del reactor	%
XMEAS (9)	Temperatura en el reactor	ºC
XMEAS (10)	Flujo de purga	kscmh
XMEAS (11)	Temperatura del producto en el separador LV	ºC
XMEAS (12)	Nivel del producto en el separador LV	%
XMEAS (13)	Presión del producto en el separador LV	kPa
XMEAS (14)	Corriente en el separador LV de producto (corriente 10)	m^3/h
XMEAS (15)	Nivel del stripper	%
XMEAS (16)	Presión del stripper	kPa
XMEAS (17)	Corriente del stripper (corriente 11)	m^3/h
XMEAS (18)	Temperatura del stripper	ºC
XMEAS (19)	Flujo de vapor del stripper	kg/h
XMEAS (20)	Potencia del compresor	kW
XMEAS (21)	Temperatura de la salida del agua de refrigeración del reactor	ºC
XMEAS (22)	Temperatura de la salida del agua de refrigerador del separador LV	ºC
XMEAS (23-28)	Análisis de concentración de la alimentación del reactor, componentes A-F. (corriente 6)	mol%
XMEAS (29-36)	Análisis de concentración del gas de purga, componentes A-H. (corriente 9)	mol%
XMEAS (37-41)	Análisis de concentración aguas abajo del stripper, componentes D-H. (corriente 11)	mol%

Tabla 2. Variables medidas del proceso [22]

En la planta se han producido diferentes tipos de fallos distribuidos a lo largo de todo el proceso y de los cuales se han obtenido datos, que pueden observarse en la tabla 3.

	FALLO O ANOMALÍA	TIPO
IDV (1)	Ratio de flujo de alimentación A/C, composición de B constante (corriente 4)	Escalón
IDV (2)	Composición de B, con ratio A/C constante (corriente 4)	Escalón
IDV (3)	Temperatura de alimentación de D (corriente 2)	Escalón
IDV (4)	Temperatura de entrada del agua de refrigeración al reactor	Escalón
IDV (5)	Temperatura de entrada del agua de refrigeración al condensador	Escalón
IDV (6)	Perdida de alimentación de A	Escalón
IDV (7)	Perdida de presión en la alimentación de C	Escalón
IDV (8)	Composición de las alimentaciones de A, B y C	Variación aleatoria
IDV (9)	Temperatura de alimentación de D	Variación aleatoria
IDV (10)	Temperatura de alimentación de C	Variación aleatoria
IDV (11)	Temperatura de entrada del agua de refrigeración al reactor	Variación aleatoria
IDV (12)	Temperatura de entrada del agua de refrigeración al condensador	Variación aleatoria
IDV (13)	Cinética de reacción	Variación lenta
IDV (14)	Válvula del agua refrigerante del reactor	Invariable
IDV (15)	Válvula del agua refrigerante del condensador	Invariable
IDV (16)	Desconocido	No especificado
IDV (17)	Desconocido	No especificado
IDV (18)	Desconocido	No especificado
IDV (19)	Desconocido	No especificado
IDV (20)	Desconocido	No especificado
IDV (21)	Desconocido	No especificado

Tabla 3. Tipos de fallo en el proceso [22]

Existen anomalías más difíciles de detectar que otras, como los fallos IDV (3), IDV (9) e IDV (15), ya que las variables con los que están relacionadas no sufren cambios significativos con respecto a un comportamiento no anómalo, denominados así, como fallos incipientes.

Los datos utilizados para este trabajo son públicos y se pueden descargar en el siguiente enlace (http://web.mit.edu/braatzgroup/links.html). Sobre estos, se aplicarán los diferentes métodos de análisis estadístico multivariante, explicados en el capítulo anterior, y de esta manera, monitorizar el proceso.

Existen dos bloques de datos diferentes, el primero corresponde a una simulación de la planta TEP con un comportamiento normal sin fallos. El siguiente bloque corresponde a las 21 simulaciones del proceso para cada uno de los fallos, de manera individual, mencionados en la tabla 3.

CAPÍTULO IV: APLICACIÓN

4.1. INTRODUCCIÓN A LA APLICACIÓN

Los datos utilizados en este estudio, como ya se ha comentado en el capítulo III, pertenecen al proceso Tennessee Eastman. Sobre ellos, aplicaremos los diferentes métodos de análisis estadístico multivariante, explicados de manera teórica en el capítulo II, con el objetivo de detectar el fallo y la pronta detección del mismo. Finalmente compararemos los diferentes métodos con el objetivo de encontrar el que mejor se adapte a este proceso, es decir, el que antes detecte el fallo, con ello ayudaremos a obtener un buen control en la calidad del proceso.

En primer lugar, utilizaremos un fichero de datos que corresponden a un comportamiento normal del proceso, es decir, sin fallos, construido por 52 variables del proceso y 960 observaciones de cada una de ellas. Sobre estos datos de entrenamiento se aplicarán cada uno de los métodos de análisis multivariante, obteniendo información necesaria para, más adelante, evaluar si el método aplicado detecta o no anomalías.

A continuación, se emplearán 21 ficheros de datos, correspondientes a los 21 fallos que se han mencionado en la tabla 3 del capítulo anterior, también con dimensiones de 52 variables y 960 observaciones cada uno de ellos.

Cada uno de estos 21 ficheros se dividen en dos tramos, las 160 primeras observaciones, son datos de comportamiento normal previo al fallo, es decir, cada fallo se registra en el instante 160. El segundo tramo va desde la observación 161 hasta la 960 donde el fallo que corresponda ya está presente. De esta manera compararemos las diferentes técnicas de análisis de fallo multivariante aplicadas y decidiremos cuál de ellas se adapta mejor a la detección de las diferentes anomalías del proceso.

Aunque ya se ha comentado en el capítulo anterior, cabe destacar que los fallos 3, 9 y 15 relacionados, los dos primeros, con la temperatura de alimentación del reactivo D, y el último con la válvula de agua de refrigeración del condensador, son denominados fallos incipientes dado que son difíciles de detectar y seguramente, alguno de los métodos que aplicaremos no lo logrará.

Para poder comparar los diferentes métodos desarrollados en este trabajo y decidir cuál es el que mejor se adapta a este proceso, se han de definir unos índices que indiquen la bondad del método. Estos índices son:

a. Porcentaje de falsas alarmas antes de la detección del fallo. Se calcula como el número de veces que cada estadística, correspondiente al método, supere su umbral sin que se haya producido ningún fallo,

dividido por el número de muestras sin fallo (160) y multiplicado por 100 para obtener el porcentaje.

b. Porcentaje de alarmas detectadas. Se calcula como el número de veces que las estadísticas superen su correspondiente umbral una vez producido el fallo, dividido por el número total de muestras con fallo (800) y multiplicado por 100 como en el índice anterior.

c. Tiempo de detección. Es el tiempo que tardan las estadísticas en superar el umbral una vez producido el fallo, es decir, el tiempo que transcurre desde que el fallo se ha producido, hasta que es detectado por el método.

Evidentemente, para que el método sea óptimo, es necesario un bajo porcentaje de falsas alarmas, que el porcentaje de alarmas después del fallo sea elevado y finalmente, que el tiempo de detección sea lo más temprano posible.

En la realización de este estudio se ha empleado el software Spyder 5.1.5 donde se ha programado el código de cada uno de los métodos aplicados en lenguaje Python 3.9.

4.2. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El primer método de análisis multivariante que utilizaremos para la detección de fallos en el proceso es el Análisis de Componentes Principales o PCA. Primero se analizarán los datos de comportamiento normal, donde calcularemos los umbrales de las estadísticas correspondientes, para más tarde, trabajar con ellos en los datos de fallo del proceso.

4.2.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON PCA

Primeramente, se importan los datos de comportamiento normal a una matriz, que tienen una dimensión de 960 observaciones y 52 variables, por lo que el espacio muestral inicial de datos es R^{960x52}. Esta matriz de datos, a la que llamaremos matriz X, tiene que ser normalizada a media cero y varianza uno, para que todos los datos estén en la misma escala y puedan compararse entre sí.

De esta nueva matriz X, obtenemos la matriz de correlación R (ecuación 2), de la que extraemos los valores propios que se almacenaran en la matriz $\Lambda \in R^{52x52}$ de manera decreciente, y los vectores propios o vectores de carga retenidos en la matriz $V \in R^{960x52}$, ordenados de manera correspondiente según al valor propio que vaya ligado.

Una vez obtenidas estas matrices, pasamos a reducir las dimensiones del espacio muestral R^{960x52}, a R^{960xa}, siendo "a" el número de componentes principales. Para ello, debemos elegir la variabilidad con la que vamos a trabajar, en este caso 90% y aplicamos un test de porcentaje de varianza (ecuación 3) para calcular los componentes principales, de manera que se va sumando la varianza de los valores propios de R, estando estos en orden decreciente, hasta igualar o superar el porcentaje de varianza elegido, en este caso 90 %. Finalmente obtenemos un total de 31 valores propios, los de mayor peso, y son albergados en la matriz diagonal $S_a \in R^{31x31}$, traduciéndose a reducir el número de variables originales a 31. De la misma manera, los vectores propios asociados a los 31 autovalores son almacenados con el mismo orden en la matriz $P \in R^{960x31}$.

Definimos la matriz de componentes principales como matriz T (ecuación 4). Esta proyecta el espacio inicial R^{960x52} a un espacio de dimensión reducida R^{960x31}.

Finalmente calculamos las estadísticas, con sus respectivos umbrales, para los datos de comportamiento normal.

Primero calculamos T^2 o estadística de Hotelling (ecuación 9), para el cual elegimos un nivel de significancia $\alpha = 0{,}01$ en este estudio. Como resultado, obtenemos una matriz de dimensiones R^{1x960}, es decir, un valor de T^2 para cada observación. Se calcula su umbral T_a^2 (ecuación 10), que se utilizará, más adelante, con los datos de fallo para comprobar si se produce o no una anomalía. El valor del umbral T_a^2 obtenido es 54,61.

Como puede observarse en la figura 10, los valores T^2 calculados con los datos de comportamiento normal, para las 960 observaciones, prácticamente no supera el umbral de T_α^2.

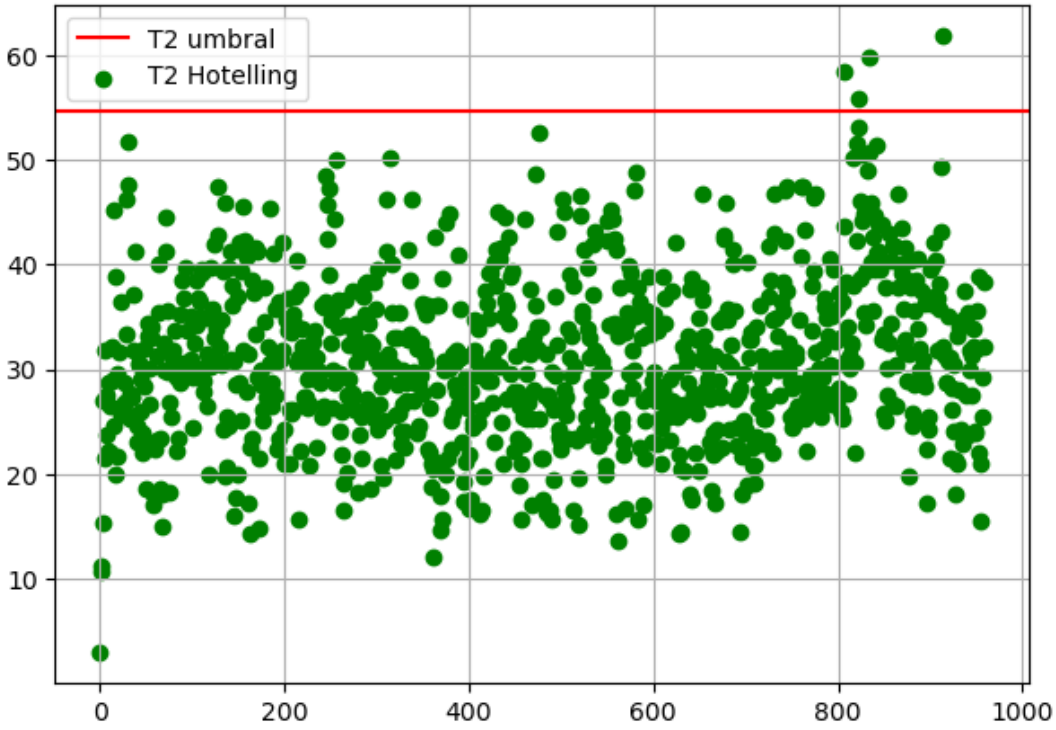

Figura 10. Estadística de Hotelling con su umbral para datos de comportamiento normal en PCA.

A continuación, se calcula la estadística Q o SPE (ecuación 11) utilizando el vector de residuos r (ecuación 12), y también su umbral Q_α (ecuación 13), que al igual que T_a^2, será empleado más adelante, para comprobar si se detecta el fallo. Como podemos observar en la figura 11, la gran mayoría de los valores de Q tampoco superan su umbral.

En este caso, se ha considerado calcular el umbral Q_α a través de un percentil donde el 99% de las observaciones queden por debajo. Sí, por el contrario, Q_α se calculase de la manera tradicional (ecuación 13), el valor de este quedaría dentro del rango de valores de las primeras 160 observaciones y como consecuencia, los porcentajes de falsas alarmas de la estadística Q serían sumamente elevados. Finalmente, el umbral Q_α tiene un valor de 11,17

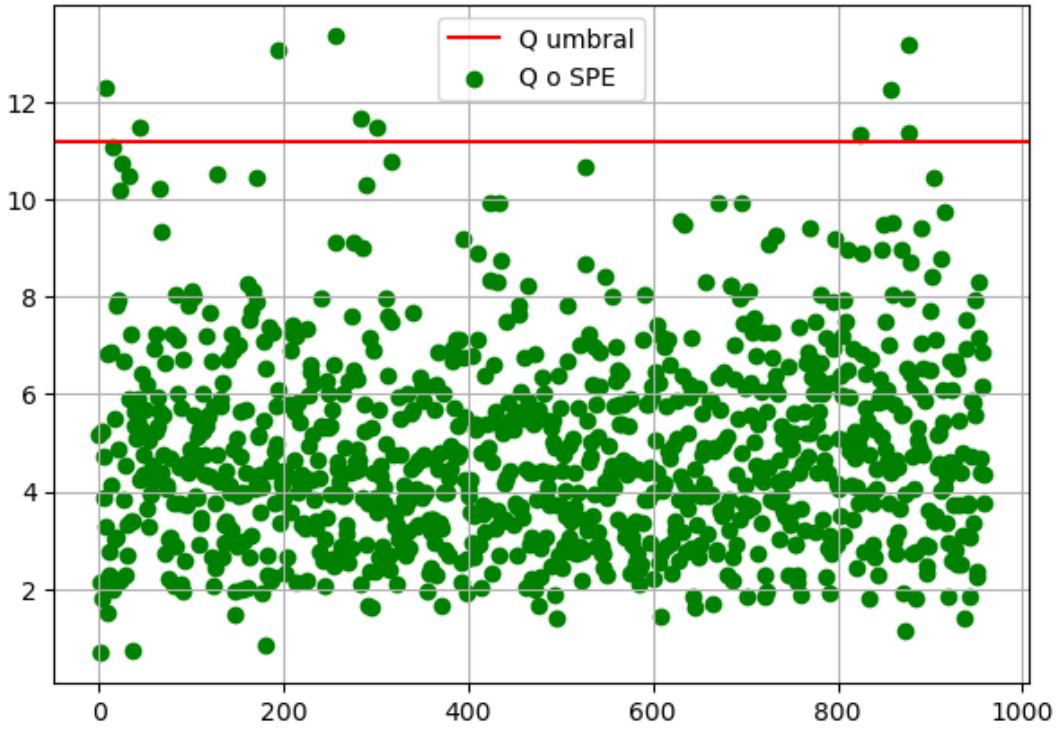

Figura 11. Estadística Q o SPE con su umbral para datos de comportamiento normal en PCA.

4.2.2. ANÁLISIS DE DATOS CON FALLO APLICANDO PCA

Una vez aplicado PCA para los datos de comportamiento normal, utilizamos este mismo método con los datos de comportamiento anómalo y de esta manera detectar los posibles fallos del sistema. Aplicamos este método a cada uno de los 21 ficheros de datos con fallo, que corresponden respectivamente con los 21 fallos del proceso mencionados en la tabla 3.

Comenzamos, al igual que en el apartado anterior, normalizando la matriz de datos inicial, pasando a tener una matriz X con 52 variables y 960 observaciones, pero en este caso, estandarizada con la media y la varianza de la matriz de datos de comportamiento normal.

Calcularemos las estadísticas T^2 y Q y las compararemos con sus respectivos umbrales ya calculados con los datos de comportamiento normal en el proceso T_α^2 y Q_α. Se producirá el fallo en el momento en que los valores de las estadísticas T^2 y Q superen su umbral un total de 8 veces consecutivas, de esta manera evitamos confundir el fallo con falsas alarmas.

Primero calculamos la estadística T^2 (ecuación 9), utilizando las matrices $P \in R^{960x31}$ y $S_a \in R^{31x31}$ obtenidas en el apartado anterior y la nueva matriz X con los datos de fallo. De la misma manera, utilizando estas matrices importadas definimos una nueva matriz de residuos r (ecuación 12), con la finalidad de obtener la estadística Q (ecuación 11).

Finalmente, contrastaremos los datos obtenidos aplicando ambas estadísticas a cada uno de los ficheros de fallo. Debido a que son una gran cantidad de datos, se mostrarán únicamente las figuras de 3 fallos, en concreto se han seleccionado los fallo 2, 9 y 17. También se mostrará la tabla 4 donde se recoge la información obtenida tras aplicar PCA a los 21 fallos, además de haberse calculado los valores medios para cada estadística del tiempo de detección, el porcentaje de alarmas y el porcentaje de falsas alarmas anteriores al fallo.

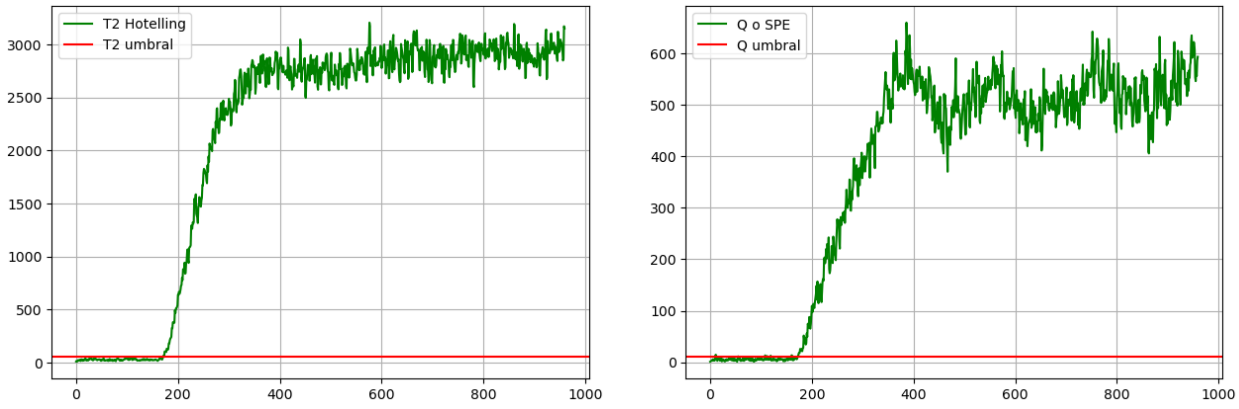

Figura 12. Detección del fallo IDV (2) con la estadística de Hotelling (izquierda) y Q (derecha) en PCA.

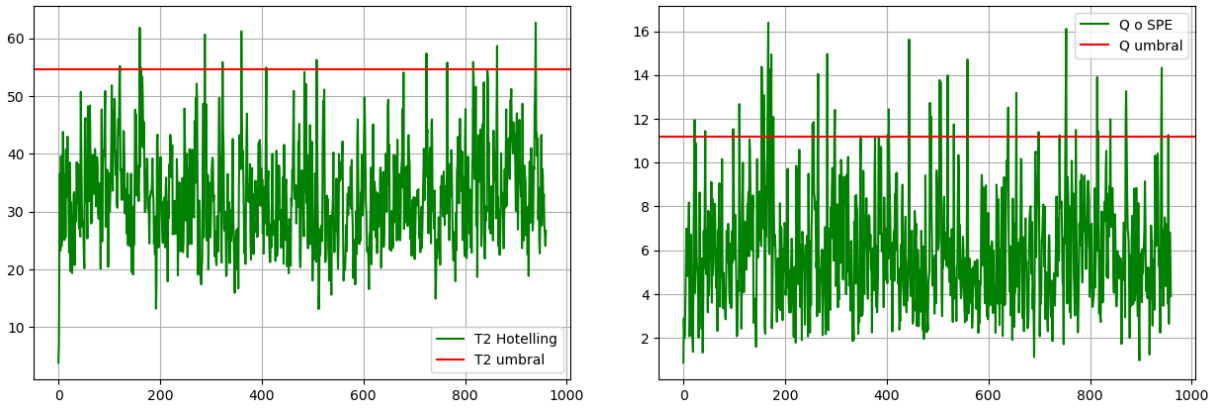

Figura 13. Detección del fallo IDV (9) con la estadística de Hotelling (izquierda) y Q (derecha) en PCA.

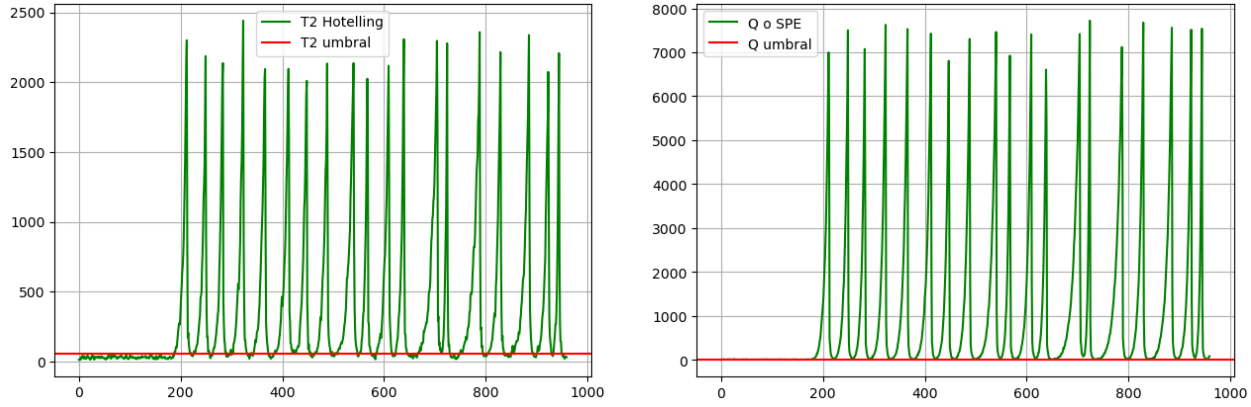

Figura 14. Detección del fallo IDV (17) con la estadística de Hotelling (izquierda) y Q (derecha) en PCA.

La línea verde corresponde a la representación de la estadística calculada para cada una de los 960 observaciones y la línea roja representa el valor del umbral correspondiente a cada estadística, que en el caso de que sea superado 8 veces consecutivas, se puede afirmar que el fallo ha sido detectado.

En la figura 12 y en la figura 14 podemos observar como el método PCA detecta el fallo 2 y 17 respectivamente, ya que ambas estadísticas, T^2 y Q, superan con creces sus correspondientes umbrales. Además, podemos decir que la detección de ambos fallos ha sido temprana, ya que sucede en torno a la observación 160, que es el instante donde se registran todos los fallos. De la misma manera, podemos ver como a partir de esta observación, la evolución de las estadísticas son similares después del fallo y varían en un rango de valores superior al umbral. Esto sucede porque después de que ocurre el fallo, el sistema cambia su comportamiento y no puede volver de manera autónoma a un estado normal.

Por el contrario, en la figura 13 se muestra representado el fallo 9, correspondiente a la temperatura de alimentación de D, este no es detectado por el método aplicado, que como ya se mencionó anteriormente, el fallo IDV (9) es un fallo incipiente o difícil de detectar. En este caso ambas estadísticas si superan el umbral, pero no lo hacen de manera consecutiva 8 veces, por lo que no se puede decir que se haya producido el fallo.

	Hotelling T^2				Q o SPE			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	166	6	99,25	0,63	162	2	99,75	3,75
IDV (2)	172	12	98,50	0,63	174	14	98,63	4,38
IDV (3)	No detecta	No detecta	1,00	0,00	No detecta	No detecta	7,50	2,50
IDV (4)	No detecta	No detecta	27,75	0,63	160	0	100,00	2,50
IDV (5)	170	10	23,75	0,63	160	0	33,88	2,50
IDV (6)	170	10	98,75	0,00	160	0	100,00	2,50
IDV (7)	160	0	100,00	0,00	160	0	100,00	2,50
IDV (8)	182	22	97,25	0,00	177	17	96,88	4,38
IDV (9)	No detecta	No detecta	1,50	0,63	No detecta	No detecta	4,50	5,00
IDV (10)	308	148	25,38	0,00	207	47	46,00	0,00
IDV (11)	543	383	45,00	0,00	166	6	69,38	5,00
IDV (12)	181	21	98,38	0,00	182	22	95,13	6,25
IDV (13)	206	46	94,50	1,25	200	40	95,13	1,25
IDV (14)	160	0	98,88	0,00	161	1	99,88	5,00
IDV (15)	No detecta	No detecta	0,63	0,00	No detecta	No detecta	6,13	3,13
IDV (16)	472	312	11,00	3,13	353	193	43,50	3,13
IDV (17)	188	28	75,88	0,00	181	21	95,63	3,13
IDV (18)	252	92	89,00	0,63	243	83	90,13	5,63
IDV (19)	No detecta	No detecta	6,25	0,00	No detecta	No detecta	21,88	4,38
IDV (20)	246	86	27,88	0,00	244	84	55,88	3,75
IDV (21)	664	504	37,50	0,00	409	249	50,13	3,75
MEDIA	430,48	270,48	55,14	0,39	349,48	189,48	67,14	3,54
MEDIA SIN IDV 3, 9 Y 15	342,22	182,22	64,16	0,42	247,72	87,72	77,32	3,54

Tabla 4. Resultados obtenidos de las estadísticas T^2 y Q, mediante la aplicación de PCA.

En la tabla anterior, se observa que la gran mayoría de los fallos son detectados por el método de análisis multivariante PCA, a excepción de los fallos incipientes 3, 9 y 15, que son difíciles de detectar y del fallo 19, que no ha podido ser detectado por ninguno de las estadísticas. Así mismo, el fallo 4 no ha sido detectado por la estadística T^2, pero si lo ha hecho la estadística Q. En definitiva, la estadística de Hotelling detecta un total de 16 anomalías, mientras que la estadística Q, identifica 17.

Si observamos las medias calculadas para ambas estadísticas es lógico que el tiempo de detección del fallo sea mejor cuando no incluimos las anomalías

incipientes, ya que estas no son detectadas en los 960 instantes disponibles. Del mismo modo, el porcentaje de alarmas o falsas alarmas tienen mejor resultado en la estadística T^2, pero el tiempo de detección es más bajo en Q.

4.3. ANÁLISIS DE COMPONENTES PRINCIPALES DINÁMICO (DPCA)

El método DPCA ("Dynamic Principal Component Analysis"), es otra técnica de análisis multivariante que utilizaremos para detectar fallos en el proceso. Este método se aplica de manera similar a PCA, lo único que cambia, son los datos iniciales, ya que estos deben de ser dinámicos.

Al igual que en PCA, también calcularemos las estadísticas T^2 y Q y sus respectivos umbrales T_a^2 y Q_α a partir de los datos de comportamiento normal, para más tarde analizar los datos de comportamiento de fallo del proceso.

4.3.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON DPCA

En primer lugar, se importan los datos de comportamiento normal y se normalizan a media cero y varianza unitaria, dando lugar a una matriz de dimensiones 960x52. A diferencia con el método anterior, no se aplicará PCA a la matriz normalizada con los datos iniciales, sino que, la matriz denominada "X" es una concatenación de matrices de datos dinámicos basada en la matriz normalizada inicial (ecuación 8).

Para obtener los datos dinámicos necesitamos la matriz normalizada inicial, pero para instantes de tiempo anteriores t-n, para ello desplazamos dicha matriz a la derecha tantas veces como n instante de tiempo, eliminando las n últimas columnas y duplicando n veces la columna inicial. Una vez calculadas las matrices hasta el instante t-n elegido concatenaríamos todas ellas y formaríamos la matriz X a la cual aplicaremos PCA.

Una vez llegados a este punto, aplicamos PCA empleando la misma metodología explicada en el apartado 4.2.1. También calcularemos las estadísticas T^2 (ecuación 9) y Q (ecuación 11) con sus respectivos umbrales T_a^2 (ecuación 10) y Q_α obtenido mediante percentil 99, para más tarde, poder evaluar los datos de fallo y ver si esta variante de PCA se ajusta al proceso. En esta caso el umbral T_a^2 tiene un valor de 56,03 y Q_α de 42,22.

En la figura 15, podemos observar como la distribución de puntos o los valores de la estadística de Hotelling aplicando DPCA es muy similar al mostrado en la figura 10, correspondiente a esta misma estadística, siendo

producto del método PCA. Además, los umbrales T_α^2 tanto de PCA como DPCA son prácticamente idénticos.

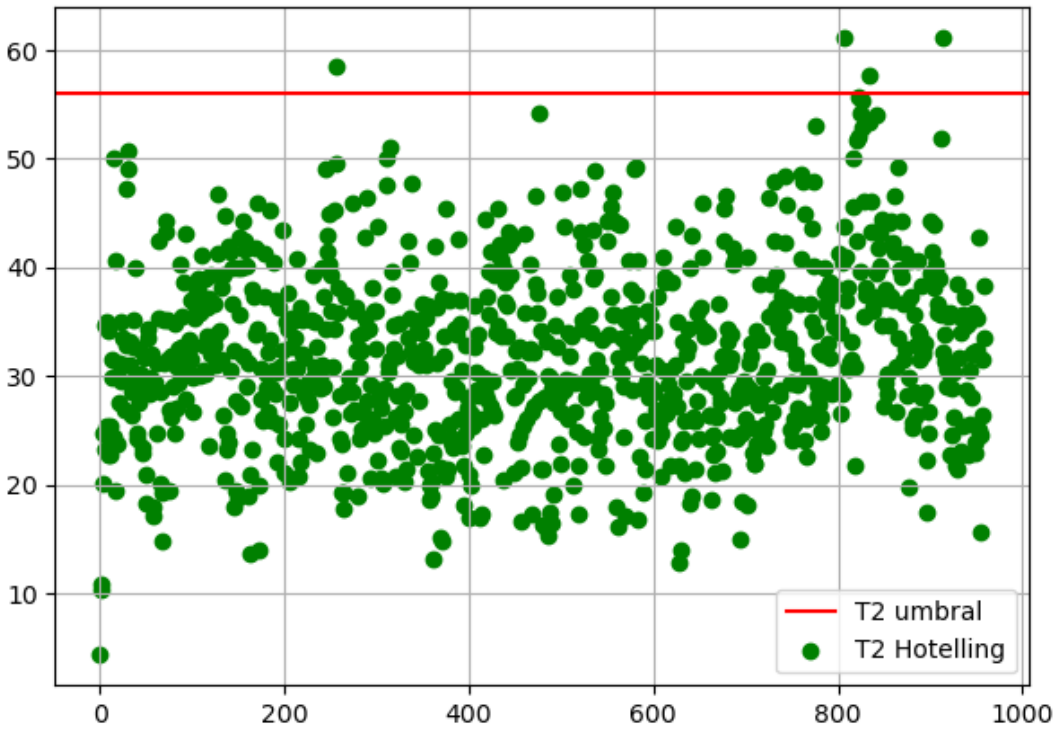

Figura 15. Estadística de Hotelling con su umbral para datos de comportamiento normal en DPCA.

Sin embargo, en la figura 16, se puede ver que, aunque la distribución de los puntos de Q sea similar a la de PCA (figura 11), los valores son más altos y como consecuencia el umbral Q_α en DPCA es superior al de PCA.

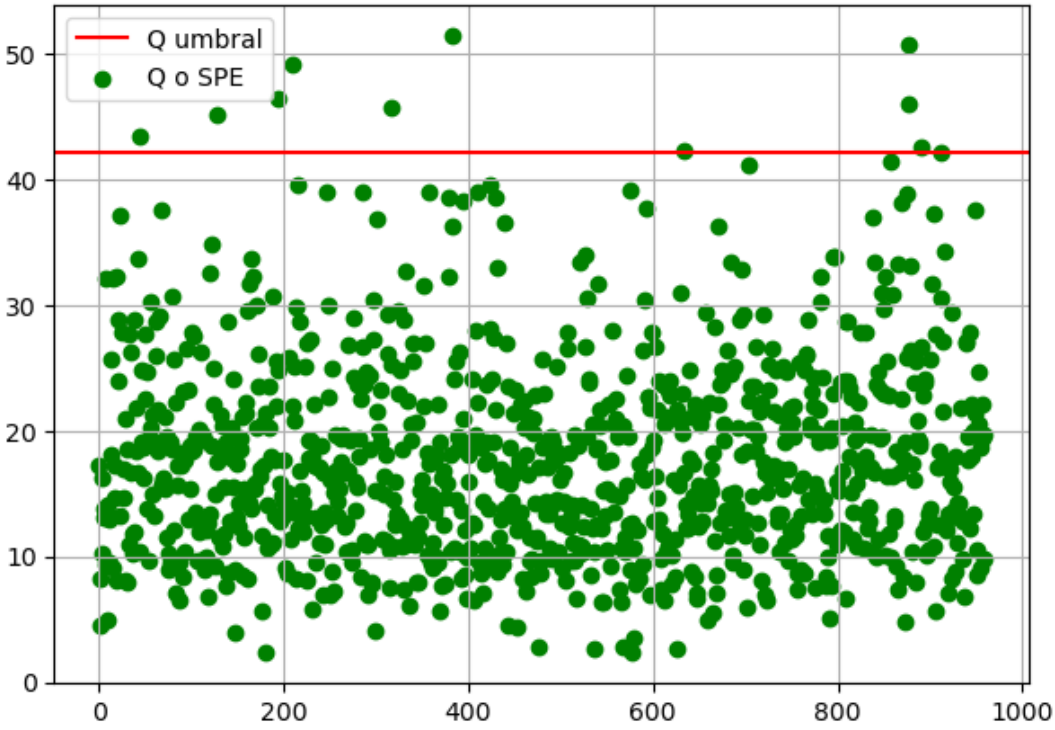

Figura 16. Estadística Q o SPE con su umbral para datos de comportamiento normal en DPCA.

4.3.2. ANÁLISIS DE DATOS CON FALLO APLICANDO DPCA

Una vez que hemos calculado DPCA con los datos normales, pasamos a implementar este método en cada una de las 21 simulaciones de fallo. Al igual que siempre, debemos estandarizar la matriz de datos inicial, en este caso, con la media y la varianza obtenida en el apartado anterior, y calcular los datos dinámicos con la misma metodología, hasta el mismo instante de tiempo que en el análisis con datos de comportamiento normal, obteniendo así una nueva matriz X.

De nuevo aplicaremos PCA a la matriz X y calcularemos los valores de T^2 (ecuación 9) y Q (ecuación 11) para cada una de las 960 observaciones. Al igual que en el método anterior, los valores de ambas estadísticas tienen que superar su correspondiente umbral un total de 8 veces consecutivas, para que el fallo sea detectado.

Como resultado, obtenemos la tabla resumen con los datos que nos proporciona cada fichero de fallo (tabla 5), y también se muestra las figuras resultantes del análisis de 3 de los fallos (figuras 17, 18 y 19).

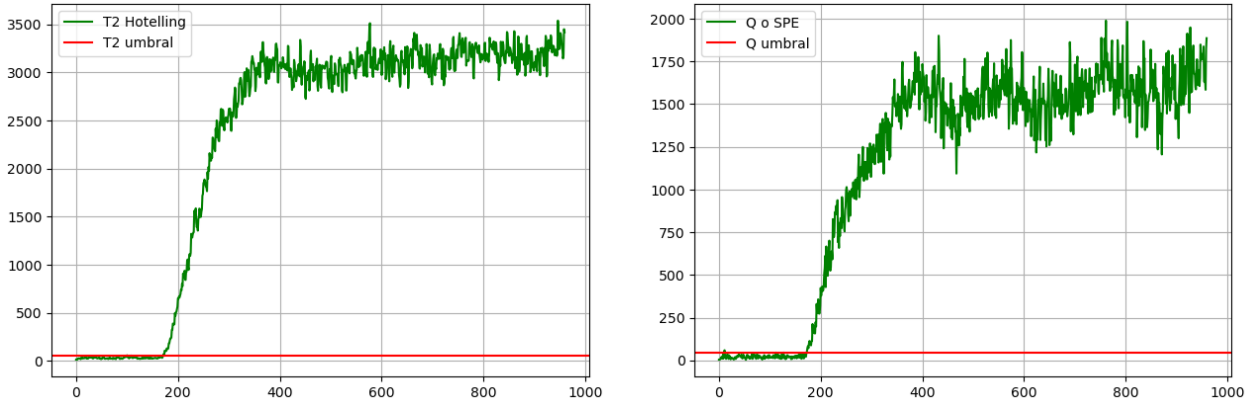

Figura 17. Detección del fallo IDV (2) con la estadística de Hotelling (izquierda) y Q (derecha) en DPCA.

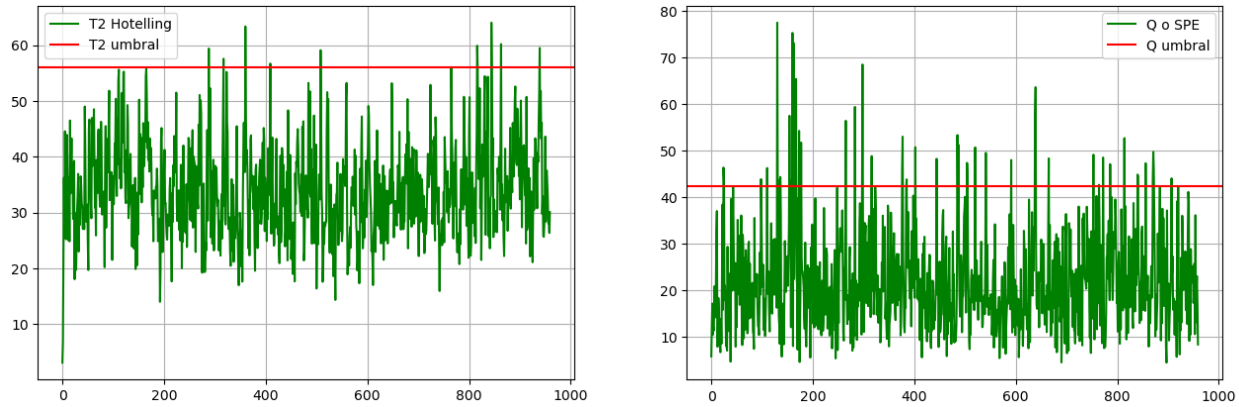

Figura 18. Detección del fallo IDV (9) con la estadística de Hotelling (izquierda) y Q (derecha) en DPCA.

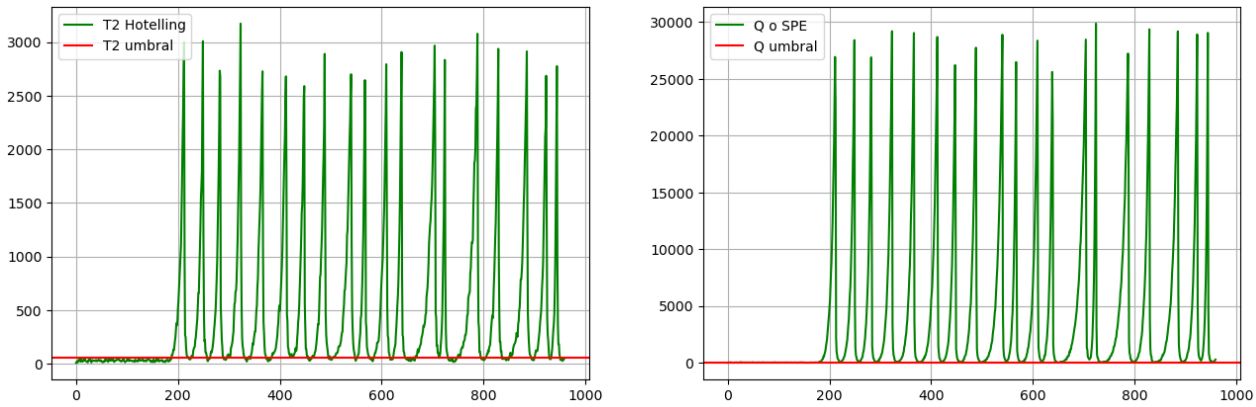

Figura 19. Detección del fallo IDV (17) con la estadística de Hotelling (izquierda) y Q (derecha) en DPCA.

Al igual que ocurre con PCA, el fallo IDV (2) (figura 17) e IDV (17) (figura 19) son detectados por el método DPCA y también lo hace de manera temprana, alrededor del instante 160, en las dos estadísticas. De la misma manera, el fallo IDV (9) (figura 18) tampoco es detectado por este método.

	Hotelling T^2				Q o SPE			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	166	6	99,25	0,00	162	2	99,75	2,50
IDV (2)	172	12	98,50	0,63	172	12	98,50	3,12
IDV (3)	No detecta	No detecta	1,00	0,00	No detecta	No detecta	5,13	6,88
IDV (4)	No detecta	No detecta	4,63	0,63	160	0	100,00	3,75
IDV (5)	170	10	23,75	0,63	160	0	29,38	3,75
IDV (6)	166	6	99,50	0,00	160	0	100,00	3,13
IDV (7)	160	0	100,00	0,00	160	0	100,00	3,13
IDV (8)	182	22	97,25	0,00	177	17	98,13	3,75
IDV (9)	No detecta	No detecta	1,38	0,00	No detecta	No detecta	4,75	6,25
IDV (10)	307	147	27,00	0,00	208	48	44,25	1,25
IDV (11)	622	462	31,63	0,00	166	6	74,88	1,88
IDV (12)	181	21	98,50	0,00	182	22	92,88	3,75
IDV (13)	205	45	94,50	1,25	197	37	95,38	1,25
IDV (14)	160	0	99,50	0,00	160	0	100,00	4,38
IDV (15)	No detecta	No detecta	1,00	0,63	No detecta	No detecta	7,75	5,00
IDV (16)	472	312	10,75	2,50	354	194	40,25	3,13
IDV (17)	188	28	79,00	0,00	181	21	96,00	3,13
IDV (18)	252	92	88,88	0,63	243	83	90,50	3,75
IDV (19)	No detecta	No detecta	5,50	0,00	No detecta	No detecta	21,63	4,38
IDV (20)	246	86	32,25	0,00	244	84	54,00	1,25
IDV (21)	660	500	42,63	0,00	660	500	44,00	5,63
MEDIA	433,76	273,76	54,11	0,33	361,24	201,24	66,53	3,57
MEDIA SIN IDV 3, 9 Y 15	346,06	186,06	62,94	0,35	261,44	101,44	76,64	3,16

Tabla 5. Resultados obtenidos de las estadísticas T^2 y Q, mediante la aplicación de DPCA.

En esta tabla podemos ver un resumen de los resultados obtenidos de ambas estadísticas aplicando DPCA. Hay un total de 5 anomalías no detectadas por este método, las mismas que tampoco muestra PCA, siendo los fallos IDV (3), IDV (9), IDV (15) e IDV (19) no detectados por ninguno de las estadísticas mientras que el fallo IDV (4), solo es localizado por la estadística Q.

Algo a tener en cuenta, y al igual que pasaba en PCA es que, los tiempos de detección de fallo son menores en la estadística Q que en la estadística de Hotelling, pero el porcentaje de falsas alarmas o de alarmas es mayor en Q que en T^2.

4.4. ANÁLISIS CANÓNICO DE CORRELACIÓN (CCA)

Otro de los métodos de análisis estadístico multivariante que se va a emplear en este trabajo, es el Análisis Canónico de Correlación o CCA. Este método, se centra en obtener la estructura de correlación de dos grupos de variables: las variables de proceso (X) y las variables de calidad (Y). Mientras que PCA solo podía obtener información de las variables de proceso (X).

4.4.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON CCA

Como hasta ahora, lo primero es importar los datos de comportamiento normal, pero en este caso los vamos a separar en dos matrices, una será X que almacenará las 22 primeras columnas y las 11 últimas, y la otra se denominará Y que llevará las columnas 35 y 36, por lo tanto, $X \in R^{960x33}$ e $Y \in R^{960x2}$. Cada matriz será normalizada individualmente a media cero y varianza unitaria.

Una vez tengamos los datos normalizados, definimos dos nuevas matrices, $A \in R^{33x33}$ (ecuación 16) y $B \in R^{2x2}$ (ecuación 17) y de cada una de ellas se obtiene los valores propios D_X y D_Y y los vectores de carga correspondientes V_X y V_Y. Con todo esto, definimos la matriz $Z \in R^{33x2}$ (ecuación 20) de la cual obtendremos sus valores singulares.

A continuación, se calculan las matrices $R \in R^{33x2}$ (ecuación 21) y $C \in R^{2x2}$ (ecuación 22), muy similares en concepto a la matriz P de PCA, cuyas dimensiones se verán reducidas debido al número de componentes principales impuesto, en este caso 2. Con estas dos matrices, obtendremos las matrices de componentes principales $T \in R^{960x2}$ (ecuación 23) para X y $U \in R^{960x2}$ (ecuación 24) para Y. Utilizaremos la matriz T para obtener otras dos matrices, $P \in R^{33x2}$ (ecuación 25) y $Q \in R^{2x2}$ (ecuación 26), con las que podremos obtener las matrices proyectadas de los valores de X e Y respectivamente (ecuaciones 27 y 28), en un espacio de trabajo de menor dimensión.

Una vez que hemos obtenido todo lo mencionado anteriormente, pasamos a construir las estadísticas. En este caso, además de la estadística de Hotelling T^2 (ecuación 33), hay dos estadísticas Q o SPE, uno con la matriz residuo de X (ecuación 27), Q_X (ecuación 34) y otro con la matriz residuo de Y (ecuación

28), Q_Y (ecuación 35). Así como también sus umbrales, dos pertenecientes a la estadística Q, $Q_{X\alpha}$ obtenido mediante percentil 99 y $Q_{Y\alpha}$ (ecuación 37) que valen 61,39 y 6,08 respectivamente, y el umbral de T^2, T_α^2 (ecuación 10) con valor 9,27.

Finalmente , se mostrará gráficamente las 3 estadísticas, y podremos comprobar que la gran parte de los valores calculados de estas no superan su correspondiente umbral.

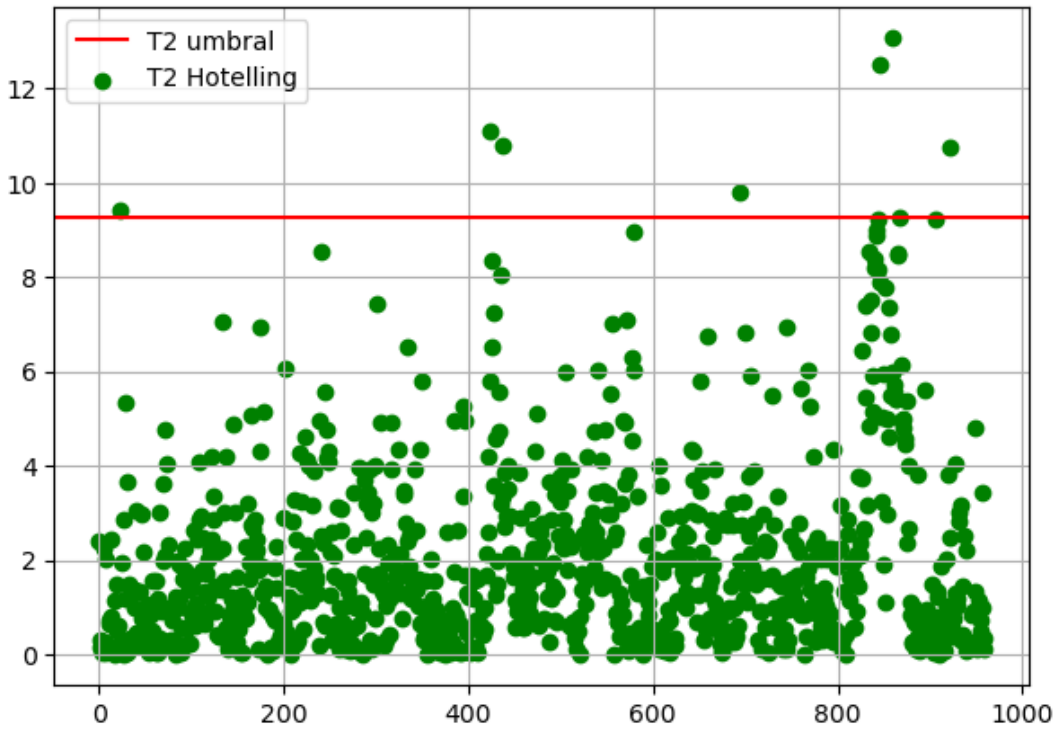

Figura 20. Estadística de Hotelling con su umbral para datos de comportamiento normal en CCA.

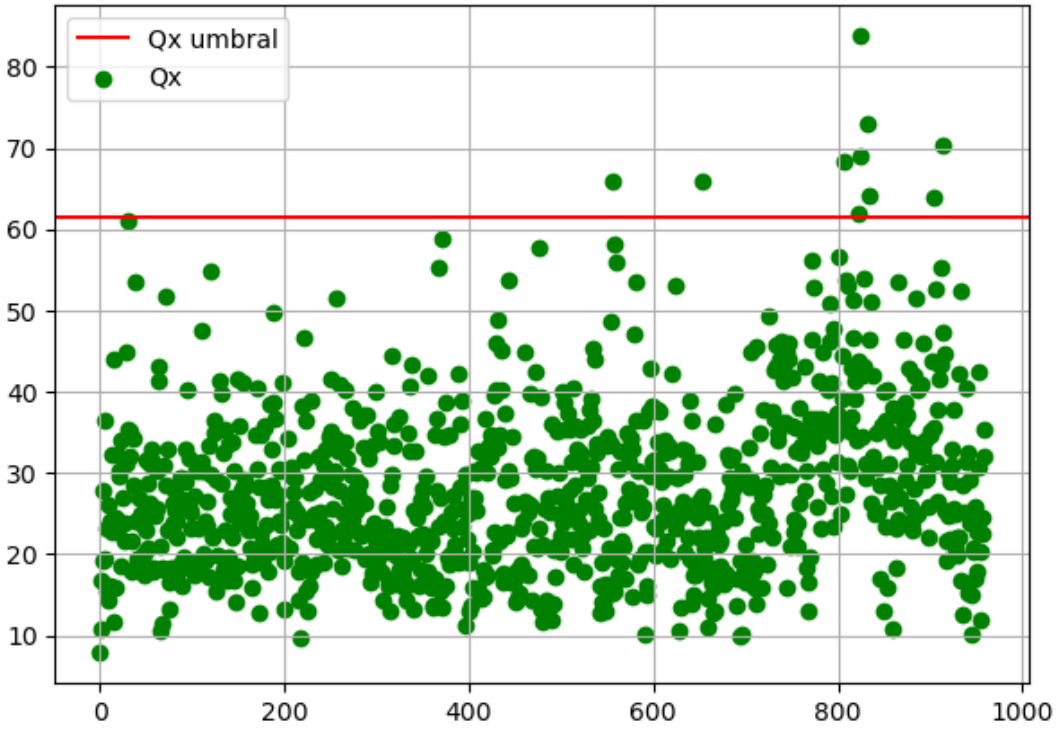

Figura 21. Estadística Q_X con su umbral para datos de comportamiento normal en CCA.

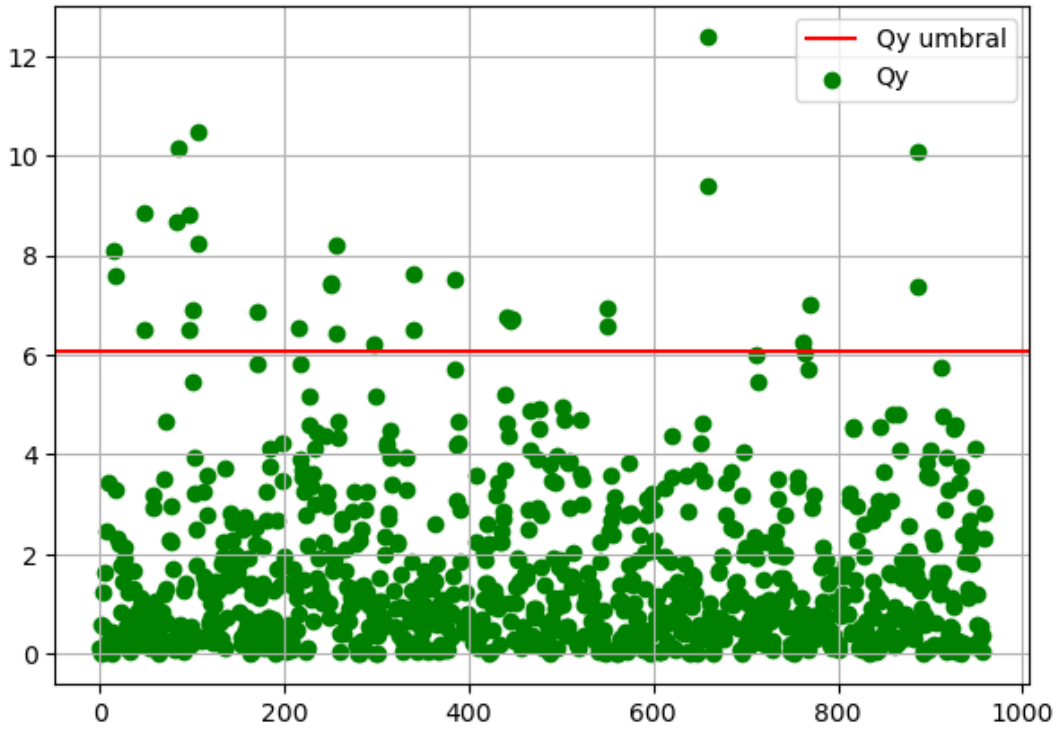

Figura 22. Estadística Q_Y con su umbral para datos de comportamiento normal en CCA.

4.4.2. ANÁLISIS DE DATOS CON FALLO APLICANDO CCA

Primero importamos los datos de fallo, dividiremos estos datos en dos matrices, la matriz X albergará las 22 primeras columnas y las 11 últimas y la matriz Y las columnas o variables 35 y 36. Estas matrices serán normalizadas de manera individual con las medias y varianzas de estas mismas matrices, pero calculadas con el fichero sin fallo del proceso.

Calcularemos la matriz T ∈ R^{960x2} (ecuación 23) utilizando la matriz X que acabamos de obtener y la matriz R ∈ R^{33x2} que obtuvimos aplicando este mismo método a los datos de comportamiento normal. Mas tarde, con esta matriz T, se calcularán los 960 valores correspondientes a la estadística T^2.

Para obtener los datos de las estadísticas Q_X y Q_Y (ecuaciones 34 y 35), necesitamos calcular la matriz residual tanto de X (ecuación 27) como de Y (ecuación 28), para ello necesitaremos importar las matrices P ∈ R^{33x2} y Q ∈ R^{2x2} calculadas en el apartado anterior.

Como hasta ahora, se utilizarán los umbrales de las estadísticas, T_α^2, $Q_{X\alpha}$ y $Q_{Y\alpha}$, calculados con los datos del proceso en condiciones normales, para poder determinar si los datos de fallo superan el umbral 8 veces consecutivas y de esa manera localizar la existencia de anomalías.

A continuación, se muestran las figuras de 3 fallos, donde veremos gráficamente la simulación de los datos y donde quedaría el umbral. También se mostrarán dos tablas donde se recogerá la información aportada por cada estadística.

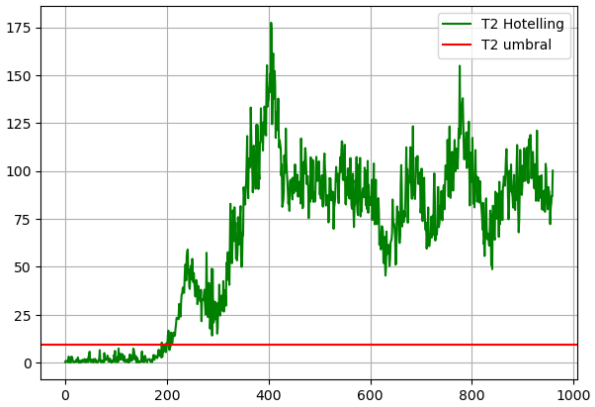

Figura 23. Detección del fallo IDV (2) con la estadística de Hotelling en CCA.

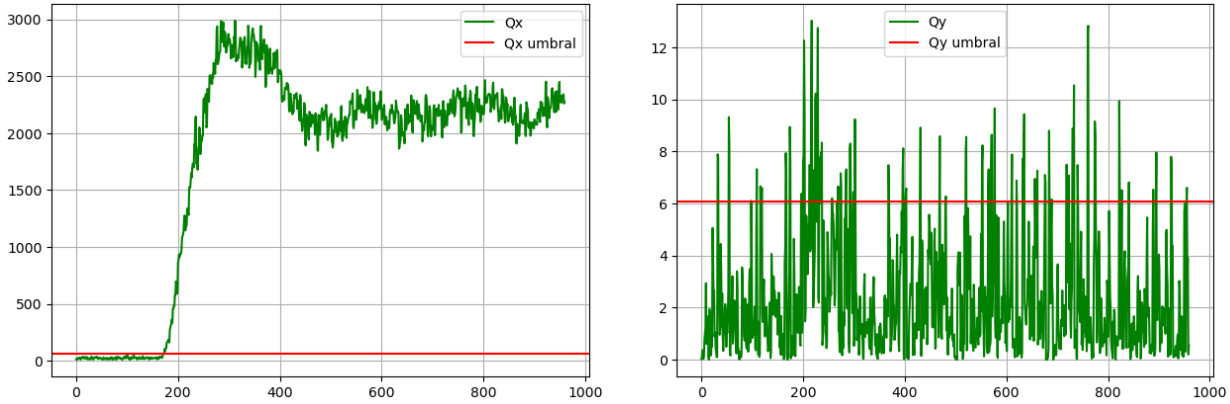

Figura 24. Detección del fallo IDV (2) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA.

En las figuras 23 y 24, donde aparecen las estadísticas del fallo IDV (2) de manera gráfica, se puede apreciar que la estadística de Hotelling y Q_X si detectan el fallo mientras que Q_Y no lo hace. Lo mismo pasa con el fallo IDV (17) (figuras 27 y 28). La estadística Q_Y, como veremos más adelante en la tabla 7, no detecta la gran mayoría de anomalías.

En las figuras 25 y 26 se muestran las estadísticas asociadas al método CCA aplicadas sobre el fallo 9, que como se ha dicho con anterioridad, es una anomalía difícil de detectar.

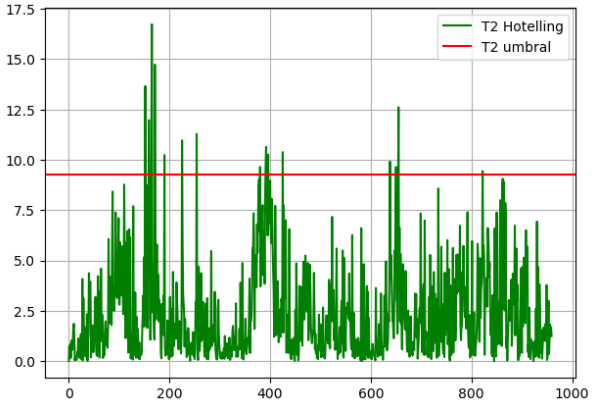

Figura 25. Detección del fallo IDV (9) con la estadística de Hotelling en CCA.

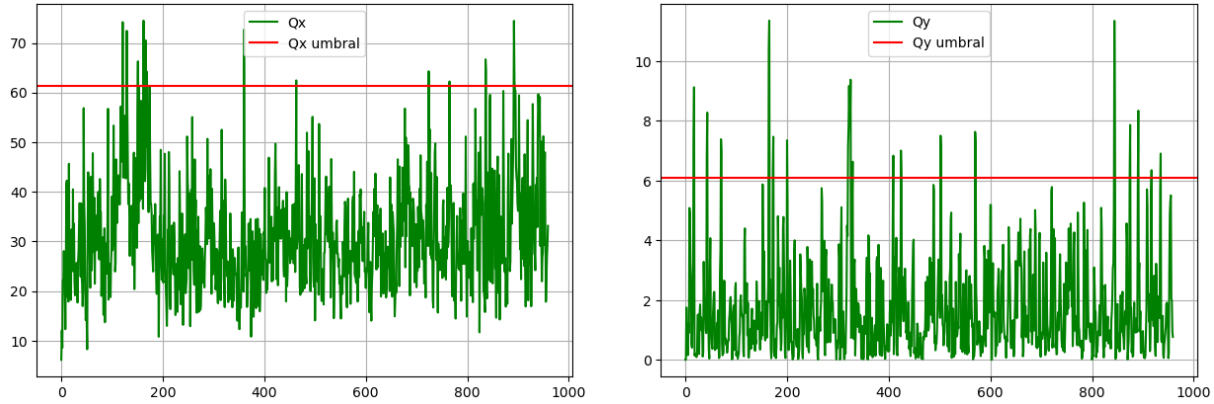

Figura 26. Detección del fallo IDV (9) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA.

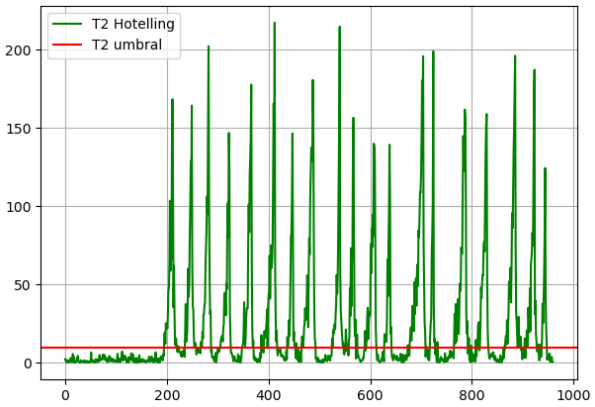

Figura 27. Detección del fallo IDV (17) con la estadística de Hotelling en CCA.

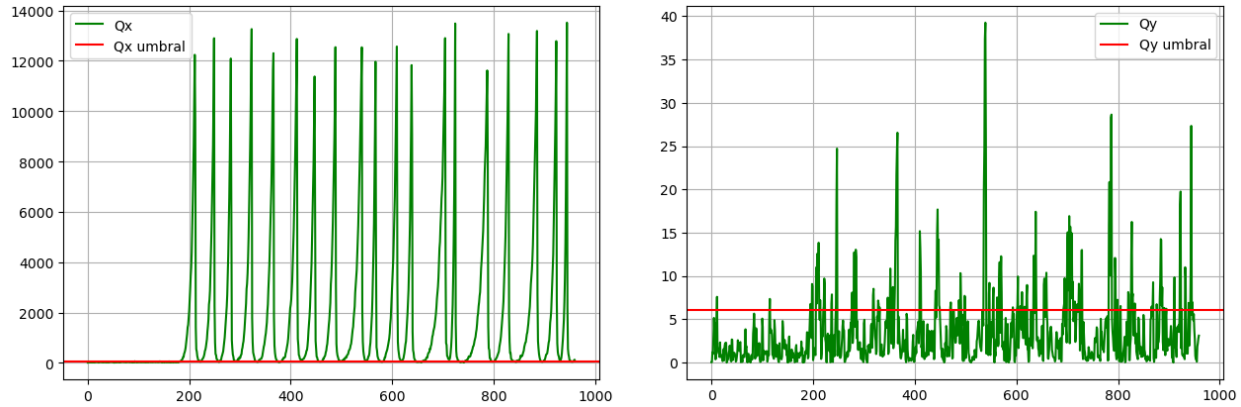

Figura 28. Detección del fallo IDV (17) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA.

	Hotelling T^2			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	171	11	98,13	0,00
IDV (2)	208	48	94,75	0,00
IDV (3)	No detecta	No detecta	2,25	1,88
IDV (4)	No detecta	No detecta	28,38	0,00
IDV (5)	160	0	100,00	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	161	1	93,88	0,00
IDV (8)	189	29	79,63	0,63
IDV (9)	No detecta	No detecta	1,88	1,25
IDV (10)	258	98	22,63	0,00
IDV (11)	306	146	18,63	0,00
IDV (12)	170	10	94,38	1,88
IDV (13)	204	44	92,00	0,63
IDV (14)	No detecta	No detecta	13,38	0,00
IDV (15)	No detecta	No detecta	2,13	0,00
IDV (16)	171	11	59,50	2,50
IDV (17)	195	35	53,63	0,00
IDV (18)	243	83	89,88	0,63
IDV (19)	598	438	46,75	0,00
IDV (20)	239	79	66,38	0,00
IDV (21)	595	435	49,63	2,50
MEDIA	420,38	260,38	57,51	0,57
MEDIA SIN FALLOS 3, 9 Y 15	330,44	170,44	66,75	0,49

Tabla 6. Resultados obtenidos de la estadística de Hotelling, mediante la aplicación de CCA.

Los datos obtenidos de la estadística T^2, se muestran en la tabla 6. Los fallos incipientes siguen sin ser detectados por este método y a mayores tampoco manifiesta las anomalías IDV (14) e IDV (4), siendo un total de 5 fallos no detectados.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	165	5	99,38	0,63	264	104	57,50	1,25
IDV (2)	171	11	98,63	0,00	No detecta	No detecta	11,25	5,00
IDV (3)	No detecta	No detecta	0,75	0,00	No detecta	No detecta	3,75	3,75
IDV (4)	162	2	82,25	0,63	No detecta	No detecta	3,75	1,25
IDV (5)	160	0	100,00	0,63	162	2	99,75	1,25
IDV (6)	160	0	100,00	0,63	160	0	100,00	1,25
IDV (7)	160	0	100,00	0,00	160	0	21,50	3,13
IDV (8)	180	20	97,63	0,00	324	164	21,25	2,50
IDV (9)	No detecta	No detecta	1,38	1,88	No detecta	No detecta	3,25	3,75
IDV (10)	226	66	37,38	0,00	No detecta	No detecta	8,13	1,25
IDV (11)	206	46	59,13	0,00	No detecta	No detecta	5,13	2,50
IDV (12)	162	2	99,75	0,00	181	21	78,13	5,63
IDV (13)	206	46	94,25	0,00	430	270	47,50	1,88
IDV (14)	160	0	100,00	0,00	No detecta	No detecta	6,13	6,25
IDV (15)	No detecta	No detecta	0,75	0,00	No detecta	No detecta	2,75	7,50
IDV (16)	355	195	38,88	4,38	No detecta	No detecta	8,38	3,13
IDV (17)	183	23	86,88	0,63	No detecta	No detecta	19,25	1,25
IDV (18)	246	86	89,50	0,00	248	88	89,38	1,25
IDV (19)	No detecta	No detecta	19,63	0,00	No detecta	No detecta	7,38	3,75
IDV (20)	249	89	46,63	0,00	294	134	29,25	3,75
IDV (21)	656	496	42,75	0,00	No detecta	No detecta	6,00	2,50
MEDIA	364,14	204,14	66,45	0,45	654,43	494,43	29,97	3,04
MEDIA SIN FALLOS 3, 9 Y 15	264,83	104,83	77,37	0,42	603,50	443,50	34,42	2,71

Tabla 7. Resultados obtenidos de las estadísticas Q_X y Q_Y, mediante la aplicación de CCA.

En la tabla 7, se muestran las estadísticas restantes de CCA, Q_X y Q_Y. En este caso, la estadística Q_X no detecta 4 anomalías, entre ellas los 3 fallos incipientes, las mismas que no detecta la estadística Q en PCA y DPCA. Las anomalías que, si son detectadas por esta estadística, en su gran mayoría, lo hacen de manera temprana, entorno al instante 160, además los porcentajes de falsas alarmas son mínimos, por debajo del 5%.

En cuanto a la estadística Q_Y, podemos ver que no detecta la gran mayoría de fallos, pero el porcentaje de alarmas de estos es bastante bajo, esto quiere decir que después del instante 160, que es cuando se produce la anomalía,

muy pocos valores de esta estadística superan el umbral, por lo que se puede deducir, que la perturbación no afecta a las variables de calidad (Y).

4.5. ANÁLISIS CANÓNICO DE CORRELACIÓN REGULARIZADO (CCA CON REGULARIZACIÓN)

CCA Regularizado se realiza con el mismo procedimiento que CCA, lo único que varía es la forma de obtener la matriz Z ϵ R^{33x2} (ecuación 20). De esta manera se solucionaría el problema de colinealidad que sufre el método CCA tradicional.

Tanto para el análisis de datos de comportamiento normal como para la detección de fallos se importan los datos correspondientes, y se dividen en dos matrices X e Y, de la misma manera que en el método CCA, y estandarizando individualmente ambas matrices. Mas tarde se calculan las matrices A ϵ R^{33x33} (ecuación 16) y B ϵ R^{2x2} (ecuación 17) de las que obtendremos sus respectivos valores singulares.

El cambio aparece a la hora de calcular las matrices $\Sigma_{xx}^{-1/2}$ (ecuación 29) y $\Sigma_{yy}^{-1/2}$ (ecuación 30) que son necesarias para calcular la matriz Z. En este método se van a definir dos valores a los que hemos denominado K_1 y K_2:

$$K_1 = 0{,}001 \; y \; K_2 = 0{,}068$$

Una vez que hemos calculado $\Sigma_{xx}^{-1/2}$ y $\Sigma_{yy}^{-1/2}$, el procedimiento a seguir tanto para el análisis de comportamiento normal como para la detección de fallos es el mismo que en el método de análisis multivariante anterior.

4.5.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON CCA REGULARIZADO

Aplicando la misma metodología empleada en CCA con los datos de comportamiento normal del proceso, calcularemos las estadísticas T^2 (ecuación 33), Q_X (ecuación 34) y Q_Y (ecuación 35) y sus respectivos umbrales T_α^2 (ecuación 10), $Q_{X\alpha}$ (percentil 99) y $Q_{Y\alpha}$ (ecuación 37). A continuación, se mostrarán las figuras resultantes de los 3 estadísticas (figuras 29, 30 y 31) y sus umbrales al aplicar este método. El umbral T_α^2 tiene como valor 9,27, $Q_{X\alpha}$ corresponde a 61,01 y $Q_{Y\alpha}$ a 6,10.

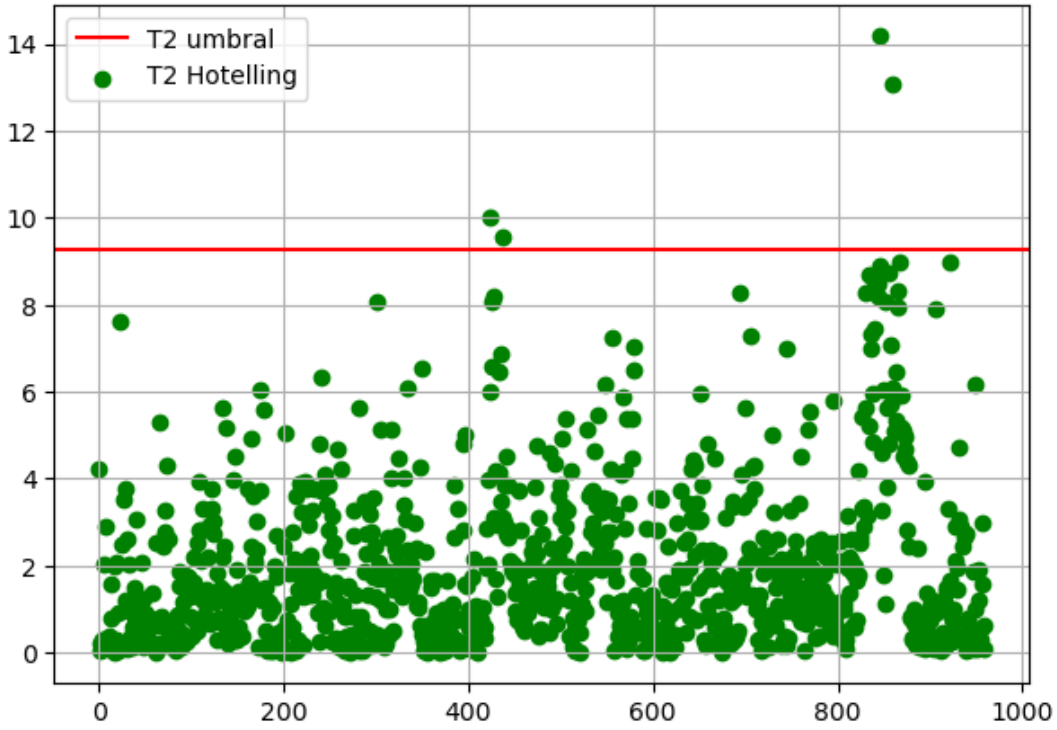

Figura 29. Estadística de Hotelling con su umbral para datos de comportamiento normal en CCA Regularizado.

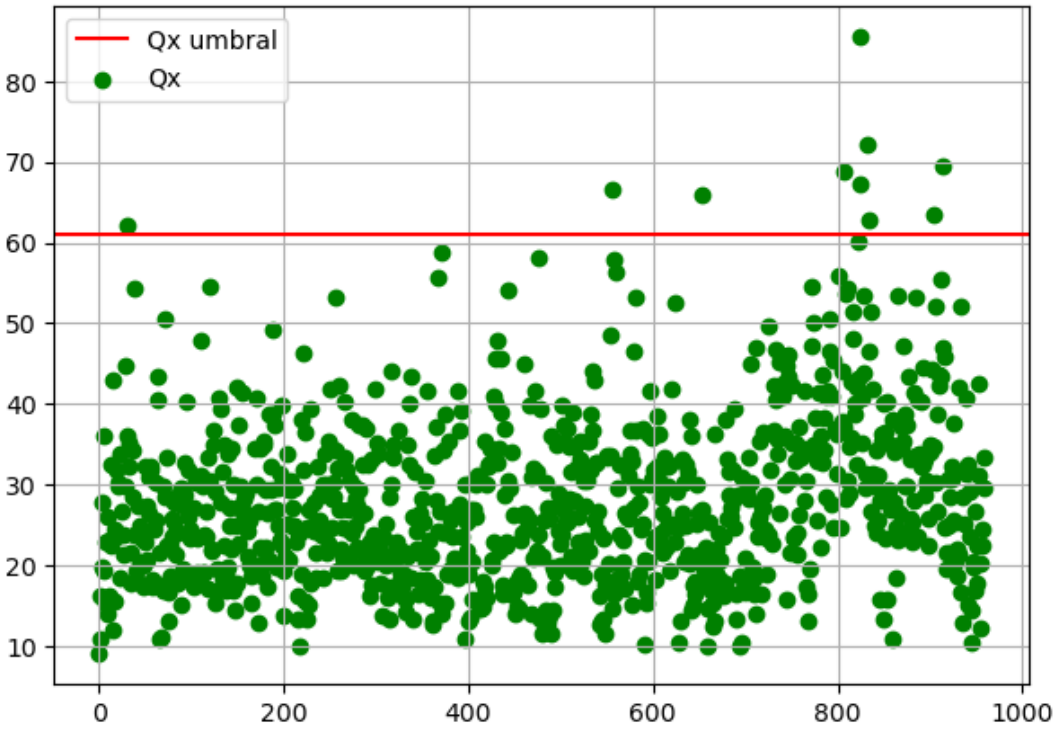

Figura 30. Estadística Q_x con su umbral para datos de comportamiento normal en CCA Regularizado.

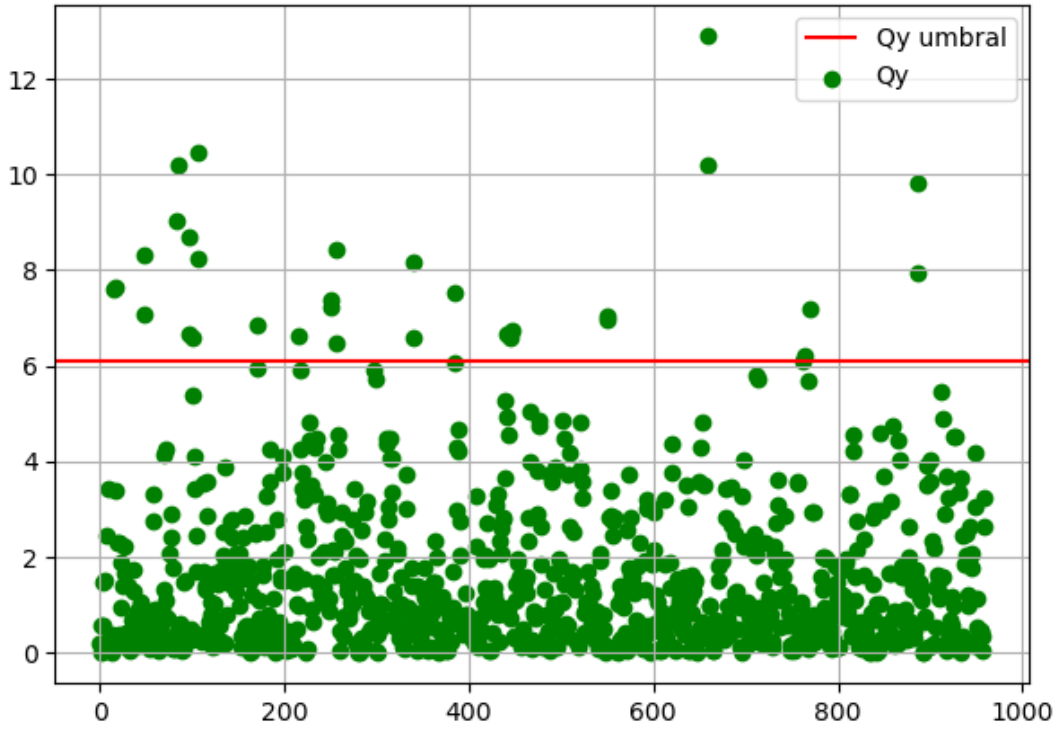

Figura 31. Estadística Q_Y con su umbral para datos de comportamiento normal con CCA Regularizado.

4.5.2. ANÁLISIS DE DATOS CON FALLO APLICANDO CCA REGULARIZADO

Aplicaremos la misma metodología empleada en CCA, exceptuando los pasos para obtener la matriz Z, ya explicados con anterioridad. Calcularemos los 960 valores de cada estadística con cada uno de los 21 ficheros de datos, utilizaremos los umbrales calculados en el análisis de comportamiento normal para poder comparar y discutir si este método de análisis de datos es el más apropiado para este proceso.

A continuación, se mostrarán las estadísticas de 3 anomalías de manera gráfica (figuras 32-37) y la información obtenida de las estadísticas al aplicar CCA regularizado a los 21 fallos, recogida en las tablas 8 y 9.

La estadística de Hotelling detecta los fallos IDV (2) (figura 32) e IDV (17) (figura 36), pero no lo hace con el fallo incipiente IDV (9) (figura 34). Además, se puede observar en la tabla 8 como esta estadística no detecta un total de 6 anomalías, una más que en el resto de los métodos de análisis multivariante explicados hasta ahora.

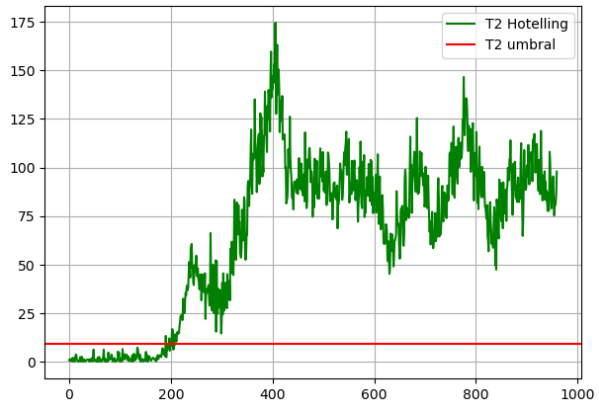

Figura 32. Detección del fallo IDV (2) con la estadística de Hotelling en CCA Regularizado.

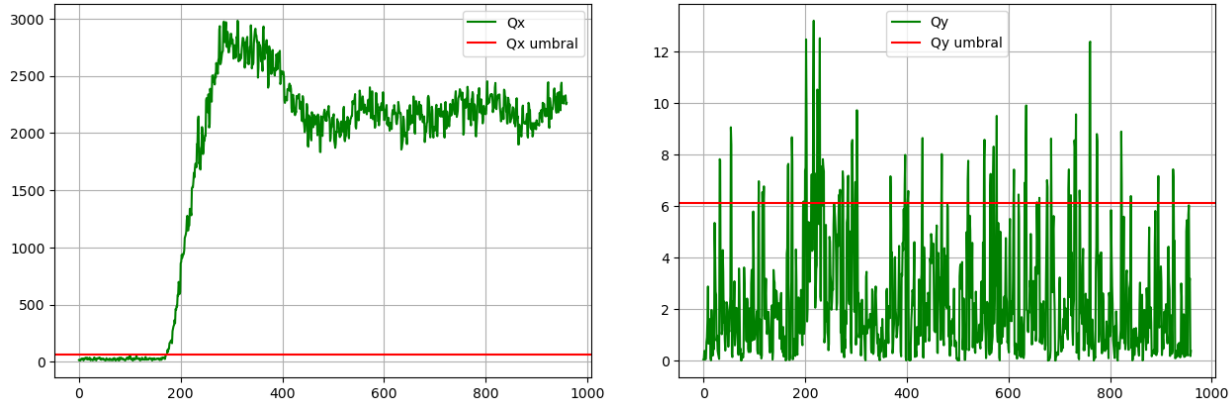

Figura 33. Detección del fallo IDV (2) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA Regularizado.

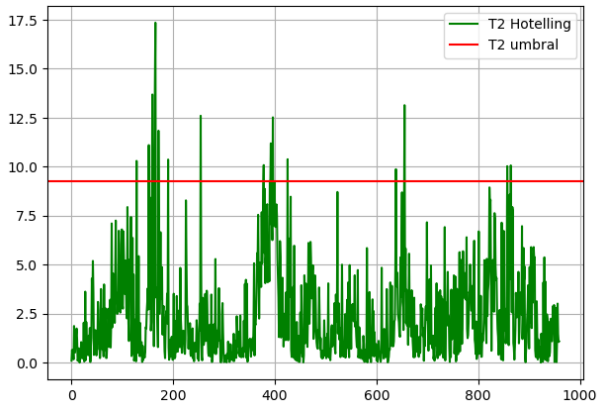

Figura 34. Detección del fallo IDV (9) con la estadística de Hotelling en CCA Regularizado.

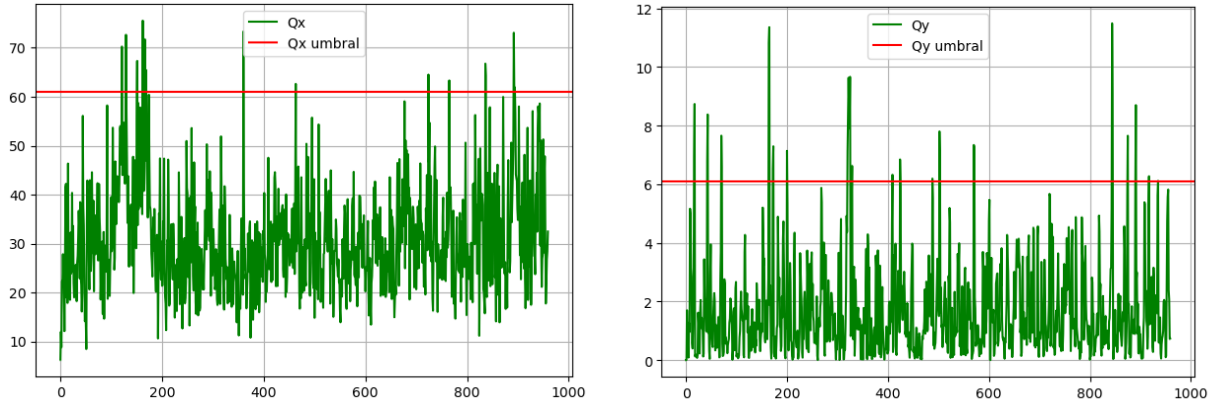

Figura 35. Detección del fallo IDV (9) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA Regularizado.

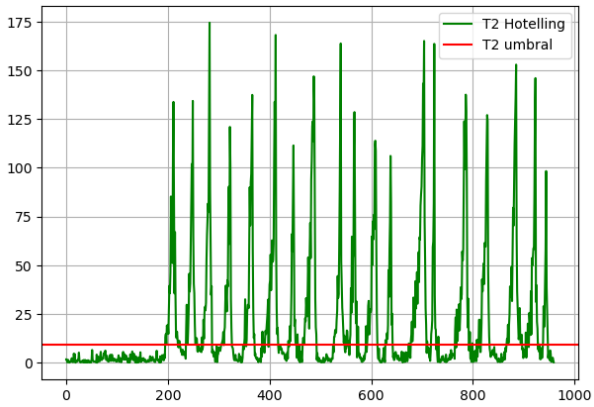

Figura 36. Detección del fallo IDV (17) con la estadística de Hotelling en CCA Regularizado.

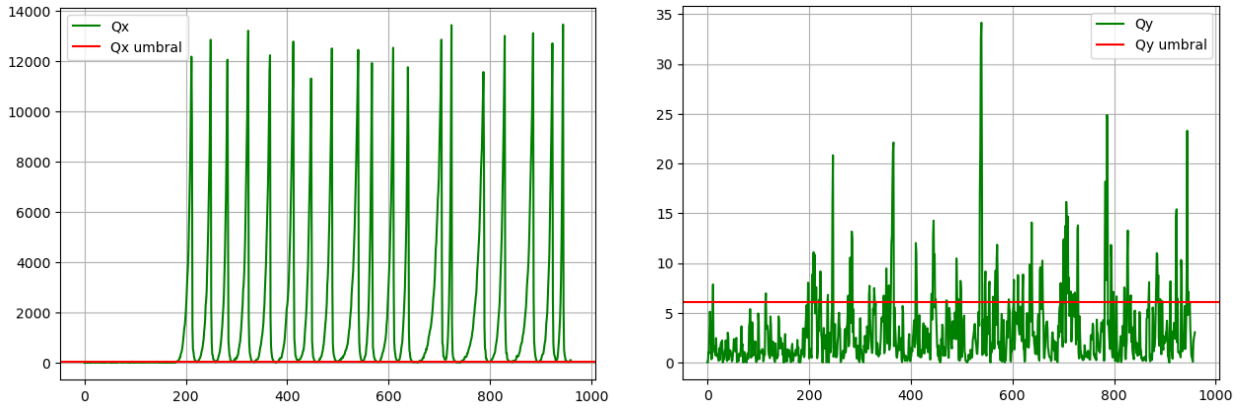

Figura 37. Detección del fallo IDV (17) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCA Regularizado.

En las figuras 33, 35 y 37 podemos observar que la estadística Q_Y no detecta ninguno de los fallos mostrados, mientras que Q_X si lo hace, excluyendo la anomalía IDV (9) difícil de localizar. La información obtenida de ambas estadísticas se muestra en la tabla 9, donde también podremos ver que, tanto el porcentaje de falsas alarmas como el tiempo de detección del fallo, tienen mejores resultados con la estadística Q_X.

	Hotelling T^2			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	171	11	98,13	0,00
IDV (2)	208	48	94,88	0,00
IDV (3)	No detecta	No detecta	2,63	3,75
IDV (4)	No detecta	No detecta	28,88	0,00
IDV (5)	160	0	100,00	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	161	1	94,00	0,00
IDV (8)	189	29	80,38	0,63
IDV (9)	No detecta	No detecta	1,63	1,88
IDV (10)	258	98	23,75	0,00
IDV (11)	306	146	19,50	0,63
IDV (12)	170	10	94,50	1,25
IDV (13)	204	44	91,75	0,63
IDV (14)	No detecta	No detecta	11,75	0,00
IDV (15)	No detecta	No detecta	2,13	0,00
IDV (16)	171	11	59,13	4,38
IDV (17)	195	35	50,50	0,00
IDV (18)	244	84	89,75	9,00
IDV (19)	No detecta	No detecta	47,88	0,00
IDV (20)	239	79	66,63	0,00
IDV (21)	577	417	50,50	2,50
MEDIA	436,81	276,81	57,54	0,74
MEDIA SIN FALLOS 3, 9 Y 15	349,61	189,61	66,77	0,56

Tabla 8. Resultados obtenidos de la estadística de Hotelling, mediante la aplicación de CCA Regularizado.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	165	5	99,38	0,63	264	104	56,50	1,25
IDV (2)	171	11	98,63	0,00	No detecta	No detecta	10,25	5,00
IDV (3)	No detecta	No detecta	1,00	0,00	No detecta	No detecta	4,50	4,38
IDV (4)	162	2	83,63	0,63	No detecta	No detecta	3,63	1,25
IDV (5)	160	0	100,00	0,63	162	2	99,75	1,25
IDV (6)	160	0	100,00	0,63	160	0	100,00	1,25
IDV (7)	160	0	100,00	0,00	160	0	22,00	3,13
IDV (8)	180	20	97,63	0,00	324	164	21,25	2,50
IDV (9)	No detecta	No detecta	1,38	1,88	No detecta	No detecta	3,13	3,75
IDV (10)	226	66	37,25	0,00	No detecta	No detecta	7,50	1,88
IDV (11)	206	46	59,25	0,00	No detecta	No detecta	5,00	2,50
IDV (12)	162	2	99,75	0,00	181	21	77,63	5,00
IDV (13)	206	46	94,25	0,00	430	270	46,13	3,13
IDV (14)	160	0	100,00	0,00	No detecta	No detecta	5,75	6,88
IDV (15)	No detecta	No detecta	1,00	0,00	No detecta	No detecta	2,38	6,88
IDV (16)	355	195	38,75	4,38	No detecta	No detecta	7,50	3,13
IDV (17)	183	23	87,25	0,63	No detecta	No detecta	16,00	1,25
IDV (18)	246	86	89,50	0,00	248	88	89,25	1,25
IDV (19)	No detecta	No detecta	21,38	0,00	No detecta	No detecta	6,75	3,75
IDV (20)	249	89	46,88	0,00	294	134	28,38	3,75
IDV (21)	656	496	42,50	0,00	No detecta	No detecta	5,38	2,50
MEDIA	364,14	204,14	66,64	0,45	654,43	494,43	29,46	3,13
MEDIA SIN FALLOS 3, 9 Y 15	264,83	104,83	77,56	0,42	603,50	443,50	33,81	2,81

Tabla 9. Resultados obtenidos de las estadísticas Q_X y Q_Y, mediante la aplicación de CCA Regularizado.

4.6. ANÁLISIS CANÓNICO DE CORRELACIÓN DINÁMICO (DCCA)

Este nuevo método de análisis estadístico multivariante, parte de la base de CCA con regularización, es decir, se aplicará el método CCA Regularizado a datos dinámicos. Los datos dinámicos serán obtenidos de manera similar a como se comentó en el método DPCA, pero en este caso se aplicará dos veces, una para la matriz X y otra para Y.

4.6.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON DCCA

Como hasta ahora, importaremos el fichero de datos de comportamiento normal en el proceso y definiremos dos matrices, la primera almacenará las 22 primeras variables y las 11 ultimas con sus 960 observaciones, y la segunda albergará las columnas 35 y 36. Estas matrices serán normalizadas a media cero y varianza unitaria. En este método, hemos variado el número de componentes principales "a", pasando de 2, que ha tenido como valor en CCA y CCA Regularizado, a 5.

A continuación, calcularemos las matrices de datos dinámicos X e Y, como hemos comentado antes, y a partir de ellas obtendremos las tres estadísticas T^2 (figura 38), Q_X (figura 39) y Q_Y (figura 40) y sus respectivos umbrales, aplicando el método CCA con regularización, con la excepción de que, en este caso tanto $Q_{X\alpha}$ como $Q_{Y\alpha}$ se calcularán mediante el percentil 99. Para este método el umbral T_α^2 tendrá un valor de 15,26, $Q_{X\alpha}$ será 243,38 y finalmente $Q_{Y\alpha}$ será 32,64.

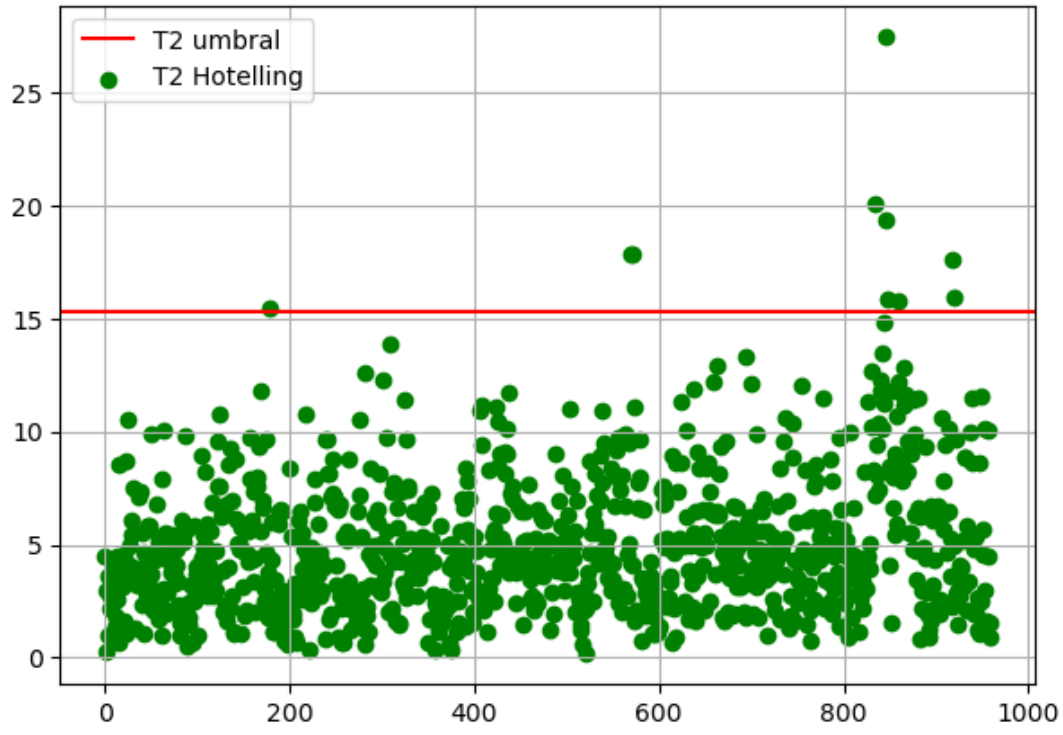

Figura 38. Estadística de Hotelling con su umbral para datos de comportamiento normal.

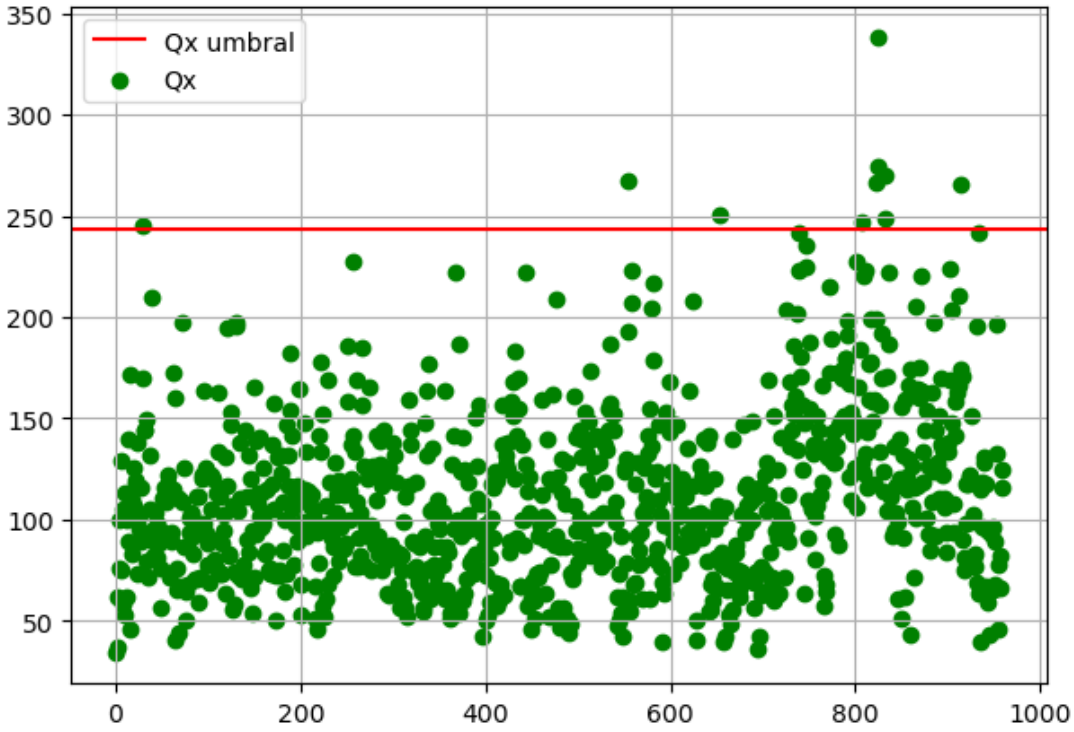

Figura 39. Estadística Q_X con su umbral para datos de comportamiento normal.

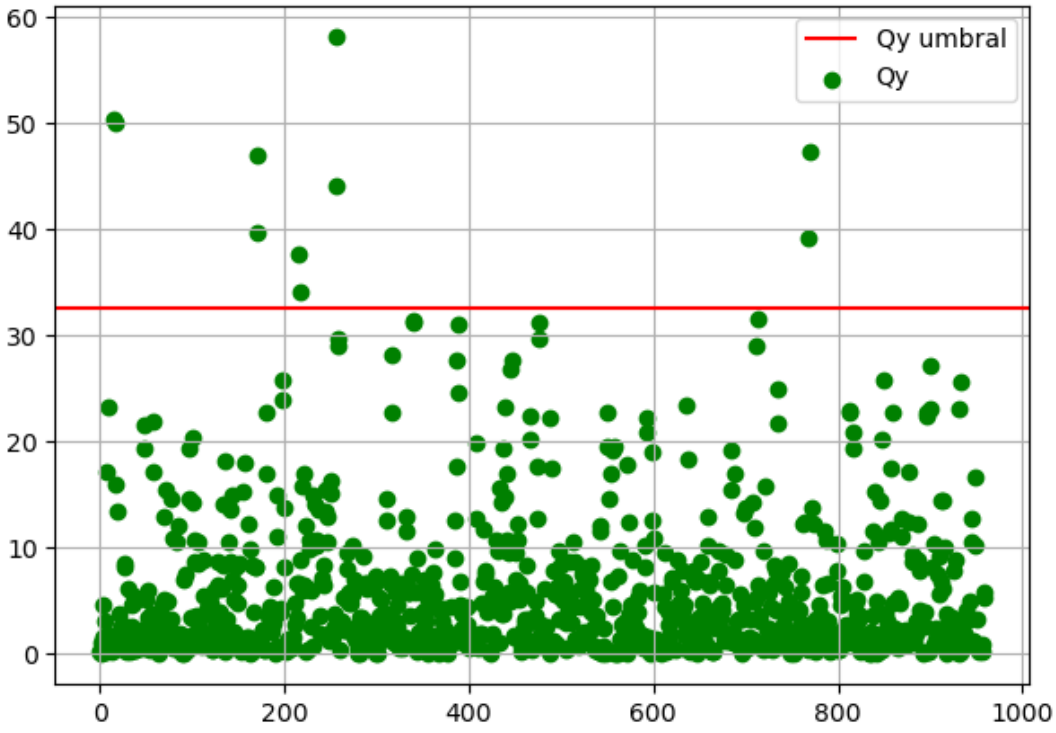

Figura 40. Estadística Q_Y con su umbral para datos de comportamiento normal.

4.6.2. ANÁLISIS DE DATOS CON FALLO APLICANDO DCCA

Aplicaremos lo explicado para datos normales del proceso en los 21 ficheros de fallo, y calcularemos las estadísticas. Los umbrales son los calculados en el apartado anterior.

Se mostrará, a continuación, los resultados de las estadísticas de forma gráfica de los fallos 2 (figuras 41 y 42), 9 (figuras 43 y 44) y 17 (figuras 45 y 46), y también un resumen de toda la información que nos aporta el método a través de las estadísticas, mostrado en las tablas 10 y 11.

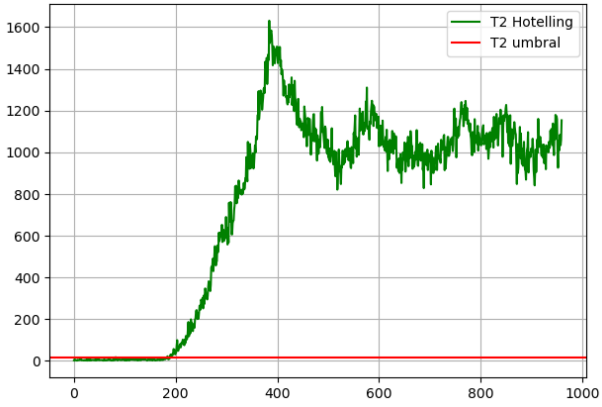

Figura 41. Detección del fallo IDV (2) con la estadística de Hotelling en DCCA.

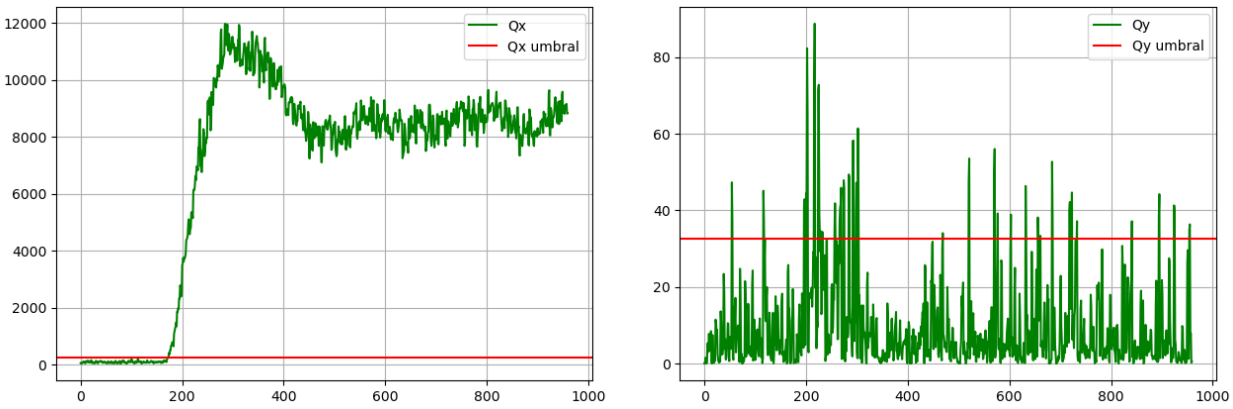

Figura 42. Detección del fallo IDV (2) con la estadística Q_X (izquierda) y Q_Y (derecha) en DCCA.

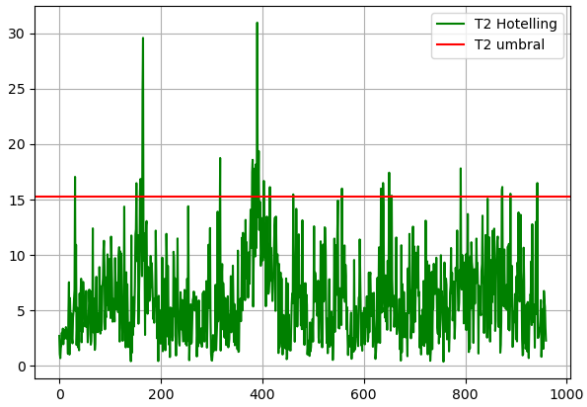

Figura 43. Detección del fallo IDV (9) con la estadística de Hotelling en DCCA.

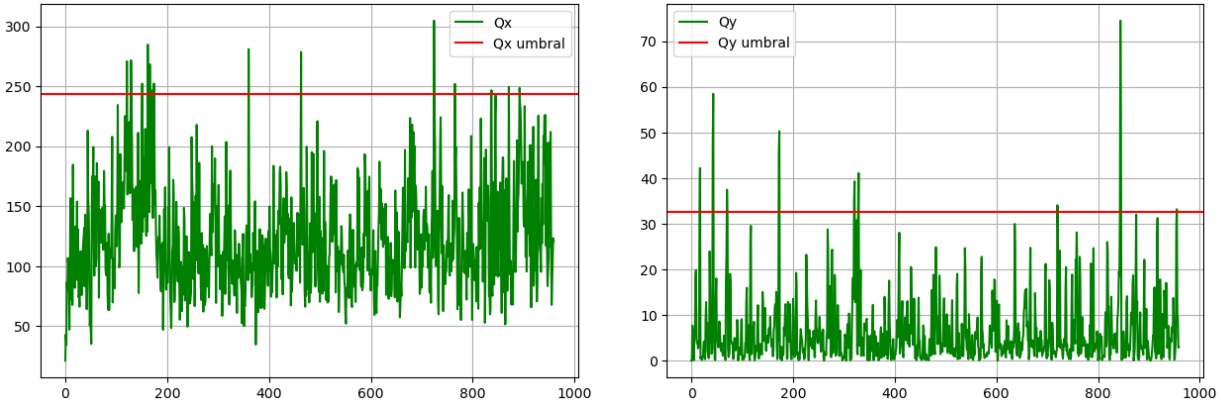

Figura 44. Detección del fallo IDV (9) con la estadística Q_X (izquierda) y Q_Y (derecha) en DCCA.

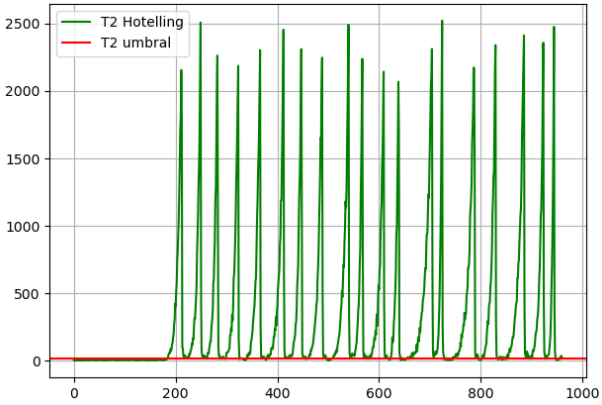

Figura 45. Detección del fallo IDV (17) con la estadística de Hotelling en DCCA.

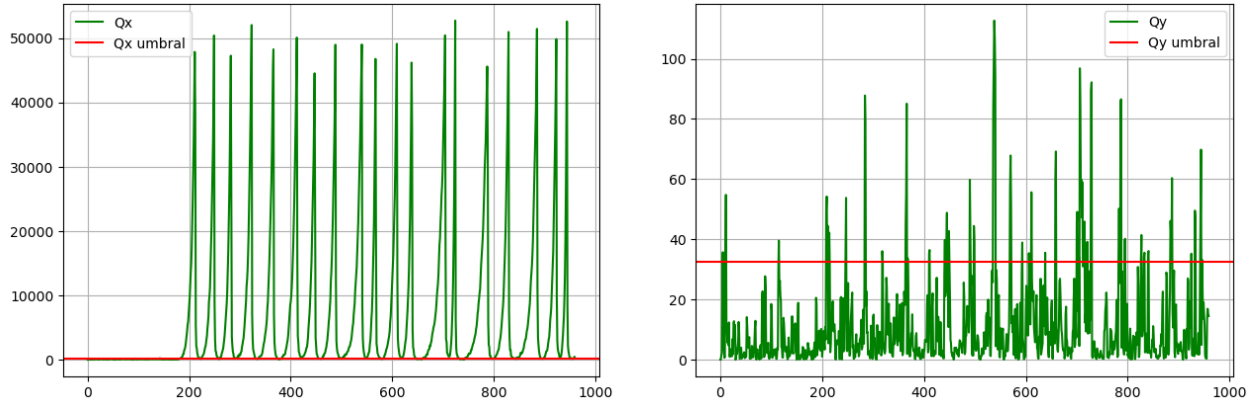

Figura 46. Detección del fallo IDV (17) con la estadística Q_X (izquierda) y Q_Y (derecha) en DCCA.

Como se puede observar en las figuras 42, 44 y 46, la estadística Q_Y no detecta ninguno de los fallos, mientras que Q_X detecta IDV (2) e IDV (17) pero no lo hace con el fallo incipiente IDV (9). Esto mismo ocurre en CCA y CCA con regularización, ya explicado con anterioridad.

La estadística de Hotelling que aparece representada en las figuras 41, 43 y 45, localiza las anomalías 2 y 17, pero no la 9, como viene haciendo hasta ahora con los fallos representados gráficamente.

Hotelling T^2

	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	162	2	99,75	1,25
IDV (2)	190	30	96,63	0,63
IDV (3)	No detecta	No detecta	3,50	8,75
IDV (4)	No detecta	No detecta	24,88	0,63
IDV (5)	160	0	100,00	0,63
IDV (6)	160	0	100,00	0,00
IDV (7)	160	0	98,88	0,00
IDV (8)	179	19	95,75	0,63
IDV (9)	No detecta	No detecta	3,00	1,88
IDV (10)	182	22	82,88	0,00
IDV (11)	306	146	22,75	1,25
IDV (12)	161	1	99,75	0,63
IDV (13)	204	44	94,50	0,63
IDV (14)	240	80	68,63	0,63
IDV (15)	No detecta	No detecta	4,88	1,88

IDV (16)	171	11	82,13	3,75
IDV (17)	185	25	85,38	0,00
IDV (18)	243	83	90,25	0,63
IDV (19)	594	434	54,63	0,63
IDV (20)	224	64	90,75	0,00
IDV (21)	563	403	53,63	5,63
MEDIA	377,33	217,33	69,17	1,43
MEDIA SIN FALLOS 3, 9 Y 15	280,22	120,22	80,06	0,97

Tabla 10. Resultados obtenidos de la estadística de Hotelling, mediante la aplicación de DCCA.

En la tabla 10, aparece toda la información que obtuvimos de la estadística T^2 aplicando DCCA al proceso. Se puede ver que en este caso no detecta 4 fallos, 3 de ellos incipientes. Hasta ahora es el método que detecta más fallos aplicando la estadística de Hotelling.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	165	5	99,38	0,00	264	104	61,00	1,88
IDV (2)	172	12	98,50	0,00	No detecta	No detecta	6,88	2,50
IDV (3)	No detecta	No detecta	0,63	0,00	No detecta	No detecta	1,25	1,88
IDV (4)	No detecta	No detecta	25,63	1,25	No detecta	No detecta	1,38	1,25
IDV (5)	160	0	100,00	1,25	162	2	99,75	1,25
IDV (6)	160	0	100,00	0,00	160	0	100,00	1,88
IDV (7)	160	0	100,00	0,00	160	0	5,88	0,63
IDV (8)	180	20	97,63	0,00	324	164	18,75	0,00
IDV (9)	No detecta	No detecta	1,38	1,88	No detecta	No detecta	1,13	2,50
IDV (10)	209	49	45,13	0,00	No detecta	No detecta	5,38	0,00
IDV (11)	210	50	49,25	0,00	No detecta	No detecta	2,25	1,88
IDV (12)	162	2	99,75	0,00	181	21	70,75	3,13
IDV (13)	206	46	94,25	0,00	432	272	35,00	1,25
IDV (14)	160	0	100,00	0,00	No detecta	No detecta	3,00	3,13
IDV (15)	No detecta	No detecta	1,38	0,00	No detecta	No detecta	0,38	3,75
IDV (16)	280	120	43,88	5,63	No detecta	No detecta	1,38	0,00
IDV (17)	183	23	87,13	1,25	No detecta	No detecta	9,50	2,50
IDV (18)	246	86	89,63	0,00	247	87	89,13	0,00
IDV (19)	No detecta	No detecta	21,00	0,00	No detecta	No detecta	6,00	0,00

IDV (20)	247	87	52,75	0,00	814	654	16,88	1,25
IDV (21)	660	500	41,63	0,00	No detecta	No detecta	4,00	0,00
MEDIA	398,10	238,10	64,23	0,54	679,24	519,24	25,70	1,46
MEDIA SIN FALLOS 3, 9 Y 15	304,44	144,44	74,75	0,52	632,44	472,44	29,83	1,25

Tabla 11. Resultados obtenidos de las estadísticas Q_X y Q_Y, mediante la aplicación de DCCA.

En la tabla 11, aparecen los datos obtenidos de las estadísticas Q_X y Q_Y. Q_X no detecta 5 anomalías, mientras que a Q_Y la sucede lo mismo que en CCA y CCA Regularizado, y es que no detecta la mayoría de los fallos. Como consecuencia, tanto los tiempos de detección del fallo como el porcentaje de falsas alarmas tienen mejor resultado en Q_X que en Q_Y.

4.7. ANÁLISIS CANÓNICO DE CORRELACIÓN CONCURRENTE (CCCA).

El ultimo método de análisis estadístico multivariante que vamos a aplicar en este trabajo es el Análisis Canónico de Correlación Concurrente o CCCA. Es un método que parte de la base de CCA con regularización y resuelve las dos deficiencias se sufre CCA, el problema de colinealidad por parte de la aplicación de CCA Regularizado y a mayores, el problema de omitir la información aportada por la variabilidad de los datos.

4.7.1. ANÁLISIS DE DATOS DE COMPORTAMIENTO NORMAL CON CCCA

Como siempre, importamos el fichero de datos normalizados de dimensiones R^{960x52}, y se definen dos matrices, de la misma manera que en los anteriores métodos de análisis canónico, la primera almacenará las 22 primeras columnas y las 11 últimas, y la segunda, las columnas 35 y 36, una vez normalizadas estas matrices a media cero y varianza la unidad, pasarán a denominarse $X \in R^{960x33}$ (ecuación 14) e $Y \in R^{960x2}$ (ecuación 15) respectivamente.

Calcularemos las matrices $A \in R^{33x33}$ (ecuación 16) y $B \in R^{2x2}$ (ecuación 17) y obtendremos los valores propios y los vectores de carga de cada una de ellas para finalmente calcular $Z \in R^{33x2}$ (ecuación 20) y sus valores singulares, de la misma manera que se ha comentado en el apartado de CCA con regularización.

A continuación, calculamos las matrices R ∈ R^{33x2} (ecuación 21), T ∈ R^{960x2} (ecuación 23), P ∈ R^{33x2} (ecuación 25) y Q ∈ R^{2x2} (ecuación 26), y las utilizaremos para obtener las matrices residuo X_C ∈ R^{960x33} (ecuación 38) e Y_C ∈ R^{960x2} (ecuación 42). Sobre estas matrices residuo, se aplica el método PCA de manera independiente, y del que obtendremos las matrices de componentes principales T_X ∈ R^{960x20} (ecuación 40) y T_Y ∈ R^{960x2} (ecuación 44). Mas tarde se calculan las matrices residuales tanto de X_C como de Y_C denominadas, X_{CC} ∈ R^{960x2} (ecuación 41) e Y_{CC} ∈ R^{960x2} (ecuación 45).

Una vez que hemos obtenido todo lo anterior, pasamos a calcular las estadísticas, en este caso tendremos 3 estadísticas pertenecientes a Hotelling y 2 pertenecientes a Q o SPE. La primera estadística será T^2 (ecuación 33), empleada para monitorizar el subespacio que hemos obtenido mediante CCA Regularizado, con su umbral T_a^2 (ecuación 10) y utilizaremos las estadísticas T_X^2 y T_Y^2 (ecuaciones 46 y 47) para monitorizar los subespacios de residuo X_C y Y_C, con sus correspondientes umbrales ${T_{X\alpha}}^2$ y ${T_{Y\alpha}}^2$ (ecuaciones 48 y 49). Aplicaremos la estadística Q o SPE para monitorizar los subespacios residuales X_{CC} e Y_{CC}, por lo que emplearemos Q_X (ecuación 50) para X_{CC}, con su umbral $Q_{X\alpha}$ en este caso calculado como se indica en la teoría (ecuación 36), y Q_Y (ecuación 51) para Y_{CC}, con su umbral $Q_{Y\alpha}$, calculado mediante percentil 99.

En este caso, el umbral T_a^2 corresponde a un valor de 9,27, ${T_{X\alpha}}^2$ y ${T_{Y\alpha}}^2$ a 38,77 y 9,27 respectivamente y, por último, el umbral $Q_{X\alpha}$ será 3,86 y $Q_{Y\alpha}$ tendrá un valor de prácticamente cero. A continuación, se mostrarán gráficamente las estadísticas aplicadas a los datos de comportamiento normal del proceso con sus umbrales, correspondiendo las figuras 47, 48 y 49 a la estadística de Hotelling y las figuras 50 y 51 a la estadística Q o SPE.

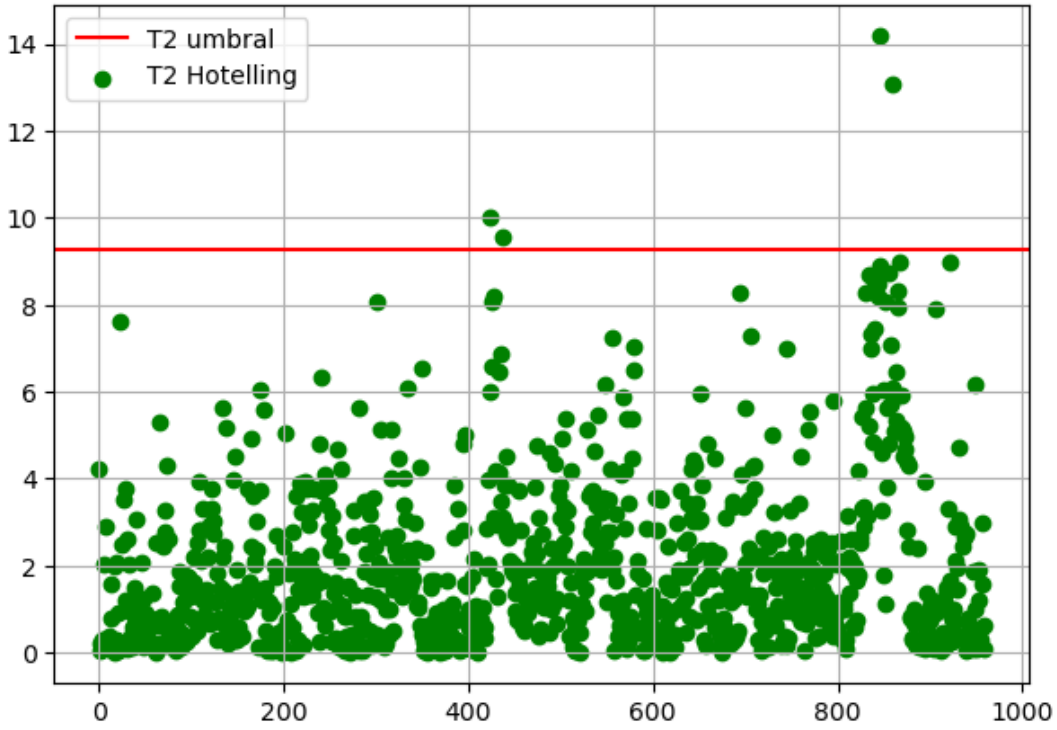

Figura 47. Estadística T^2 con su umbral para datos de comportamiento normal en CCCA.

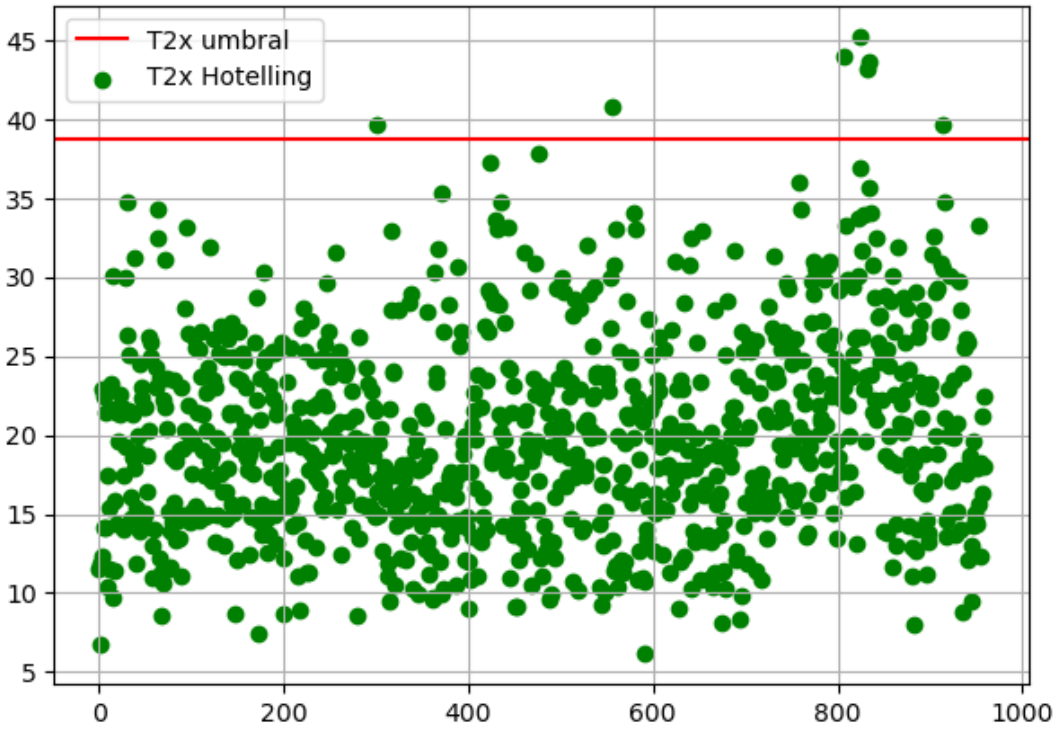

Figura 48. Estadística T_X^2 con su umbral para datos de comportamiento normal en CCCA.

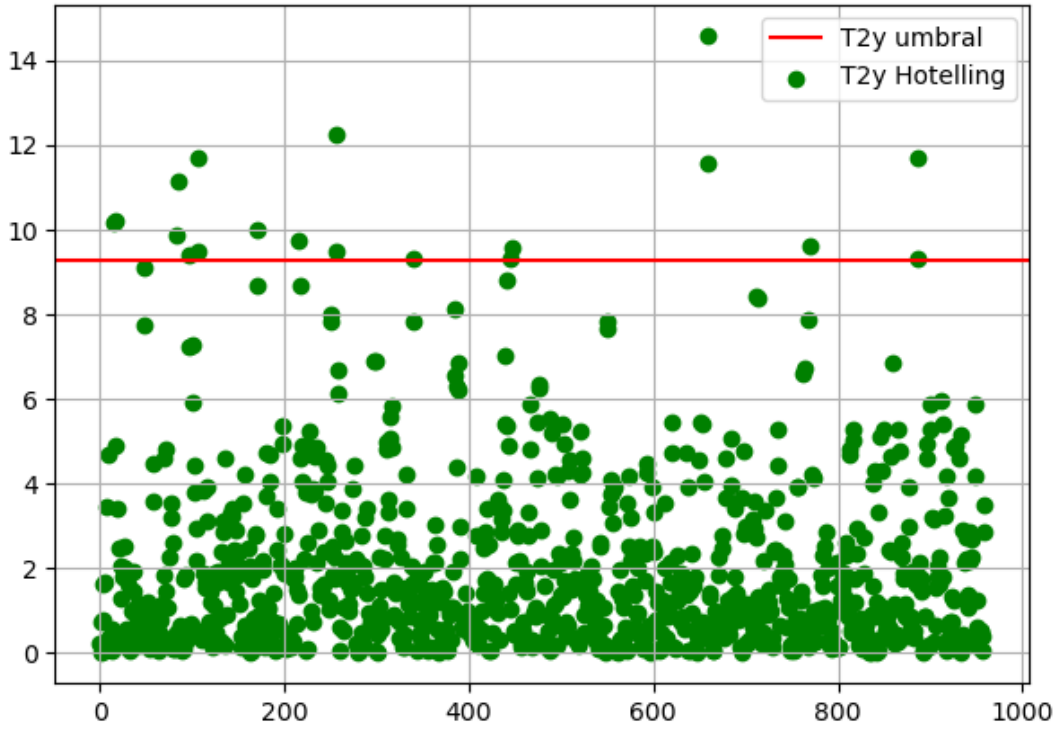

Figura 49. Estadística T_Y^2 con su umbral para datos de comportamiento normal en CCCA.

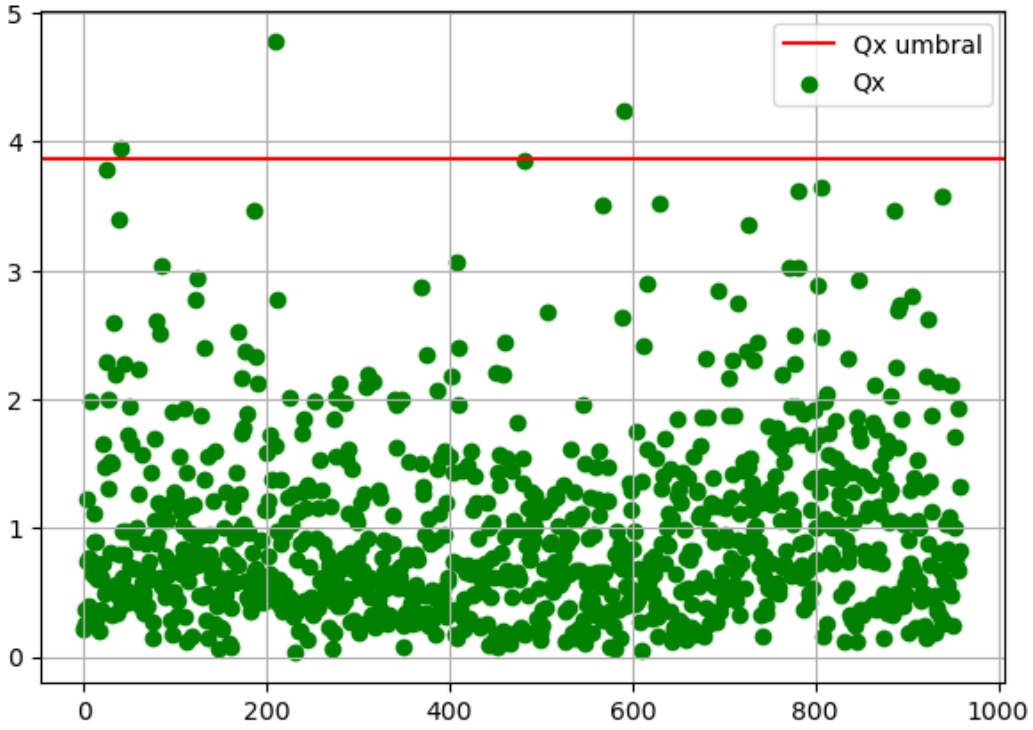

Figura 50. Estadística Q_X con su umbral para datos de comportamiento normal en CCCA.

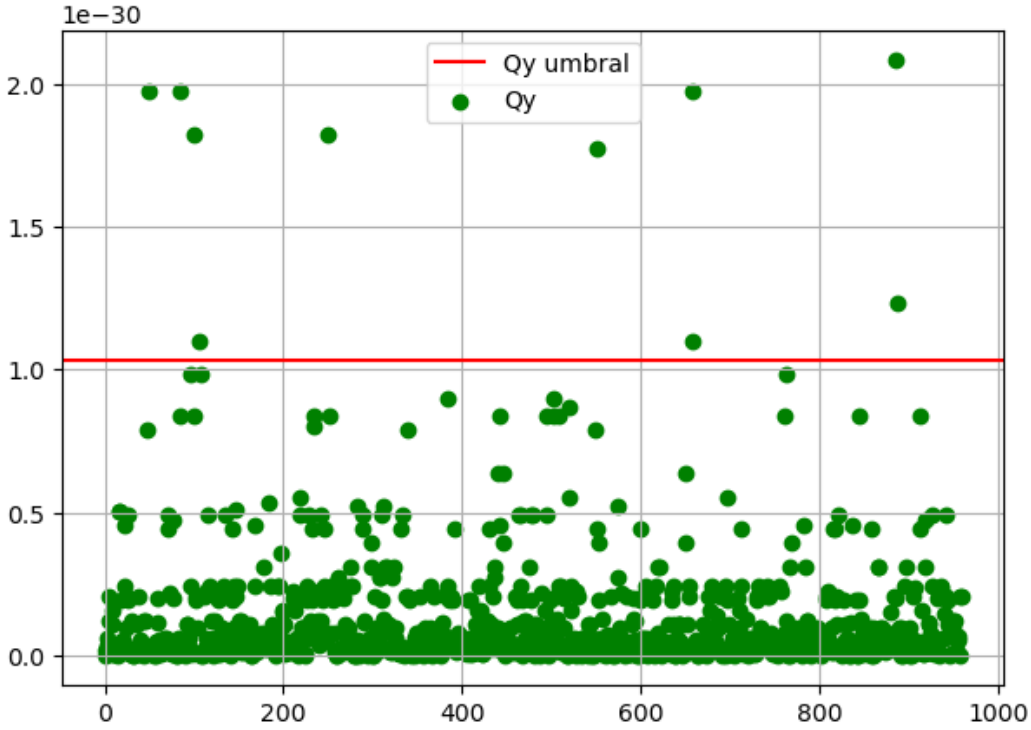

Figura 51. Estadística Q_Y con su umbral para datos de comportamiento normal en CCCA.

4.7.2. ANÁLISIS DE DATOS CON FALLO APLICANDO CCCA

Para la detección de los fallos, como hasta ahora, lo primero de todo es importar los 21 archivos de fallo. A cada fichero se le divide en dos matrices X e Y, como se ha explicado con anterioridad y se normalizarán de manera individual utilizando la media y la varianza de cada una, pero calculadas en condiciones normales del proceso.

En este caso, no aplicaremos CCA Regularizado exactamente, sino que, directamente calculamos la matriz T (ecuación 23) de componentes principales donde nos haría falta la matriz $R \in R^{33x2}$, siendo esta importada del apartado anterior. También calcularemos las matrices residuo $X_C \in R^{960x33}$ (ecuación 38) e $Y_C \in R^{960x2}$ (ecuación 42), así como, $T_X \in R^{960x20}$ (ecuación 40) y $T_Y \in R^{960x2}$ (ecuación 44) utilizando $R^{\dagger} \in R^{2x33}$, $Q \in R^{2x2}$ $P_X \in R^{33x20}$ y $P_Y \in R^{33x2}$ obtenidas con los datos del proceso en condiciones normales. Finalmente recalcularemos las matrices residuales $X_{CC} \in R^{960x2}$ e $Y_{CC} \in R^{960x2}$.

Una vez que hemos obtenido los 5 subespacios para datos de fallo, calculamos las 5 estadísticas, ya mencionadas en el apartado anterior, y como siempre, los umbrales se importarán del mismo.

A continuación, se mostrarán las estadísticas de manera grafica para los tres fallos de ejemplo elegidos, IDV (2) (figuras 52, 53 y 54), IDV (9) (figuras 55, 56 y 57) e IDV (17) (figuras 58, 59 y 60), y finalmente las tablas resumen de las 5 estadísticas donde se mostrará toda la información obtenida de estas.

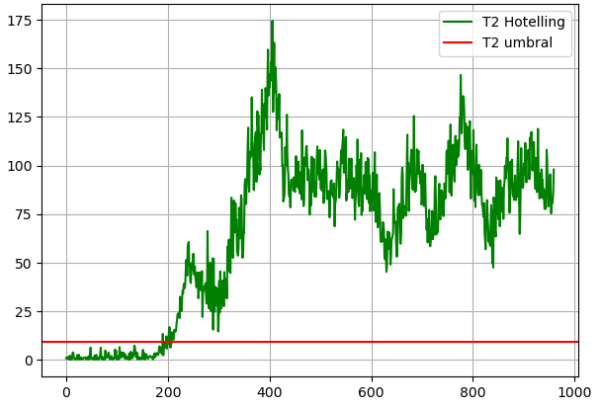

Figura 52. Detección del fallo IDV (2) con la estadística T^2 en CCCA.

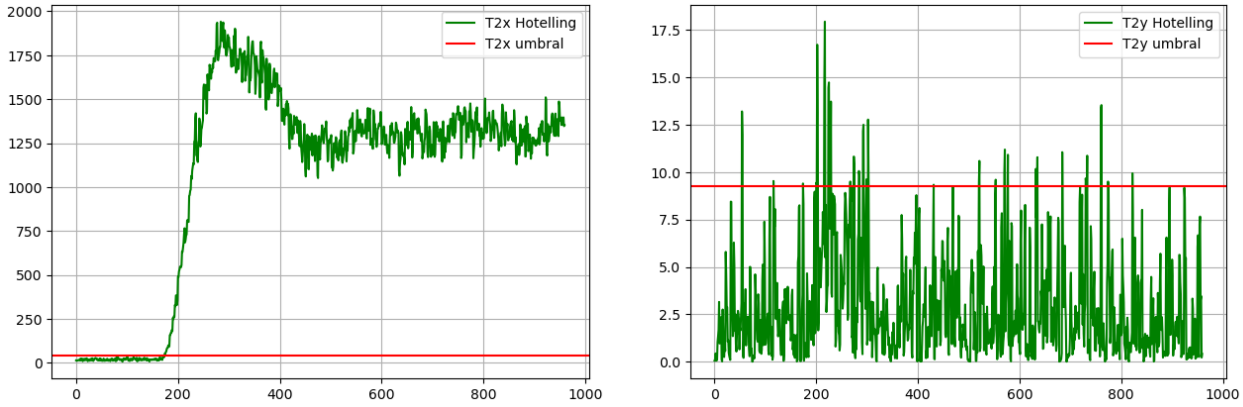

Figura 53. Detección del fallo IDV (2) con la estadística T_X^2 (izquierda) y T_Y^2 (derecha) en CCCA.

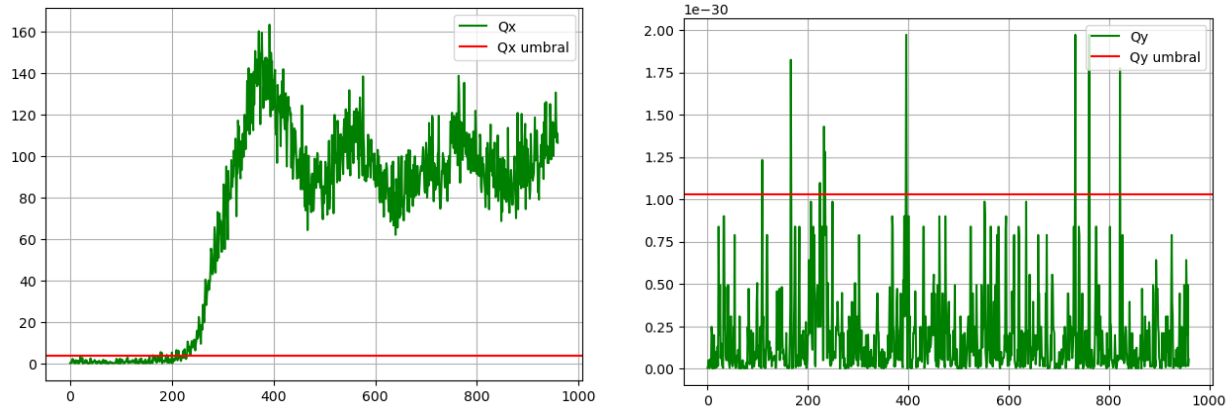

Figura 54. Detección del fallo IDV (2) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCCA

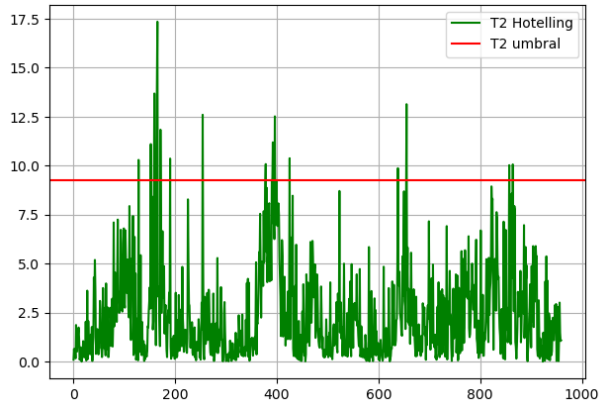

Figura 55. Detección del fallo IDV (9) con la estadística T^2 en CCCA.

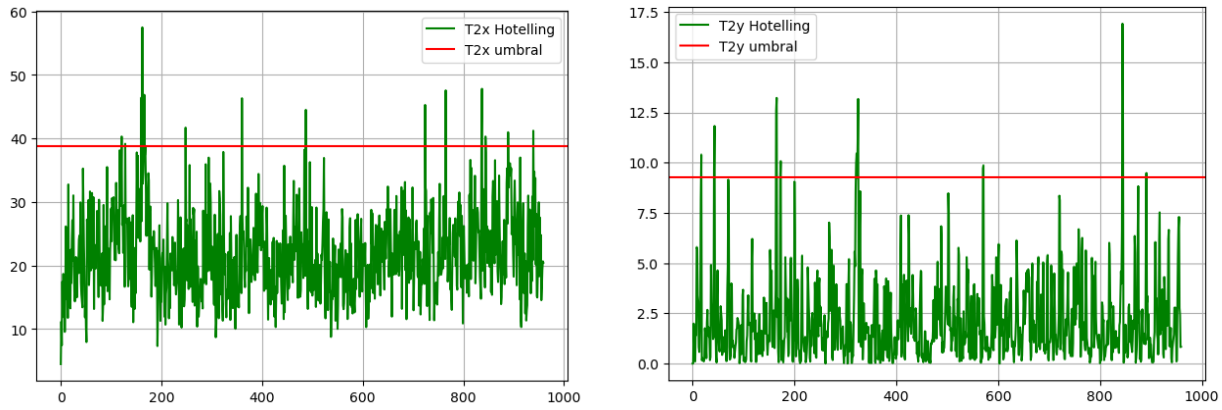

Figura 56. Detección del fallo IDV (9) con la estadística T_X^2 (izquierda) y T_Y^2 (derecha) en CCCA.

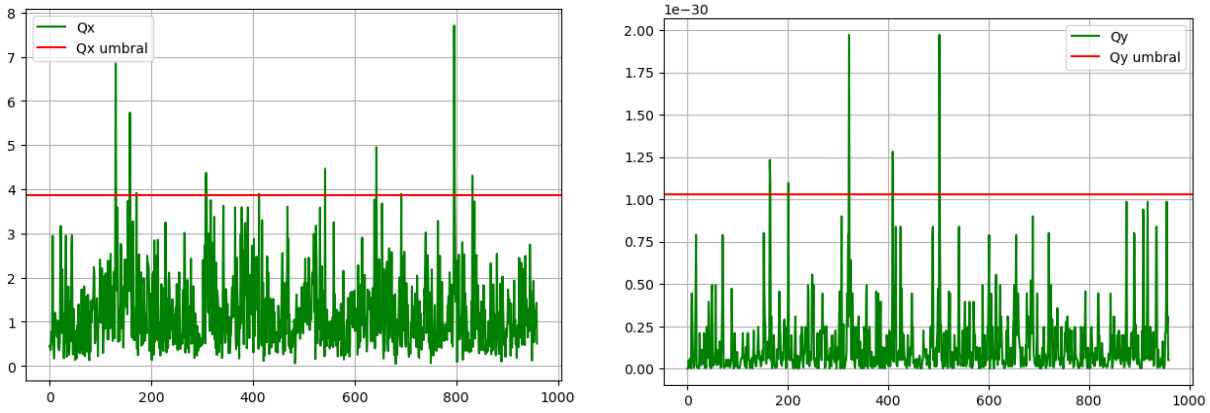

Figura 57. Detección del fallo IDV (9) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCCA.

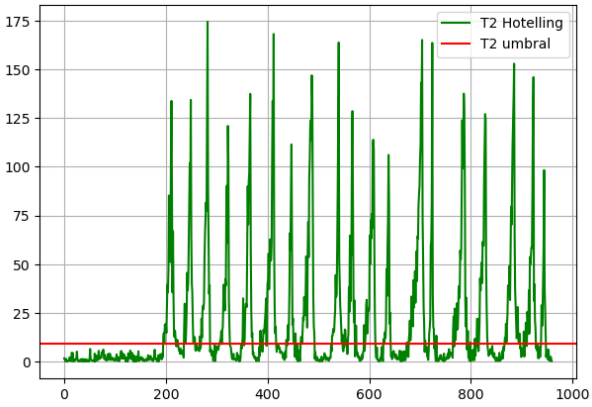

Figura 58. Detección del fallo IDV (17) con la estadística T^2 en CCCA.

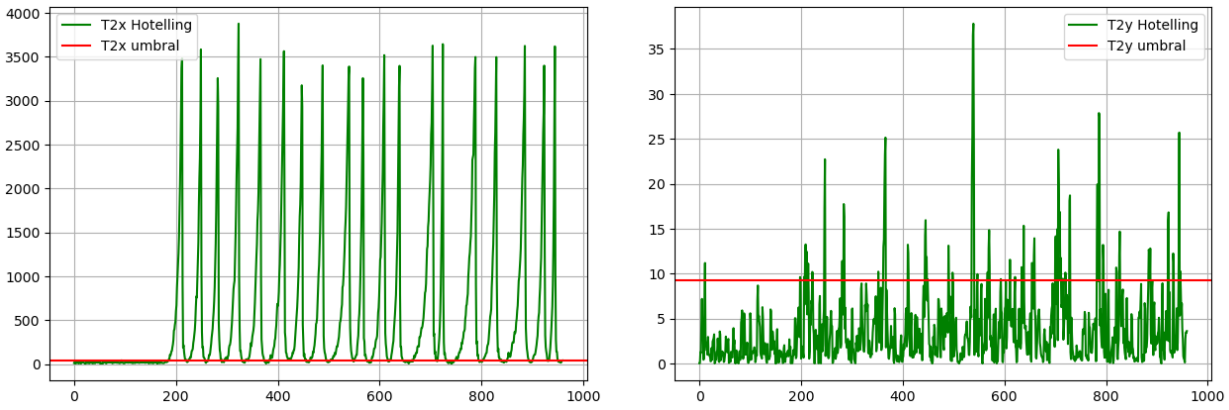

Figura 59. Detección del fallo IDV (17) con la estadística T_X^2 (izquierda) y T_Y^2 (derecha) en CCCA.

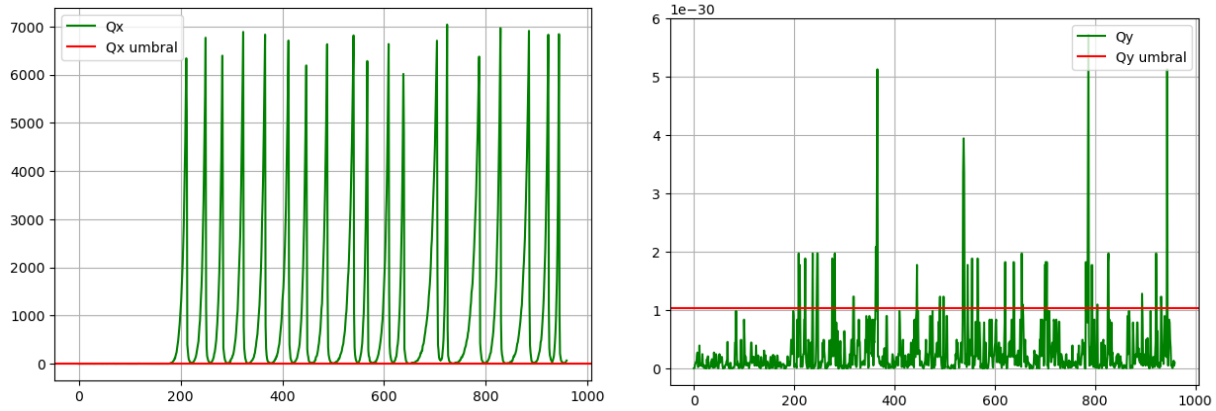

Figura 60. Detección del fallo IDV (17) con la estadística Q_X (izquierda) y Q_Y (derecha) en CCCA.

Hotelling T^2

	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	171	11	98,13	0,00
IDV (2)	208	48	94,88	0,00
IDV (3)	No detecta	No detecta	2,63	3,75
IDV (4)	201	41	28,88	0,00
IDV (5)	160	0	100,00	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	161	1	94,00	0,00
IDV (8)	189	29	80,38	0,63
IDV (9)	No detecta	No detecta	1,63	1,88
IDV (10)	258	98	23,75	0,00
IDV (11)	306	146	19,50	0,63
IDV (12)	161	1	94,50	1,25
IDV (13)	204	44	91,75	0,63
IDV (14)	No detecta	No detecta	11,75	0,00
IDV (15)	No detecta	No detecta	2,13	0,00
IDV (16)	171	11	59,13	4,38
IDV (17)	195	35	50,50	0,00
IDV (18)	244	84	89,75	0,00
IDV (19)	278	118	47,88	0,00
IDV (20)	239	79	66,63	0,00
IDV (21)	577	417	50,50	2,50
MEDIA	367,76	207,76	57,54	0,74

MEDIA SIN FALLOS 3, 9 Y 15	269,06	109,06	66,77	0,56

Tabla 12. Resultados obtenidos de la estadística T^2, mediante la aplicación de CCCA.

	Hotelling T_X^2				**Hotelling T_Y^2**			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	163	3	99,63	0,00	264	104	50,00	0,00
IDV (2)	174	14	98,38	0,00	No detecta	No detecta	4,75	1,88
IDV (3)	No detecta	No detecta	1,63	0,00	No detecta	No detecta	1,50	2,50
IDV (4)	228	68	52,25	0,63	No detecta	No detecta	1,50	1,25
IDV (5)	166	6	24,00	0,63	162	2	99,75	1,25
IDV (6)	168	8	99,13	0,63	160	0	100,00	1,25
IDV (7)	160	0	100,00	0,00	161	1	14,50	2,50
IDV (8)	180	20	97,63	0,63	324	164	16,75	1,25
IDV (9)	No detecta	No detecta	1,75	1,25	No detecta	No detecta	1,50	1,88
IDV (10)	258	98	30,13	0,63	No detecta	No detecta	4,63	0,00
IDV (11)	210	50	53,25	0,63	No detecta	No detecta	2,75	0,00
IDV (12)	166	6	98,88	1,25	174	14	73,50	3,13
IDV (13)	204	44	94,63	0,00	430	270	39,13	1,25
IDV (14)	160	0	100,00	0,00	No detecta	No detecta	2,75	1,25
IDV (15)	No detecta	No detecta	2,13	0,00	No detecta	No detecta	1,00	3,75
IDV (16)	470	310	15,25	0,63	No detecta	No detecta	3,13	0,63
IDV (17)	186	26	82,00	1,25	706	546	10,63	0,63
IDV (18)	246	86	89,50	1,25	248	88	89,25	1,25
IDV (19)	No detecta	No detecta	13,38	0,00	No detecta	No detecta	4,75	1,88
IDV (20)	246	86	43,25	0,00	294	134	23,25	1,88
IDV (21)	631	471	45,25	2,50	No detecta	No detecta	3,13	1,25
MEDIA	374,10	214,10	59,14	0,57	642,05	482,05	26,10	1,46
MEDIA SIN FALLOS 3, 9 Y 15	276,44	116,44	68,69	0,59	589,06	429,06	30,23	1,25

Tabla 13. Resultados obtenidos de las estadísticas T_X^2 y T_Y^2, mediante la aplicación de CCCA.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	160	0	100,00	2,50	No detecta	No detecta	3,25	0,63
IDV (2)	230	70	92,63	0,00	No detecta	No detecta	1,13	0,63
IDV (3)	No detecta	No detecta	2,25	2,50	No detecta	No detecta	0,88	0,00
IDV (4)	161	1	99,88	0,63	No detecta	No detecta	1,00	0,63
IDV (5)	161	1	18,25	0,63	169	9	99,25	0,63
IDV (6)	160	0	100,00	0,00	208	48	95,00	0,00
IDV (7)	160	0	36,75	0,00	No detecta	No detecta	6,13	1,88
IDV (8)	179	19	80,50	1,25	No detecta	No detecta	4,38	0,63
IDV (9)	No detecta	No detecta	1,13	1,25	No detecta	No detecta	0,75	0,00
IDV (10)	192	32	48,75	0,00	No detecta	No detecta	0,88	0,00
IDV (11)	172	12	67,38	0,63	No detecta	No detecta	0,75	0,63
IDV (12)	182	22	85,13	0,00	No detecta	No detecta	52,25	1,25
IDV (13)	195	35	95,75	0,00	671	511	13,50	0,00
IDV (14)	162	2	91,63	0,00	No detecta	No detecta	0,88	1,88
IDV (15)	No detecta	No detecta	3,13	0,00	No detecta	No detecta	0,38	1,25
IDV (16)	174	14	48,88	3,13	No detecta	No detecta	2,00	0,63
IDV (17)	179	19	97,50	0,63	No detecta	No detecta	5,50	0,00
IDV (18)	237	77	90,38	0,00	260	100	86,50	0,63
IDV (19)	348	188	34,25	1,25	No detecta	No detecta	0,38	0,00
IDV (20)	240	80	54,88	0,63	328	168	13,00	1,88
IDV (21)	416	256	50,00	6,25	No detecta	No detecta	0,88	0,63
MEDIA	313,71	153,71	61,86	1,01	809,33	649,33	18,51	0,65
MEDIA SIN FALLOS 3, 9 Y 15	206,00	46,00	71,81	0,97	784,22	624,22	21,48	0,69

Tabla 14. Resultados obtenidos de las estadísticas Q_X y Q_Y mediante la aplicación de CCCA.

Como podemos ver en las tablas 12, 13 y 14, la estadística que más fallos detecta es Q_X, además lo hace de manera temprana. Únicamente no localiza las anomalías incipientes, y como consecuencia, es la estadística que tiene de media menos tiempo de detección. T_X^2 es la que menor porcentaje de falsas alarmas tiene de todas las estadísticas calculadas con CCCA, pero no detecta 4 de los fallos, al igual que la estadística T2. La estadística Q_Y solamente detecta 5 fallos, y aunque tenga muy bajo porcentaje de falsas alarmas, no es una buena estadística para monitorizar el proceso. Lo mismo ocurre con T_Y^2, que no detecta la mitad de las anomalías.

4.8. IMPLEMENTACIÓN DE LA FUNCIÓN LLE EN LOS MÉTODOS DE ANÁLISIS ESTADÍSTICO MULTIVARIANTE

Como ya se explicó en el capítulo II, LLE o "Locally Linear Embedding" es una herramienta que reduce las dimensiones originales de los datos manteniendo la mayor cantidad de información posible.

Para los datos de comportamiento normal, primeramente, se importan y se normalizan a media cero y varianza unidad, como se ha venido haciendo hasta ahora. Una vez que tenemos la o las matrices de datos normalizados, dependiendo del método de análisis multivariante que posteriormente se aplique, llamamos directamente a la función LLE en Python.

Esta función necesita que se le indique los k vecinos más próximos, k es un numero arbitrario que como hemos mencionado antes, indica puntos de alta dimensión, más cercanos al punto de interés de baja dimensión. Este parámetro, se especifica dentro del modelo, al igual que el número de componentes, es decir, las dimensiones o número de variables a los que queremos reducir los datos originales. Para los métodos que solo trabajen con la matriz X, PCA y DPCA, el número de vecinos elegido es 13 y las componentes finales 36.

Para los métodos CCA, DCCA y CCCA, donde los datos iniciales se dividen en variables de proceso (matriz X) y variables de calidad (matriz Y), el número de vecinos será 13 para X y 5 para Y, y el número de variables finales será 30 para X y 2 para Y.

Una vez que tenemos la o las matrices obtenidas mediante LLE, procedemos a aplicar el método de análisis estadístico multivariante correspondiente sobre los datos de comportamiento normal del proceso, tal y como se ha explicado en los apartados anteriores. Pero antes necesitamos calcular la matriz de pesos tanto para X como para Y , la cual utilizaremos más adelante para reducir las dimensiones en los datos de fallo del proceso.

$$W_X = X^{\dagger} \cdot X_{LLE} \tag{54}$$

$$W_Y = Y^{\dagger} \cdot Y_{LLE} \tag{55}$$

Donde $X^{\dagger} = (X^t \cdot X)^{-1} \cdot X^t$,y de la misma manera, $Y^{\dagger} = (Y^t \cdot Y)^{-1} \cdot Y^t$. X_{LLE} e Y_{LLE} hacen referencia a la matriz resultante de aplicar la función LLE sobre los datos originales.

Una vez obtenidos los umbrales y estadísticas correspondientes del método aplicado, procedemos a analizar los datos de fallo del proceso. Como

siempre, primero importamos los 21 ficheros, estandarizaremos X, o X e Y con la media y la varianza obtenida de los datos de comportamiento normal.

Para reducir las dimensiones originales mediante LLE y obtener la o las matrices con las que monitorizar el proceso a través de las estadísticas, emplearemos las matrices de peso calculadas con los datos normales del proceso y las nuevas matrices X e Y con los datos de fallo.

$$X_{LLE} = X \cdot W_X \tag{56}$$

$$Y_{LLE} = Y \cdot W_Y \tag{57}$$

Finalmente aplicaremos el método de análisis multivariante correspondiente, y los umbrales serán los mismos que los calculados con los datos sin anomalías del proceso, como se ha hecho hasta ahora.

4.8.1. PCA CON LLE

Aplicamos el método PCA con LLE tanto a los datos de comportamiento normal como de fallo, utilizando la metodología explicada anteriormente. Primeramente, obtendremos la estadística T^2 (ecuación 9), donde el valor de su umbral, T_a^2 (ecuación 10) corresponde a 61,71, y finalmente, calcularemos la estadística Q o SPE (ecuación 11) donde su umbral Q_α, en este caso calculado mediante un percentil del 90%, tiene un valor de 0,20. Se ha optado calcular el umbral Q_α con un percentil al 90% porque aplicando la ecuación 13 el umbral sería sumamente pequeño para los valores de Q, y la opción que se ha optado hasta ahora del percentil 99, daba como resultado un umbral demasiado alto. También se mantiene la norma de superar 8 veces consecutivas el umbral para que el fallo sea detectado.

A continuación, se mostrará la tabla 15 que recoge toda la información obtenida de ambas estadísticas y donde podremos ver que el porcentaje de falsas alarmas tanto para T^2 como para Q es de cero, pero el número de fallos no detectados para T^2 es 7, dos más que en PCA, y para Q son 8 anomalías, el doble que en PCA.

	Hotelling T^2				Q o SPE			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	166	6	99,25	0,00	168	8	78,13	0,00
IDV (2)	184	24	97,00	0,00	183	23	97,13	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,13	0,00

IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (5)	162	2	99,75	0,00	161	1	99,88	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	161	1	28,50	0,00	162	2	25,88	0,00
IDV (8)	182	22	89,25	0,00	182	22	77,88	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	209	49	31,88	0,00	264	104	21,88	0,00
IDV (11)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,13	0,00
IDV (12)	173	13	94,75	0,00	181	21	83,75	0,00
IDV (13)	210	50	91,63	0,00	213	53	86,25	0,00
IDV (14)	No detecta	No detecta	14,50	0,00	No detecta	No detecta	12,88	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,25	0,00
IDV (16)	357	197	23,63	0,00	No detecta	No detecta	4,50	0,00
IDV (17)	190	30	55,75	0,00	190	30	54,88	0,00
IDV (18)	252	92	88,50	0,00	253	93	88,63	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (20)	244	84	32,63	0,00	244	84	34,13	0,00
IDV (21)	789	629	21,88	0,00	839	679	18,88	0,00
MEDIA	483,76	323,76	46,14	0,00	518,10	358,10	42,15	0,00
MEDIA SIN FALLOS 3, 9 Y 15	404,39	244,39	53,83	0,00	444,44	284,44	49,15	0,00

Tabla 15. Resultados obtenidos de las estadísticas T^2 y Q mediante la aplicación de PCA con LLE.

4.8.2. DPCA CON LLE

El método DPCA, que se ha explicado con anterioridad en este capítulo, implica trabajar con datos dinámicos, por lo que se implementará la herramienta LLE a la matriz X resultado de la concatenación de los datos dinámicos hasta el instante t-3.

Primeramente, se aplicará LLE a los datos, en este caso dinámicos, de comportamiento normal, calculando a mayores la matriz de pesos W (ecuación 54), que utilizaremos más tarde en los ficheros de datos con anomalías. Finalmente obtendremos las estadísticas con sus respectivos umbrales T_a^2 y Q_α, este último calculado mediante percentil 90.

A continuación, utilizaremos la matriz de pesos W para aplicar LLE sobre los datos dinámicos, calculados a partir de los datos con fallo, aplicando la ecuación 57 donde X seria la matriz concatenada y X_{LLE} la matriz X resultante al implementar la herramienta propiamente dicha. Finalmente calcularemos las estadísticas T^2 y Q para compararlas con los umbrales obtenidos con los

datos de comportamiento normal en el proceso y comprobar si el fallo es detectado. En este caso el umbral T_a^2 tiene un valor de 61,72 y Q_α de 0,08.

	Hotelling T²				Q o SPE			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	167	7	99,13	0,00	168	8	99,00	0,00
IDV (2)	181	21	97,50	0,00	190	30	96,25	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (5)	165	5	99,38	0,00	167	7	99,13	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	161	1	29,13	0,00	161	1	20,25	0,00
IDV (8)	186	26	92,75	0,00	189	29	78,63	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	260	100	21,75	0,00	No detecta	No detecta	3,75	0,00
IDV (11)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (12)	180	20	93,38	0,00	181	21	84,25	0,00
IDV (13)	216	56	93,00	0,00	216	56	91,25	0,00
IDV (14)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (16)	357	197	23,13	0,00	357	197	20,88	0,00
IDV (17)	194	34	46,75	0,00	197	37	37,38	0,00
IDV (18)	254	94	88,25	0,00	253	93	88,38	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (20)	249	89	30,13	0,00	254	94	23,88	0,00
IDV (21)	834	674	17,63	0,00	No detecta	No detecta	2,50	0,00
MEDIA	489,71	329,71	44,38	0,00	530,14	370,14	40,26	0,00
MEDIA SIN FALLOS 3, 9 Y 15	411,33	251,33	51,77	0,00	458,50	298,50	46,97	0,00

Tabla 16. Resultados obtenidos de las estadísticas T^2 y Q mediante la aplicación de DPCA con LLE.

En la tabla 16 podemos ver la información aportada por las estadísticas calculadas a partir del método DPCA con LLE. Los porcentajes de falsas alarmas en ambas estadísticas es cero, los tiempos de detección de T^2 son más tempranos que en la estadística Q. Además, esta última, tiene un total de 9 fallos no detectados, mientras que T^2 no localiza las mismas anomalías que en PCA con LLE.

4.8.3. CCA CON LLE

Como hemos explicado con anterioridad, CCA divide las variables en dos matrices, X donde almacena las variables de proceso (22 primeras columnas y las 11 ultimas) e Y las variables de calidad (columnas 35 y 36). Aplicaremos de manera individualizada la herramienta LLE, tanto a la matriz X como a Y, en los datos de comportamiento normal. De la misma manera se calcularán las matrices de pesos W_X (ecuación 54) y W_Y (ecuación 55).

Sobre las matrices producto de la aplicación de LLE, calculamos las estadísticas correspondientes a CCA, mediante dicho método. Finalmente obtenemos los umbrales que emplearemos más tarde en los datos de comportamiento con fallo para saber si este es detectado.

En este caso el umbral T_a^2 (ecuación 10) tiene un valor de 9,27, el mismo que en CCA sin aplicar LLE. Los umbrales pertenecientes a la estadística Q o SPE, $Q_{X\alpha}$ y $Q_{Y\alpha}$, se han obtenido mediante un percentil del 90% dando como resultado un valor de 0,13 para $Q_{X\alpha}$ y 0,002 para $Q_{Y\alpha}$.

En los ficheros de fallo, como hasta ahora, utilizaremos la matriz de peso para implementar la herramienta LLE. En este caso al haber dos matrices de datos inicial X e Y, aplicaremos LLE de manera individualizada a cada matriz utilizando la matriz de peso que corresponda W_X para X y W_Y para Y. Finalmente tendremos como resultado dos matrices X_{LLE} e Y_{LLE} sobre las que aplicaremos el método CCA, para obtener las estadísticas T^2 (ecuación 33), Q_X (ecuación 34) y Q_Y (ecuación 35) que serán comparadas con sus correspondientes umbrales, calculados con los datos de comportamiento normal del proceso, y si estos son superados 8 veces consecutivas habría existencia de fallo. Los datos recogidos de las estadísticas se muestran en las tablas 17 y 18.

Hotelling T^2

	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	173	13	98,38	0,00
IDV (2)	216	56	93,25	0,00
IDV (3)	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00
IDV (5)	168	8	99,00	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	164	4	8,50	0,00
IDV (8)	344	184	16,63	0,00

IDV (9)	No detecta	No detecta	0,00	0,00
IDV (10)	263	103	15,25	0,00
IDV (11)	No detecta	No detecta	0,00	0,00
IDV (12)	183	23	47,00	0,00
IDV (13)	231	71	61,50	0,00
IDV (14)	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00
IDV (16)	356	196	18,63	0,00
IDV (17)	No detecta	No detecta	0,00	0,00
IDV (18)	264	104	81,38	0,00
IDV (19)	No detecta	No detecta	0,00	0,00
IDV (20)	254	94	21,00	0,00
IDV (21)	No detecta	No detecta	1,00	0,00
MEDIA	543,62	383,62	31,50	0,00
MEDIA SIN FALLOS 3, 9 Y 15	474,22	314,22	36,75	0,00

Tabla 17. Resultados obtenidos de la estadística T^2 mediante la aplicación de CCA con LLE.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	167	7	99,13	0,00	173	13	98,38	0,00
IDV (2)	193	33	95,88	0,00	216	56	93,88	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (5)	168	8	99,13	0,00	168	8	99,00	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	161	1	34,00	0,00	216	56	10,75	0,00
IDV (8)	187	27	65,63	0,00	287	127	16,13	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	259	99	20,88	0,00	259	99	20,75	0,00
IDV (11)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (12)	182	22	76,63	0,00	183	23	50,25	0,00
IDV (13)	218	58	90,00	0,00	231	71	69,13	0,00
IDV (14)	No detecta	No detecta	9,38	0,00	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (16)	280	120	29,63	0,00	283	123	24,38	0,00

IDV (17)	194	34	47,50	0,00	No detecta	No detecta	0,13	0,00
IDV (18)	259	99	87,63	0,00	269	109	80,75	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (20)	247	87	32,75	0,00	254	94	19,63	0,00
IDV (21)	916	756	12,63	0,00	No detecta	No detecta	1,38	0,00
MEDIA	491,00	331,00	42,89	0,00	539,95	379,95	32,60	0,00
MEDIA SIN FALLOS 3, 9 Y 15	412,83	252,83	50,04	0,00	469,94	309,94	38,03	0,00

Tabla 18. Resultados obtenidos de las estadísticas Q_X y Q_Y mediante la aplicación de CCA con LLE.

En las tablas 17 y 18 observamos que T^2 no detecta 9 fallos, exactamente los mismos que tampoco localiza Q_Y, mientras que Q_X son 7 las anomalías que no detecta.

La estadística Q_Y da mejores resultados que en CCA convencional, ya que eran 12 los fallos que no detectaba, además en este caso los porcentajes de falsas alarmas son cero. Por el contrario, las estadísticas T^2 y Q_X han empeorado los resultados en comparación con CCA sin LLE, ya que detectan menos fallos a pesar de que el porcentaje de falsas alarmas es nulo.

4.8.4. DCCA CON LLE

El procedimiento para incluir LLE en DCCA, es el mismo que el explicado en el apartado anterior para CCA con LLE. La particularidad en este método, como ya sabemos, es que parte de CCA con regularización, por lo que habría que tener en cuenta, además de los datos dinámicos, la manera de calcular las matrices $\Sigma_{xx}^{-1/2}$ (ecuación 29) y $\Sigma_{yy}^{-1/2}$ (ecuación 30), utilizando los valores K_1 y K_2 mencionados anteriormente en el apartado 4.5.

En este método, los umbrales obtenidos a partir de los datos de comportamiento normal tienen un valor de 9,27 para T_a^2 (ecuación 9), el mismo valor que en CCA, CCA Regularizado y CCA con LLE. Mientras que $Q_{X\alpha}$ y $Q_{Y\alpha}$, calculados mediante un percentil del 90 %, tienen un valor 0,08 y 0,001 respectivamente.

Se impondrá LLE a los datos de fallo utilizando las matrices de pesos W_X y W_Y (ecuaciones 54 y 55), previamente calculadas con los datos de comportamiento normal del proceso. Finalmente, al aplicar DCCA a las matrices resultantes de LLE, en este caso X_{LLE} e Y_{LLE}, obtendremos las estadísticas T^2, Q_X y Q_Y, mostradas en las tablas 19 y 20.

	Hotelling T^2			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	247	87	11,25	0,00
IDV (2)	208	48	95,00	0,00
IDV (3)	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00
IDV (5)	172	12	98,63	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	195	35	8,88	0,00
IDV (8)	406	246	15,88	0,00
IDV (9)	No detecta	No detecta	0,00	0,00
IDV (10)	No detecta	No detecta	0,13	0,00
IDV (11)	No detecta	No detecta	0,00	0,00
IDV (12)	186	26	55,63	0,00
IDV (13)	253	93	51,88	0,00
IDV (14)	No detecta	No detecta	3,25	0,00
IDV (15)	No detecta	No detecta	0,00	0,00
IDV (16)	No detecta	No detecta	4,63	0,00
IDV (17)	192	32	50,13	0,00
IDV (18)	254	94	86,88	0,00
IDV (19)	No detecta	No detecta	0,00	0,00
IDV (20)	252	92	21,00	0,00
IDV (21)	No detecta	No detecta	0,00	0,00
MEDIA	577,38	417,38	28,72	0,00
MEDIA SIN FALLOS 3, 9 Y 15	513,61	353,61	33,51	0,00

Tabla 19. Resultados obtenidos de la estadística T^2 mediante la aplicación de DCCA con LLE.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	169	9	98,88	0,00	201	41	19,25	0,00
IDV (2)	188	28	96,50	0,00	204	44	94,63	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00

IDV (5)	168	8	99,00	0,00	165	5	99,38	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	160	0	23,38	0,00	177	17	17,88	0,00
IDV (8)	187	27	79,63	0,00	237	77	43,13	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	261	101	9,25	0,00	No detecta	No detecta	4,75	0,00
IDV (11)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,38	0,00
IDV (12)	184	24	87,63	0,00	184	24	70,50	0,00
IDV (13)	224	64	87,88	0,00	240	80	72,25	0,00
IDV (14)	485	325	44,75	0,00	No detecta	No detecta	15,38	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,25	0,00
IDV (16)	356	196	20,13	0,00	409	249	16,00	0,00
IDV (17)	190	30	56,00	0,00	194	34	52,75	0,00
IDV (18)	253	93	88,38	0,00	253	93	88,63	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,25	0,00
IDV (20)	246	86	29,50	0,00	256	96	25,50	0,00
IDV (21)	886	726	13,38	0,00	886	726	12,00	0,00
MEDIA	470,33	310,33	44,49	0,00	535,52	375,52	34,90	0,00
MEDIA SIN FALLOS 3, 9 Y 15	388,72	228,72	51,90	0,00	464,78	304,78	40,70	0,00

Tabla 20. Resultados obtenidos de las estadísticas Q_X y Q_Y mediante la aplicación de DCCA con LLE.

Como se puede ver en las tablas 19 y 20, y como venimos viendo hasta ahora con la incorporación de LLE, los porcentajes de falsas alarmas para las 3 estadísticas son nulos. La estadística que más bajo presenta el tiempo de detección es Q_X, con 6 fallos no localizados, y la estadística que peores resultados muestra sería T^2, con 10 anomalías no detectadas, más del doble que en el método DCCA convencional.

4.8.5. CCCA CON LLE

Se incorporará LLE al método CCCA de la misma manera que hemos venido haciendo hasta ahora. Para los datos de comportamiento normal, se aplica la función LLE a la matriz X de variables del proceso, ya normalizada a media cero y varianza unitaria, indicándole que el número de componentes deseado es 30 y el número de vecinos más cercanos es de 13. De igual modo se aplicará LLE a la matriz Y normalizada, con 5 vecinos próximos y 2 componentes finales. También se calcularán las correspondientes matrices de pesos W_X y W_Y (ecuaciones 54 y 55)

Una vez obtenidas las matrices resultado de la función LLE, aplicamos CCCA sobre ellas, tal y como se ha descrito en el apartado 4.7.1. Obteniendo finalmente, los umbrales de las 5 estadísticas que se emplean en este método, los umbrales T_{α}^{2} y ${T_{Y\alpha}}^{2}$ tienen un valor de 9,27, ${T_{X\alpha}}^{2}$ corresponde a 94,51 y los umbrales $Q_{X\alpha}$ y $Q_{Y\alpha}$, calculados mediante un percentil del 90%, tienen un valor de 0,042 y 0,002 respectivamente.

Para implementar LLE a los datos de fallo, se utilizarán las matrices de pesos W_X y W_Y, y se emplearán las ecuaciones 56 y 57 para obtener las matrices a las que aplicaremos CCCA para los ficheros de fallo, de la misma manera que lo explicado en el apartado 4.7.2. Finalmente calculamos las 5 estadísticas y las compararemos con sus respectivos umbrales, obtenidos con datos normales del proceso. La información resultante de cada una de las estadísticas aparece reflejada en las tablas 21, 22 y 23.

Hotelling T^2

	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	173	13	98,38	0,00
IDV (2)	216	56	93,25	0,00
IDV (3)	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00
IDV (5)	168	8	99,00	0,00
IDV (6)	160	0	100,00	0,00
IDV (7)	164	4	8,50	0,00
IDV (8)	344	184	16,63	0,00
IDV (9)	No detecta	No detecta	0,00	0,00
IDV (10)	263	103	15,25	0,00
IDV (11)	No detecta	No detecta	0,00	0,00
IDV (12)	183	23	46,88	0,00
IDV (13)	231	71	61,38	0,00
IDV (14)	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00
IDV (16)	356	196	18,50	0,00
IDV (17)	No detecta	No detecta	0,00	0,00
IDV (18)	264	104	81,38	0,00
IDV (19)	No detecta	No detecta	0,00	0,00
IDV (20)	254	94	21,00	0,00
IDV (21)	No detecta	No detecta	1,00	0,00

MEDIA	543,62	383,62	31,48	0,00
MEDIA SIN FALLOS 3, 9 Y 15	474,22	314,22	36,73	0,00

Tabla 21. Resultados obtenidos de la estadística T^2 mediante la aplicación de CCCA con LLE.

	Hotelling T_X^2				**Hotelling T_Y^2**			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	172	12	47,88	0,00	182	22	22,50	0,00
IDV (2)	199	39	95,13	0,00	266	106	87,00	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (5)	176	16	98,00	0,00	316	156	15,25	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	164	4	17,63	0,00	No detecta	No detecta	0,00	0,00
IDV (8)	223	63	35,38	0,00	No detecta	No detecta	0,88	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	No detecta	No detecta	0,13	0,00	No detecta	No detecta	0,00	0,00
IDV (11)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (12)	346	186	47,63	0,00	No detecta	No detecta	1,13	0,00
IDV (13)	259	99	70,00	0,00	250	90	26,25	0,00
IDV (14)	173	13	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (16)	No detecta	No detecta	0,38	0,00	No detecta	No detecta	0,00	0,00
IDV (17)	196	36	39,50	0,00	No detecta	No detecta	0,00	0,00
IDV (18)	259	99	87,63	0,00	341	181	74,38	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (20)	257	97	16,50	0,00	No detecta	No detecta	1,63	0,00
IDV (21)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
MEDIA	534,48	374,48	31,23	0,00	757,86	597,86	15,67	0,00
MEDIA SIN FALLOS 3, 9 Y 15	463,56	303,56	36,43	0,00	724,17	564,17	18,28	0,00

Tabla 22. Resultados obtenidos de las estadísticas T_X^2 y T_Y^2 mediante la aplicación de CCCA con LLE.

	Q_X				Q_Y			
	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)	Instante de detección	Tiempo de detección	Alarmas (%)	Falsas alarmas (%)
IDV (1)	170	10	98,75	0,00	169	9	99,00	0,00
IDV (2)	180	20	97,75	0,00	206	46	94,75	0,00
IDV (3)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (4)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (5)	172	12	98,50	0,00	165	5	99,38	0,00
IDV (6)	160	0	100,00	0,00	160	0	100,00	0,00
IDV (7)	161	1	23,50	0,00	161	1	36,75	0,00
IDV (8)	215	55	80,75	0,00	187	27	37,63	0,00
IDV (9)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (10)	261	101	13,75	0,00	208	48	27,75	0,00
IDV (11)	No detecta	No detecta	1,00	0,00	No detecta	No detecta	0,00	0,00
IDV (12)	182	22	85,25	0,00	182	22	61,63	0,00
IDV (13)	237	77	87,38	0,00	218	58	73,50	0,00
IDV (14)	No detecta	No detecta	81,88	0,00	No detecta	No detecta	0,00	0,00
IDV (15)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,00	0,00
IDV (16)	400	240	14,38	0,00	194	34	36,25	0,00
IDV (17)	190	30	59,88	0,00	No detecta	No detecta	0,50	0,00
IDV (18)	258	98	87,88	0,00	262	102	83,75	0,00
IDV (19)	No detecta	No detecta	0,00	0,00	No detecta	No detecta	0,13	0,00
IDV (20)	249	89	27,38	0,00	249	89	30,63	0,00
IDV (21)	880	720	7,88	0,00	916	756	12,75	0,00
MEDIA	496,90	336,90	45,99	0,00	521,76	361,76	37,83	0,00
MEDIA SIN FALLOS 3, 9 Y 15	419,72	259,72	53,66	0,00	448,72	288,72	44,13	0,00

Tabla 23. Resultados obtenidos de las estadísticas Q_X y Q_Y mediante la aplicación de CCCA con LLE.

Como podemos ver en las tablas anteriores (tablas 21, 22 y 23), la estadística que peores resultados muestra es T_Y^2, ya que son 15 los fallo que no detecta. Por el contrario, Q_X solamente no localiza 7 de las 21 anomalías totales.

4.9. COMPARACIÓN DE MÉTODOS

La finalidad de este trabajo es ver cuál de los métodos que hemos estudiado hasta ahora se adapta mejor al proceso, y para ello se debe realizar una serie de comparativas entre ellos analizando los datos recopilados.

Uno de los factores más importante a la hora de elegir cual es el método más conveniente para el proceso, es el número de fallos detectados (tabla 24). El método ideal sería aquel que alguna de sus estadísticas detectase cada uno de los 21 fallos, pero eso no va a ser posible, al menos en este caso, ya que los fallos 3, 9 y 15 son incipientes, como ya hemos comentado con anterioridad, y no son detectados por ninguno de los métodos.

MÉTODO	ESTADÍSTICA	FALLOS NO DETECTADOS	FALLOS DETECTADOS	PROMEDIO DE FALLOS DETECTADOS POR MÉTODO
PCA	T^2	5	16	16
	Q	4	17	
PCA CON LLE	T^2	7	14	13
	Q	8	13	
DPCA	T^2	5	16	16
	Q	4	17	
DPCA CON LLE	T^2	7	14	13
	Q	9	12	
CCA	T^2	5	16	14
	Q_X	4	17	
	Q_Y	12	9	
CCA CON LLE	T^2	9	12	12
	Q_X	7	14	
	Q_Y	9	12	
CCA REGULARIZADO	T^2	6	15	13
	Q_X	4	17	
	Q_Y	12	9	
DCCA	T^2	4	17	14
	Q_X	5	16	
	Q_Y	12	9	
DCCA CON LLE	T^2	10	11	13
	Q_X	6	15	
	Q_Y	8	13	
CCCA	T^2	4	17	13
	T_X^2	4	17	
	T_Y^2	11	10	
	Q_X	3	18	
	Q_Y	16	5	
CCCA CON LLE	T^2	9	12	11
	T_X^2	9	12	

T_Y^2	15	6
Q_X	7	14
Q_Y	8	13

Tabla 24. Fallos detectados y no detectados por cada estadística y método.

Si comparamos la misma estadística en un método y el equivalente, pero implementando LLE, podemos ver, en la tabla anterior, que el número de fallos no detectados es mayor en la estadística obtenida con LLE, a excepción de Q_Y que aparentemente mejora sus resultados, aunque también hay que tener en cuenta que generalmente es la estadística que detecta menos fallos. La estadística que más fallos detecta es Q o SPE en PCA y DPCA, o en su defecto, Q_X en CCA, CCA Regularizado, DCCA y CCCA y los métodos que más fallos detectan son PCA y DPCA.

Otro de los factores clave y que tiene continuación con el número de fallos es el tiempo de detección de estos. Para que el método sea eficiente, la detección del fallo debe ser lo más temprana posible.

Para poder realizar una comparación este factor entre los diferentes métodos se calcula el promedio del tiempo de detección de todos los fallos por cada estadística y método, como se muestra en la tabla 25, teniendo o no en cuenta los fallos incipientes.

El tiempo de detección empieza a contar a partir del instante 160, ya que como se ha mencionado con anterioridad, es en este instante donde se produce el fallo.

MÉTODO	ESTADÍSTICA	TIEMPO DE DETECCIÓN DEL FALLO	TIEMPO DE DETECCIÓN DEL FALLO (SIN 3, 9 Y 15)	PROMEDIO DE TIEMPO POR MÉTODO (SIN 3, 9 Y 15)
PCA	T^2	270,48	182,22	134,97
	Q	189,48	87,72	
PCA CON LLE	T^2	323,76	244,39	264,42
	Q	358,10	284,44	
DPCA	T^2	273,76	186,06	143,75
	Q	201,24	101,44	
DPCA CON LLE	T^2	329,71	251,33	274,92
	Q	370,14	298,50	
CCA	T^2	260,38	170,44	239,59
	Q_X	204,14	104,83	
	Q_Y	494,43	443,50	

CCA CON LLE	T^2	383,62	314,22	292,33
	Q_X	331,00	252,83	
	Q_Y	379,95	309,94	
CCA REGULARIZADO	T^2	276,81	189,61	245,98
	Q_X	204,14	104,83	
	Q_Y	494,43	443,50	
DCCA	T^2	217,33	120,22	245,70
	Q_X	238,10	144,44	
	Q_Y	519,24	472,44	
DCCA CON LLE	T^2	417,38	353,61	295,70
	Q_X	310,33	228,72	
	Q_Y	375,52	304,78	
CCCA	T^2	207,76	109,06	264,96
	T_X^2	214,10	116,44	
	T_Y^2	482,05	429,06	
	Q_X	153,71	46,00	
	Q_Y	649,33	624,22	
CCCA CON LLE	T^2	383,62	314,22	346,08
	T_X^2	374,48	303,56	
	T_Y^2	597,86	564,17	
	Q_X	336,90	259,72	
	Q_Y	361,76	288,72	

Tabla 25. Promedio del tiempo de detección del fallo por cada estadística y método.

Evidentemente el promedio, por cada estadística, del tiempo de detección del fallo sin tener en cuenta las anomalías IDV (3), IDV (9) e IDV (15), es mucho menor que el mismo promedio realizado con todos los fallos, por lo que nos fijaremos en las columnas que ignoran los fallos incipientes para comparar los diferentes métodos y estadísticas.

Al igual que sucede en el número de fallos detectados, los métodos con LLE tienen peor resultados que los métodos convencionales, es decir, tardan más en detectar el fallo. La estadística Q_Y es la excepción, ya que experimenta una mejoría al aplicarse LLE, pero desde el punto de vista grupal, el conjunto de estadísticas de un mismo método tiende a dar peores resultados. Por el contrario, la estadística Q_X o en su defecto Q, es la que, en general, menos tiempo de detección del fallo ofrece. Los métodos que antes detectan el fallo son PCA y DPCA, como podemos ver en la tabla 25.

El último factor que vamos a comparar va a ser el número de alarmas después de que se produzca el fallo, es decir, después del instante 160. En la tabla 26 que aparece a continuación se muestra el promedio de alarmas después del fallo por cada método y estadística, así como teniendo o no en cuenta los fallos 3, 9 y 15.

En el análisis de esta factor sucede lo mismo que hemos comentado como en el número de fallos detectados como en el tiempo de detección de este, y es que los métodos con LLE obtienen peores resultados que su método equivalente convencional, ya que una vez ocurrido el fallo en el instante 160, el número de alarmas debe de ser alto. Finalmente, podemos afirmar que introducir LLE a un método, lejos de mejorarlo, lo empeora.

MÉTODO	ESTADÍSTICA	ALARMAS (%)	ALARMAS (SIN 3, 9 Y 15) (%)	PROMEDIO DE ALARMAS POR MÉTODO (SIN 3, 9 Y 15) (%)
PCA	T^2	55,14	64,16	70,74
	Q	67,14	77,32	
PCA CON LLE	T^2	46,14	53,83	51,49
	Q	42,15	49,15	
DPCA	T^2	54,11	62,94	69,79
	Q	66,53	76,64	
DPCA CON LLE	T^2	44,38	51,77	49,37
	Q	40,26	46,97	
CCA	T^2	57,51	66,75	59,51
	Q_X	66,45	77,37	
	Q_Y	29,97	34,42	
CCA CON LLE	T^2	31,50	36,75	41,61
	Q_X	42,89	50,04	
	Q_Y	32,60	38,03	
CCA REGULARIZADO	T^2	57,54	66,77	59,38
	Q_X	66,64	77,56	
	Q_Y	29,46	33,81	
DCCA	T^2	69,17	80,06	61,55
	Q_X	64,23	74,75	
	Q_Y	25,70	29,83	
DCCA CON LLE	T^2	28,72	33,51	42,04
	Q_X	44,49	51,90	
	Q_Y	34,90	40,70	
CCCA	T^2	57,54	66,77	51,80
	T_X^2	59,14	68,69	

	$\mathbf{T_Y^2}$	26,10	30,23	
	$\mathbf{Q_X}$	61,86	71,81	
	$\mathbf{Q_Y}$	18,51	21,48	
CCCA CON LLE	$\mathbf{T^2}$	31,48	36,73	37,85
	$\mathbf{T_X^2}$	31,23	36,43	
	$\mathbf{T_Y^2}$	15,67	18,28	
	$\mathbf{Q_X}$	45,99	53,66	
	$\mathbf{Q_Y}$	37,83	44,13	

Tabla 26. Promedio de alarmas después del fallo por cada estadística y método.

En la tabla 26 podemos observar que, generalmente, la estadística que tiene mayor número de alarmas y por tanto ofrece mejor resultado, es la estadística Q_X o Q en el caso de PCA y DPCA. Estos dos métodos son los que mayor número de alarmas después del fallo obtienen en conjunto, pero es la estadística Q_X de CCA regularizado la que mayor promedio de alarmas consigue de todas las estadísticas Q o SPE.

CAPÍTULO V: CONCLUSIONES Y TRABAJO FUTURO

5.1. CONCLUSIONES

Debido a la gran cantidad de datos que se recogen a lo largo de los procesos industriales, gracias a la nueva industria 4.0, surge la necesidad de aplicar técnicas y métodos capaces de procesar dicha información. En este trabajo se han empleado diferentes métodos estadísticos, en el ámbito del control de calidad, que analizan y disminuyen las dimensiones de estos datos, haciendo que su procesamiento sea más sencillo. La finalidad de la implementación de estos métodos es la detección de fallos a lo largo del proceso, permitiendo una pronta respuesta de actuación y toma de decisiones sobre este. Los datos que se han utilizado en la parte experimental proceden de la planta química Tennessee Eastman.

Los métodos empleados en este estudio han sido métodos de análisis estadístico multivariante como PCA, DPCA, CCA y su variante regularizada, así como DCCA y también CCCA. En todos ellos se ha realizado un primer análisis estadístico con datos normales del proceso, es decir, datos en los que no existe fallo, de esta manera obtener los umbrales de las diferentes estadísticas aplicadas. Mas tarde se realiza un segundo análisis, pero esta vez con cada uno de los 21 ficheros de fallo, en este caso se obtienen una serie de valores estadísticos que comparados con los umbrales obtenidos en el primer análisis sabremos si el fallo es o no detectado.

El primer método empleado ha sido el Análisis de Componentes Principales (PCA), el cual ha logrado detectar una media de 16 fallos con un tiempo promedio de 134,97 después de producirse el fallo. El siguiente método fue DPCA o Análisis de Componentes Principales Dinámico, también captó 16 fallos de media, pero tardo algo más de tiempo en detectarlos, 143,75 instantes después del fallo.

Luego se aplicó el Análisis Canónico de Correlación o CCA, el cual detectó una media de 14 fallos al igual que su versión dinámica o DCCA, mientras que CCA regularizado únicamente localizo 13 de los 21 fallos. Los tiempos de detección de estos 3 métodos son bastante más tardíos que en el caso de PCA y DPCA. Finalmente, se aplicó el método de Análisis Canónico de Correlación Concurrente (CCCA), pero no dio mejores resultados, ya que el promedio de fallos localizados es 13 y el tiempo promedio de detección es de 264,96 instantes después del fallo.

Por último, se implementó una herramienta llamada LLE (Locally Linear Embedding en inglés) en alguno de los métodos mencionados anteriormente, con la esperanza de mejorar los resultados obtenidos. Esta técnica, es un algoritmo de aprendizaje no supervisado que reduce las altas dimensiones de los datos de entrada, calculando puntos de interés en un espacio de baja

dimensionalidad manteniendo la información del vecindario en el espacio de datos inicial. En concreto, este algoritmo se ha implementado a los métodos PCA, DPCA, CCA, DCCA y CCCA y ninguno de los casos se han mejorado los resultados en relación con su método convencional, sino todo lo contrario.

Cabe mencionar que ninguno de los métodos mencionados con o sin la implementación de LLE ha logrado detectar alguno de los 3 fallos incipientes, IDV (3), IDV (9) e IDV (15).

5.2. TRABAJO FUTURO

Otros trabajos interesantes que pueden ser realizados en un futuro y que potencialmente puedan mejorar los resultados obtenido en este, sería la experimentación y desarrollo de un método que fuese capaz de detectar los fallos incipientes o también se podría explorar otra manera de calcular los umbrales de las diferentes estadísticas que se adapte de manera óptima al proceso.

También se podrían probar otras herramientas o algoritmos basados en manifold learning, que lejos de empeorar los métodos de análisis estadístico multivariante mencionados en este trabajo, los mejore.

Como ampliación a este estudio, se podrían experimentar una vez que se ha detectado el fallo, cuáles son las variables que producen esta alteración en el proceso.

BIBLIOGRAFÍA

[1] Pequeño Alonso, Á. (2020). Mejora del control de calidad de un proceso mediante técnicas de aprendizaje automático. Trabajo Fin de Grado. Universidad de Valladolid.

[2] Cubillos Rodríguez, M. C., & Rozo Rodríguez, D. (2009). El concepto de calidad: Historia, evolución e importancia para la competitividad. *Revista de la Universidad de la Salle*, *2009*(48), 80-99. https://ciencia.lasalle.edu.co/cgi/viewcontent.cgi?article=1170&context=ruls#:~:text=En%201924%20el%20matem%C3%A1tico%20Walter,costos%20m%C3%A1s%20econ%C3%B3micos%20que%20los [Último acceso: agosto 2022]

[3] Evolución de la Calidad - Gestión de Calidad - Asignaciones. (s. f.-b). https://sites.google.com/site/gestiondecalidadasign/home/tareas/tarea-2/evolucion-de-la-calidad [Último acceso: Septiembre 2022]

[4] Carro, R., & González Gómez, D. A. (2012). Control estadístico de procesos. http://nulan.mdp.edu.ar/1617/ [Último acceso: Septiembre 2022]

[5] Pérez Franco, I. (2021). Control de calidad de un proceso mediante la detección y diagnóstico de anomalías usando técnicas de control estadístico de procesos. Trabajo Fin de Grado. Universidad de Valladolid.

[6] Valenzuela, L. (2000). Diagrama de Ishikawa. Santiago de Chile, Chile: UNAB. https://www.academia.edu/31609684/Diagrama_de_Ishikawa [Último acceso: Septiembre 2022]

[7] Castro, C. H., Alonso, P. B., & González, J. I. A. (2005). Aplicación de los gráficos de control en el análisis de la calidad textil. Pecunia: revista de la Facultad de Ciencias Económicas y Empresariales, (1), 125-148.

[8] Riu, J. (2005). Gráficos de control de Shewhart. Técnicas de Laboratorio, 306, 1016.Grupo de Quimiometría, Cualimetría y Nanosensores. Universitat Rovira i Virgili. Tarragona. http://www.quimica.urv.es/quimio/index.php?option=com_content&view=article&id=20%3Atutoriales&catid=9%3Aarticulosdivulgacion&Itemid=22&lang=es [Último acceso: Septiembre 2022]

[9] Peirats de Castro, I. (2016). Gráficos de control univariantes según el promedio y la desviación típica del rango. Trabajo Fin de Grado. Universidad Carlos III. https://e-archivo.uc3m.es/handle/10016/27237?show=full [Último acceso: Septiembre 2022]

[10] López, C. P. (2004). *Técnicas de análisis multivariante de datos*. Pearson Educación.

[11] Rodrigo, J. A. (s. f.). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. https://www.cienciadedatos.net/documentos/35_principal_component_analysis#interpretaci%C3%B3n_geom%C3%A9trica_de_las_componentes_principales [Último acceso: Septiembre 2022]

[12] Sánchez Fernández, Á. (2020). Métodos de detección y diagnóstico de fallos mediante aproximaciones distribuidas: modelos, métodos y computación. Tesis Doctoral. Universidad de Valladolid.

[13] Garcia-Alvarez, D., & Fuente, M. J. (2011). Estudio comparativo de técnicas de detección de fallos basadas en el Análisis de Componentes Principales (PCA). *Revista Iberoamericana de Automática e Informática Industrial RIAI*, *8*(3), 182-195.

[14] Zhu, Q., Liu, Q., & Qin, S. J. (2016). Concurrent canonical correlation analysis modeling for quality-relevant monitoring. *IFAC-PapersOnLine*, *49*(7), 1044-1049.

[15] Calabuig, J. M., García Raffi, L. M., & Sánchez-Perez, E. A. (2015). Álgebra lineal y descomposición en valores singulares. *Modelling in Science Education and Learning*, *8*(2), 133-144

[16] Zhu, Q., Liu, Q., & Qin, S. J. (2017). Concurrent quality and process monitoring with canonical correlation analysis. Journal of Process Control, 60, 95-103.

[17] Valencia-Aguirre, J., Daza-Santacoloma, G., Acosta, C. D., & Castellanos-Domínguez, G. (2010). Comparación de métodos de reducción de dimensión basados en análisis por localidades. TecnoLógicas.

[18] Dobilas, S. (2022). LLE: Locally Linear Embedding—A Nifty Way to Reduce Dimensionality in Python. Medium. Geraadpleegd op, 27. https://towardsdatascience.com/lle-locally-linear-embedding-a-nifty-way-to-reduce-dimensionality-in-python-ab5c38336107 [Último acceso: Septiembre 2022]

[19] tok.wiki. (n.d.). Reducción de dimensionalidad no lineal Métodos de descomposición lineal relacionados y Aplicaciones de NLDR. https://hmong.es/wiki/Nonlinear_dimensionality_reduction. [Último acceso: Octubre 2022]

[20] Wu, P., Lou, S., Zhang, X., He, J., & Gao, J. (2020). Novel quality-relevant process monitoring based on dynamic locally linear embedding concurrent canonical correlation analysis. *Industrial & Engineering Chemistry Research*, *59*(49), 21439-21457.

[21] González Velázquez, M. (2020). Mejora de la calidad de un proceso mediante la detección de anomalías basada en datos. Trabajo Fin de Grado. Universidad de Valladolid.

[22] Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers & chemical engineering*, *17*(3), 245-255.

[23] Ricker, N. L., & Lee, J. (1995). Nonlinear modeling and state estimation for the Tennessee Eastman challenge process. *Computers & chemical engineering*, *19*(9), 983-1005.