TRABAJO FIN DE MÁSTER

Análisis estadístico del desgaste de la herramienta de mecanizado en la planta de motores Ford

Máster universitario en ingeniería de análisis de datos, mejora de procesos y toma de decisiones.

Alumna: Inés Castillo Santamaría

Tutora: Ana María Debón Aucejo

Cotutora: Susana Barceló Cerdá

Septiembre 2018

Resumen

En sectores como el de la automoción, generar un producto final de alta calidad y a la primera es fundamental. Para ello, es necesario aplicar técnicas estadísticas y técnicas de monitorización online para la detección precoz de fallos que ayuden a obtener una producción estable y libre de defectos. En la planta de motores Ford se han instalado unos sensores en una máquina de mecanizado capaces de capturar datos en continuo sobre las variables del proceso.

En primer lugar, sobre estos datos recopilados se han aplicado diversas técnicas estadísticas como PCA, modelos *clustering*, *Random Forest* y CART y los resultados obtenidos concuerdan y muestran que la variable torque es la variable más correlacionada con el desgaste de la herramienta de mecanizado. Sobre la variable torque se le ha aplicado el modelo Lee-Carter que nos ha permitido entender el comportamiento del torque a lo largo del tiempo.

En segundo lugar, se han monitorizando las variables del proceso mediante los gráficos de control de calidad multivariantes T^2 y SPE. También se han aplicado los gráficos MCUSUM y MEWMA para monitorizar el torque de los distintos ejes.

Todas estas herramientas estadísticas nos han permitido pasar del mantenimiento preventivo que se realiza actualmente en la factoría sobre las herramientas de mecanizado a un mantenimiento predictivo.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. Introducción

El objetivo principal del trabajo es predecir y monitorizar el desgate de la herramienta para tratar de alargar su vida útil, reduciendo de ese modo el número de cambios de la misma, produciendo piezas de manera óptima y reduciendo los costes de fabricación. Para ello es fundamental conocer que variables del proceso están relacionadas con el desgaste de la herramienta de mecanizado de la máquina OP110, encargada del mandrilado los cilindros del motor. El mandrilado es la operación de mecanizado que se realiza en agujeros de piezas ya realizados para obtener mayor precisión dimensional.

Para conseguir este objetivo principal ha sido necesario recopilar datos a los que se ha aplicado diversas técnicas de mineria de datos como PCA, modelos *clustering*, *Random Forest* y CART. Además de estas técnicas, más clasicamente utilizadas en este ámbito, hemos importado modelos como Lee-Carter para entender el comportamiento de la variable torque a lo largo del tiempo. Finalmente, se ha monitorizando las variables del proceso mediante los gráficos de control multivariantes T^2 y SPE. También se ha aplicado los gráficos MCUSUM y MEWMA para monitorizar el torque de los distintos ejes.

Hemos de subrayar que todas estas técnicas se han aplicado utilizando el software libre R, implementando cada uno de los ajustes de la herramientas estadísticas en la librería o R-paquete específico (R Core Team, 2017). Aunque

el algunas ocasiones tambien se ha utilizado Python (van Rossum, 1995) y Aspen ProMV.

El trabajo se divide en varios capítulos. En este capítulo, se hace una pequeña introducción sobre la empresa, la motivación y los objetivos del trabajo, además también se explica el funcionamiento de la máquina OP110. A continuación, en el segundo capítulo se introduce teóricamente la metodología estadística empleada. En el tercer capítulo se explica la compleja estructura de la base de datos. En el capítulo 4 se detallan los resultados obtenidos tras aplicar el software R, Python y Aspen ProMV. Para finalizar, en el quinto capítulo se muestran las conclusiones más importantes, la bibliografía empleada y los anexos.

1.1 Introducción a la empresa

La empresa americana Ford Motor Company fue fundada en 1903 por Herny Ford. Esta empresa es una multinacional que se dedica a la fabricación y venta de automóviles. Henry Ford es conocido mundialmente por su idea revolucionaria del sistema de producción en serie de manera masiva. La optimización de la cadena de montaje permitió aumentar el número de vehículos fabricados y disminuir el precio unitario de fabricación.

En 1976, se inauguró la fábrica Ford de Almussafes, Valencia. En los primeros años, la factoría valenciana se dedicaba exclusivamente a la fabricación del Ford Fiesta. En la actualidad se fabrican los modelos Mondeo, Galaxy, Kuga, Tansit Connect y S-Max. En la Figura 1.1 se muestra una imagen satélite de toda la factoría Ford en Almussafes.

Figura 1.1 Factoría Ford Almussafes (España)

Décadas después, en 2002 se construyó la planta de motores Ford España S. L. que recibió el nombre de *Valencia Engine Plant* (VEP). La planta de motores se divide en dos partes: mecanizado y montaje. En el mecanizado se reciben las piezas del motor en bruto y se mecanizan en las diferentes líneas: culatas, cigüeñales, árbol de levas y bloques. En el montaje se ensamblan las piezas ya mecanizadas y otras, que llegan ya terminadas por diferentes proveedores.

En *Valencia Engine Plant* se aplican las técnicas de *Fabricación Lean*, basadas en la participación de los grupos de trabajo, sistemas de mantenimiento de las instalaciones y un flujo sincronizado de piezas con mínimos inventarios, todo ello con una cultura de mejora continua en la calidad, minimización de costes y plazos de entrega.

1.2 Motivación del trabajo

En este proyecto se analizan los datos de una máquina fresadora de la línea de bloques denominada OP110. En esta máquina se mecanizan los 4 cilindros del bloque motor.

En los cilindros se producen explosiones constantes de combustible, lo que somete la pieza a un gran trabajo en condiciones extremas. Además, el cilindro es la parte del motor por donde se desplaza el pistón lo que produce desgaste superficial. Para disminuir la fatiga superficial se lubrica el cilindro con aceite.

La OP110 es una máquina que históricamente mecaniza numerosas unidades de forma incorrecta. Si las paredes del cilindro carecen de un buen acabado se producirá una alta fatiga superficial entre cilindro y pistón. Estos motores tendrán problemas de fuga de aceite y con el tiempo acabarán provocando fallos mecánicos. La Figura 1.2 muestra una imagen del bloque motor con los 4 cilindros.

Figura 1.2 Imagen bloque motor

Por todo ello, es de vital importancia para el buen funcionamiento del motor, que la geometría de estos cilindros sea muy precisa. Detectar cuando no se está mecanizando correctamente evitaría el trabajo excesivo en la producción de esta pieza, lo que disminuiría los costes de fabricación y aumentaría la capacidad productiva de la factoría.

Dado que se sabe que el desgaste de la herramienta de mecanizado produce un mal acabado superficial, los ingenieros de la panta de motores Ford han estimado un tiempo de vida medio de la herramienta de 3000 ciclos.

Actualmente, para evitar mecanizar cilindros incorrectos cada 3000 ciclos de mecanizado el operario debe parar la máquina y cambiar la herramienta. Este mantenimiento preventivo conlleva algunos problemas: se desechan muchas herramientas sin que hayan llegado realmente al fin de su vida útil y además supone paradas continuas de la producción.

1.3 Objetivos del trabajo

En primer lugar, como se tiene una base de datos muy grande, uno de los objetivos será reducir la dimensionalidad del problema y para ello se ha empleado análisis de componentes principales (PCA).

En segundo lugar, se desea caracterizar los ciclos de mecanizado para identificar patrones y de está forma conocer que ciclos son buenos y cuales no y para ello se han utilizado distintos métodos de *clustering*. Para saber qué variables son las más importantes a la hora de clasificar un ciclo se ha empleado *Random Forest* y CART. Una vez conocidas las variables más importantes se les ha aplicado el modelo Lee-Carter que nos ha permitido descomponerlas para obterner más información. La aplicación de estas técnicas estadísticas nos ha permitido conocer y entender mejor el funcionamiento de la máquina.

Finalmente, el trabajo persigue monitorizar el proceso y así saber cuando la máquina está mecanizando incorrectamente y es por ello que se han construido diversos gráficos de control multivariante. Esta monitorización de las variables del proceso nos permitiría pasar de un mantenimiento preventivo a realizar un mantenimiento predictivo y asi conseguir el objetivo principal, predecir cuando debería pararse la máquina para sustituir la herramienta.

1.4 Descripción de la OP110

La OP110 consta de diferentes estaciones de mecanizado. Este trabajo se centra en el estudio de los datos de la estación 7 de la máquina OP110. En esta estación se mecanizan 2 cilindros de 2 motores, es decir, se mecanizan 4 cilindros. Cada cilindro se mecaniza con un husillo o herramienta de mecanizado y se denominan H1, H3, H5, H7. La Figura 1.3 muestra la vista de planta de 2 bloques motor y las circunferencias en color rojo indican que 4 cilindros se mecanizan en la estación 7. Los 4 cilindros que no se mecanizan en la estación 7 se mecanizan en la estación simétrica siguiente denominada, estación 8.

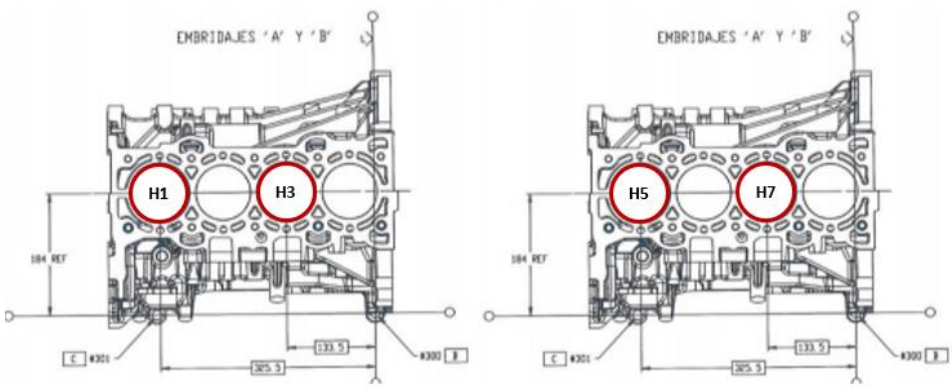

Figura 1.3 Dos bloques motor, vista de planta

La Figura 1.4 es una fotografía de uno de los 4 husillos también denominados en este trabajo como herramientas de mecanizado. Los husillos son los cabezales donde se colocan las plaquitas de mecanizado. Además, realizan dos movimientos simultáneos: se desplazan en el eje vertical (movimiento de avance) y giran sobre si mismos (movimiento rotatorio). Además, cada husillo consta de 5 plaquitas de mecanizado, 3 para el semiacabado, 1 para el acabado y 1 para el chaflán.

Figura 1.4 Husillo OP110

En la Figura 1.5 se muestra la posición del eje de avance de un ciclo. El eje de avance o eje Z es el encargado de realizar el movimiento vertical de los 4 husillos a la vez. Nótese que en la Figura 1.5 el eje de ordenas indica los milímetros que descienden los husillos y el eje de las abcisas indica los 8192 instantes de tiempo que dura un ciclo.

Un ciclo de mecanizado consta de 6 diferentes fases: aproximación, semiacabado, chaflán, expansión de la herramienta, acabado y retroceso (ver Figura 1.5). En la primera fase los husillos descienden desde la cota superior hasta que se aproximan a la pieza, esta fase la denominaremos aproximación. A continuación, los husillos siguen descendiendo hasta que se introducen en los cilindros del bloque motor para realizar el semiacabado. Los últimos 5 milímetros del cilindro se denomina chaflán. Es importante que para mecanizar esta parte la velocidad del eje Z sea inferior. Después de realizar el chaflán los husillos cambian de geometría para prepararse para el acabado. Tanto en el acabado como en el semiacabado se mecanizan los cilindros pero la diferencia entre ellos es que en el semiacabado se quita más material y es por ello por lo que se emplean 3 plaquitas duras. En el acabado se produce un mecanizado más fino y solamente se emplea 1 plaquita. Al finalizar el acabado los husillos ascienden y retornan a su posición inicial.

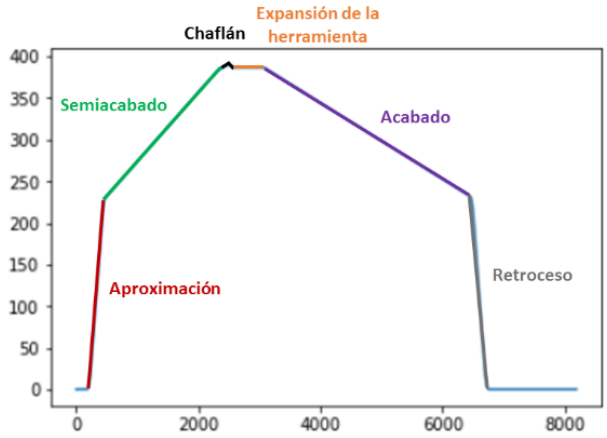

Figura 1.5 Gráfica posición eje Z

De las 6 fases descritas solamente en el semiacabado, chaflán y acabado se mecanizan los cilindros y por tanto se produce desgaste de la herramienta.

En la Figura 1.6 muestra un esquema del funcionamiento de la estación 7 de la OP110. Se han instalado 4 sensores en cada uno de los ejes: Z, S1, S2 y S3. Tal y como se muestra en la Figura 1.6 los husillos H3 y H5 comparten el mismo eje de giro, el eje S2. El eje S1 y S3 también son ejes de giro unidos cada uno a los husillos H1 y H7 respectivamiente. En cambio, el eje Z es un eje de avance que se encarga de hacer descender y ascender los 4 husillos a la vez.

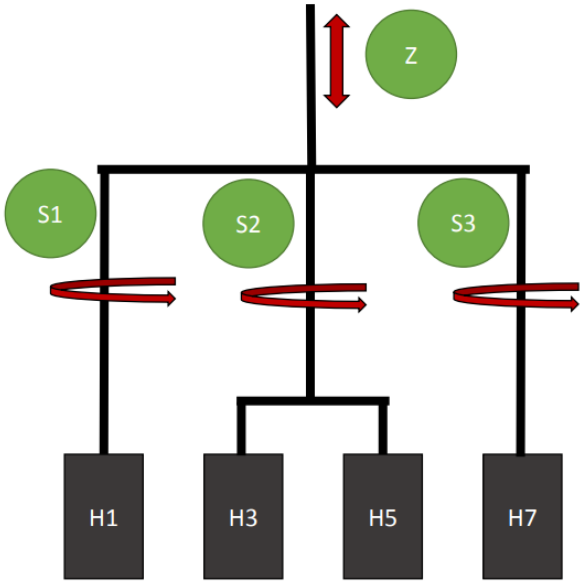

Figura 1.6 Esquema funcionamiento de la OP110

CAPÍTULO 2

2. Metodología

Para lograr alcanzar el objetivo principal ya planteado, en este segundo capítulo se describe detalladamente la metodología que se usará. En este capítulo, se describirán todas las herramientas y técnicas estadísticas utilizadas para poder preparar, analizar y validar la información.

Comenzaremos por los métodos de reducción de la dimensionalidad para después continuar con los métodos *cluster* que nos ayudarán a encontrar patrones ocultos en los datos a simple vista.

2.1 PCA

Principal Component Analisys (PCA) es una técnica estadística multivariante utilizada por muchos científicos en distintas disciplinas para la reducción de dimensionalidad en los datos. El PCA consiste en analizar conjuntos de datos de observaciones compuestas por muchas variables dependientes, que en están normalmente, correlacionadas entre sí. El objetivo es extraer la información importante y expresar esta información como un conjunto de nuevas variables ortogonales entre sí denominadas componentes principales. Además, el PCA es también una buena herramienta para comprimir los datos, tomando solamente la información más importante. De esta forma

simplificamos la estructura de los datos facilitando la interpretación de la relación entre variables.

Las componentes principales se obtienen mediante la combinación lineal de las variables originales. La primera componente principal será aquella con mayor varianza explicada, es decir, aquella que explique la máxima parte de la inercia de los datos originales. La segunda componente se compondrá bajo la restricción de ser ortogonal a la primera componente y que tenga la máxima inercia posible. Las siguientes componentes se calcularán de la misma forma. Los valores que toman las observaciones en el nuevo sistema de referencia definido por las componentes se denominan *factor scores.*

Antes de extraer las componentes principales debemos realizar un preprocesado de los datos. En primer lugar, se debe centrar los datos, esto quiere decir substraer la media de cada variable a cada una de las observaciones, lo que facilita la interpretación de los resultados. Además, y dado que el PCA es un método que calcula las componentes determinando las direcciones de máxima varianza, si las distintas variables no se miden en la misma escala debemos escalar los datos dividiéndolos por sus correspondientes desviaciones típicas, para que teniendo todas la misma varianza, ninguna variable predomine a priori en la formación de las componentes. La combinación de centrar y escalar los datos se denomina autoescalado

El resultado de un modelo PCA se compone de 4 partes: los datos, los *scores*, los *loadings* y los residuos, y su expresión matemática es la siguiente:

$$X = TP^T + E \qquad (2.1)$$

Dónde X es la matriz de datos originales formada por i observaciones o filas y j variables o columnas. La matriz de *scores* (T) es una matriz de i filas y componentes o columnas. Los *scores* se pueden interpretar geométricamente como las proyecciones de las observaciones en los componentes principales. La matriz de *loadings* (P) tiene una dimensión de A filas y J columnas. Los *loadings* indican la correlación entre una componente y una variable original. Y finalmente, la matriz de residuos(E), de iguales dimensiones que la matriz de datos es la parte de la variabilidad no explicada por las componentes principales. La figura 2.1 muestra un esquema de la ecuación (2.1).

Figura 2.1 Esquema PCA

Para un mayor detalle de la técnica de componentes principales puede consultarse en Abdi & Williams (2010) y en Jackson (2005).

2.2 *Cluster*

El *cluster* es un conjunto de técnicas no supervisadas, según terminología por *machine learning*, cuya finalidad es encontrar patrones o grupos dentro de un conjunto de observaciones. Las particiones se establecen de forma que, las observaciones que están dentro de un mismo grupo, son similares entre ellas y distintas a las observaciones de otros grupos. Se trata de un método no supervisado, ya que el proceso ignora la variable respuesta que indica a que grupo pertenece realmente cada observación (si es que existe tal variable). Esta característica diferencia al *clustering* de las técnicas estadísticas conocidas como análisis discriminante, que emplean un *set* de entrenamiento en el que se conoce la verdadera clasificación.

Dentro de los *clustering* pueden diferenciarse tres tipos de métodos principales:

- *Partitioning Clustering*: Este tipo de algoritmos requieren que el usuario especifique de antemano el número de *clusters* que se van a crear (*Kmeans*, *Kmedoids*, *CLARA*).
- *Hierarchical Clustering*: Este tipo de algoritmos no requieren que el usuario especifique de antemano el número de *clusters*. (*agglomerative clustering, divisive clusterig*).
- Métodos que combinan o modifican los anteriores (*hierarchical Kmeans*, *fuzzy clustering*, *model based clustering* y *density based clustering*).

A continuación, se describen teóricamente los 4 métodos de *clusters* que se emplean en este proyecto: *fuzzy clustering, Hierarchical Clustering* Ward, *Hierarchical Clustering* Complete y *Kmeans*.

2.2.1 *Fuzzy clustering*

El algoritmo *fuzzy clustering* asigna a cada muestra un valor de pertenencia dentro de cada uno de los *clusters* por lo que un individuo puede pertenecer parcialmente a más de un *cluster*. Este algoritmo realiza una partición suave del conjunto de datos, en tal partición las muestras pertenecen en algún grado a todos los *clusters.*

La partición suave restringida, es aquella que la suma de los grados de pertenencia de un punto específico en todos los *clusters* es 1.

$$\sum_j \mu c_j(x_i) = 1 \qquad \forall x_i \in X \tag{2.2}$$

EL *fuzzy clustering* produce una partición suave restringida y para hacer esto se introduce en la función objetivo los grados de pertenencia difusos de

cada observación al *cluster*. El parámetro m es un peso que determina el grado en el cuál los miembros parciales del cluster afectan al resultado.

$$J_m\,(P,V) = \sum_{i=1}^{k}\sum_{x_k}(\,\mu c_j(x_k))^m\,\|x_k - v_i\|^2 \qquad (2.3)$$

El *fuzzy clustering* intenta buscar una buena partición mediante la búsqueda de los prototipos o centros v_i que minimicen la función objetivo J_m y también debe buscar las funciones de pertenencia μc_j que minimicen a J_m . En este trabajo utilizamos la versión *fuzzy* del algoritmo de *clustering kmeans*, así como su actualización (Pal, Bezdek, & Hathaway, 1996) implementada en el paquete de R llamado e1071 (Meyer, Dimitriadou, Hornik, & Weingessel, 2017).

Con el fin de conocer el número de *clusters* óptimo se analiza la validez del agrupamiento. La validez se estima mediante una serie de índices (Rezaee, Lelieveldt, & Reiber, 1998) y un criterio objetivo para ver cuán buena es la partición generada por el algoritmo. Las medidas de validez de una partición suave restringida se dividen en tres categorías: medidas basadas en el grado de pertenencia, medidas basadas en el desempeño y medidas basadas en la geometría. Estas últimas medidas hacen referencia a que cuanto más compactos y separados estén los *clusters* mejor es su partición. Existen una serie de índices que ayudan a cuantificar lo compactos que son los grupos. Los índices *partition entropy* (pe), *fukuyama sugeno* (fs) y *xie beni* (xb) indican que cuanto más bajos sean sus valores mejor es la agrupación. En cambio, el índice *partition coefficient* (pc) los valores mayores presentarán una mejor partición.

2.2.2 *Cluster* jerárquico método Ward

El método de Ward es un criterio aplicado en los métodos jerárquicos de análisis *cluster*. El objetivo de los métodos jerárquicos es agrupar *clusters* para formar uno nuevo o para separar alguno existente y dar origen a otros dos. De esta forma, si sucesivamente se va realizando este algoritmo se minimice o maximice alguna medida de similitud.

Los métodos jerárquicos se dividen en: aglomerativos y disociativos. El método de Ward es una estrategia empleada en los métodos jerárquicos aglomerativos. Los métodos aglomerativos comienzan el análisis con tantos grupos como individuos haya. Posteriormente, a partir de estos grupos iniciales se van formando grupos de forma ascendente. Al final del proceso, todos los individuos estarán englobados en un mismo conglomerado. En el método de Ward en cada etapa se fusionan los dos grupos que tengan el menor incremento en el valor total de la suma de los cuadrados de las diferencias de cada individuo al centroide de su *cluster*.

$$E_k = \sum_{i=1}^{n_k}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 \qquad (2.4)$$

Donde E_k es la suma de cuadrados de los errores del *cluster* k, es decir, la distancia euclídea al cuadrado entre cada individuo del *cluster* k a su centroide. Si suponemos que hay h *clusters* la expresión a minimizar será la siguiente:

$$E = \sum_{k=1}^{k} E_k \qquad (2.5)$$

donde E es la suma de cuadrados de los errores para todos los *clusters*.

2.2.3 *Cluster* jerárquico método Complete

El método Complete es otro método jerárquico aglomerativo también conocido como el procedimiento de *amalgamamiento completo*. Este método considera que la distancia o similitud entre dos *clusters* debe medirse de acuerdo a sus elementos más dispares. Es decir, la distancia o similitud viene dada por la máxima distancia o similitud entre sus componentes.

Si tenemos ya formados K *clusters*, la distancia y similitud entre los *clusters* C_i y C_j (con n_i y n_j elementos respectivamente) será:

$$\begin{aligned} d(C_i C_j) &= Max\{d(x_l x_m)\} \quad l = 1 \dots n_i \;\; m = 1 \dots n_j \\ s(C_i C_j) &= Min\{s(x_l x_m)\} \quad l = 1 \dots n_i \;\; m = 1 \dots n_j \end{aligned} \qquad (2.6)$$

En el caso de emplear las distancias, se unirán los *clusters* C_i y C_j si los miembros mas dispares de cada cluster es la mínima.

$$Min\left\{Max\{d(x_l x_m)\}\right\} \qquad (2.7)$$

En el caso de emplear similitudes, se unirán los *clusters* C_i y C_j si la similitud entre los individuos más dispersos de cada *cluster* es la máxima.

$$Max\left\{Min\{s(x_l x_m)\}\right\} \qquad (2.8)$$

2.2.4 *Cluster Kmeans*

El algoritmo *Kmeans* es una técnica propia de agrupamiento no supervisado que tiene como objetivo encontrar patrones que sirvan para la clasificación de los diversos individuos. *Kmeans* realiza una partición dura del conjunto de datos lo que significa que cada observación pertenece exclusivamente a un *cluster*. Además, todos los datos tienen que pertenecer a alguno de los *clusters*. El número de *clusters* debe ser definido para inicializar el algoritmo.

El algoritmo *Kmeans* trata de encontrar el centro de cada *cluster* (centroide) y determinar cual es el único *cluster* al que pertenece cada individuo. Para hallar el centroide existen diversos criterios. Uno de esos criterios es minimizar la suma de la distancia entre los puntos de cada *cluster* y su centro:

$$J(P,V) = \sum_{j=1}^{c} \sum_{x_i} \left\| x_i - v_j \right\|^2 \quad (2.9)$$

Donde v_j es el vector de los centros de cada *cluster*. Se debe minimizar J. En primer lugar, el algoritmo *Kmeans* calcula la partición actual con base a los prototipos actuales usando un método de optimización para minimizar la función objetivo J. Posteriormente, el paso anterior se repite iterativamente hasta alcanzar algún criterio de parada. Normalmente este criterio de parada es la diferencia de los prototipos entre dos ciclos consecutivos. Cuando el algoritmo alcanza su criterio de parada significa que la función J llegó a un mínimo local.

2.3 Árbol de decisión

En general tenemos dos tipos de árboles: de clasificación y de regresión por esto se conocen como CART que es su acrónimo en ingles (*Classification and Regression Trees*). El objetivo de los árboles de clasificación es dado un conjunto de datos previamente etiquetados construir un modelo en forma de árbol para clasificar los nuevos casos. Una descripción detallada de CART se puede encontrar en Hastie, Tibshirani, & Friedman (2001). En este trabajo se describen los árboles de clasificación, cuya variable independiente es de tipo categórica.

Un árbol de decisión es una forma gráfica y analítica de representar los sucesos. Esta forma gráfica comienza por un único nodo y luego se ramifica en resultados posibles. Cada uno de estos resultados crea nuevos nodos y se ramifica en otras posibilidades.

Se tiene un *dataset* con n observaciones y C clases. El modelo parte esas observaciones en C grupos y a cada uno de estos grupos se le asigna una clase predicha. Si separamos un nodo A en dos ramas A_L será la ramificación izquiera y A_R la ramificación derecha. Siendo f una función de impureza se define la impureza de un nodo A como:

$$I(A) = \sum_{i=1}^{C} f(p_i A) \quad (2.10)$$

Dónde $p_i A$ es la proporción de A que pertenece a la clase i para futuras muestras. El nodo A es puro si I(A)=0 y para ello $f(0) = f(1) = 0$. Dos candidatos para f son el índice de Gini $f(p) = p(1-p)$ y el índice de información $f(p) = -plog(p)$. Utilizando la separación con máxima reducción de impureza:

$$\Delta I = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \quad (2.11)$$

El árbol de clasificación empleado ha sido el utilizado por la librería de R denominada *rpart* (Therneau & Atkinson, 2018).

2.4 Random forest

El Random Forest es un conjunto de árboles de clasificación o regresión creados mediante muestras *bootstrap* o *bagging* fueron propuestos por Liaw & Wiener (2002). En lugar de emplear el conjunto total de variables para la construcción de un único árbol, en el Random Forest se construyen múltiples árboles de decisión empleando para cada uno un subconjunto de variables predictoras elegidas al azar. Por tanto, esta técnica añade aleatoriedad al proceso de *bagging*. Con este modelo evitamos los problemas de sobreajuste del *training set*.

El método de *bootstrapping* permite realizar mejores modelos porque desciende la varianza del modelo haciéndolo más robusto. Si entrenamos varios árboles con el mismo *set* de entrenamiento obtendríamos árboles altamente correlacionados. El muestreo *bootstrap* en una forma de descender la correlación entre los múltiples árboles.

Además, el *Random Forest* nos proporciona una medida de importancia de las variables predictoras que puede ser utilizada como técnica de reducción de la dimensionalidad. Esta técnica fue descrita por Breiman (2001). Para saber que variables son las más influyentes se utiliza el *Mean Decrease Accuracy y el Mean Decrease Gini.*

Los modelos *Random Forest* son populares debido a su precisión, su robustez y su fácil uso. Una descripción más detallada de los CART puede encontrarse en Hastie, Tibshirani, & Friedman (2001). En este trabajo se ha empleado la librería *randomForest* que implementada en el software R (Liaw & Wiener, 2002).

2.5 Monitorización multivariante del proceso

La mejora de la calidad implica el incremento de la productividad y la disminución de costes y por tanto, un aumento del beneficio. El control estadístico de calidad es una de las herramientas que se utiliza para mejorar la calidad. Con los gráficos de control se detecta la aparición de salidas de control o cambios importantes en el proceso. Cada vez es más frecuente plantearse en la industria controlar simultáneamente en un mismo producto o pieza varias características del proceso. En nuestro caso, se tiene p=12 variables del proceso.

Si observamos el comportamiento de estas p variables, veremos que, al pertenecer a un mismo producto, están correlacionadas. Es decir, su

comportamiento no es independiente, y cuando una de las variables cambia de magnitud las demás tienden a modificar también sus valores.

La monitorización multivariante de procesos por lotes se basa en modelar las causas comunes de la variabilidad presentes en un conjunto de observaciones representativo de las condiciones normales de operación. En los enfoques multivariantes, la variación en las trayectorias entre los lotes representativos de las condiciones normales de operación (variación debida a causas comunes) se sintetiza en un reducido espacio de vectores latentes mediante la aplicación de métodos de proyección multivariante.

2.5.1 Gráfico T^2 de Hotelling

La principal aportación al control estadístico de procesos multivariante fue la realizada por el Profesor Harold Hotelling (1947), quien propuso un gráfico de control multivariante basado en la distancia de Mahalanobis (1936).

El gráfico de control multivariante T^2 de Hotelling monitoriza la distancia entre el vector de promedios esperado y el vector de promedios observados, considerando su matriz de covarianzas estable. MacGregor y Kourti (1995) adaptaron este gráfico al subespacio de A dimensiones definidas por el modelo PCA. Solo tiene un límite de control el cual se basa en un valor umbral que si se supera por el valor observado indica que la distancia entre los dos vectores es lo suficientemente grande para declarar al proceso como fuera de control. Dicho límite se calcula mediante la Ecuación 2.12 .

$$UCL(T^2)_\alpha = \frac{A(m^2-1)}{m(m-A)} F_{(A,(m-A)),\alpha} \qquad (2.12)$$

dónde:

A : Número de componentes principales
m : Número de observaciones del modelo de referencia
$F_{\alpha,(A,N-A)}$: es el percentil 100(1- α)% de una distribución *F* con $(A, N-A)$ grados de libertad

En este trabajo se va a realizar un gráfico de control multivariante T_A^2 utilizando los *scores* que son los valores que toman las variables en el nuevo subespacio tras la transformación realizada durante el PCA (Nomikos & Macgregor, 1995). La suma de los cuadrados de los scores tipificados (divididos por los valores propios, λ_a, para que la varianza sea 1) se denomina estadístico T_A^2 de Hotelling.

$$T_A^2 = \sum_a \frac{t_{ia}^2}{\lambda_a} \qquad (2.13)$$

Dónde A se refiere a las componentes principales extraídas en el PCA. Luego por cada observación se tendrá un valor de dicho estadístico y al representar estos valores sobre un gráfico se monitoriza el proceso.

2.5.2 Gráfico Squared Prediction Error, SPE

El estadístico *Squared Prediction Error* (SPE) se calcula a partir de la Ecuación 2.16 .

$$SPE = e_i^T e_i = (y_i - \overline{y_i})^T (y_i - \overline{y_i}) \quad (2.14)$$

Tal y como su nombre indica, el SPE es el error de predicción al cuadrado. Así, según la fórmula anterior, y_i es el valor original de la variables e $\overline{y_i}$ el valor predicho por el modelo matemático generado a partir del PCA. A esta diferencia entre el valor original y el valor predicho se le denomina residuo. En este caso, a cada nueva observación le corresponde un valor SPE, que dibujado en una gráfica forma el gráfico de control de SPE. Este gráfico de control también tiene su respectivo límite de control, que viene dado por la Ecuación 2.15:

$$UCL(SPE)_\alpha = \frac{K-A}{c^2} s_0^2 F_{(K-A,(m-A-1)),\alpha} \quad (2.15)$$

dónde:

K : Número de variables originales
A : Número de componentes principales
c : Factor de corrección
s_0: Desviación típica
$F_{(K-A,(m-A-1)),\alpha}$: Percentil 100(1-α)% de la correspondiente distribución

En el cálculo de ambos límites de control (UCL, Upper Control Limit) se utiliza el valor F que hace referencia a la distribución F de Snedecor.

2.5.3 Gráfico MEWMA

Los gráficos EWMA y los gráficos CUSUM son apropiados para detectar pequeños desajustes y derivas precozmente. Son denominados gráficos con memoria ya que su representación gráfica se basa en la acumulación de información, de forma que cada instante se considera información histórica. En el caso del MEWMA, la ponderación que se le da a cada período para el cálculo de la estadística decrece en forma exponencial, a medida que se aleja del período actual. En el MCUSUM, a cada periodo se le asigna la misma ponderación para el cálculo de la estadística.

. El gráfico de medias móviles ponderadas exponencialmente para el caso multivariante se denomina MEWMA(*Multivariate Exponentially Weighted Moving*

Average) es una extensión a esta nueva situación del correspondiente gráfico univariante. La extensión multivariante, propuesta por Lowry (1992), adopta la forma:

$$Z_i = \Lambda X_i + (I - \Lambda) Z_{i-1} \qquad (2.16)$$

donde: X_i es el vector de medias muestrales y Λ es la matriz diagonal formada por los valores λ para las distintas variables. Los λ_j marcan la profundidad de la memoria para cada variable. A mayor valor de λ_j , menor profundidad. I es la matriz identidad y se considera como valor inicial $Z_0 = 0$.

La información que proporcionan los Z_i se recoge en el estadístico:

$$T_i^2 = Z_i^t \sum_{Z_i}^{-1} Z_i \qquad (2.17)$$

Donde $\sum_{Z_i}^{-1}$ es la inversa de la matriz de varianzas-covarianzas de los Zi . La señal de salida de control se produce cuando T_i^2 supera un cierto valor h, (h > 0) seleccionado de manera tal que se logre un cierto valor de ARL cuando el proceso está bajo control. Si no existe a priori ninguna razón para ponderar en forma diferente las observaciones pasadas de cada una de las p variables (como generalmente sucede), entonces se considera $\lambda_1 = \ldots = \lambda_p$.

La matriz $\sum_{Z_i}$ puede obtenerse a partir de los elementos de la matriz de varianzas-covarianzas correspondiente a las variables analizadas mediante la expresión:

$$\sum_{Z_i} = \frac{\lambda}{2-\lambda}[1 - (1-\lambda)^{2i}] \sum_x \qquad (2.18)$$

donde $\sum_x$ es la matriz de varianzas-covarianzas original. Cuando r = 1, el gráfico MEWMA coincide con el gráfico de control T^2 dado que el valor asintótico de la matriz de varianza-covarianza de Z_i es:

$$\sum_{Z_i} = \frac{\lambda}{2-\lambda} \sum_x \qquad (2.19)$$

En lo que respecta al límite de control (superior) empleado, Runger y Montgomery (1996) sugieren una aproximación mediante cadenas de Markov, que permite estudiar el funcionamiento del gráfico referente al ARL. Proporcionan además una serie de recomendaciones para la selección de los parámetros del gráfico. En lo que se refiere al parámetro λ, Montgomery (1991) recomienda que su valor está comprendido entre 0.05 y 0.25. Cuanto mayor sea el valor del parámetro, menor importancia se le estará dando a los valores más alejados en el tiempo. En aplicaciones prácticas se elige 0.1 como el valor del parámetro λ.

2.5.4 Gráfico MCUSUM

El término CUSUM procede del inglés *cumulative-sum*, que significa suma acumulada. Estos gráficos se basan en la representación de la acumulación de las desviaciones de cada observación respecto a un valor de referencia. Los gráficos CUSUM pueden extenderse también al caso multivariante, aunque no hay una única forma de hacerlo. Los primeros en realizar estudios en la materia fueron Woodall & Ncube (1985), usando un esquema basado en múltiples (p) CUSUM univariantes.

Otras propuestas de gráficos MCUSUM son los de Crosier (1988) o los de Pignatello & Runger (1990). En este proyecto se utiliza el paquete de R denominado MSQC (Santos-Fern, 2013) el cual implementa el primer procedimiento de Crosier que reduce cada observación multivariada a un escalar y luego construye el estadístico CUSUM con los escalares. Crosier considera el siguiente estadístico:

$$T_i = \sqrt{T_i^2} = \sqrt{n(\bar{X}_\iota - \mu_0)^t(\bar{X}_\iota - \mu_0)} \qquad (2.20)$$

donde:

μ_0: vector de valores objetivos
i=1, 2,...,m es el número de muestra
$\bar{X}_\iota$: es el vector de observaciones o de promedios de subgrupos de tamaño n, de la i- ésima muestra.

Crosier (1988) considera el estadístico T_i porque de esa manera, se acumulan distancias en lugar de distancias al cuadrado. La estadística MCUSUM se calcula de la siguiente manera:

$$S_i = \max(0, S_{i-1} + T_i - k) \qquad (2.21)$$

donde $S_0 \geq 0$ (en general S_0 = 0) y k>0.

El gráfico MCUSUM da una señal de fuera de control cuando el valor de S_i es mayor que un cierto valor h que depende del valor del ARL deseado cuando el proceso está funcionando en el valor objetivo [S_i > h].

Para determinar los valores de h y k para p = 2, 5, 10 y 20 y para valores de ARL de 200 y de 500 cuando el proceso está bajo control Crosier empleó procesos de Markov. Estas tablas se diseñaron para detectar un cambio d = 1 en el vector de medias, donde d = λ , con λ parámetro de no centralidad. En este caso especial los valores de k óptimos se aproximan a p.

Tanto en los gráficos MCUSUM y MEWMA se ha dividido en control de calidad en dos etapas. En la fase 1 el objetivo es estimar la media y la varianza

del proceso, y al mismo tiempo, saber si son estables. La fase 1 es una fase de estimación y comprobación. Si en la fase 1 podemos admitir que la media y la varianza son constantes entonces podemos empezar la fase 2 realizando el control con los valores estimados en esa primera fase.

2.6 Modelo Lee- Carter

El modelo Lee-Carter es un algoritmo utilizado para la predicción de la mortalidad y la esperanza de vida (Lee & Carter, 1992). En este proyento vamos a emplear este modelo para la predicción del desgaste de la herramienta de mecanizado que debe reflejarse en la variables torque en el eje Z, torque en el eje S2 y torque en el eje S3. El modelo basa su aproximación en la proyección de la tendencia histórica presentada por las variables, a través de series de tiempo para generar análisis sobre el comportamiento futuro.

La aportación más importante de Lee y Carter en su articulo es la descomposición de la información en dos dimensiones, es decir, el desgaste de los ciclos a través de periodos de tiempo puede descomponerse en varios parámetros. Concretamente, el modelo asume que la dinámica del desgaste puede explicarse a traves de tres parametros generados por la regresión. Entre esos tres parametros está el parámetro dependiente del tiempo, el cual identifica la dinámica creciente o decreciente del desgaste. A partir de este índice se puede proyectar el comportanmiento del desgaste usando un modelo clásico de series de tiempo como es Box-Jenkins.

El modelo Lee-Carter consiste en ajustar la siguiente ecuación a cada una de las variables del proceso de mecanizado, v_{ct} , donde c indica el ciclo y t indica el periodo

$$v_{ct} = \exp(a_c + b_c k_t + E_{ct}), \qquad (2.22)$$

o de forma equivalente:

$$\ln(v_{ct}) = a_c + b_c k_t + E_{ct}. \qquad (2.23)$$

En dichas expresiones obtemos dos parámetros dependientes de los ciclos, a_c y b_c, y un tercero que depende del tiempo. Los parámetros de la ecuación tienen la siguiente interpretación:

- a_c : Comportamiento general de la variable a lo largo de los ciclos.
- b_c : Sensibilidad de cada ciclo a la evolución general a lo largo del tiempo.
- k_t : Evolución de la variable a lo largo del tiempo.
- E_{ct} : Influencia de cada especifico ciclo no capturada por el modelo.

El modelo sufre problemas de identificabilidad lo que supone que la solución no es única. Para evitar problemas de identificabilidad al obtener los valores de a_c, b_c y k_t, Lee y Carter (1992) proponen que la suma del los b_c sea

igual a 1 y que la suma de los k_t sea igual a 0. El enfoque de Lee-Carter propone que para estimar los parámetros del modeloa_c, b_c y k_t, se emplee el primer término de la descomposición de valores singulares (SVD) de la matriz correspondiente a las variables.

La estimación de a_c es,

$$\breve{a}_c = \frac{\sum_t v_{ct}}{T} \qquad (2.24)$$

Con T número de instantes de tiempo, en nuestro caso 2050 milisegundos. Los valores de $\breve{b_c}$ y $\breve{k_t}$ se estiman con el primer térnimo de SVD .

$$v_{ct} - \breve{a}_c = \sum_{i=1}^{\min(n,T)} s^i u_c^i v_t^i \qquad (2.25)$$

dónde n indica el numero de individuos (en nuestro caso n=500) y s^i, u_c^i y v_t^i son, de izquierda a derecha, los valores singulares y los vectores singulares (Renshaw & Haberman, 2003).

$$\breve{k}_t = s^1 v_t^1 \qquad \breve{b}_c = u_c^1 \qquad (2.26)$$

Los valores de $\breve{b_c}$ y $\breve{k_t}$ se deben normalizar con la siguiente transformación:

$$\breve{k}_t = c\breve{k}_t \qquad \breve{b}_c = \frac{\breve{b}_c}{c} \qquad (2.27)$$

Los valores estimados $\breve{b_c}$ y $\breve{k_t}$ pueden presentar discrepancias entre el valor predicho y el valor estimado. Para evitar este problema k_t se debe reestimar empleando $\breve{k}_t$ como un valor inicial.

$$D_t = \sum_c \left(E_{ct} \frac{\exp(\breve{a}_c + \sum_i k_t \widetilde{b_c})}{1 + \exp(\breve{a}_c + \sum_i k_t \widetilde{b_c})} \right. \qquad (2.28)$$

Por último, la solución final requiere de una translación.

$$\widehat{k_t} = \widetilde{k_t} - mean(\widetilde{k_t}) \qquad \widehat{a_c} = \widetilde{a_c} + \widehat{b_c} mean(\widetilde{k_t}) \qquad (2.29)$$

CAPÍTULO 3

3. Descripción de la base de datos.

La base de datos que vamos a analizar presenta una estructura de un proceso por lotes. En cada ciclo (i) se recogen 8192 datos por variables (k=8192). Por tanto, se tiene una base de datos inicial enorme. Para reducir la dimensionalidad se ha decidido dividir la base de datos según las fases descritas en el apartado 1.4.

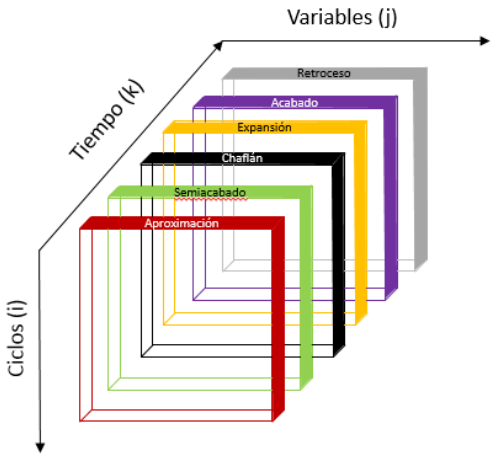

Figura 3.1 Estructura base de datos separada por fases

En la Figura 3.1 muestra un esquema de la estructura de la base de datos. El cubo se puede dividir en diferentes fases y se ha decididao trabajar con la fase del semiacabado representado en la Figura 3.1 y en la Figura 1.5 de color verde.

El número de ciclos o filas de la base de datos es **i=500.** Cada fila corresponde con un ciclo de mecanizado.

En las columnas tenemos las variables velocidad, aceleración, torque y posición para cada uno de los 4 ejes (ver Figura 1.6). Sin embargo, los datos del eje S1 están mal guardados en la base de datos y no se han podido incluir en este estudio. De esta forma tendremos **j**=4*3=**12 variables**: Velocidad_Z, Aceleración_Z, Posición_Z, Torque_Z, Aceleracion_S2, Velocidad_S2, Torque_S2, Posicion_S2, Aceleracion_S3, Velocidad_S3, Torque_S3, Posicion_S3. Nótese que la variable torque es la palabra inglesa utilizada para decir par motor y en este trabajo se emplean ambos términos de forma sinónima.

En número de instantes de tiempo de cada ciclo para la fase de semiacabado es **k=2050**.

Como la estructura de la base de datos es por lotes debemos desdoblar el cubo para poder analizarla. En la Figura 3.2 se muestra un esquema del despliegue del cubo tipo Batch Wise.

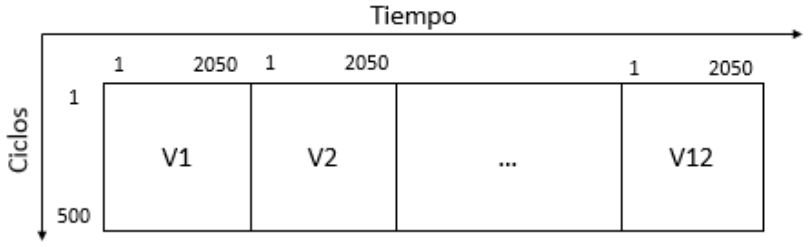

Figura 3.2 Desdoble de lotes (Batch Wise)

Los datos están almacenados en una base de datos NoSQL (Cassandra). Tanto la extracción de los datos de Cassandra (Apache Software Foundation , 2010) como para la separación del *dataset* en las distintas fases se ha empleado Python (van Rossum, 1995).

CAPÍTULO 4

4. Resultados

Para el análisis de los datos se han empleado diversos softwares: R (R Core Team, 2017) , Python (van Rossum, 1995) y Aspen ProMV.

En primer lugar, hemos realizado un PCA introduciendo los datos según la estructura de la Figura 3.2. A continuación, en la Figura 4.1 se ha representado el *score plot* de la primera frente a la segunda componente principal y podemos ver que que hay 9 puntos agrupados en 3 grupos (amarillo, verde y azul) que se encuentran alejados del resto de individuos. En los gráficos T^2 de hotelling y SPE (Figura 4.2 y Figura 4.3 respectivamente) vemos como estos 9 puntos sobresalen del límite de control lo que nos hace pensar que puede tratarse de *outliers* o datos anómalos. A continuación, vamos a analizar estos 9 puntos.

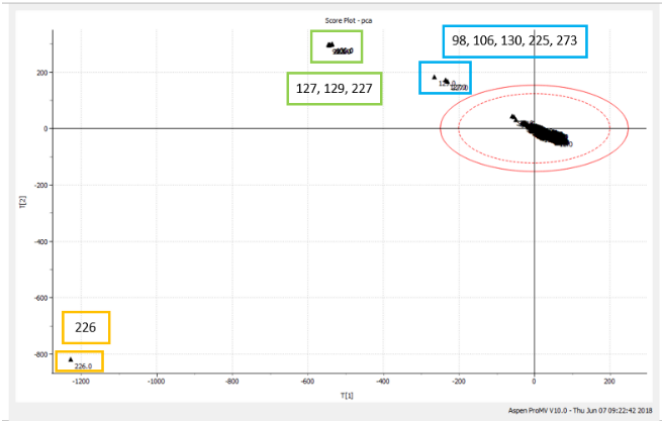

Figura 4.1 *Score plot*

4.1 Análisis de los *outliers*

Los datos anómalos los podemos clasificar en el 3 grupos. En el grupo amarillo se encuentra solamente el ciclo 226, en el grupo verde se encuentra los ciclos 127, 129 y 227 y en el grupo azul se encuentra 98, 106, 130, 225 y 273. Estos 3 grupos de colores se ha realizado para una mejor interpretación de los datos anómalos debido a que los individuos de cada grupo presentarán comportamientos muy parecidos.

4.1.1 Grupo amarillo, punto 226

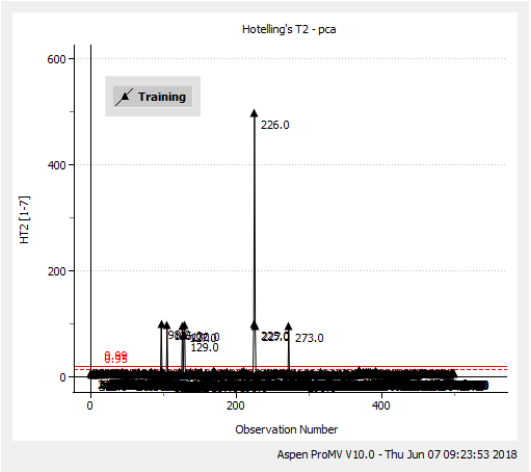

Figura 4.2 Gráfico T^2 Hotelling

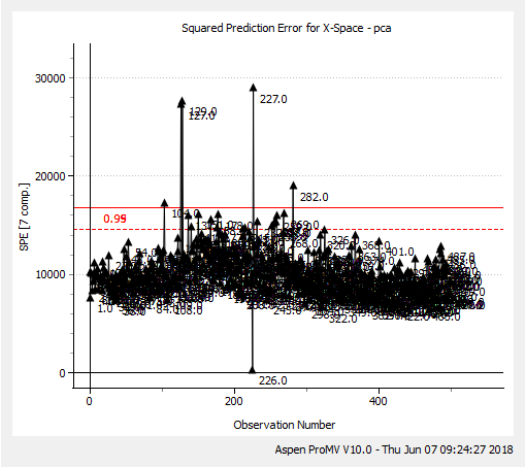

Figura 4.3 Gráfico SPE

El punto 226 es el de posición más elevada en el gráfico de la T^2 de Hotelling (ver Figura 4.2). Lo que quiere decir que es un *outlier* severo y podría forzar la formación de una componente principal. Para conocer que variables han sido las causantes de esta salida de control observamos el gráfico de

contribuciones. En la Figura 4.4 vemos contribuciones muy negativas en la variable posición del eje Z y en la velocidad del eje Z.

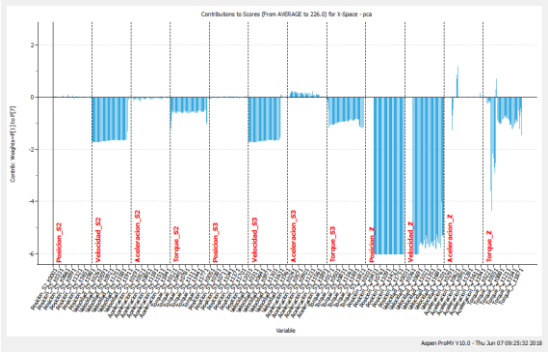

Figura 4.4 Contribution plot, ciclo 226

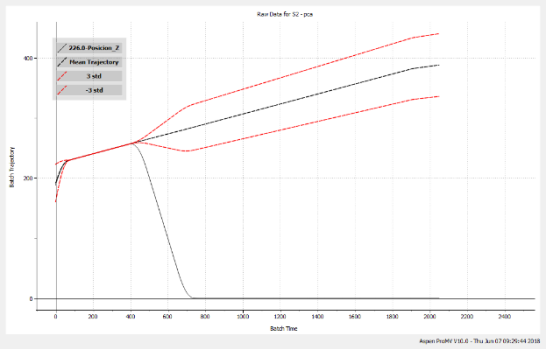

Figura 4.5 Posición Z, ciclo 226

La Figura 4.5 es la gráfica de la posición del eje Z del ciclo 226 y el eje de ordenadas indica los milímetros que desciende la herramienta.En la Figura 4.5 se observa como la herramienta en los primeros instantes del mecanizado ha regresado a la posición inicial sin apenas mecanizar el cilindro. En lugar de descender para mecanizar el cilindro asciende hasta la posición inicial.

Hay ocasiones en las que los operarios realizan pruebas en la OP110 sin piezas en su interior. Debido a que la trajectoria del eje de avance no corresponde con la trayectoria autoprogramada de la máquina nos hace pensar que el ciclo 226 lo podríamos clasificar como un ciclo dónde se ha realizado alguna prueba.

Los ciclos anómalos de los grupos verde y azul también se han clasifcado como ciclos en vacío. En anexo 1 y 2 se detalla el comportamiento de estos ciclos.

4.2 PCA bajo *Normal Operation Conditions*

A continuación se ha realizado un PCA NOC (*Normal Operation Conditions*) es decir, sin los *outliers* descritos en el apartado anterior.

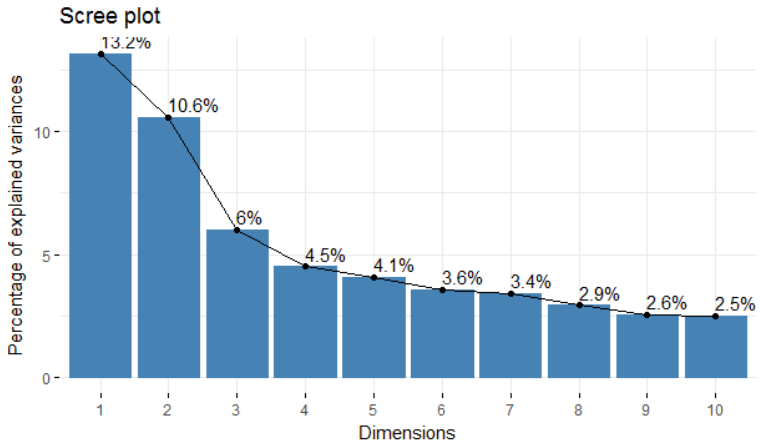

Figura 4.6 Scree plot

4.2.1 Primera componente principal

La primera componente principal explica el 13.2% de la variabilidad. En la Figura 4.7 vemos que las variables que más explican la primera componente son el torque de los 3 ejes. El par motor del eje S3 es inversamente proporcional a la primera componente mientras que el par del eje Z y S2 están directamente relacionados.

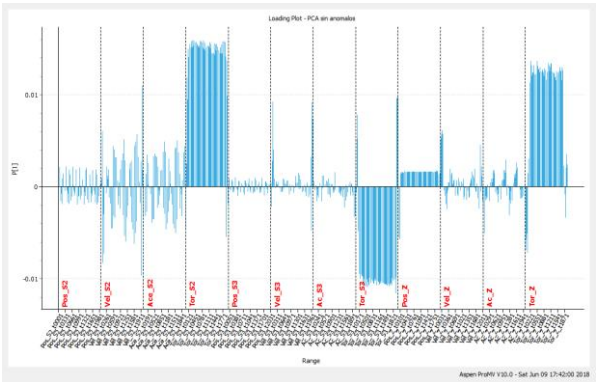

Figura 4.7 P1 Bar, 1º componente principal

4.2.2 Segunda componente principal

La segunda componente principal explica el 10.6% de la variabilidad. En el gráfico de contribuciones, Figura 4.8, observamos que la segunda componente está muy explicada por la variable posición del eje Z. Por tanto, existe una correlación posisitiva entre posición del eje Z y la segunda componente principal.

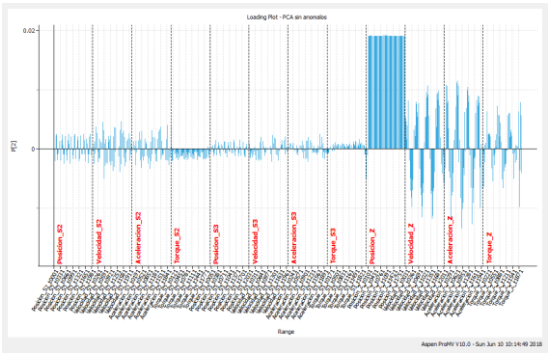

Figura 4.8 P2 Bar, 2ª componente principal

4.2.3 Tercera componente principal

La tercera componente principal explica el 6% de la variabilidad explicada. La variables más correlacionadas con esta componente son las variables del eje S2 y el torque S3. El torque del eje S2 y S3 están correlacionadas positivamente con la tercera componente.

En el gráfico de contribuciones se muestran las 12 variables en sus 2050 instantes de tiempo.

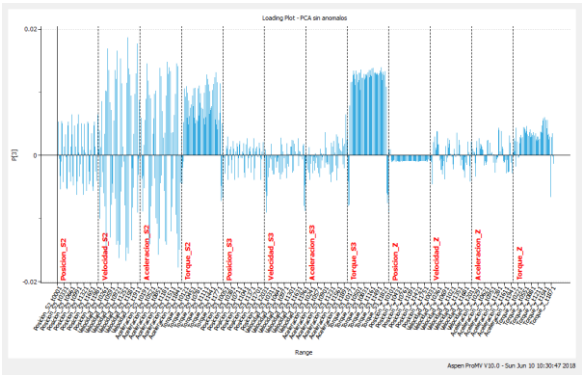

Figura 4.9 P3 Bar, 3º componente principal

Uno de los aspectos mas llamativos de la Figura 4.9 es que las variables velocidad y aceleración del eje S2 también están muy correlacionadas con la tercera componente principal. Sin embargo no podemos definir esta correlación como una correlación positiva o negativa debido a que presenta correlaciones negativas y positivas en instantes de tiempo consecutivos. Por tanto, la dinámica de estas dos variables se representa gráficamente de forma sinusoidal. El fenómeno de rozamiento podría explicar la variación del signo de las contribuciones.

Cuando dos superficies se ponen en contacto, el movimiento de una respecto a la otra genera fuerzas tangenciales llamadas fuerzas de fricción, las

cuales tienen sentido contrario al movimiento. La máquina está configurada para que durante el semiacabado se mecanice a una velocidad constante. Esto es prácticamente imposible debido a que cuando el husillo entra dentro del cilindro se encuentra con una fueza de rozamiento la cuál provoca un descenso de la velocidad. Entonces, la OP110 aumenta la aceleración para alcanzar el *setpoint* establecido. El rozamiento provoca la inestabilidad de la aceleración y de la velocidad. Por tanto, los individuos que se clasifiquen cerca del la tercera componente principal serán individuos con alto torque en el eje S2 y S3 y además con velocidades y aceleraciones más inestables, es decir, con alta varianza a lo largo del mecanizado en el eje S2.

4.2.4 Descripción de los individuos.

Para interpretar los resultados es necesario saber que en el ciclo 100 ha habido un cambio de herramienta en el husillo H7 (cuyo eje de giro es el S3) y H1(cuyo eje de giro es el S1, excluido del análisis) además, en el ciclo 400 también se ha cambiado los husillos H5 y H3 (correspondiente a cada uno de los dos husillos del eje S2).

Se ha dividido los ciclos de mecanizado en 5 grupos. El grupo 1 son los ciclos comprendidos del 1 al 100, son los ciclos que se han mecanizado con una herramienta nueva en el eje S3 y S1. Los ciclos del grupo 1 se han representado en color negro. En el grupo 2 se han clasificado los ciclos del 101 al 200, son los ciclos anteriores al cambio de herramienta del eje S1 y S3 y por tanto la herramienta de sus correspondientes husillos, H1 y H7, estará muy desgastada. Los ciclos del grupo 2 se representan en el gráfico con el color naranja. Los ciclos del 201 al 300 se clasifican en el grupo 3(pintados de color azul en el gráfico) y se piensa que serán ciclos con desgaste medio tanto en el eje S2 como en el eje S3. En el grupo verde se clasifican los ciclos del 301 al 400, son ciclos con las herramientas del eje S2 (husillo H3 y husillo H5) nuevas, por tanto, son ciclos con poco desgaste. En el grupo 5, representado de color morado, tenemos los ciclos del 401 al 500, estos ciclos son los que se han mecanizado antes del cambio de las herramientas del eje S2 por ello mecanizarán con mucho desgaste ese eje. En la Figura 4.10 se muestra un esquema aclaratorio de la división de los ciclos descrita anteriormente.

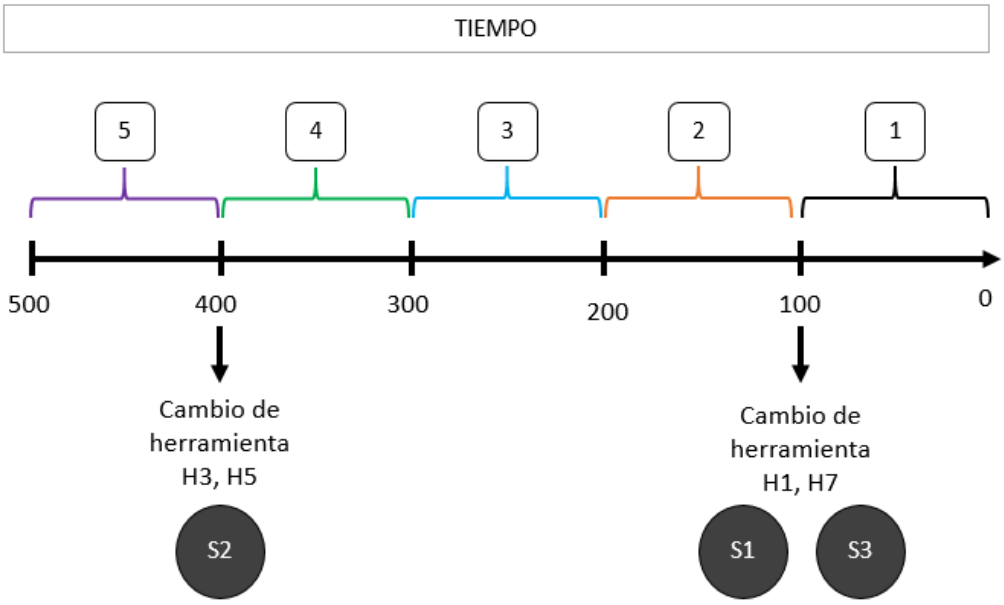

Figura 4.10 Esquema gráfico división de los ciclos

4.2.5 Gráfico *scores*, primera CP frente a la segunda CP.

Al realizar el gráfico de *scores* o *score plot* hemos pintado los ciclos según la clasificación descrita anteriormente.

En la Figura 4.11 se obserava como la primera componente principal separa los ciclos en dos grandes grupos. En la derecha del *score plot* se agrupan los ciclos de los grupos 1 y 5 y en la izquierda se encuentran los ciclos de los grupos 2, 3 y 4, por tanto, la primera componente separa los grupos donde el desgaste la herramienta S3 es menor (grupo de la derecha) de los que el desgaste es mayor (grupo de la izquierda). Además también podríamos decir que el grupo de la derecha presenta mayor torque en el eje S2 que en el grupo de la izquierda. Recordemos que la primera componente principal explica el par motor o torque de los dos ejes de giro(eje S2 y eje S3) y del eje de avance (eje Z). Es lógico que la variable más correlacionada con el desgaste de la herramienta sea el par motor. El par es la energía que consume el motor para realizar el semiacabado de los cilindros. Un ciclo con un par motor muy muy alto indica que el motor ha necesitado mucho esfuerzo para mecanizar. Este sobreesfuerzo es debido a que la herramienta con la que se está mecanizando está muy desgastada. Como la herramienta está desgastada el motor debe hacer más esfuerzo para mecanizar el cilindro y realizar un diámetro adecuado.

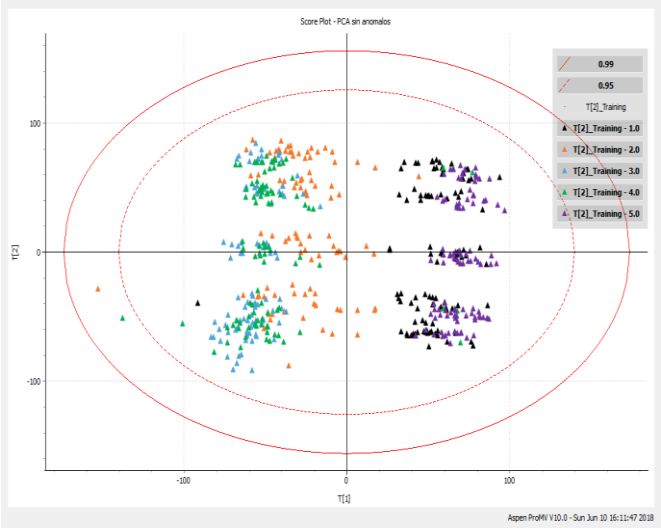

Figura 4.11 *Score plot*, 1CP y 2CP

Por todo esto, podríamos decir el desgaste de la herramienta está muy correlacionado con la primera componente principal.

En cuanto a la segunda componente principal cabe recordar que la variable que más explica es la posición en el eje Z (el eje de avance). En la Figura 4.12 el eje de ordenadas correspone con la segunda componente, por tanto, cuanto más alto se situe un ciclo, mayor posición tendrá y viceversa. Si nos fijamos en la Figura 4.12 vemos como los ciclos se nos clasifican en 3 grupos. El grupo que se encuentra más arriba será el que mayor posición Z tenga. En el tercer cuadrante se situan los ciclos con menor posición en el eje Z. En la Figura 4.13 hemos gráficado la posición del eje Z en estos ciclos y hemos visto diferencias con la media en los primeros instantes. Estos ciclos son ciclos en los que la herramienta ha estado parada unos instantes antes de entrar. Existen varias hipótesis acerca de esta parada, como por ejemplo, puede que la herramienta esté esperando taladrina para empezar el mecanizado.

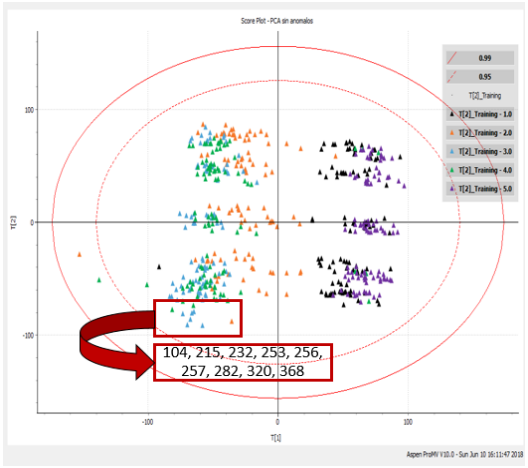

Figura 4.12 *Score Plot* 2, 1CP y 2CP

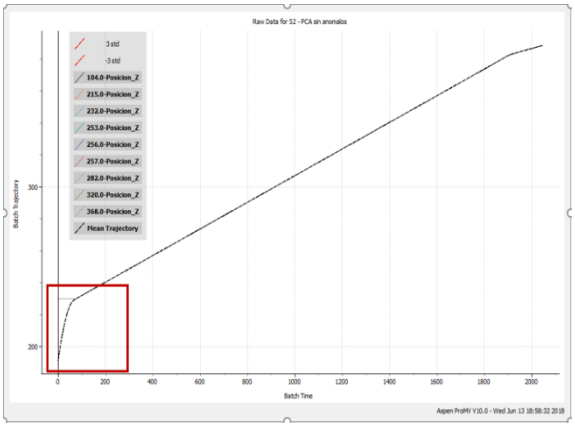

Figura 4.13 Posicion Z (varios ciclos)

4.2.6 Gráfico *scores*, primera CP frente a la tercera CP

En la Figura 4.14 vemos como los grupos 1(de color negro) y 5(de color morado) se diferencian según la tercera componente principal, representada en el eje de ordenadas. Como hemos explicado anteriormente la tercera componente principal explica una alta inestabilidad de la aceleración y velocidad del eje S2. Esto podría explicarse de la siguiente forma: en el periodo de tiempo en el que se ha mecanizado los ciclos del 400 al 100 se han registrado problemas de vibraciones en el eje S2. Los operarios para solventar este problema de vibraciones han cambiado el husillo H3 (perteneciente al eje S2) en 3 instantes de tiempo distinto. Todo esto provoca que el husillo H5 mecanice con una herramienta más desgastada que el husillo H3 (ambos pertenecientes al eje S2). Si uno de los husillos esta mucho más desgastado que el otro esto provocaría dificultades en la máquina para mantener la velocidad específica de mecanizado.

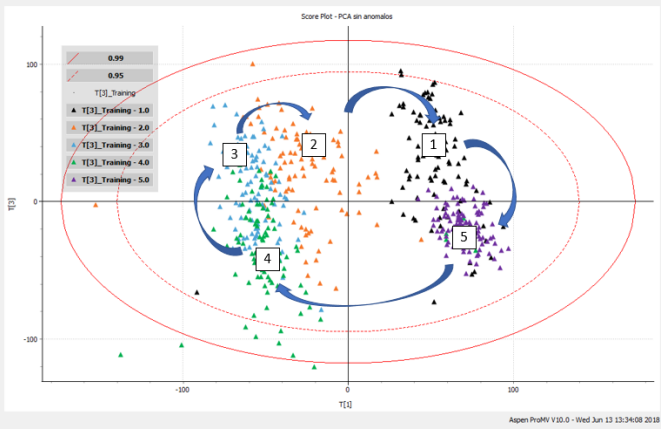

Figura 4.14 *Score plot* 2, 1CP y 3CP

Los ciclos con mayor inestabilidad en el eje S2 serán aquellos que se sitúen más arriba en el gráfico. La mayoría de estos ciclos pertenecen al grupo 1, 2 y 3. Por tanto, en estos grupos la diferencia de desgaste de las herramientas del eje S2 (husillo H3 y H5) serán altas.

En resumen, la primera componente principal explica el desgaste de los husillos S2 y S3. La segunda componente principal explica la variabilidad de la trayectoria del eje Z o eje de avance. La tercera componente principal explica la diferencia entre el desgaste de los dos husillos del eje S2, H3 y H5.

4.3 *Clusters*

En primer lugar hemos realizado un nuevo PCA. Este PCA se ha realizado sin los datos anómalos analizados en el apartado 4.1. Como el objetivo del PCA es reducir la dimensionalidad se ha establecido que se deben escoger tantas componentes principales como expliquen el 80% de la varianza explicada. Es por ello que se va a trabajar con 41 componentes principales.La Figura 4.15 muestra en azul que 41 componentes principales explican el 80.20% de la varianza explicada.

comp 32	121.757841	0.494950572	76.77738
comp 33	109.475783	0.445023510	77.22240
comp 34	108.253589	0.440055240	77.66246
comp 35	103.697158	0.421533161	78.08399
comp 36	98.262729	0.399441987	78.48343
comp 37	90.878749	0.369425810	78.85286
comp 38	87.725860	0.356609186	79.20947
comp 39	84.563390	0.343753617	79.55322
comp 40	82.104893	0.333759726	79.88698
comp 41	78.198151	0.317878663	80.20486
comp 42	75.391765	0.306470588	80.51133
comp 43	70.267807	0.285641494	80.79697
comp 44	68.417737	0.278120883	81.07509

Figura 4.15 Varianza explicada

A continuación realizaremos con estas 41 componentes principales cuatro tipos de *clusters*: *fuzzy clustering*, *kmeans*, jerárquico Complete y jeráquico Ward.

4.3.1 *Fuzzy clustering*

El *fuzzy clustering* trabaja con una partición suave del conjunto de datos, en tal partición todos los individuos pertecenen algún grado a todos los *cluster*. Con el fin de validar con que partición quedarnos, se analiza una serie de indices. El índice de *xie-beni* (xb), *partition entropy*(pe) y *fukuyama sugeno* (fs) muestran que la agrupación es mejor cuanto más bajos sean sus valores. En cambio, cuanto mayor sea el índice *partition coeficient* (pc) mejor será la partición. Se han calculado estos índices para el*fuzzy clustering* de 2 grupos, 3 grupos, 4 grupos y 5 grupos.

	pe	xb	fs	pc
2 GRUPOS	6,93E-01	9,48E+03	9,87E+03	5,00E-01
3 GRUPOS	1,10E+00	2,17E+04	6,58E+03	3,33E-01
4 GRUPOS	1,39E+00	3,79E+05	4,93E+03	2,50E-01
5 GRUPOS	1,61E+00	2,83E+05	3,95E+03	2,00E-01

Tabla 4.1 Índices pe, xb, fs, pc

En la Tabla 4.1 se muestra como el número de grupos óptimos para el *fuzzy clustering* es 2. De esta forma los individuos quedan clasificados tal y como se muestra en la Figura 4.16: a la izquierda y de color azul oscuro se clasifican los ciclos del *cluster* 1 y a la derecha y de color azul claro los ciclos del *cluster* 2.

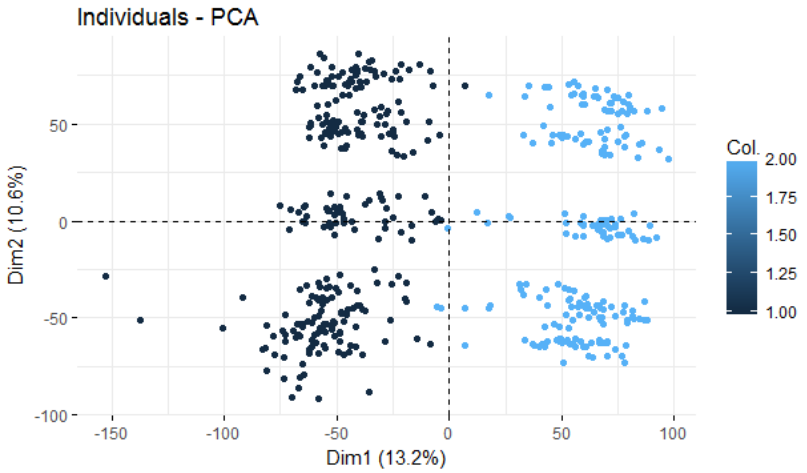

Figura 4.16 *Score plot* y *fuzzy clustering*

4.3.2 *Cluster* jerárquico Complete

Se va a realizar un *cluster* jerárquico con el método Complete. Nos vamos a ayudar del dendograma que se muestra en la Figura 4.17 para decirdir el número de *clusters* óptimo.

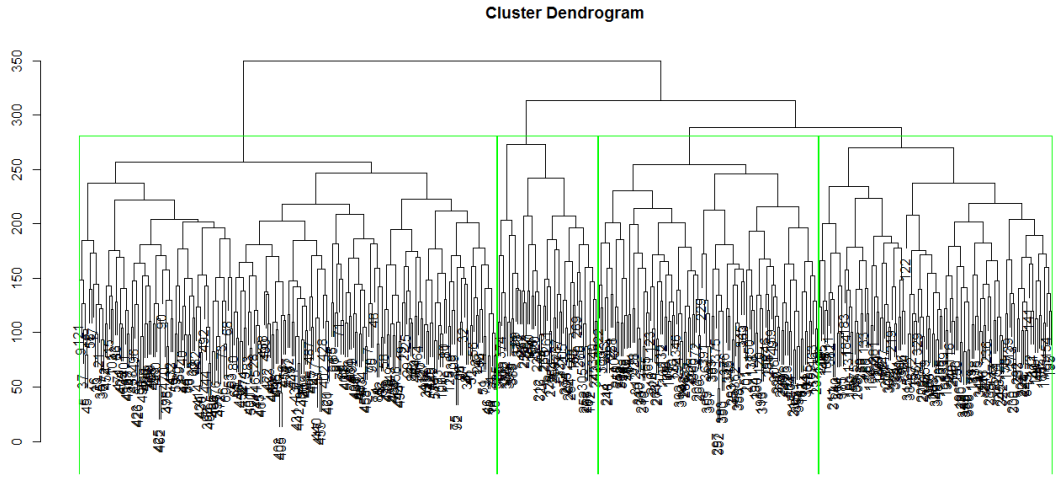

Figura 4.17 Dendograma método Complete

En base a este algoritmo el número óptimo es 4 por lo que se van a realizar 4 *clusters*. El número de individuos por *cluster* se detalla en la Tabla 4.2.

	Cluster 1	***Cluster 2***	***Cluster 3***	***Cluster 4***
Num. individuos	211	111	51	118

Tabla 4.2 Número de individuos por cluster

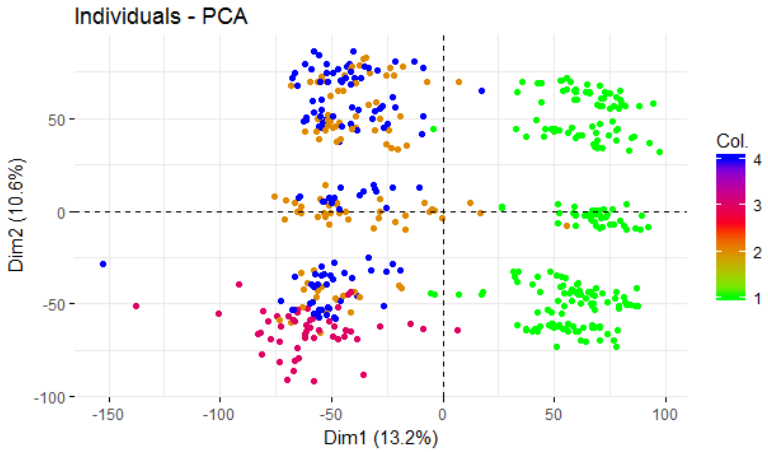

Figura 4.18 *Score plot* y *cluster* Complete

En la Figura 4.18 se muestra en el *score plot* la clasificación de los ciclos en 4 grupos según el método Complete. Los ciclos en color verde pertenecen al *cluster* 1, en color naranja al *cluster* 2, en color fucsia pertenecen al *cluster* 3 y en color azul pertenecen al *cluster* 4.

4.3.3 *Cluster* jerárquico método de Ward

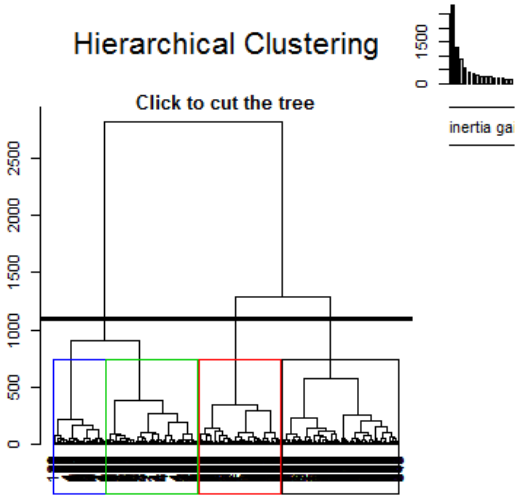

Figura 4.19 Dendograma cluster Ward

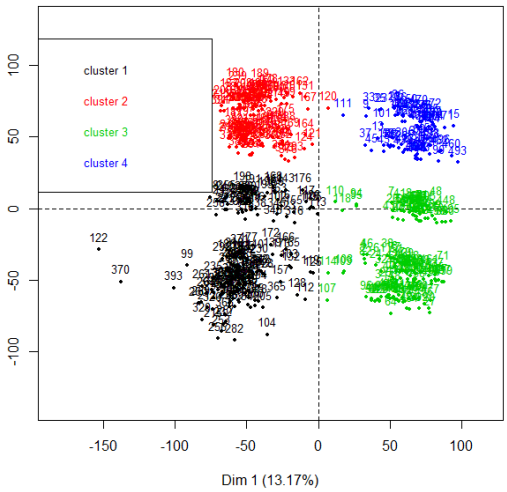

Figura 4.20 *Score plot* y *cluster* Ward

Con ayuda del dendograma, ver Figura 4.19, se ha escogido realizar una partición de 4 grupos.

En la Figura 4.20 se muestra en el *score plot* la clasificación de los ciclos en 4 grupos según el método Ward. Los ciclos en color negro pertenecen al *cluster* 1, en color rojo al *cluster* 2, en color verde pertenecen al *cluster* 3 y en color azul pertenecen al *cluster* 4. Nótese que la primera y la segunda componente principal separan los diferentes *clusters*.

4.3.4 *Cluster Kmeans*

A continuación, vamos a realizar un *cluster Kmeans* con las 41 componentes principales mediante Python. El número de *clusters* elegido es 4 tal y como en el *cluster* jerárquico.

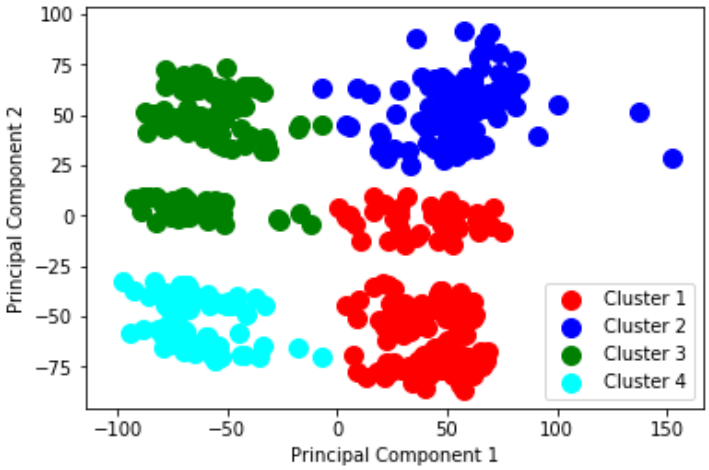

Figura 4.21 *Score plot* y *cluster Kmeans*

Nótese que el gráfico de la Figura 4.21 el *score plot* que nos muestra Python es el mismo que R (R Core Team, 2017) pero girado 180 grados.

En resumen, en las 4 metodologías *clusters* que hemos empleado se ha dividido los individuos en 2 grandes grupos, uno dónde se ha mecanizado con bajo torque en el eje S3 y por tanto con poco desgaste en la herramienta H7(grupo de la derecha) y el otro donde se ha mecanizado con bajo torque en el eje S2 y por tanto con poco desgaste en las herramientas H3 y H5 (grupo de la izquierda). También se observa como el método *Kmeans* nos muestra agrupaciones muy parecidas a las agrupaciones del método Ward. En ambas clasificaciones se han establecido 4 grupos bien separados por la primera y segunda componente principal, ver Figura 4.20 y Figura 4.21. Recordemos que la segunda componente principal estaba correlacionada con la variable posición del eje Z o eje de avance, es decir, los ciclos que se encuentren en la parte superior del *score plot* tendrán una posición mayor y por tanto descenderán más. La diferencia entre los ciclos con una posición alta y los ciclos con una posición baja es de 40 micras. Esta baja diferencia se piensa que es debida a errores de precisión de los sensores en la captura de los datos.

En la clasificación del método Complete, Figura 4.18, vemos en el grupo de la izquierda se subdivide en tres grupos y sería interesante estudiar el porqué de esta clasificación debido a que el grupo de la izquierda ha presentado en tres ocasiones problemas de vibraciones. Es por ello que se ha decidido trabajar con la clasificación del *cluster* Complete para la realización del *Randon Forest* que se detalla en el siguiente apartado.

4.4 *Random Forest*

Para realizar una buena caracterización de los *clusters* es necesario saber qué variables son las más influyentes en la separación de los individuos en estos grupos. Por ello, se ha decidido realizar un *Random Forest* tomando como variable dependiente el grupo de pertenencia de cada individuo. En este caso se ha elegido trabajar con la clasificación del *cluster* jerárquico Complete por ser la más coherente.

Se ha realizado una reducción de la dimensionalidad resumiendo cada una de las 12 variables en 4 estadísticos: media, desviación típica, máximo y mínimo. De esta forma se tiene un total de 48 variables. El motivo de esta reducción ha sido para estudiar la influencia la tendencia y la variabilidad de las variables originales en la formación de los grupos.

Para conocer cuál de estas 48 variables son las más importantes en la separación de los distintos *clusters* vamos a emplear un gráfico donde se representan el *Mean Decrease Gini* de cada una de las variables, ver Figura 4.22. El índice de Gini es una media de cuanto contribuye a la clasificación de los ciclos de mecanizado cada variable.

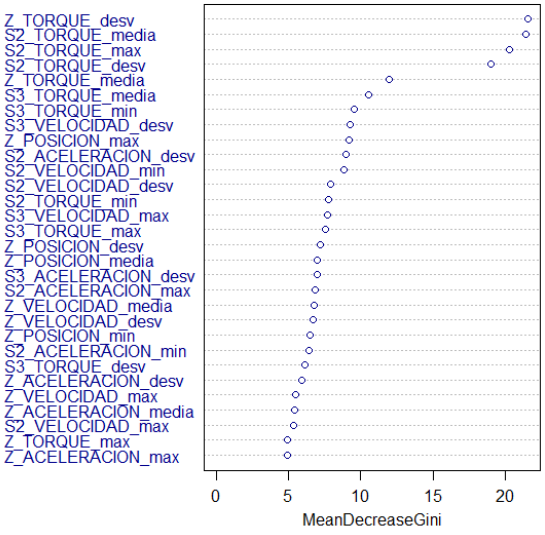

Figura 4.22 Índice de Gini

En la Figura 4.22 muestra que la desviación típica del torque del eje de avance (eje Z) es la variable más significativa seguida de la media, el máximo y la desviación típica de el torque en el eje S2. Finalmente, la media del torque de Z y la media y el mínimo torque del eje S3 también son variables importantes.

A grandes rasgos podríamos resumir que la variable torque en los distintos ejes resulta la variable más importante. En el análisis de las componentes principales habíamos visto como la primera componente principal (componente que explica la máxima inercia) las variables que más se explicaban era el torque de los diferentes ejes. Por tanto, los resultados obtenidos en el PCA concuerdan con los resultados obtenidos en el *Random Forest*.

4.5 Árbol de clasificación

Con motivo de caracterizar los individuos de los diferentes *clusters* vamos a realizar un árbol de clasificación con las 11 variables más importantes obtenidas en el *Random Forest*. La variable a predecir es el *cluster* de pertenencia de cada individuo. A continuación, se muestra el árbol obtenido:

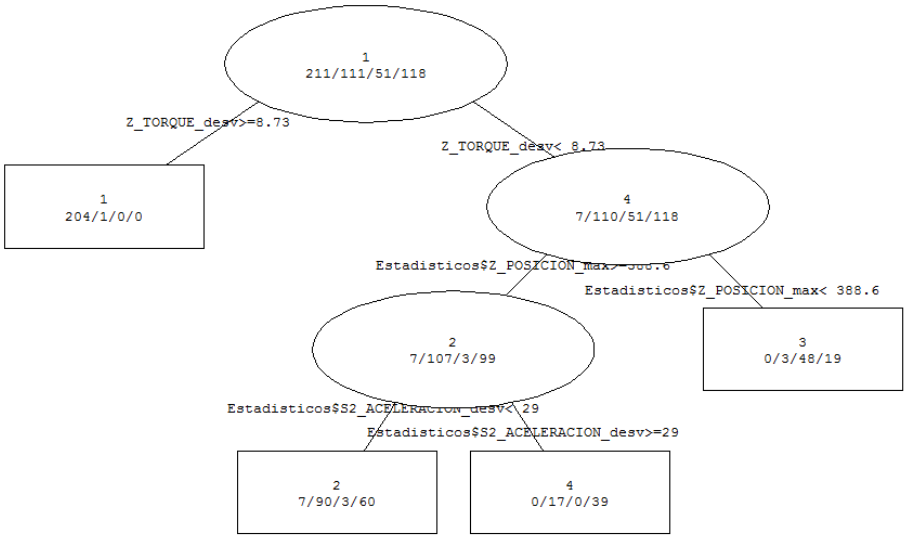

Figura 4.23 Árbol de clasificación

En la Figura 4.23 se observa como la variable desviación típica del torque Z separa los ciclos del *cluster* 1 (grupo de la derecha) del resto de ciclos. La desviación típica del torque del eje Z es mayor en los miembros del cluster 1.

El *cluster* 3, representado en la Figura 4.18 de color fucsia, se separa del cluster 2 y 4 por la variable posición máxima del eje Z. La posición máxima del eje Z de los individuos del grupo 3 es menor que en el *cluster* 2 y 4. Además este grupo de ciclos ya se había analizado y descrito en el apartado 4.2.5 y concuerda con los resultados del árbol de clasificación.

Finalmente, el grupo 2 y 4 se separan debido a la variable desviación típica de la aceleración del eje S2. Los ciclos pertenecientes al *cluster* 4 tienen en general una mayor desviación típica de la aceleración del eje S2 que los ciclos del *cluster* 2. Una alta desviación típica en el eje S2 indica que a lo largo del mecanizado la aceleración no se ha podido mantener correctamente en el *set point* establecido. Esto podría deberse a que uno de los dos husillos del eje S2 esté más desgastado que el otro. Esta hipótesis cobra sentido debido a que en el *dataset* analizado, el husillo H3, ha sido reemplazado en tres ocasiones mientras que el husillo H5 solamente en una, lo que quiere decir que ha habido instantes de tiempo en los que se ha mecanizado con diferentes estados de desgaste de los husillos del eje S2. Resumiendo la información anterior podriamos decir que los ciclos del *cluster* 4 se han mecanizado con diferentes estados de desgaste de las herramientas del eje S2.

4.6 Modelo Lee-Carter

Del análisis anterior sabemos que el desgaste de la herramienta está muy correlacionado con las variables de torque de cada uno de los ejes analizados. Para profundizar el análisis de estas variables se han realizado 3 modelos Lee-Carter, uno para cada una de las distintas variables de torque mediante la librería de R gnm (Turner & Firth, 2018). Recordemos que tenemos 3 variables que hacen referencia al torque, una para cada uno de los ejes que se están analizando. Para la realización del modelo Lee-Carter se han eliminado los 9 datos anómalos analizados en el apartado 4.1.

Parámetro a_c: A continuación, en la Figura 4.24 se muestran los valores de a_c propios de la descomposición del torque del eje Z, del torque del eje S2 y del torque del eje S3.

A rasgos generales el torque o energía necesaria para mecanizar el cilindro el el eje Z es menor que en los demás ejes. Además esta energía se mantiente estable a lo largo del tiempo. En la Figura 4.24 observamos que el torque más alto lo presenta el eje S3. Esto quiero decir que el motor encargado del giro del eje S3 es el que más energía necesita para mecanizar, lo que tiene sentido debido a que el motor de eje S3 presenta una potencia inferior al resto de motores. El motor del eje S2 es distinto, como necesitará más fuerza para mecanizar 2 cilindros tiene un motor con potencia mayor al S3.

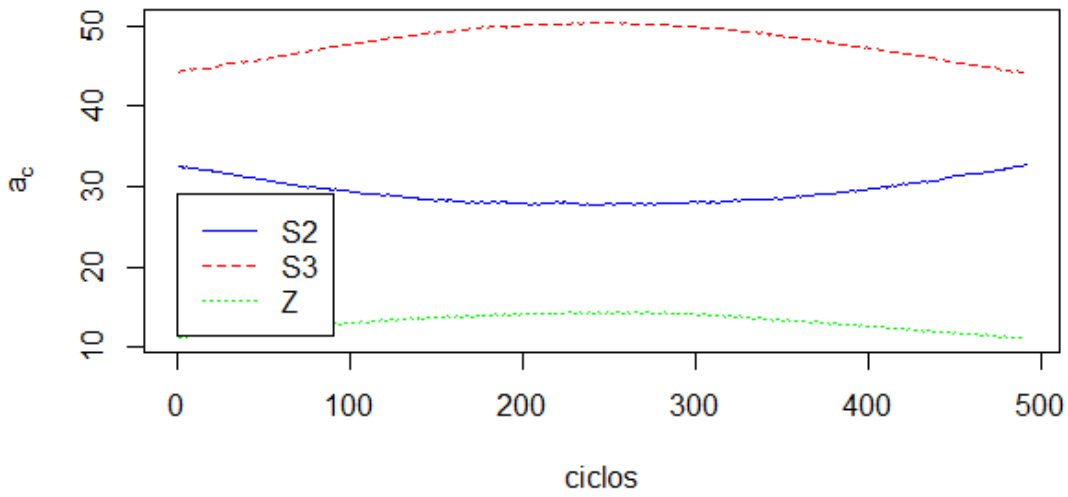

Figura 4.24 Parámetro a_c

Parámetro b_c: A continuación, en la Figura 4.25 se muestran los valores de b_c propios de la descomposición del torque del eje Z, del torque del eje S2 y del torque del eje S3.

Recordemos que el parámetro b_c indica la aportación de cada ciclo a la evolución general a lo largo del tiempo. En la Figura 4.25 como la curva del torque del eje Z y del eje S3 presenta comportamiento similar mientras que la curva del torque del eje S2 tiene una pendiente inferior. Esto es debido a que en el eje S2 se han producido 3 cambios de herramienta en el husillo H3 lo que provoca que el torque se mantenga más constante. Si mecanizamos con una herramienta nueva el motor necesitaría menos energía para eliminar el material del cilindro.

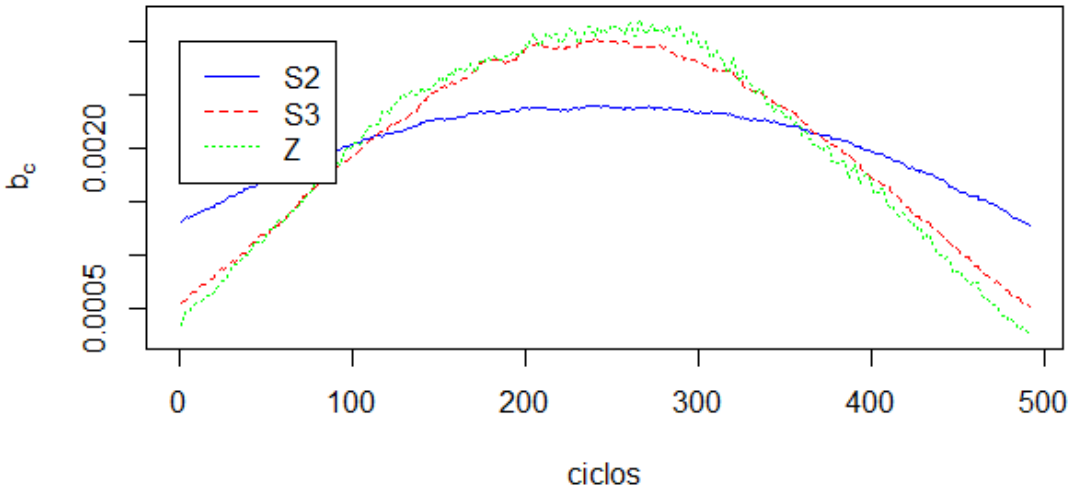

Figura 4.25 Parámetro b_c

Parámetro k_t: A continuación, en la Figura 4.26 se muestran los valores de k_t propios de la descomposición del torque del eje Z, del torque del eje S2 y del torque del eje S3.

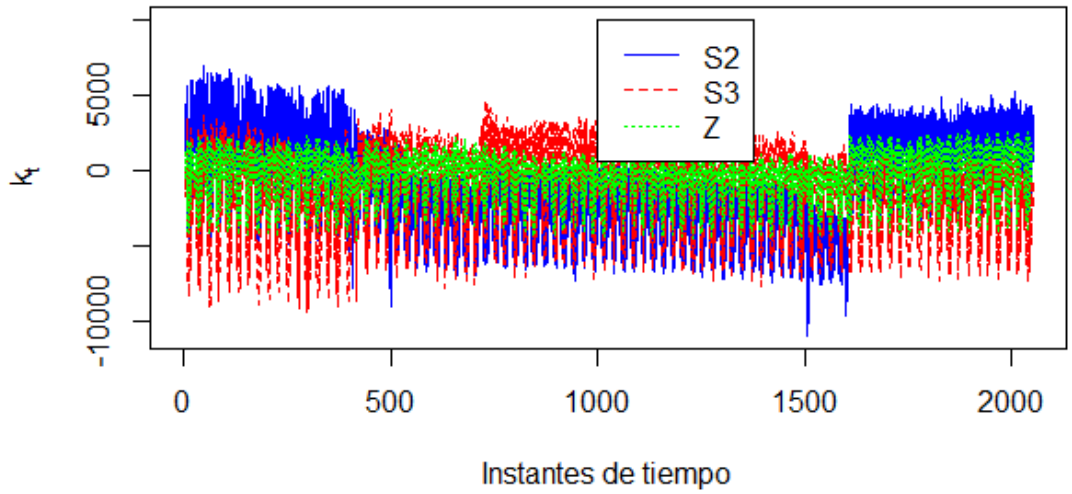

Figura 4.26 Parámetro k_t

De la Figura 4.26 podemos destacar un comportamiento periódico. Cada 40 instantes de tiempo se registra un patrón similar. Esto puede ser debido a la geometría de la pieza, es decir, el radio del cilindro a un punto p a cierta altura x

estará correlacionada con el radio de la circunferencia a cierta altura x+1. Además, en la Figura 4.26 también se observa que en los primeros y en los últimos 400 instantes de tiempo el torque de los ejes de giro S2 y S3 es más alto. Esto quiere decir que se ha necesitado más energía a la hora de mecanizar la parte superior e inferior del cilindro lo que podría indicar que la forma del cilindro era más ancha en los laterales y más estrecha en el centro de la circunferencia.

4.7 Monitorización del proceso

Resulta de interés poder disponer de un sistema de monitorización capaz de detectar salidas de control en el mecanizado de los cilindros. Es por ello que se va a aplicar un control multivariante SPC sobre las variables del proceso.

4.7.1 Gráficos SPE y T_A^2

En primer lugar, hemos realizado los gráficos SPE y T_A^2 con las componentes principales. Para establecer los límites de control debemos escoger ciclos de mecanizado en los que se hayan operado bajo condiciones normales. Para construir el *training set* se han tomado 74 ciclos al azar. En la Figura 4.27 y 4.28 se muestan los gráficos T_A^2 y SPE empleando los 74 ciclos del *trainig set*.

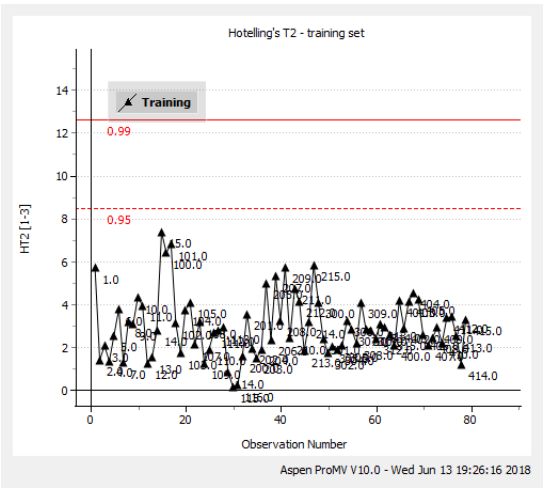

Figura 4.27 T^2 *training set*

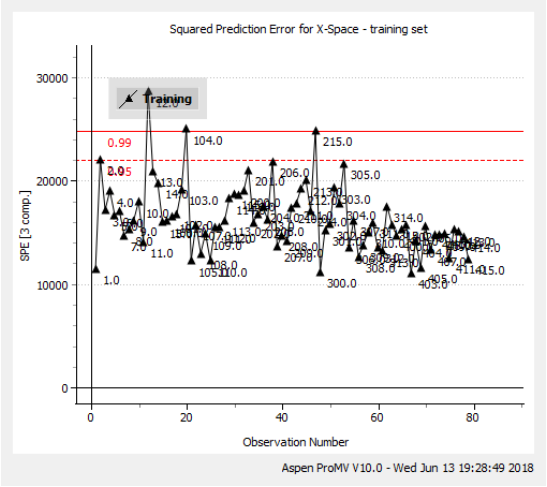

Figura 4.28 SPE *training set*

Una vez establecido estos límites de control, vamos a simular una salida de control. Para ello se ha construido un *set* de predicción con ciclos diferentes al *training set* donde se ha incluido un ciclo anómalo. Las Figuras 4.29 y 4.30 se observan los gráficos T_A^2 de Hotelling y SPE respectivamente con el set de predicción y vemos como el ciclo 227 presenta valor muy alto en el gráfico SPE.

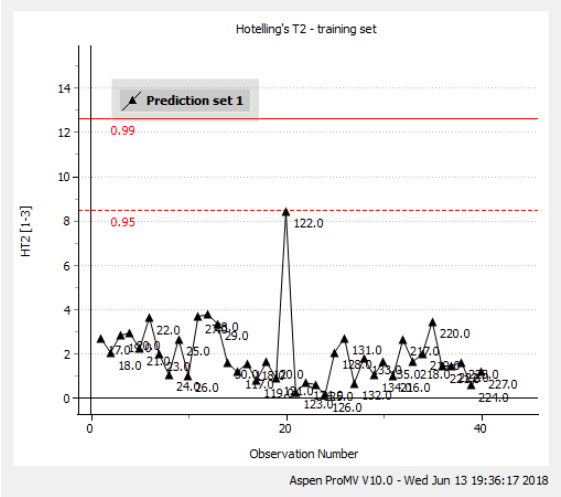

Figura 4.29 T^2 *prediction set*

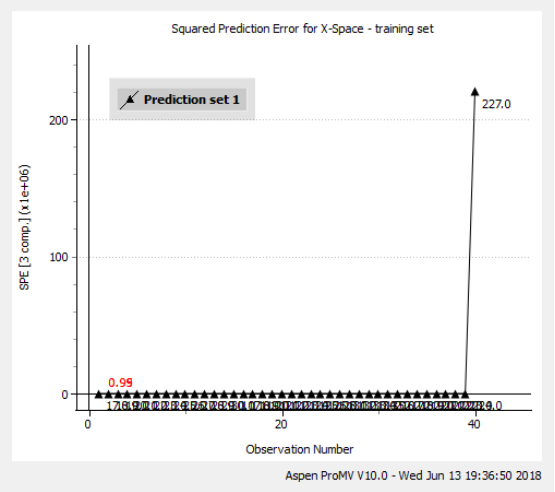

Figura 4.30 SPE *prediction set*

A continuación, vamos a mirar que variables han contribuido más en esta salida de control. Para ello debemos visualizar el gráfico de contribuciones, Figura 4.31.

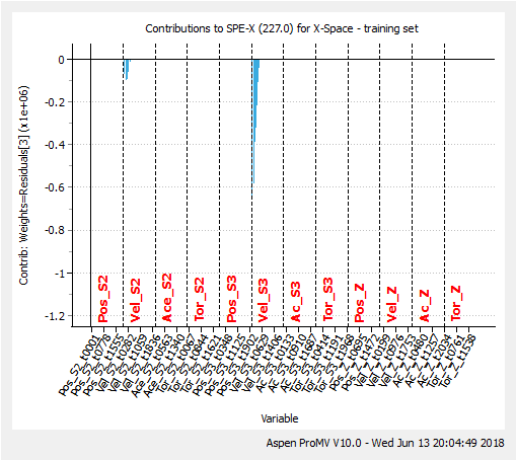

Figura 4.31 Contribution plot, ciclo 227

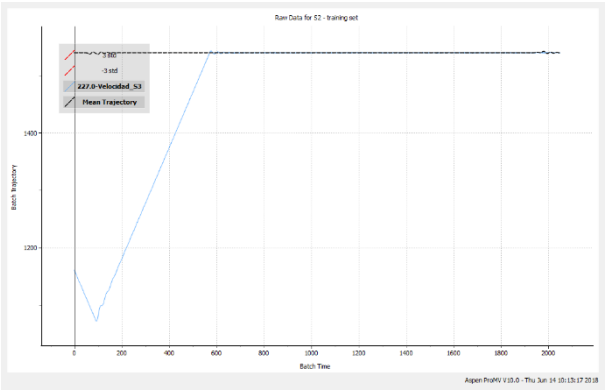

Figura 4.32 Velocidad S3, ciclo 227

En la Figura 4.31 vemos como la velocidad del eje S3 en el ciclo 227 presenta una alta contribución en los primeros instantes de tiempo. En la Figura 4.32 se ha gráficado la variable velocidad del eje S3, la línea en color azul indica la velocidad del ciclo 227 mientras que la línea discontinua negra representa la velocidad media de los ciclos.

Normalmente, los ciclos que presentan velocidades tan bajas en los ejes de giro son propios de ciclos de puesta en marcha de la máquina. A lo largo de la jornada laboral la máquina se para por distintos motivos. Cuando se vuelve a encender se realiza un ciclo en vacío (sin pieza) para ajustar la máquina a las

condiciones operativas programadas. El ciclo 227 es uno de estos ciclos de mecanizado en vacío.

4.6.1 Gráfico MCUSUM y MEWMA

Para la elaboración de los gráficos MCUSUM y MEWMA se ha empleado la librería de R MSQC (Santos-Fern, 2013). Los gráficos MCUSUM y MEWMA son gráficos con memoria que pueden considerarse una extensión de las CUSUM y EWMA respectivamente.

En el apartado anterior el gráfico T_A^2 que se ha realizado ha sido estimando un modelo PCA a partir de todas las variables del proceso, en cambio ahora se van a monitorizar las medias de los 3 torques. Para comparar los gráficos T^2 de hotelling con los gráficos MEWMA y MCUSUM se ha construido un nuevo gráfico T^2 con la librería MSQC con las 3 variables torque.

Se debe dividir el control de calidad en dos etapas. En la primera fase el objetivo es estimar la media de cada variable y la varianza del proceso, y al mismo tiempo, saber si son estables. Para la primera fase se han seleccionado 50 muestras. Las variables que se están monitorizando son las tres variables de torque para cada uno de los tres ejes. Se ha seleccionado estas variables debido a que en los análisis anteriores hemos concluido que el torque es la variable más correlacionada con el desgaste de la herramienta. Se ha decidido trabajar con la media del torque de los 2050 instantes de tiempo debido a que un ciclo de mecanizado dura 23 segundos y por tanto, en el caso de detectar una salida de control se dispone de muy poco tiempo de reacción para cambiar la herramienta de mecanizado.

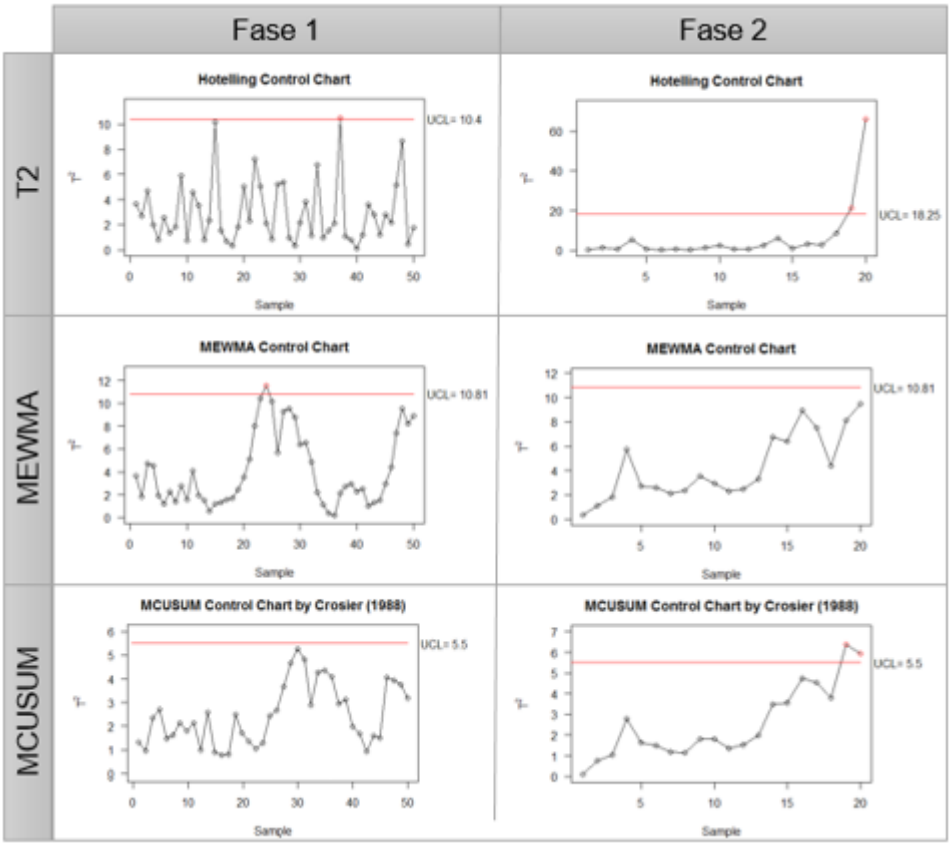

Figura 4.33 Gráficos T^2, MEWMA, MCUSUM

En la Figura 4.33 se observa a la izquierda los gráficos de control en la fase 1 y como todas las observaciones se encuentran por debajo de los límites de control (exceptuando unas pocas) aceptamos que el proceso está bajo control y pasamos a la fase 2.

En la fase 2 empieza el control de calidad y para esta segunda fase se han tomado 20 muestras distintas a las empleadas en la fase 1. Además, para comprobar el funcionamiento de los gráficos de control la observación 20 se ha simulado que ha sido un ciclo con un alto torque en el eje S2. En la Figura 4.33 se observa a la derecha los gráficos de control en la fase 2. Los gráficos T_A^2 de Hotelling y MCUSUM detectan la salidad de control en el ciclo 20 mientras que el gráfico MEWMA no es capaz de detectar la salida de control. A pesar que tanto el gráfico T_A^2 como el gráfico MCUSUM son capaces de detectar cuando se ha mecanizado con alto torque la salida de control es mucho más pronunciada en el gráfico T_A^2 de Hotelling.

Los gráficos MCUSUM y MEWMA presentan una limitación adicional que es que no consideran autocorrelación entre las variables y como hemos visto en el PCA existe una alta correlación entre las variables de torque. Para considerar el efecto de autocorrelación están siendo desarrollados en la actualidad varios métodos para el control estadístico de procesos, principalmente basados en análisis de series temporales.

Para finalizar, solamente remarcar que en el caso de cambiar una de las herramientas de mecanizado por una nueva, es decir, cambiar las condiciones operativas de la máquina, se deben recalcular los límites de control de los gráficos de calidad multivariantes.

CAPÍTULO 5

5. Conclusiones

Debido a la gran cantidad de datos que los sensores son capaces de capturar se empiezan a tener problemas de almacenamiento en la planta de motores Ford. Gracias al estudio de las componentes principales sabemos que no todas las variables son igual de explicativas del proceso de mecanizado. Se deben dejar de capturar las variables aceleración, velocidad y posición y centrarse en una captura estable de la variable torque. En el análisis de las componentes principales también se ha concluido que las 3 variables de torque para cada uno de los 3 ejes están altamente correlacionadas entre sí. Además el PCA es una buena herramienta para reducir la dimensionalidad de los datos. En este trabajo hemos pasado de tener 24600 variables del proceso a 41 componentes principales. Estas 41 variables latentes explican un 80% de la variabilidad total.

En segundo lugar, la obtencion de los *clusters* nos ha permitido saber el estado en el que se encuentra la herramienta. Los ciclos de mecanizado se clasifican en dos grandes grupos. Un primer grupo caracterizado por tener poco desgaste en el eje S3 y el segundo grupo caracterizado por tener poco desgaste en el eje S2. Cada uno de estos dos grandes grupos se dividen en subgrupos de forma diversa dependiendo del método de *clustering* que se aplique.

Mediante el *Random Forest* y el árbol de clasificación hemos concluido que las variables más importantes en la clasificación de los ciclos según su

estado de desgaste es la desviación típica del torque del eje Z. Estas conclusiones concuerdan con las optenidas en el análisis de componentes principales.

Aplicando el modelo Lee-Carter a las 3 variables de torque hemos conocido que la herramienta que más torque presenta es la del eje S3 lo que indica que el husillo H7 será el que mayor desgaste presente en sus plaquitas de mecanizado. Además el modelo Lee-Carter nos proporciona información sobre la geometría de los cilindros del bloque motor.

Actualmente, para evitar tener piezas con diámetros incorrectos, cada 3000 ciclos se para la máquina y se cambia la herramienta. Sin embargo, se están desechando muchas herramientas sin que hayan llegado realmente al final de su vida útil. Este problema se ha solventado con los gráficos T_A^2, SPE y MCUSUM que controlan el proceso de mecanizado y permiten conocer cuando un cilindro se ha mecanizado de forma incorrecta. En el gráfico de contribuciones veríamos que variable está provocando la salida de control. Si la variable torque presenta altas contribuciones en el eje x indicaría un gran desgaste de la herramienta en el husillo correspondiente al eje x. Además, también hemos conocido que el gráfico MEWMA no es capaz de detectar cuando la OP110 deja de trabajar bajo condiciones normales.

Nos gustaria subrayar que este es un trabajo con gran utilidad práctica en el mundo de la empresa por el ahorro que supone tanto en almacenamiento de datos como en ahorro de costes en la sustitución de la herrramienta pues conseguiriamos aproximarnos de forma más prescisa a su vida útil.

Finalmente, para completar este estudio en un futuro sería interesante disponer de los datos del eje S1 para monitorizar y conocer el estado del husillo H1. Asi como proponer un gráfico adapatdo a este proceso y basado en la característica señalada por el random forest, la desviación típica del torque Z.

6. Bibliografía

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics, 2(4)*, 433-459.

Apache Software Foundation . (2010). *Hadoop*. Retrieved from https://hadoop.apache.org

Breiman, L. (2001). Random forests. . *Machine learning, 45(1)*, 5-32.

Crosier, R. B. (1988). Multivariante Generalization of Cumulative Sum Quality Control Schemes. *Technometrics, 30(3)*, 291-303.

Flores, M. (2016). *qcr: Quality Control Review*. Retrieved from R package version 1.0: https://CRAN.R-project.org/package=qcr

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of Statistical Learning. Data Mining, Inference, and Prediction. *Springer, 1*(10).

Jackson, J. E. (2005). A user's guide to principal components. *John Wiley & Sons, 587*.

Lee, R., & Carter, L. (1992). Modelling and forecasting U.S. mortality. *Journal of America Statistical Association, 87(419)*, 659-671.

Liaw, A., & Wiener, M. (2002). Classification an regression by randomforest. *R news, 2(3)*, 18-22.

Lowry, C., Woodall, W., Champ, C., & Rigdon, S. (1992). Multivariate Exponentially Weighted Moving Average Control Chart. *Technometrics, 34(1)*.

MacGregor, J., & Kourti, T. (1995). Statistical Process Control of Multivariante Processes. *Control Engineering Practice, 3*, 403-414.

Meyer, D., Dimitriadou, E., Hornik, K., & Weingessel, A. (2017). *e1071: Misc Functions of the Department of Statistics, Probability.* Retrieved from R package version 1.6-8: https://CRAN.R-project.org/package=e1071

Nomikos, P., & Macgregor, J. (1995). Multivariante SPC charts for monitoring batch processes. *Technometrics, 37(1)*, 41-59.

Pal, N., Bezdek, J., & Hathaway, R. (1996). Sequential competitive learning and the fuzzy c-means. *Neural Networks, 9(5)*, 787-796.

Pignatello, J., & Runger, G. (1990). Comparison of multivariante CUSUM charts. *Journal of Quality Technology, 22(3)*, 173-186.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Retrieved from R Foundation for Statistical Computing: https://www.R-project.org/

Renshaw, A., & Haberman, S. (2003). Lee-Carter mortality forecasting with age specific enhancement. *Insurance: Mathematics and Ecomomics, 33(2)*, 255-272.

Rezaee, M., Lelieveldt, B., & Reiber, J. (1998). A new cluster validity for the fuzzy c-mean. *Pattern recognition letters, 19(3-4)*, 237-246.

Runger, G., & Montgomery, D. (1996). Contributors to a multivariante SPC chart singal. *Communications in Statistics- Theory and Methods, 25(10)*, 2203-2213.

Santos-Fern, E. (2013). *Multivariante Statistical Quality Control Using R.* Retrieved from http://www.springer.com/statistics/computational+statistics/book/978-1-4614-5452-6

Therneau, T., & Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees.* Retrieved from R package version 4.1-13: https://CRAN.R-project.org/package=rpart

Turner , H., & Firth, D. (2018). *Generalized nonlinear models in R: An overview of the gnm package.* Retrieved from R package version 1.1-0: https://cran.r-project.org/package=gnm

van Rossum, R. (1995, May). Python tutorial. *Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI)*. Retrieved from Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI).

Woodall, W., & Ncube, M. (1985). Multivariante CUSUM quality control procedures. *Technometrics, 27(3)*, 285-292.

7. Anexos

Anexo 1. GRUPO 2, Punto 127, 129 y 227

Los ciclos del grupo 2 son los más altos en el gráfico SPE(ver Figura 4.2), lo que quiere decir que, son outliers moderados que rompen la estructura de correlación. Vamos a analizar el punto 127. Para conocer que variables han sido las causantes de esta salida de control realizamos el gráfico de contribuciones(ver Figura 7.2). Ahí vemos contribuciones muy negativas en el torque de los ejes de giro (S3 y S2). Otra cosa que nos llama la atención es que los primeros instantes de la velocidad S2 y S3 tambien tenemos contribuciones altas.

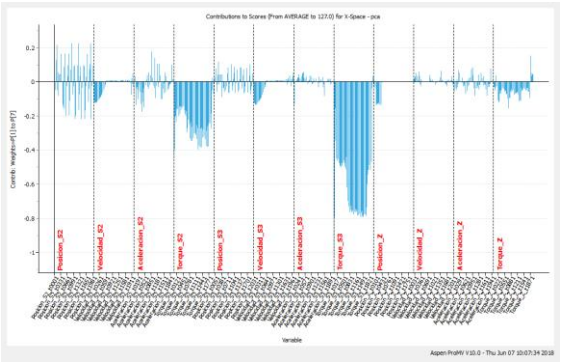

Figura 7.2 Contribution Plot, ciclo 127

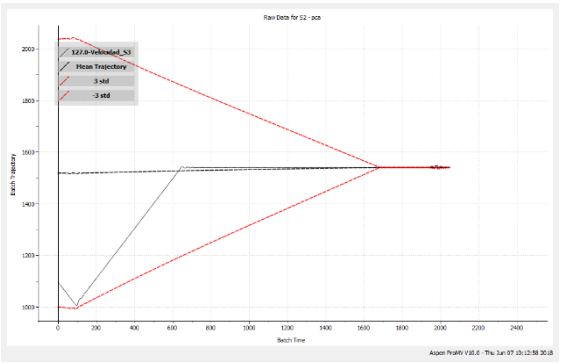

Figura 7.1 Velocidad S3, ciclo 127

El punto 127 tiene un par motor muy bajo y muy corto. Además en cuanto a la velocidad (ver Figura 7.1), podemos decir que no escontante a lo largo del tiempo sino que en los primeros instates disminuye y a continuación asciende hasta mantenerse sobre los 1500rpm (velocidad a la que se ha configurado la máquina). El par motor bajo es debido a la energía que consume la herramienta para emperzar a girar y mantenerse a la velocidad adecuada. Por tanto, el par

que estamos viendo no corresponde con un mecanizado del ciclindro, corresponde con el aumento de la velocidad de los husillos. Es por ello que se piensa que este ciclo podríamos clasificarlo como ciclo de ajuste a las condiciones operativas de la máquina. No se ha producido mecanizado, por tanto, ha sido un ciclo sin pieza, un ciclo en vacío.

Anexo 2. GRUPO 3, Punto 98, 106, 130, 225 y 273

Los ciclos del grupo 3 tienen un T2 de hotelling alto. Lo cual quiere decir que son outliers severos y podrían forzar la formación de una componente principal. Vamos a analizar el ciclo 98. En el gráfico de contribuciones que se muestra en la Figura 7.3 se observa una contribución muy negativa en las variables velocidad y torque de los dos ejes de giro (S2 y S3).

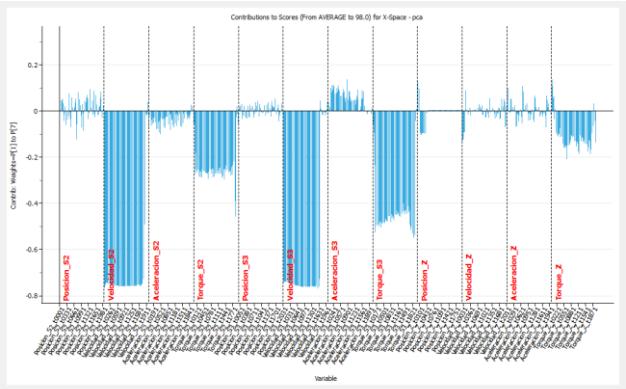

Figura 7.3 Contribution plot, ciclo 98

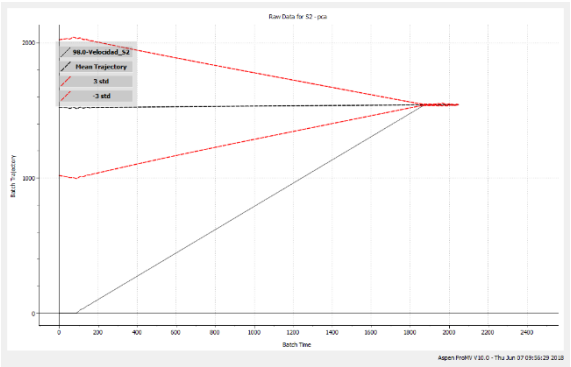

Figura 7.4 Velocidad S2, ciclo 98

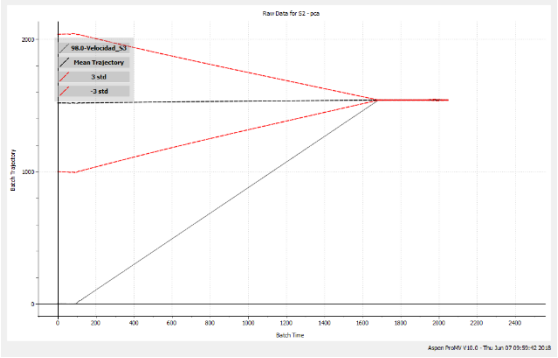

Figura 7.5 Velocidad S3, ciclo 98

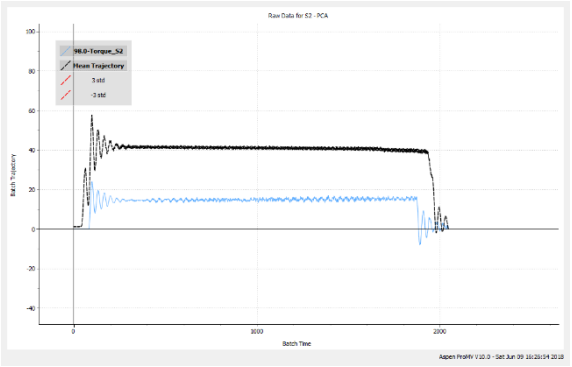

Figura 7.7 Torque S2, ciclo 98

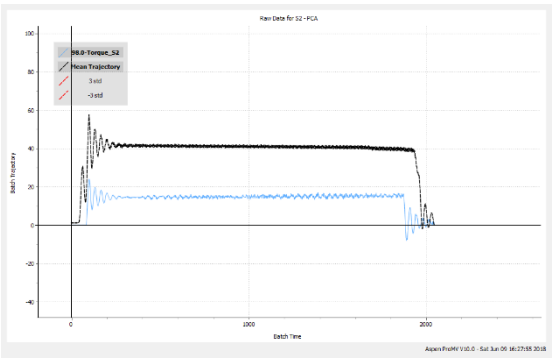

Figura 7.6 Torque S3, ciclo 98

En la Figura 7.4 y Figura 7.5 vemos como en los primeros instantes de tiempo la velocidad es 0, después se incrementa la velocidad hasta que finalmente se mantiene constante. La Figura 7.7 y 7.6 se representa gráficamente el torque de los ejes S2 y S3. En color azul se muestra el torque del ciclo 98 mientras que en color negro se muestra el torque medio de los 500 ciclos. Al igual que pasaba con el grupo 2 el toque más bajo que la media y más corto. Este par es debido a la energía que se consume para alcanzar la velocidad programada. Por tanto, en estos ciclos tampoco se está mecanizando los cilindros.

Muchas veces la OP110 se queda bloqueada debido a que la máquina siguiente es más lenta. Cuando se vuelve a poner en marcha la máquina el primer ciclo que se realiza es un ciclo en vacio. El grupo 3 se corresponden con los ciclos posteriores a una parada de la máquina.

Como los outliers corresponden a ciclos en vacio y por tanto sin desgaste de la herramienta se han eliminado del estudio.