Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

ESTUDIO DEL CONSUMO ESPECÍFICO DE COMBUSTIBLE EN EL HORNO DE UNA UNIDAD DE DESTILACIÓN DE CRUDO

Máster en Ingeniería del Análisis de Datos, Mejora de Procesos y Toma de Decisiones

Héctor Rodríguez López
9-9-2018

Contenido

Resumen

El trabajo propuesto se basa en el estudio del consumo específico de combustible en el horno de una unidad de destilación de crudo de una empresa petroquímica. Por motivos de confidencialidad no se revelarán datos relativos a la empresa, al proceso estudiado y a las variables analizadas, por lo que las variables serán definidas con una nueva nomenclatura. Son muchas las variables que intervienen en el refino del petróleo. Algunas de ellas no se podrán manipular mientras que en otras sí que será posible modificar su comportamiento. Teniendo en cuenta esta complejidad, este trabajo persigue tres objetivos.

Por una parte, se quiere desarrollar un modelo empírico que permita conocer la eficiencia máxima (consumo específico mínimo) dadas unas condiciones de trabajo del proceso. Esto permitiría saber en todo momento si, dadas unas condiciones de proceso, este se está operando de forma eficiente o no.

Por otra, se desea obtener información que puedan usar los operarios del proceso para saber cómo tienen que manipular ciertas variables del proceso para tratar de operar el mismo de la forma más eficiente posible.

Por último, se pretende establecer un modelo predictivo que informe de cuál va a ser el valor del consumo específico en la siguiente medida. De este modo, se puede prever qué comportamiento va a seguir el proceso y poder tomar medidas con antelación.

Para abordar estos objetivos se hará uso de técnicas estadísticas de análisis multivariante estudiadas en el máster. En procesos industriales en los que existen una gran cantidad de variables, estos métodos no son solamente una manera de interpretar mejor los datos, sino que, además permiten extraer conclusiones muy útiles, lo que no sería posible usando enfoques estadísticos más tradicionales.

Palabras clave: Análisis Multivariante, PCA, PLS, Monitorización.

1 Introducción

Hoy en día, la industria petrolera tiene un peso vital en la economía mundial. El principal producto de esta son los combustibles, pero el petróleo también hace las veces de materia prima para la producción de plásticos y productos químicos como disolventes o fármacos. Así pues, es entendible que en una sociedad tan industrializada y con tanta demanda de energía como la actual, a pesar del auge de las energías renovables, el campo de la industria petroquímica sea todavía un sector estratégico (Kulcsar, Koncz, Balaton, Nagy, & Abonyi, 2014).

Por consiguiente, no es extraño pensar que en una industria tan grande y que mueve tanto capital, la optimización de cada una de sus etapas se considere como algo crítico. Mejorar alguna de sus etapas, aunque sea en una pequeña proporción puede conllevar ganancias económicas muy sustanciales.

Entre los procesos globales de este ámbito, se incluyen la exploración, extracción, refino, transporte y mercadotecnia de los productos del petróleo.

Es la etapa de refino, o de destilación, la que se trata en este trabajo. En ella se consigue un fraccionamiento y unas transformaciones químicas del petróleo a fin de obtener derivados que se puedan consumir.

Por tanto, se trata de un problema en el cuál las variables a estudiar son de carácter químico-físico. Con la cantidad de datos (es decir, potencial información), que es posible producir y manejar actualmente, la manera de plantear el estudio resulta determinante. No será posible un enfoque estadístico con técnicas tradicionales dado el carácter multivariante del problema. Así pues, es necesario el uso de herramientas más avanzadas.

En cuanto a la organización del trabajo, en el apartado *Material y Métodos* se describen el tipo de datos que se han empleado, así como su estructura. También se presentan los métodos estadísticos y de inteligencia artificial que se han utilizado y el *software* empleado.

A continuación, en *Resultados* se muestran las tablas y gráficos producidos explicando cómo se han obtenido y qué se deduce de ellos.

Por último, en el apartado de *Conclusiones*, se discuten los resultados más relevantes y qué novedades aporta el estudio realizado.

2 Material y métodos

La base de datos analizada consta de 15420 registros horarios desde enero de 2014 hasta octubre de 2015. En dicho periodo se registró una parada por limpieza que abarca los meses de septiembre y octubre de 2014.

En dicha parada se limpiaron los tubos del horno y se realizaron operaciones que afectaron a la eficiencia energética de la unidad. Debido a esto es de esperar que, para unas mismas condiciones de proceso, los consumos en el periodo posterior a la parada de septiembre de 2014 sean inferiores a los del periodo previo. Se cree por otra parte, que los datos tomados durante la parada no son representativos.

Las variables del proceso que se van a estudiar son características químicas, físicas y termodinámicas. La variable respuesta es el consumo específico de combustible en el horno, que mide el consumo de energía por tonelada de crudo procesado. Las 89 variables del proceso registradas se han subdividido en 3 grupos:

- variables "*drivers*": variables no manipulables por los operarios pero que pueden afectar a la variable respuesta. (*Identificadas con la letra D*)
- variables operativas: variables manipulables por parte de los operarios. (*Identificadas con la letra X*)
- variables "dependientes": están correlacionadas con otras que se consideran más representativas para el proceso. Éstas serán tratadas del mismo modo que la variable de la que dependan.

En relación con el software utilizado, se han empleado tres programas. En primer lugar, a fin de tratar los datos, organizarlos y poder así llevar a cabo el pretratamiento, se ha optado por el *Microsoft Excel 2016*. La herramienta principal para los análisis estadísticos ha sido el software *MVA-GIEM*, complementado con el programa *Statgraphics*.

En este trabajo se han empleado diversas técnicas de análisis de datos procedentes tanto del ámbito estadístico como del de la inteligencia artificial.

La clasificación de estas técnicas se suele hacer en función de si se conoce previamente el objetivo buscado. Si los individuos están previamente clasificados o etiquetados, se tendrá un *Aprendizaje Supervisado*. En cambio, si se

desconocen las variables respuesta o etiquetas asociadas a ellos, se habla de *Aprendizaje No Supervisado*.

Así, dentro de los métodos *No Supervisados,* el Análisis de Componentes Principales (PCA) (Wold, Esbensen, & Geladi, 1987) es una técnica estadística muy potente que se emplea para reducir la dimensionalidad de un conjunto de datos. Consiste en buscar la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. De este modo, las variables originales correlacionadas se resumen en unas nuevas variables incorrelacionadas llamadas componentes principales que explican las principales fuentes de variabilidad.

Una opción que puede resultar interesante a partir de los resultados del PCA es aplicar la técnica de *Análisis Clúster* o *de Conglomerados* para clasificar un conjunto heterogéneo de elementos en una serie de grupos que reflejen las relaciones existentes entre los mismos. De esta manera, esta técnica genera grupos formados por elementos similares entre sí. El enfoque clásico de este tipo de análisis asume que cada individuo pertenece a un solo clúster. Por el contrario, el *Fuzzy Clustering* se caracteriza por el hecho de que los individuos pueden pertenecer a más de un grupo. Esa asociación a cada grupo está medida por un nivel de pertenencia (partición difusa).

Por otra parte, en el apartado de Aprendizaje Supervisado, una de las herramientas estadísticas más sencillas es la Regresión Lineal Múltiple. Sin embargo, en contextos como los de este trabajo, con muchas variables potencialmente explicativas altamente correlacionadas, esta técnica sufre de inestabilidad en la estimación de los parámetros del modelo (por problemas de mal condicionamiento de la matriz de covarianzas de los regresores), así como de problemas en la interpretación del modelo predictivo ajustado. Es decir, se pueden obtener modelos predictivos con diferentes regresores, pero con una bondad de ajuste similar. Esto provoca que no sea posible saber si los regresores no están incluidos porque no deben estarlo (no son significativos a la hora de explicar variabilidad de la respuesta) o porque están fuertemente relacionados con otros regresores que sí se han incluido ya en el modelo.

Una técnica de aprendizaje supervisado que funciona muy bien en situaciones de fuerte colinealidad es la Regresión en Mínimos Cuadrados Parciales (PLS, Partial Least Squares) (Wold, Sjöström, & Eriksson, 2001). Esta es una técnica estadística multivariante muy potente que permite relacionar dos estructuras de datos: la de los datos del proceso **X** y la de las variables respuesta a estudiar **Y**. Si bien el PCA trata de maximizar la varianza de **X**, el PLS intenta explicar la covarianza entre **X** e **Y**, es decir, las fuentes de variabilidad de *X* que estén relacionadas con las de **Y**. Una modificación del PLS es el PLS-DA, usado en contextos en los que la variable a estudiar es cualitativa y se pretende obtener un modelo de clasificación.

Otra técnica de *Aprendizaje Supervisado* es la del *Vecino más próximo*. Esta técnica no obtiene un modelo como tal, sino que trata de predecir el valor de una observación basándose en el valor de las observaciones más cercanas.

Cuando no se obtiene una buena predicción o separación entre clases usando la técnica anterior, las *Máquinas de Soporte Vectorial* (*SVM* por sus siglas en inglés) son una buena alternativa. Este conjunto de algoritmos construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta que puede ser utilizado en problemas de clasificación o regresión altamente no lineales.

Por último, una técnica más simple pero no por ello menos útil es el *Árbol de decisión*. En lugar de crear un modelo matemático como tal, crea una serie de reglas que hagan posible la clasificación de los nuevos elementos. Si bien no es un método tan exacto, su simplicidad hace que en ocasiones sea más intuitivo y aplicable en determinados problemas.

La mayoría de estas técnicas serán aplicadas en este trabajo con el fin de comprobar cuál de ellas se adapta mejor a la problemática del caso.

No obstante, hay que tener presente el objetivo del problema, y plantear entonces las técnicas más apropiadas para este fin. En este sentido, lo que se va a buscar no es simplemente una clasificación de los individuos, sino un enfoque discriminatorio. Esto es, saber discernir qué variables son aquellas que nos van a proporcionar información útil a la hora de clasificar, es decir, entender

qué variables manipulables del proceso tienen comportamientos distintos entre los periodos de alta y baja eficiencia del proceso.

Esto es posible en el caso del PLS. Por el contrario, otras técnicas se pueden considerar como “cajas negras”, ya que su función se limita a clasificar individuos, sin que el usuario realmente pueda interpretar qué sucede dentro del proceso. No es posible saber qué variables son responsables de las diferencias entre grupos.

Por tanto, las últimas técnicas mencionadas anteriormente como el *Vecino más próximo* o las *Máquinas de soporte vectorial* no van a ser abordadas puesto que su planteamiento no se adapta a las necesidades del problema.

Tres han sido, principalmente, las técnicas multivariantes empleadas en el análisis: Análisis de Componentes Principales (PCA) (Bro & Smilde, 2014), Regresión en Mínimos Cuadrados Parciales (PLS) (Wold, Sjöström, & Eriksson, 2001) y Análisis Discriminante mediante PLS (PLS-DA) (Prats-Montalbán, Ferrer, Malo, & Gorbeña, 2005). Mediante el *PCA* se ha realizado un estudio descriptivo multivariante de las variables con fines exploratorios. La técnica *PLS* ha sido usada para la construcción de modelos predictivos con predictores altamente correlacionados. Por último, el *PLS discriminante* (*PLS-DA)* se ha utilizado para discriminar entre periodos de consumo eficiente y no eficiente.

Además, se ha aplicado otra técnica con el objetivo de comprobar su utilidad en este trabajo y estudiar las diferencias con las primeras. Los Árboles de Decisión han permitido comparar con las técnicas iniciales si la discriminación entre variables daba los mismos resultados.

A continuación, se explica más a fondo las dos técnicas principales empleadas, el PCA y el PLS.

El *Análisis de Componentes Principales* (PCA) es una herramienta estadística multivariante que busca direcciones de máxima variabilidad. El objetivo principal es comprimir la información de la matriz de datos **X** a analizar.

La estructura de los datos sería la mostrada en la *Ilustración 1*:

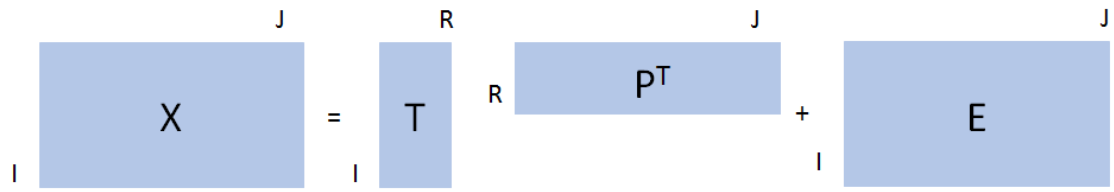

Ilustración 1.- Estructura de los datos del PCA.

Donde:

- **X**: Matriz de datos de partida: *I* individuos (filas) x *J* variables (columnas).
- **T**: Scores, valores que toman los *I* individuos en las nuevas *R* variables latentes.
- **P**: Loadings, relaciones entre las variables *J* originales y las *R* variables latentes.
- **E**: Residuos, variabilidad no explicada por el modelo.

El primer paso es el preprocesado necesario para hacer los datos comparables entre sí y el modelo más fácilmente interpretable. El tipo de preprocesado más común es el autoescalado, es decir centrado y escalado (en el caso en que las variables estén medidas en unidades distintas).

Para entender el funcionamiento del PCA, se puede recurrir a la *Descomposición en Valores Singulares*, *SVD*. Aplicándola, se descompone la matriz original **X** en un producto de matrices:

$$SVD\ (\mathbf{X}) = \mathbf{U} * \mathbf{\Sigma} * \mathbf{V^T}$$

Donde:

- **U**: matriz de autovectores de **XX'** asociados a los valores singulares de $\sqrt{\lambda_i}$.
- **Σ**: matriz diagonal con los valores singulares.
- **V**: matriz de autovectores de **X'X** asociados a los valores singulares $\sqrt{\lambda_i}$.

Si se autoescala **X**, la matriz **X'X** es proporcional a la matriz de correlaciones de **X**.

Calcular **Σ** y **V** equivale por tanto al cálculo de los vectores y valores propios de la matriz **X'X**.

$$\mathbf{X'X} * \vec{\mathbf{u}} = \lambda * \vec{\mathbf{u}}$$

$$(\mathbf{X'X} - \lambda * \mathbf{I}) * \vec{\mathbf{u}} = 0$$

$$\vec{\mathbf{u}} = 0 \; es\; solución\; trivial$$

$$|\mathbf{X'X} - \lambda * \mathbf{I}| = 0$$

Así, *λ* indica cómo de fuertes son las direcciones de máxima variabilidad de **X'X** y **v** sus direcciones. El autovector 1 ($\mathbf{v_1}$) es el vector director del eje para la primera componente principal (loading). Se expresa respecto al sistema de ejes original.

De esta manera se consiguen componentes principales con las siguientes características:

- Ortogonales
- Decrecientes en cuanto a la variabilidad explicada
- Combinaciones lineales de las variables originales

Hay varios criterios para obtener las componentes principales, pero todos representan matemáticamente lo mismo:

- Subespacio que mejor se ajusta a la nube de modo que los residuos sean mínimos.
- Subespacio que minimiza la deformación de la nube proyectada.
- Predicción óptima de las *J* variables originales a partir de *R* combinaciones lineales de las mismas.

Si bien ya se ha explicado que el SVD es un procedimiento para calcular las componentes principales, lo cierto es que en muchos casos no es necesario obtenerlas todas, sino solo las más importantes. En ese caso puede recurrirse a algoritmos secuenciales como el NIPALS, planteado en la *Ilustración 2*:

1. Iniciar con $\mathbf{t}=\mathbf{x}_k$ (columna de **X** con mayor varianza)
2. Regresión lineal múltiple (RLM) de las columnas de **X** sobre **t**

$$\underset{N,K}{\mathbf{X}} = \underset{K,1}{\mathbf{t}}\underset{1,N}{\mathbf{b}^T} \Rightarrow \times(\mathbf{t}^T\mathbf{t})^{-1}\mathbf{t}^T \Rightarrow \hat{\mathbf{b}}^T = (\mathbf{t}^T\mathbf{t})^{-1}\mathbf{t}^T\mathbf{X} = \mathbf{t}^T\mathbf{X}/(\mathbf{t}^T\mathbf{t}) = \underset{1,K}{\mathbf{p}^T} \Rightarrow \text{normalizar } \mathbf{p} \Rightarrow \|\mathbf{p}\| = 1$$

$(p_k = \mathbf{t}^T\mathbf{x}_k / \mathbf{t}^T\mathbf{t}$ C.L. de N individuos con pesos $t_s^{'})$

3. RLM de las filas de **X** sobre **p**

$$\underset{K,N}{\mathbf{X}^T} = \underset{K,1}{\mathbf{p}}\underset{1,N}{\mathbf{b}^T} \Rightarrow \times(\mathbf{p}^T\mathbf{p})^{-1}\mathbf{p}^T \Rightarrow \hat{\mathbf{b}}^T = (\mathbf{p}^T\mathbf{p})^{-1}\mathbf{p}^T\mathbf{X}^T = \mathbf{p}^T\mathbf{X}^T/(\mathbf{p}^T\mathbf{p}) = \mathbf{t}^T \Rightarrow \mathbf{t} = \mathbf{X}\mathbf{p}/(\mathbf{p}^T\mathbf{p}) = \mathbf{X}\mathbf{p}$$

$(t_i = \mathbf{x}_i^T\mathbf{p}$ C.L. de K variables con pesos $p_s^{'})$

4. En la convergencia **p** es vector propio de $\mathbf{X}^T\mathbf{X}$ con valor propio $\lambda = \mathbf{t}^T\mathbf{t}$ (SC de la nueva variable **t**)
   Calcular residuos: $\mathbf{E}=\mathbf{X}-\mathbf{t}\mathbf{p}^T$
   Volver a (2) haciendo $\mathbf{X}=\mathbf{E}$ y repetir el proceso para a=1,2,...A componentes

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \ldots\ldots + \mathbf{t}_A\mathbf{p}_A^T + \mathbf{E} \quad (\mathbf{X} \text{ centrado})$$

   Una dimensión cada vez ($\mathbf{t}_A$, $\mathbf{p}_A$), significación mediante diversos procesos

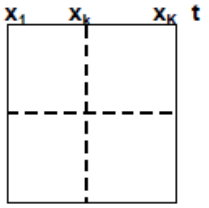

Ilustración 2.- Algoritmo NIPALS para el PCA.

En el caso de la *Regresión por Mínimos Cuadrados Parciales* (PLS) se busca maximizar las fuentes de variabilidad de los dos espacios X (matriz de variables explicativas o regresores) e Y (matriz de variables respuesta) que estén correlacionadas, por lo que la función objetivo a maximizar es la covarianza:

$$cov(x, y) = s_x * s_y * r_{xy}$$

Se busca una regresión lineal mediante la proyección de las variables observables y las variables predictivas a un nuevo espacio, en vez de buscar hiperplanos de máxima varianza entre las variables independientes y la variable respuesta.

De este modo, esta técnica se aplica a fin de hallar las relaciones entre las dos matrices (**X** e **Y**). Así, se modela con variables latentes la estructura de covarianza en los dos espacios originales.

Se lleva a cabo de nuevo el algoritmo NIPALS como explica la *Ilustración 3*:

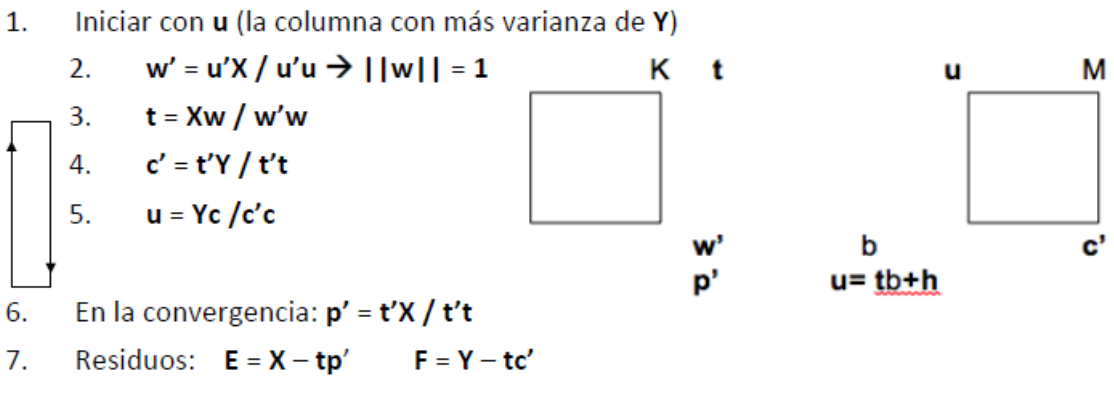

Ilustración 3.- Algoritmo NIPALS para el PLS.

La *Regresión por Mínimos Cuadrados Parciales-Análisis Discriminante* (PLS-DA) se da cuando la matriz **Y** a estudiar con tiene variables de carácter binario.

3 Resultados

3.1 Pretratamiento de los datos

Uno de los primeros pasos que hay que dar a la hora de analizar una base de datos es comprobar la calidad de esta. En cierto sentido, una buena base de datos contendrá poca proporción de datos faltantes.

Para solventar este problema existen varias vías. La más simple de ellas es eliminar completamente cada una de las observaciones en las cuales exista algún dato faltante. No obstante, teniendo en cuenta que se tienen 89 variables, no sería extraño que en muchas de las observaciones falte alguna variable por medir, así que realizar este procedimiento conllevaría una gran pérdida de información potencial. Una opción más apropiada es usar métodos de imputación de datos faltantes explotando las correlaciones entre las variables (Folch-Fortuny, Arteaga, & Ferrer, 2015 y 2017)

Así, se calculan para los tres periodos (*Fase previa*, *Parada* y *Fase posterior*) los porcentajes de datos faltantes para cada variable.

Se observa que para los periodos de antes y después de la parada, los datos faltantes no suponen ningún problema pues la proporción de estos no llega

apenas al 2% en ninguna variable. Sin embargo, en el periodo de parada por limpieza estas cantidades sí que son importantes, pudiendo llegar hasta el 88%. A modo de ejemplo, en la *Tabla 1* se muestra el porcentaje de datos faltantes de 5 variables en los tres periodos.

Tabla 1.- Porcentaje de datos faltantes por periodo.

	X6	X12	X13	X18	X61
PREVIO	0.02%	0.13%	0.22%	0.81%	0.15%
PARADA	**31.30%**	**67.31%**	**84.72%**	**63.68%**	**88.03%**
POST	0.01%	0.01%	0.01%	0.01%	0.01%

La *Ilustración 4* muestra la estructura de datos faltantes en el periodo de parada. Tras consultar con los expertos de la empresa se decidió no usar los datos del periodo de parada, sino solo los de la fase previa y posterior a la limpieza del horno de destilación.

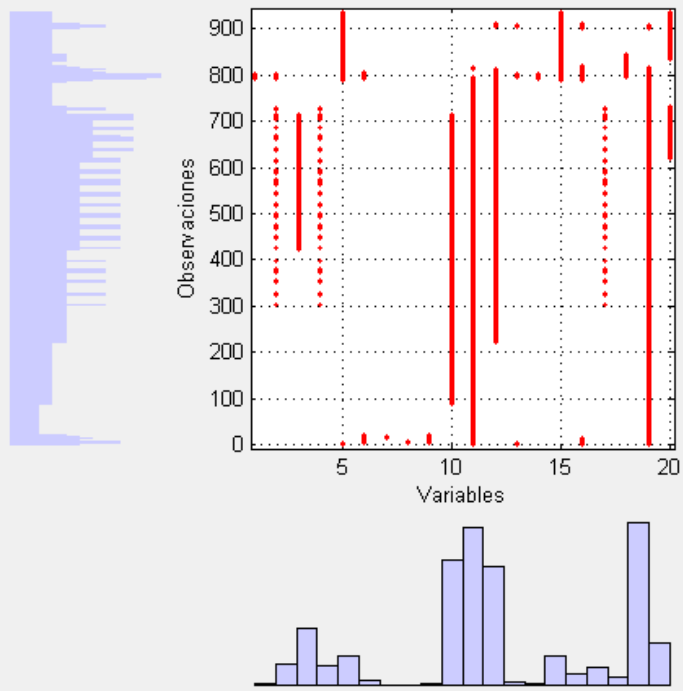

Ilustración 4.- Estructura de los datos faltantes en la fase de parada.

La siguiente cuestión estudiada fue si la limpieza tuvo un efecto sobre el consumo específico. La *Ilustración 5* muestra la serie de tiempo del consumo

específico antes (en azul) y después (en rojo) de la parada. En la *Tabla 2* se muestran la media y la desviación típica muestrales en ambos periodos.

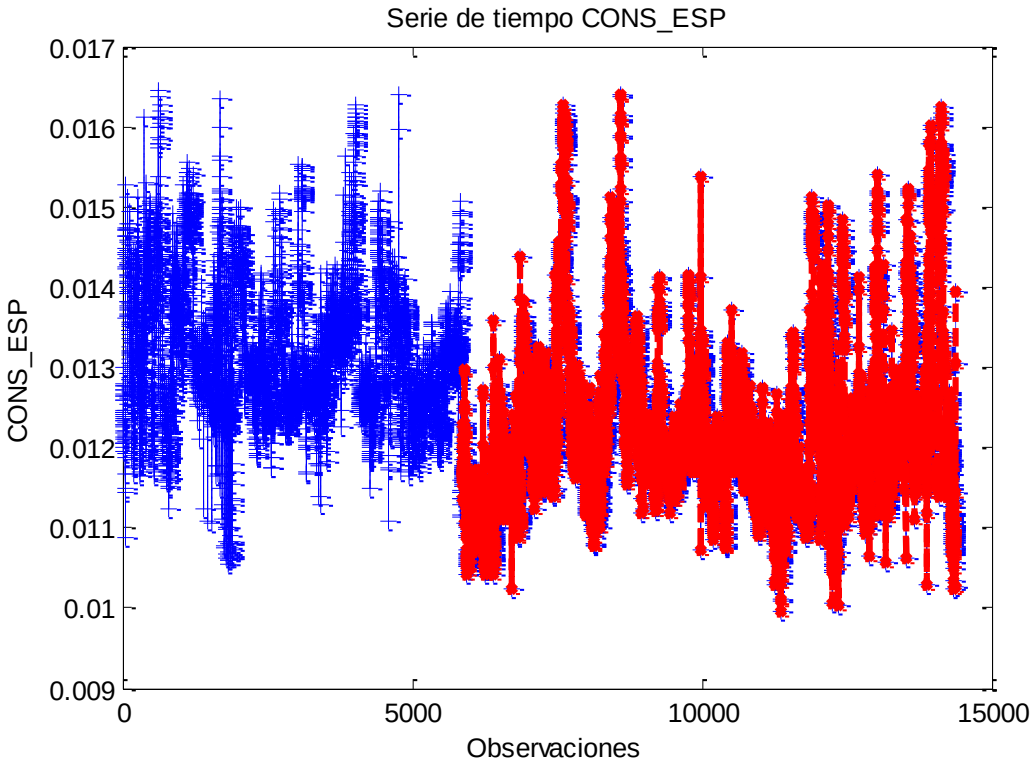

Ilustración 5.- Serie de tiempo del consumo específico: antes (azul) y después (rojo) de la parada por limpieza.

Tabla 2.- Media y desviación típica del consumo específico para las fases previa y posterior.

	Media	Desv. Típica
Fase previa	0.0132	0.0012
Fase posterior	0.0122	9,7256E-04

Para confirmar si ha habido un cambio en la varianza del consumo específico medio antes y después de la parada, se ha realizado un contraste de hipótesis planteando como hipótesis nula que dicha varianza no ha cambiado. Los resultados obtenidos con el *Statgraphics* se muestran a continuación en la *Ilustración 6*:

Pruebas de Hipótesis

Desviaciones estándar muestrales = 0,0012 y 0,00097256
Tamaños de muestra = 5951 y 8533

Intervalos de confianza del 95,0% para el cociente de varianzas: [1,45397;1,59682]
Hipótesis Nula: cociente de varianzas = 1,0
Alternativa: no igual
Estadístico F calculado = 1,5224
Valor-P = 1,31006E-14
Rechazar la hipótesis nula para alfa = 0,05.

Ilustración 6.- Prueba de hipótesis sobre el cociente entre las varianzas.

Con un *p*-valor menor del 0,05, se rechaza la hipótesis de que la varianza se mantiene constante. El intervalo de confianza para el cociente de varianzas indica que es razonable concluir que la varianza antes de la limpieza es entre 1,45 y 1,59 veces la varianza después de la misma, por lo que el proceso, tras la parada, ha reducido la variabilidad de su consumo específico, lo que conlleva una mejora en la calidad.

Para estudiar un posible cambio en la media del consumo específico antes y después de la parada, se ha realizado también un contraste de hipótesis para la diferencia de medias. Los resultados obtenidos con el *Statgraphics* son los mostrados en la *Ilustración 7*:

Pruebas de Hipótesis

Medias muestrales = 0,0132 y 0,0122
Desviaciones estándar muestrales = 0,0012 y 0,00097256
Tamaños de muestra = 5951 y 8533

Intervalo aproximado del Intervalos de confianza del 95,0% para la diferencia entre medias: 0,001 +/- 0,0000368153 [0,000963185;0,00103682]
Hipótesis Nula: diferencia entre medias = 0,0
Alternativa: no igual
Estadístico t calculado = 53,2378
Valor-P = 0,0
Rechazar la hipótesis nula para alfa = 0,05.

(No asumiendo varianzas iguales).

Ilustración 7.- Prueba de hipótesis sobre el cociente entre las medias.

Del cual se concluye (con un riesgo de primera especie del 5%) que tampoco la media se mantiene constante. De hecho, el intervalo de confianza para la diferencia de medias indica que es razonable asumir que el consumo específico

medio tras la parada es entre [0.000963185; 0.00103682] unidades menores que antes de la parada.

Una vez se ha demostrado que el segundo periodo del proceso es más eficiente, se aplicarán las técnicas multivariantes en este periodo, que representa las condiciones actuales tras la parada por limpieza.

En dicho sentido, se va a proceder, por tanto, a una imputación de datos faltantes, necesaria para poder llevar a cabo los análisis multivariantes.

3.2 Análisis de Componentes Principales (PCA)

El primer análisis realizado, a modo de análisis exploratorio de los datos, ha sido un PCA. Inicialmente, éste se va a ejecutar únicamente sobre las variables llamadas *Drivers* ya que no es posible manipularlas y puede resultar interesante observar su comportamiento. Al no ser una cantidad demasiado grande de variables, 16, es de esperar que con pocos componentes se explique una proporción grande de la variabilidad del proceso original, como se puede observar en la *Tabla 3*:

Tabla 3.- Estadísticos R^2 y Q^2 para el PCA.

R2/Q2 acumulada

A	R2	Q2
1	0.62324	0.6231
2	0.72634	0.72602
3	0.79841	0.79798
4	0.85221	0.85177
5	0.9019	0.90156
6	0.94353	0.94335
7	0.97246	0.97235
8	0.99057	0.99055

Efectivamente, con cuatro componentes principales ya se explica aproximadamente el 85% de la variabilidad de los datos tanto en bondad de ajuste (R^2) como en bondad de predicción (Q^2).

Lo primero que hay que hacer una vez creado el modelo, es validarlo. Para ello, se estudia primero el gráfico de control de la suma de cuadrados residual (SPE) y a continuación el de la T^2 de Hotelling. Ambos gráficos se construyen fijando un límite de control superior (LCS) asociado a un cierto percentil de la distribución del estadístico correspondiente. En este trabajo se usará el percentil 95%, por lo que cada gráfico tendrá una tasa de falsas alarmas del 5%. Para conseguirlo, una vez construido cada gráfico, se calculará el número medio de observaciones que se espera caigan por encima del LCS estando el proceso bajo control estadístico (N×0,05), siendo N el número de datos representados en el gráfico. Se identificarán las observaciones que superen el LCS (95%), N_{fuera}. De estas se considerarán anómalas aquellas cuyo valor del estadístico sea 3 veces el LCS. Del resto, también tendrán esta consideración las N_{fuera} - N×0,05 observaciones con valores más grandes del estadístico. En el gráfico SPE a estas observaciones anómalas se les llamará *outliers* moderados, mientras que en el gráfico T^2-Hotelling recibirán el nombre de *outliers* severos.

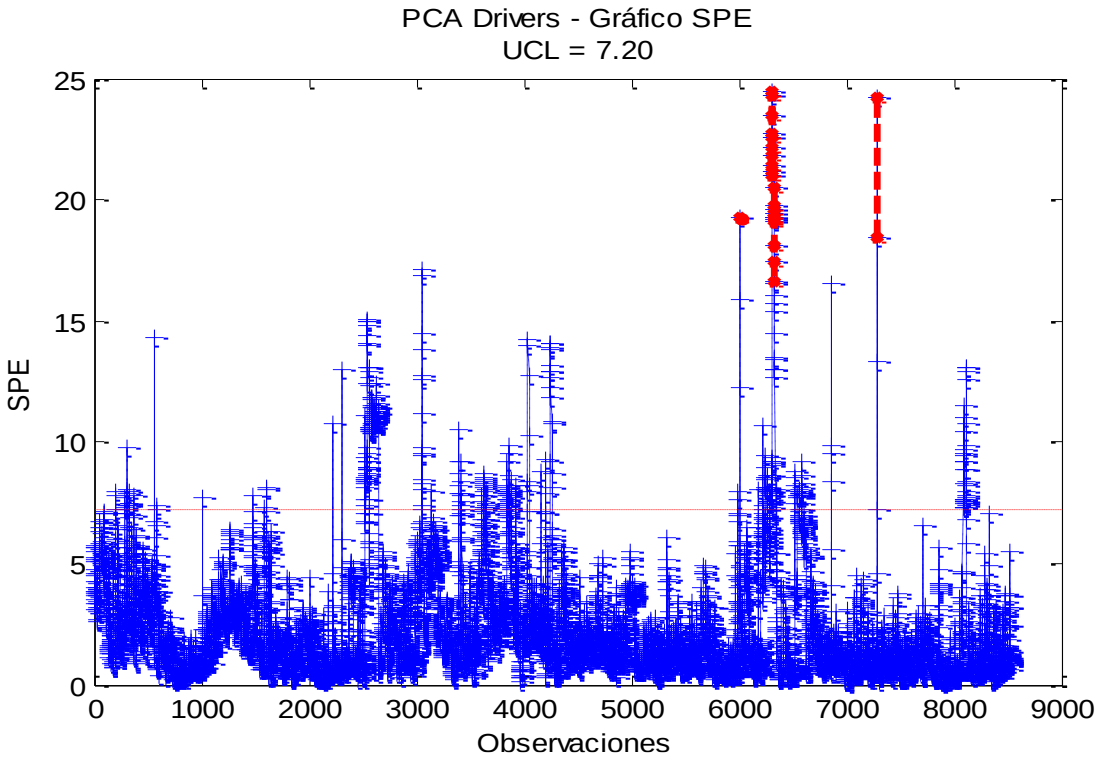

Ilustración 8.- Suma de cuadrados residual (SPE) para el PCA.

En el gráfico SPE (ver *Ilustración 8*) se pueden observar los *outliers* moderados (en rojo), es decir, observaciones atípicas que tienen una distancia euclídea

grande a su proyección en el subespacio de componentes principales. Normalmente están asociados a observaciones que rompen la estructura de correlación de las variables del modelo.

El SPE no es más que el producto escalar del vector de residuos **e** para cada observación:

$$SPE = \boldsymbol{e^T e}$$

Así, se puede entender el SPE como una suma de contribuciones de las variables del modelo. De este modo, se pueden estudiar las contribuciones de cada variable original en cada observación atípica. En este caso esto se realiza con las observaciones más alejadas, las señaladas en la *Ilustración 5* en color rojo. A modo de ejemplo, las *Ilustraciones 9* y *10* muestran el gráfico de contribución a la SPE de dos observaciones atípicas (13199 y 14181). En ambos casos se puede apreciar la variable que más influye al hecho de que estas observaciones sean atípicas: la D15 en la observación 13199 y la D16 en la observación 14181.

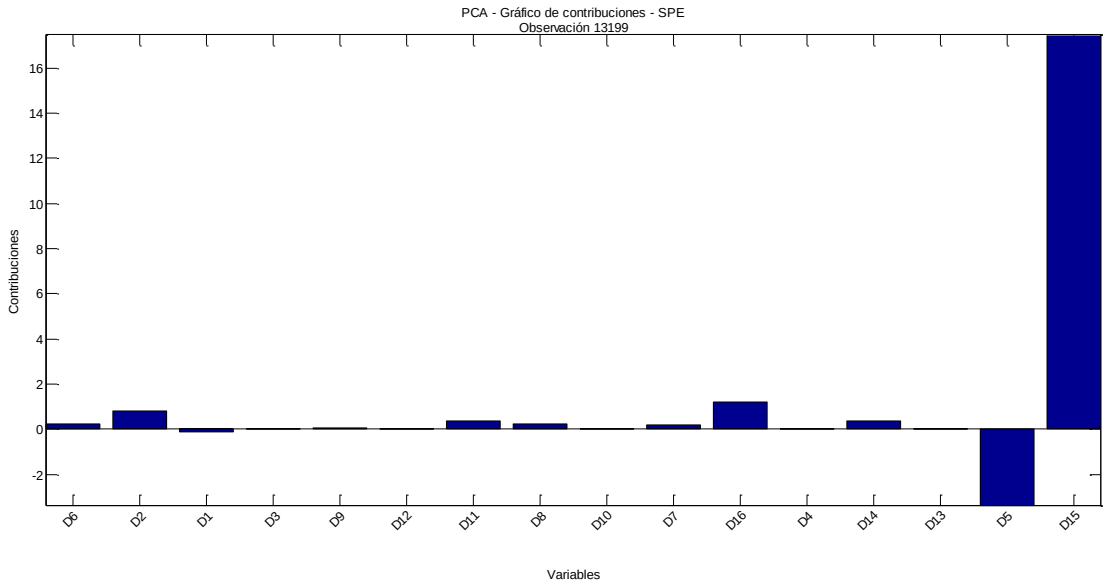

Ilustración 9.- Gráfico de contribuciones para la observación 13199.

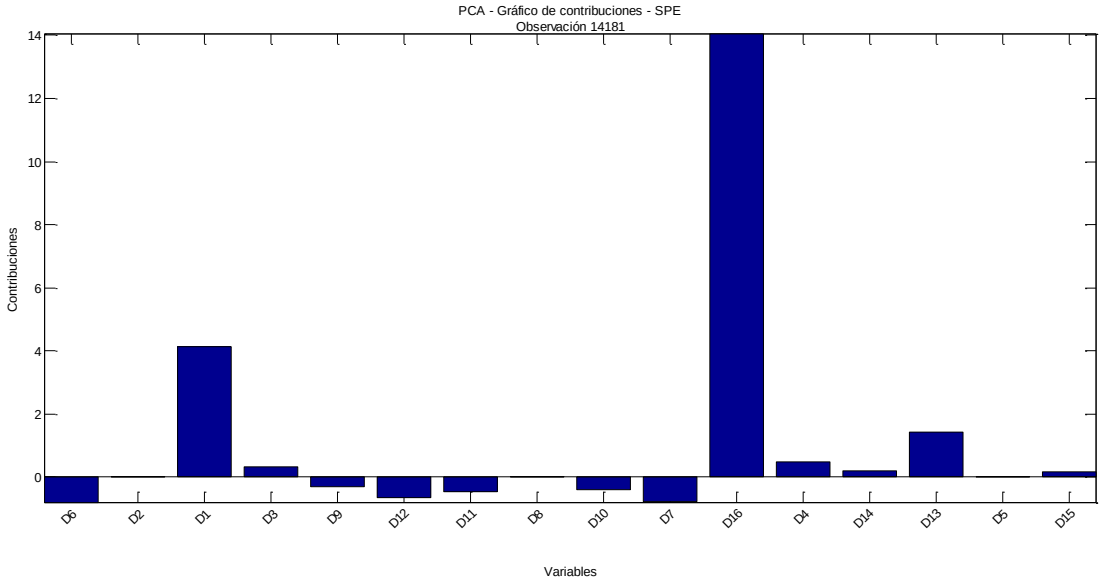

Ilustración 10.- Gráfico de contribuciones para la observación 14181.

De forma más precisa, en la observación *14181* se observa que la variable *D16* provoca que ésta tenga un comportamiento extremo. Esto se corrobora con una serie de tiempo en la cual efectivamente se aprecia un valor anómalo para esta observación (marcada en rojo en la *Ilustración 11*):

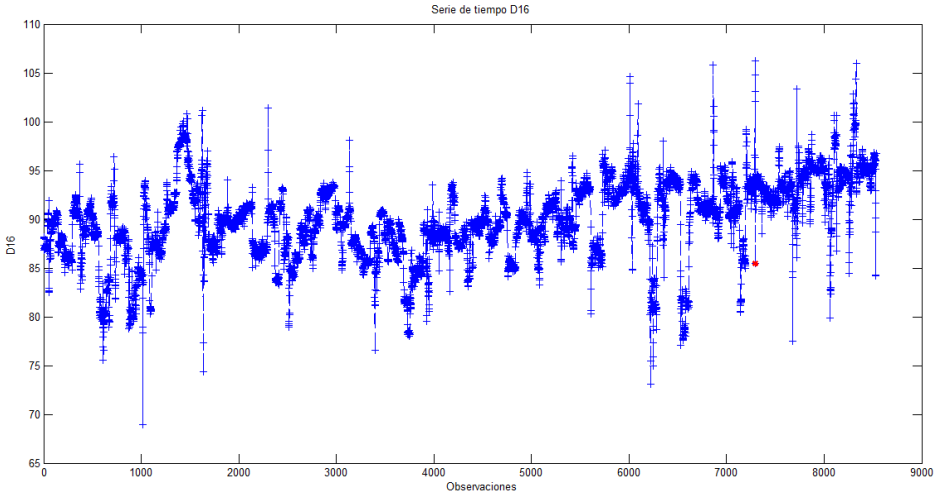

Ilustración 11.- Serie de tiempo para la variable D16.

Ahora se puede calcular el estadístico T^2 de Hotelling y observar aquí los outliers severos. Estos son observaciones extremas que, si bien no rompen la estructura de correlación (en el caso en que no tengan un SPE elevado), pueden significar un cambio en las condiciones de operación.

En cuanto al estadístico, se calcula mediante la fórmula:

$$T^2 = \boldsymbol{\tau}^T \boldsymbol{\Theta}^{-1} \boldsymbol{\tau}^{-1}$$

Donde $\boldsymbol{\tau}$ es el vector de scores para cada individuo y $\boldsymbol{\Theta}$ es la matriz diagonal que contiene los valores propios.

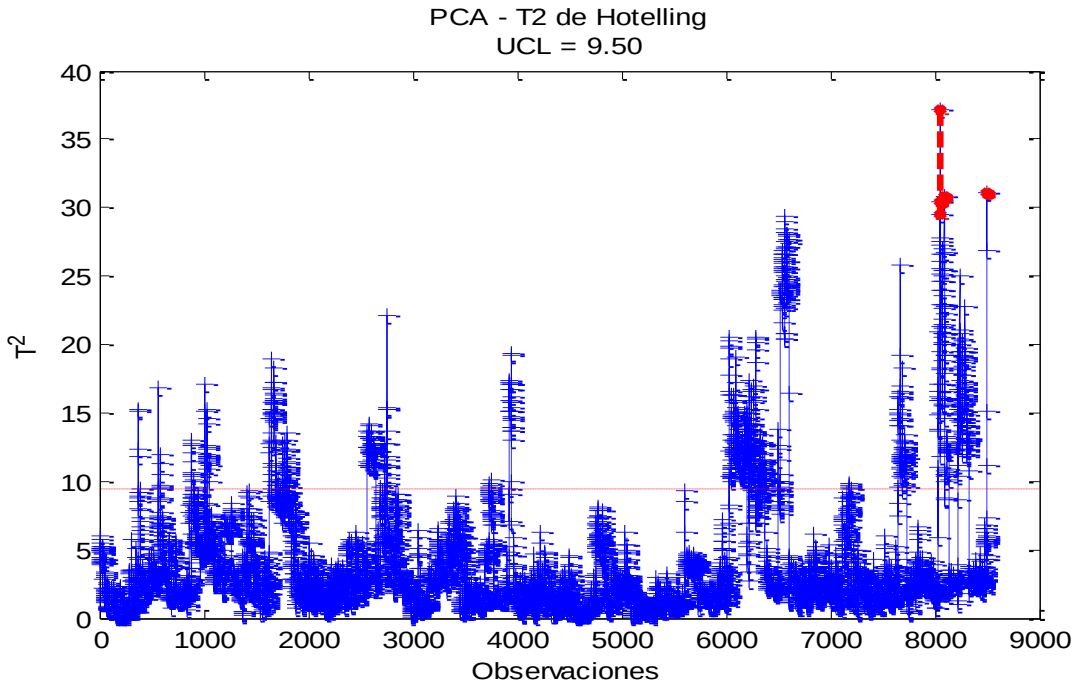

Ilustración 12.- Gráfico T^2 de Hotelling para el PCA.

En la *Ilustración 12* se observa que en algunas de las observaciones el valor del T^2 es demasiado elevado (marcadas con color rojo). Estudiando cuáles son las componentes que hacen que tome este valor tan alto y en ellas cuáles son las variables originales responsables, se puede detectar qué es lo que provoca esta anomalía. A modo de ejemplo, la *Ilustración 13* muestra los gráficos de contribución a la T^2 de Hotelling para una observación anómala (14964). Se observa que los problemas aparecen en la primera componente, habiendo bastantes variables Drivers implicadas en la anomalía.

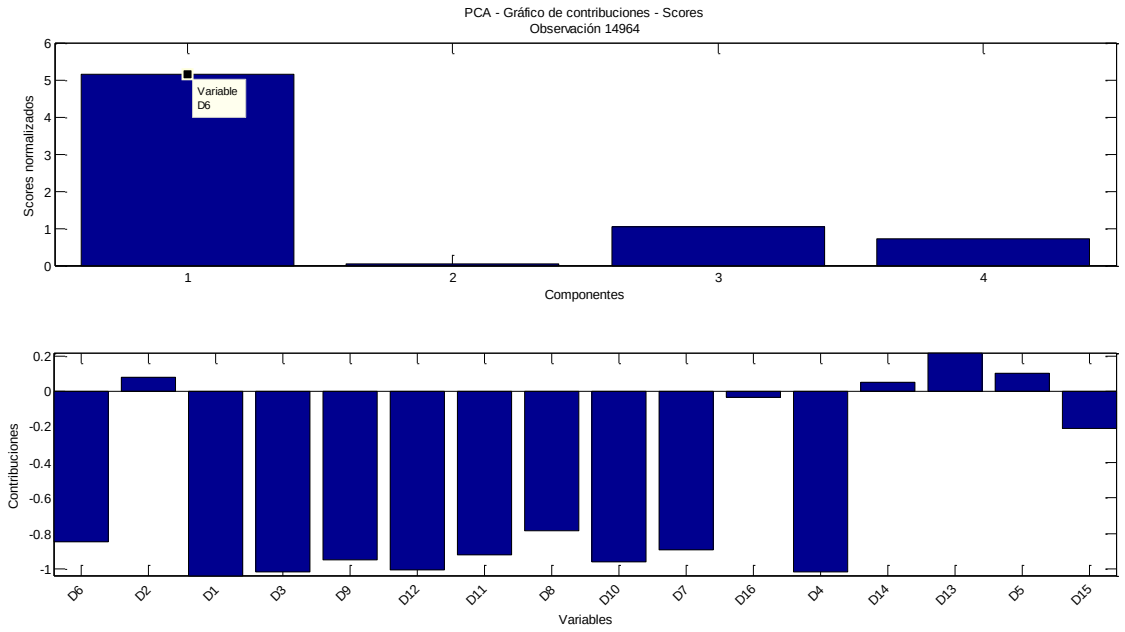

Ilustración 13.- Gráfico de contribuciones para la observación 14964.

Tras consultar con los expertos del proceso, se decide eliminar las observaciones atípicas detectadas, ajustando un nuevo modelo PCA con las observaciones restantes. En este modelo ya no se detectaron observaciones anómalas, por lo que se pasó a su interpretación.

En el gráfico de *loadings* (como el de la *Ilustración 14*) pueden observarse las relaciones entre las variables originales. Aquellas que estén correlacionadas positivamente tendrán valores similares en las dos primeras componentes, y aparecerán situadas cerca, pero lejos del origen del gráfico. Si la correlación es negativa, las variables aparecerán en situaciones antipódicas, también lejos del origen. Por último, si las variables no tienen relación aparecerán formando un ángulo recto entre las líneas imaginarias que, partiendo de cada variable, pasan por el origen.

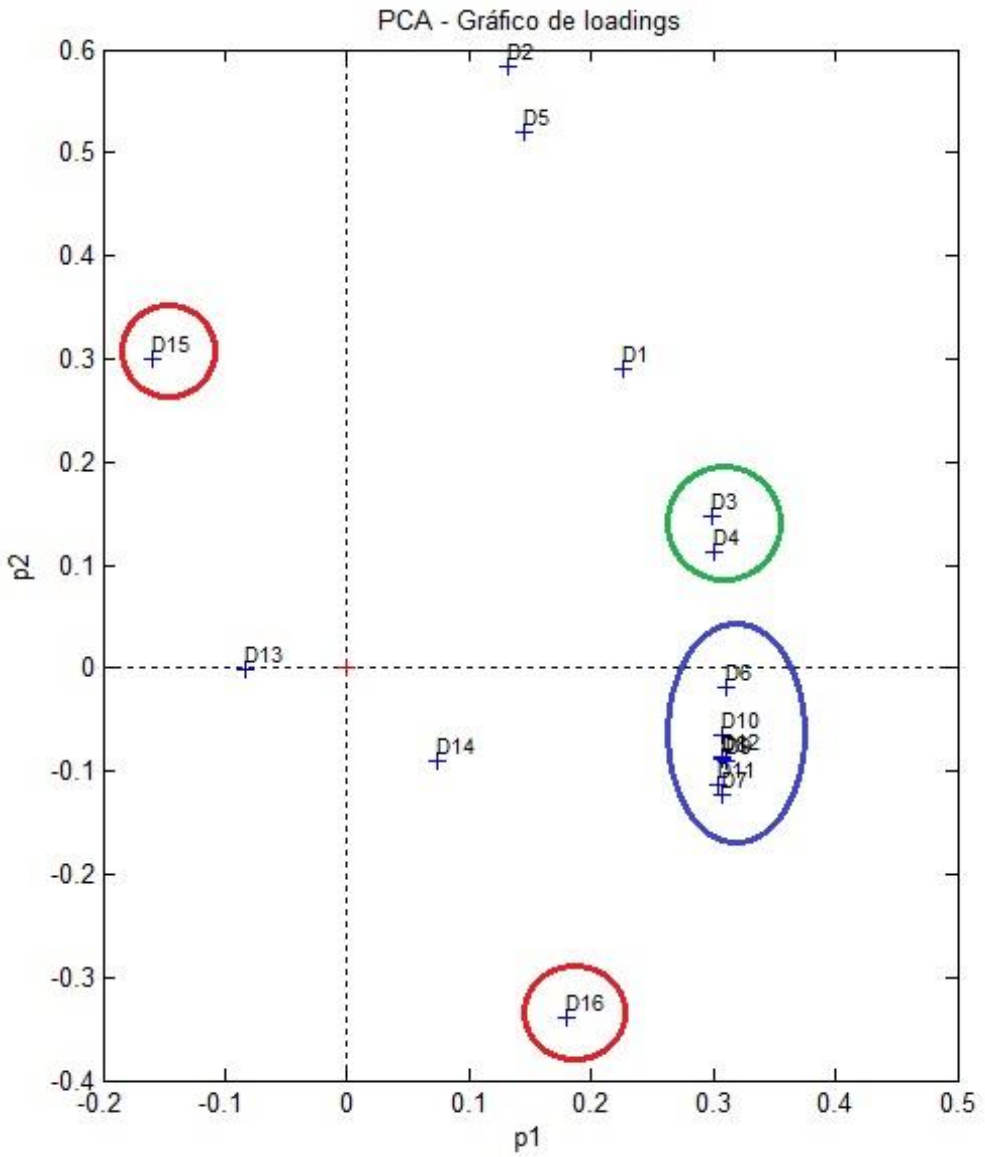

Ilustración 14.- Gráfico de loadings para el PCA.

De este modo, puede verse que hay dos grupos de variables altamente relacionados (señaladas por los círculos verde y azul) mientras que las variables *D15* y *D16* están también correlacionadas, pero negativamente (señaladas en rojo).

En cuanto al gráfico de *scores* (ver *Ilustración 15*), es posible extraer información sobre los individuos. Es decir, cómo se distribuyen las observaciones en el plano definido por las dos primeras componentes principales.

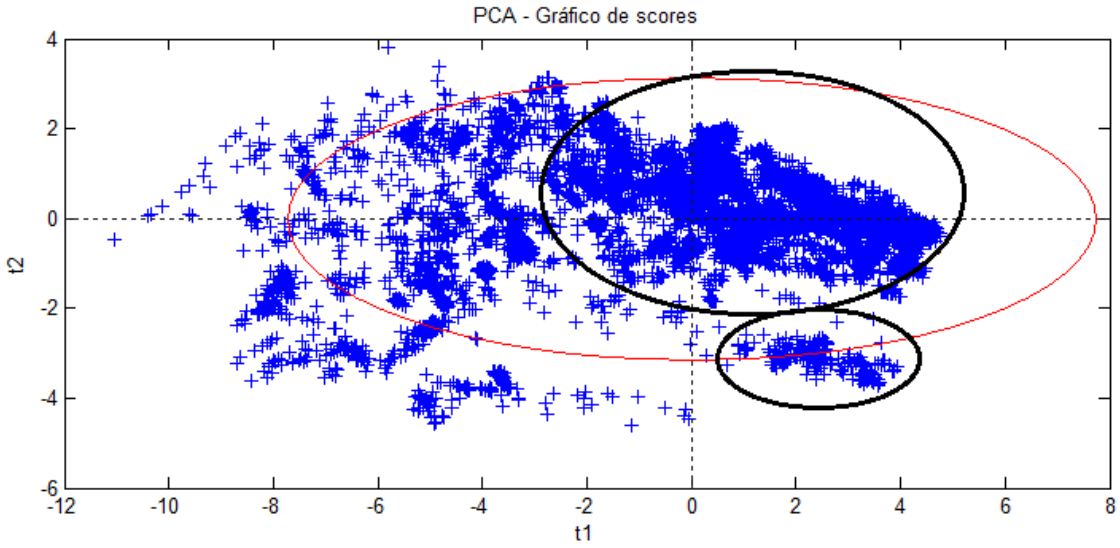

Ilustración 15.- Gráfico de scores para el PCA.

Como se observa en el gráfico, hay varias zonas (fundamentalmente las destacadas en las dos elipses negras) en las que el proceso ha tenido un comportamiento diferente.

3.3 PLS (Partial Least Squares)

Para conseguir el objetivo del trabajo: desarrollar un sistema que permita, en base a los datos registrados, monitorizar y evaluar la eficiencia energética con que está operando la planta, se seguirá el siguiente procedimiento:

En un primer paso se ajustará un modelo PLS (llamado PLS-I) usando como variable respuesta el consumo específico, y como variables predictoras las variables no manipulables (*Drivers*) indicadas por los expertos de la empresa, y se calcularán los residuos. Residuos positivos indican consumos específicos observados mayores que los predichos, por tanto, periodos de baja eficiencia; mientras que residuos negativos indican consumos específicos observados menores que los predichos, por tanto, periodos de operación eficiente.

Posteriormente, se ajustará un segundo modelo PLS (llamado PLS-II) usando como variable respuesta los residuos del modelo anterior, y considerando como variables predictoras, las variables manipulables (el resto de las variables disponibles que no sean “no manipulables”). El objetivo de este segundo modelo

es conocer qué variables manipulables están más relacionadas con los residuos del modelo anterior y, por tanto, con los periodos de baja/alta eficiencia energética. Esto permitirá conocer qué combinación de valores de las variables manipulables proporcionan un manejo eficiente de la planta en términos de consumo específico.

De este modo, se tendrán dos modelos PLS:

$$PLS - I \equiv CONS_{ESP} = f(Drivers) + e$$

$$PLS - II \equiv e = f(No\ Drivers) + e'$$

Como paso previo para realizar el PLS-I, se va a crear un conjunto de entrenamiento (*training set*) con el 75% de los datos disponibles, para después comprobar con el 25% restante (conjunto de validación) que el modelo representa bien el proceso. Como ilustra la *Tabla 4*, extrayendo tres componentes latentes la bondad de predicción por validación cruzada (Q^2) es superior al 60%. La *Ilustración 16* muestra que añadir una cuarta componente prácticamente no mejora sustancialmente el Q^2 del ajuste, por lo que se decide extraer únicamente 3 componentes PLS.

Tabla 4.- Estadísticos $R^2(X)$, $R^2(Y)$ y Q^2 para el PLS-I.

A	R2(X)	R2(Y)	Q2
1	0.57867	0.36041	0.35984
2	0.70319	0.56355	0.56256
3	0.76316	0.60588	0.60462

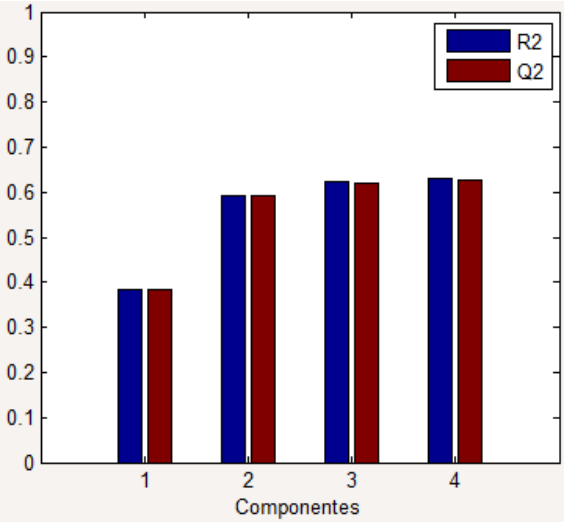

Ilustración 16.- Acumulado de los estadísticos R^2 y Q^2 para el PLS-I.

Siguiendo el mismo procedimiento que en el PCA, para validar el modelo se eliminan los valores anómalos (marcados en rojo) tanto del gráfico SPE (*Ilustración 17*) como de la T^2 de Hotelling (*Ilustración 18*):

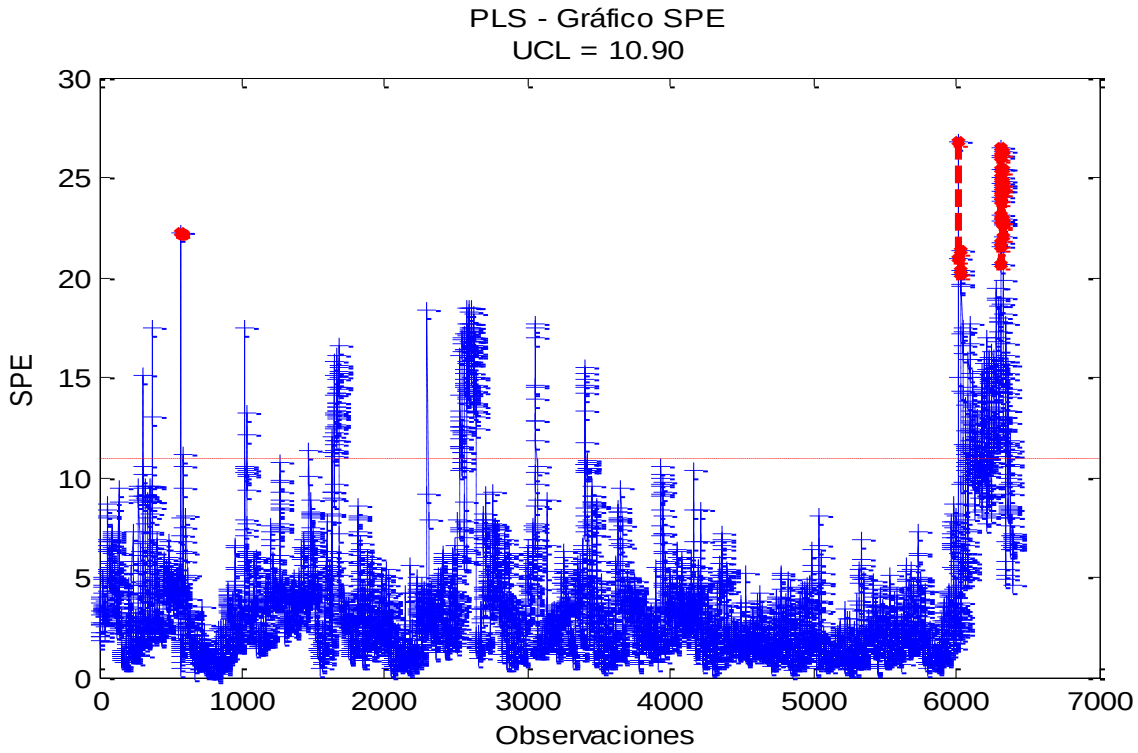

Ilustración 17.- Gráfico SPE para el PLS-I.

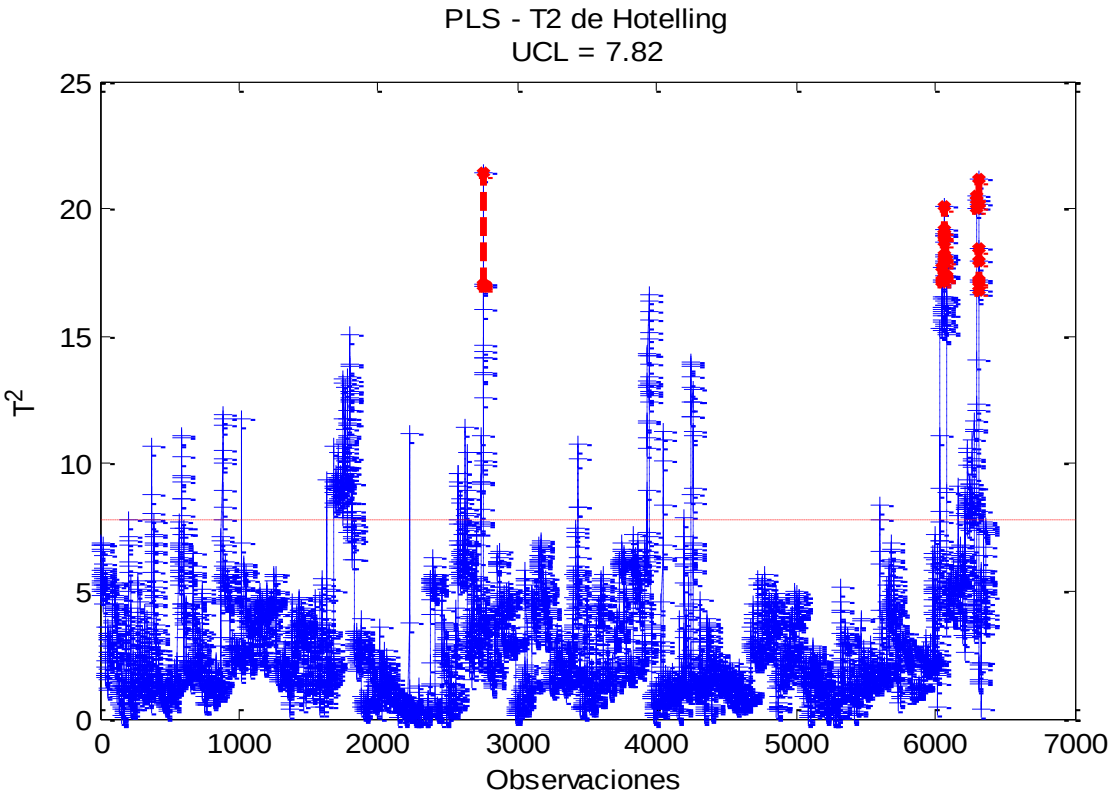

Ilustración 18.- Gráfico T^2 de Hotelling para el PLS-I.

Una vez eliminadas las observaciones anómalas (con el visto bueno de los expertos del proceso) se ajusta un nuevo modelo PLS.

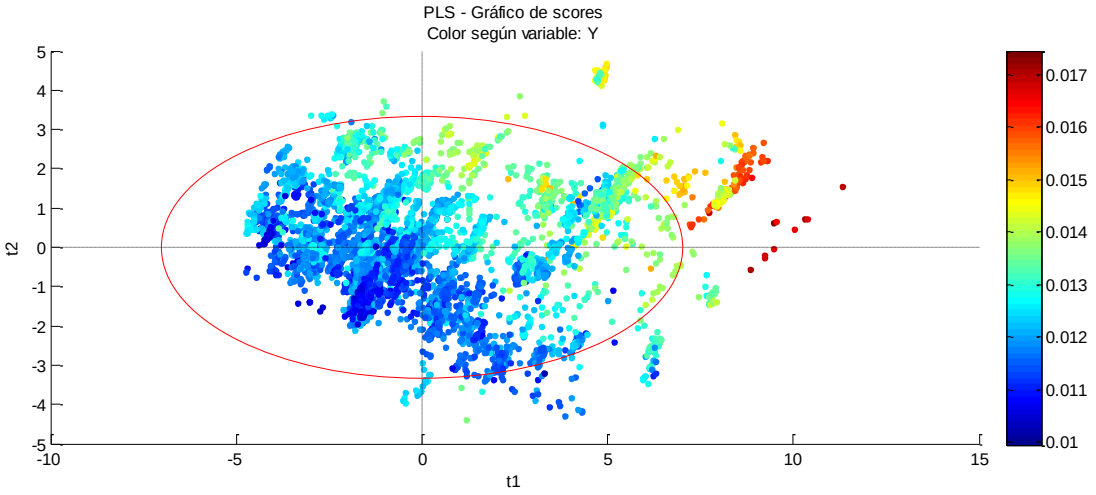

Ilustración 19.- Gráfico de scores para el PLS-I coloreado según el consumo específico.

En la *Ilustración 19* se puede observar cómo se distribuyen las observaciones a lo largo del espacio creado por la primera y la segunda componente PLS. El gráfico de *scores* t1/t2 es como una ventana en la que se muestran los individuos proyectados en el plano definido por dos componentes. Las coordenadas de cada individuo serán el vector de *scores* de cada componente. La elipse que se muestra incluye a todas las observaciones que no sobrepasen los límites de la T^2 de Hotelling con un 95% de confianza. El gráfico de los scores t1/t2 se ha coloreado según el consumo específico. Se observa un claro gradiente en la dirección Noreste, lo que indica que en los datos analizados ha existido una clara variación del consumo específico en esa dirección.

La *Ilustración 20* muestra que la relación interna entre los *scores* de ambos espacios, t y u, es lineal, como asume el modelo PLS.

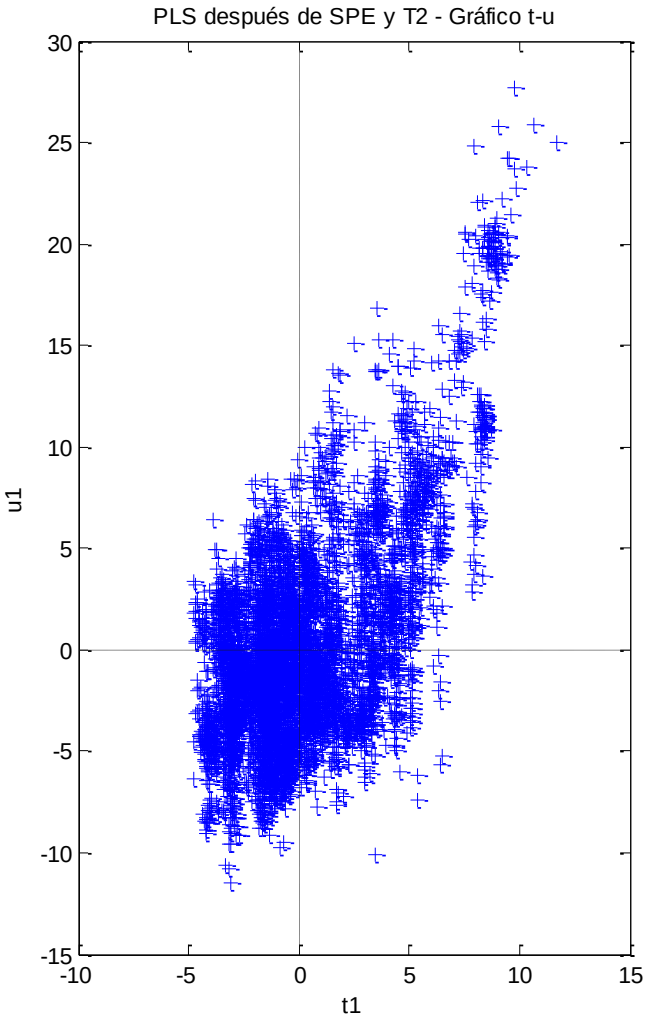

Ilustración 20.- Gráfico t-u para PLS-I.

La correlación entre los *drivers* y el consumo específico en las dos primeras componentes se muestra en el gráfico de *weightings* la *Ilustración 21*.

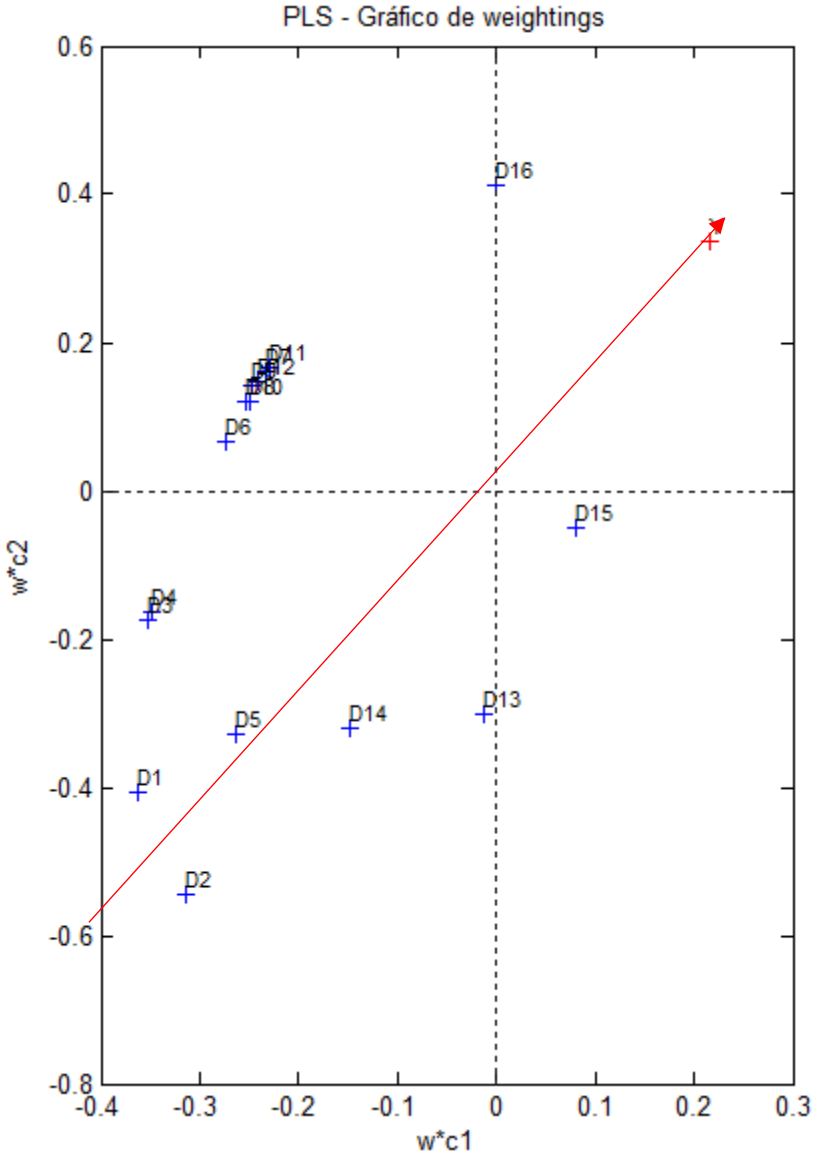

Ilustración 21.- Gráfico de weightings para el PLS-I.

En este gráfico de *weightings* se superponen las relaciones entre los drivers y la variable respuesta (consumo específico). Esto da información de la estructura de correlación entre X e y, es decir, cómo las variables se combinan para formar la relación cuantitativa entre X e y.

De este modo, se puede entender qué variables son importantes en la relación y cuáles proporcionan la misma información. En este gráfico se observa que la variable respuesta consumo específico está en el cuadrante noreste, de ahí el

gradiente observado en el gráfico de scores coloreado por el valor del consumo específico de la *Ilustración 19*. La mayoría de los Drivers están en el cuadrante suroeste (en situación antipódica) lo que indica que estarán relacionados negativamente con el consumo específico, por lo que, en periodos de alto consumo específico, sus valores tenderán a estar por debajo de su media (y viceversa). Esto se confirma con la *Ilustración 22*, que muestra los coeficientes de regresión PLS escalados, que estiman el efecto que tiene cada *Driver* en el consumo específico.

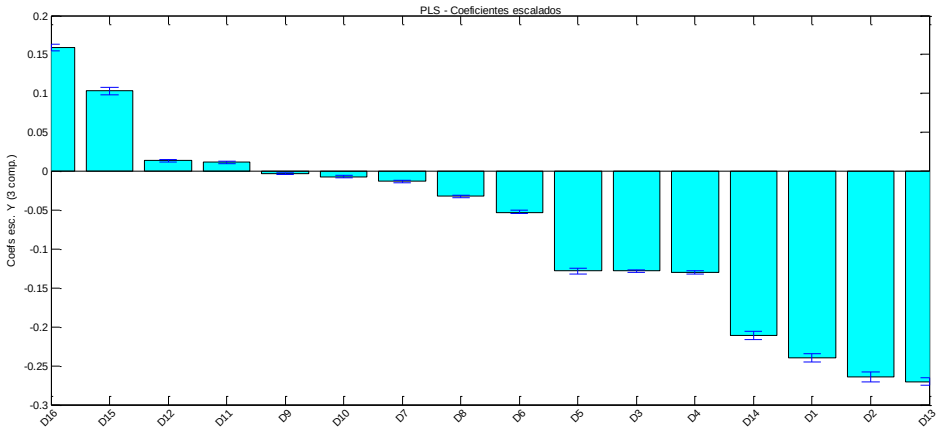

Ilustración 22.- Coeficientes escalados para el PLS-I.

La *Ilustración 23* muestra el diagrama de dispersión entre uno de los drivers con coeficiente negativo y el consumo específico, corroborando la correlación negativa entre ambos.

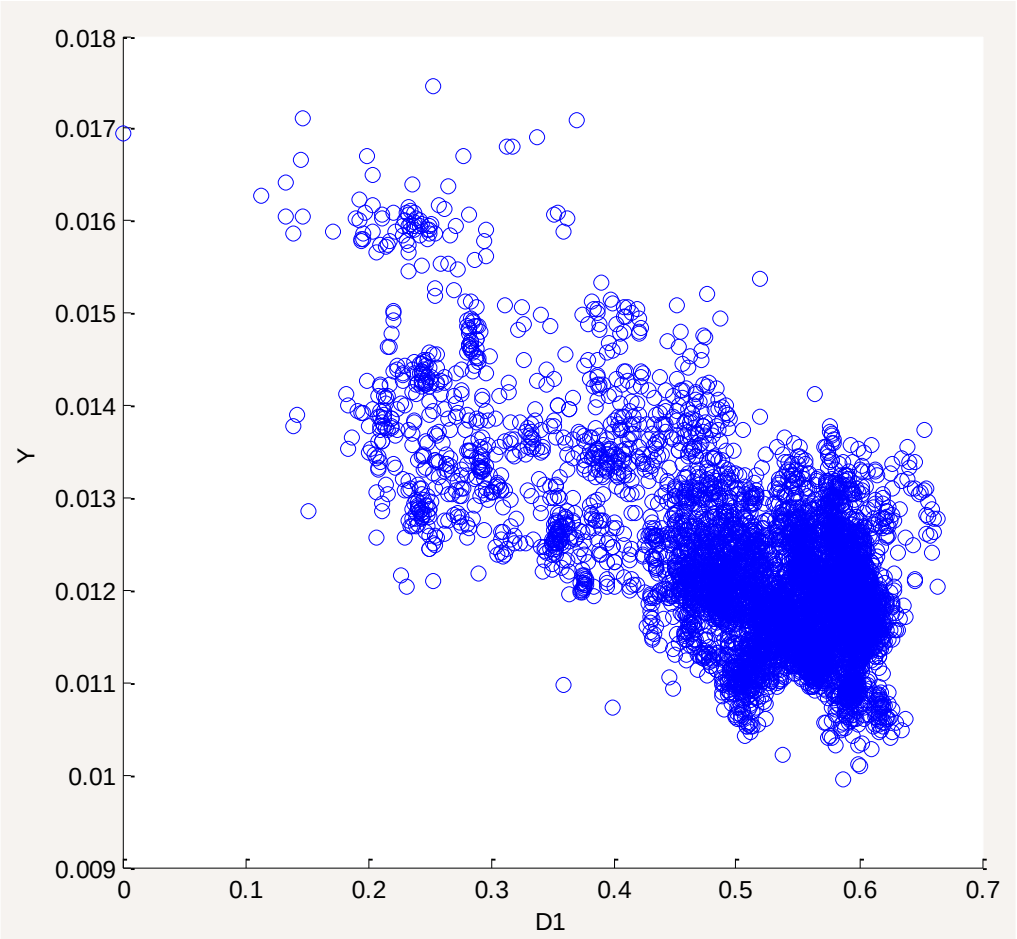

Ilustración 23.- Gráfico de dispersión Y vs. D1.

Para evaluar la calidad del modelo PLS ajustado con el conjunto de entrenamiento, se proyectan sobre el mismo las observaciones del conjunto de validación (25% de los datos). Se obtienen los siguientes gráficos (*Ilustraciones 24 y 25*):

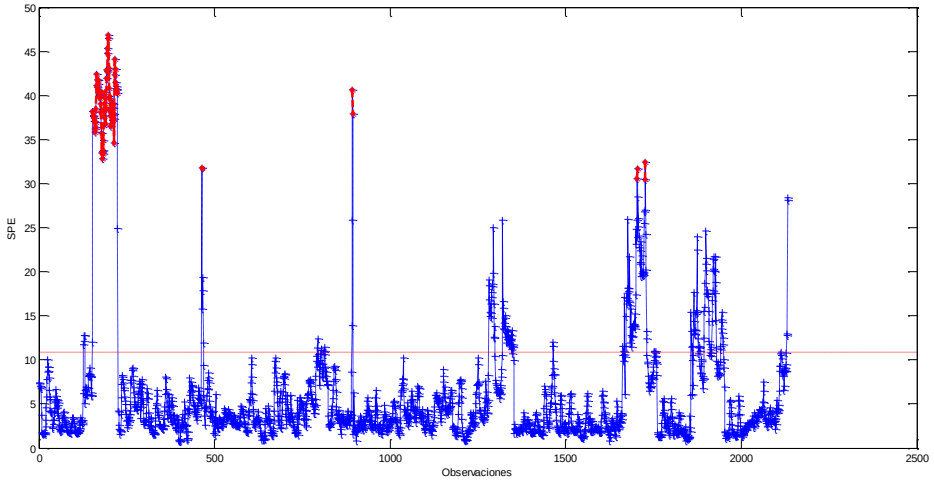

Ilustración 24.- Gráfico SPE para el set de validación del PLS-I.

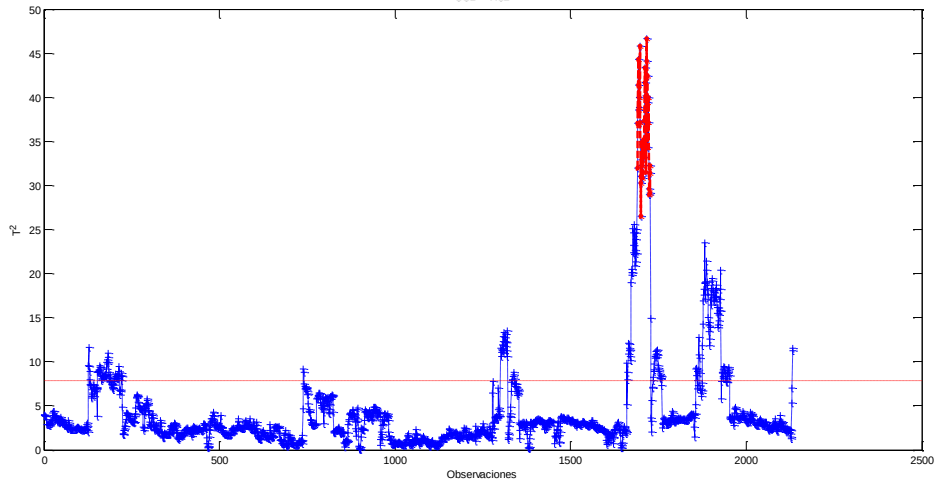

Ilustración 25.- Gráfico de T^2 para el set de validación del PLS-I.

La *Ilustración 26* muestra los valores observados y predichos:

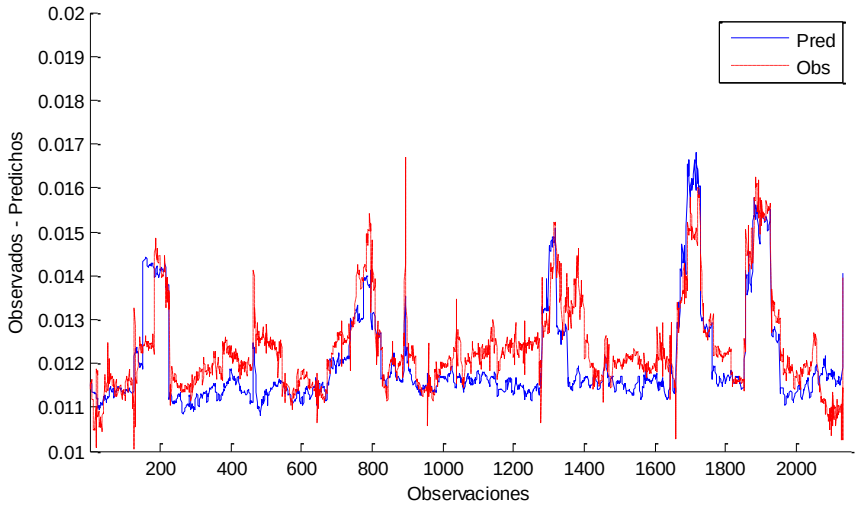

Ilustración 26.- Valores observados frente a predichos en el conjunto de validación para el modelo PLS-I.

En las *Ilustraciones 24* y *25* se muestran algunos periodos de observaciones anómalas y extremas en el conjunto de validación (marcadas en rojo). Excluyendo estos periodos en la *Ilustración 26* se observa que, en general hay periodos con discrepancias entre el consumo específico predicho por el modelo PLS-I (que usa solo los Drivers del proceso) y el real. Estas discrepancias pueden deberse a la distinta manera que los operarios del proceso han manipulado el resto de las variables (No Drivers).

En el Apéndice se muestran los resultados de los ajustes obtenidos mediante la regresión lineal múltiple y la regresión stepwise, y se comparan con los obtenidos mediante el PLS. Como se discute, la existencia de correlación en los regresores genera discrepancias en los signos y la significación estadística de algunos de los coeficientes, lo que dificulta la interpretación de los modelos de regresión.

A continuación, y del mismo modo que para el PLS-I, se lleva a cabo el PLS-II.

El modelo PLS-I explicaba un 60% de la variabilidad del consumo específico. La idea de construir un nuevo modelo PLS usando como variable respuesta los residuos del modelo PLS-I es tratar de comprobar qué variables manipulables (No-Drivers) son capaces de explicar el 40% de la variabilidad del consumo específico no explicado por los Drivers. De esta forma se podrían dar pistas a los

operadores del proceso sobre cómo operar las variables manipulables para aumentar la eficiencia energética del proceso de destilación.

La *Tabla 5* y la *Ilustración 27* muestran la bondad de ajuste y de predicción por validación cruzada para distinto número de componentes del modelo.

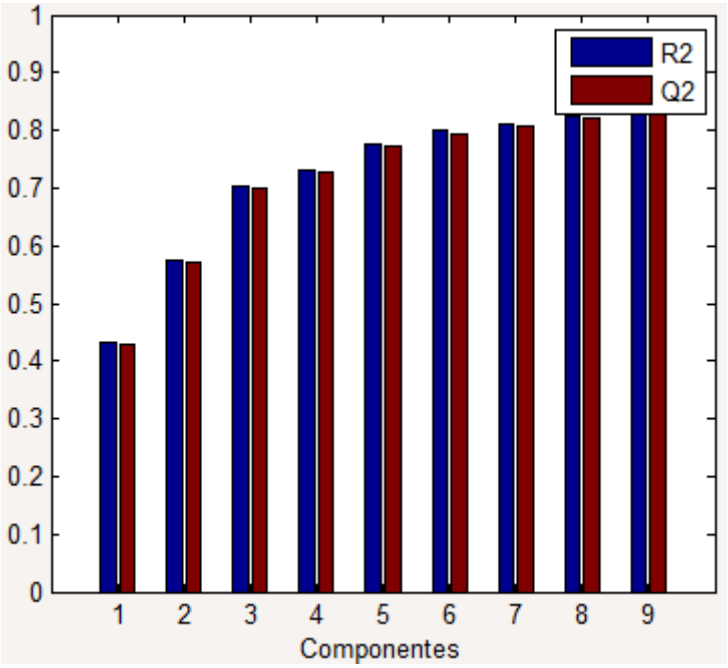

Ilustración 27.- Acumulado de los estadísticos R^2 y Q^2 para el PLS-II.

Tabla 5.- Estadísticos R2(X), R2(Y) y Q2 para el PLS-II.

R2/Q2 acumulada

A	R2(X)	R2(Y)	Q2
1	0.12938	0.21607	0.21194
2	0.23967	0.3922	0.3864
3	0.3184	0.50036	0.49429
4	0.4149	0.55825	0.55163
5	0.45019	0.62597	0.61865
6	0.49332	0.66091	0.65408
7	0.52419	0.69226	0.68629
8	0.54987	0.71901	0.71271

Si se escogen ocho componentes, una vez validado el modelo se consigue explicar más de un 70% de la variabilidad de la nueva variable respuesta: residuos del consumo específico.

A pesar de que el modelo tiene ocho componentes, con las dos primeras, que explican un 40% de la variabilidad, ya se puede apreciar una clara relación entre los residuos del consumo específico y algunas variables manipulables por los operarios (No-Drivers), como se aprecia en el gráfico de weightings de la *Ilustración 28* y en el de scores de la *Ilustración 29*:

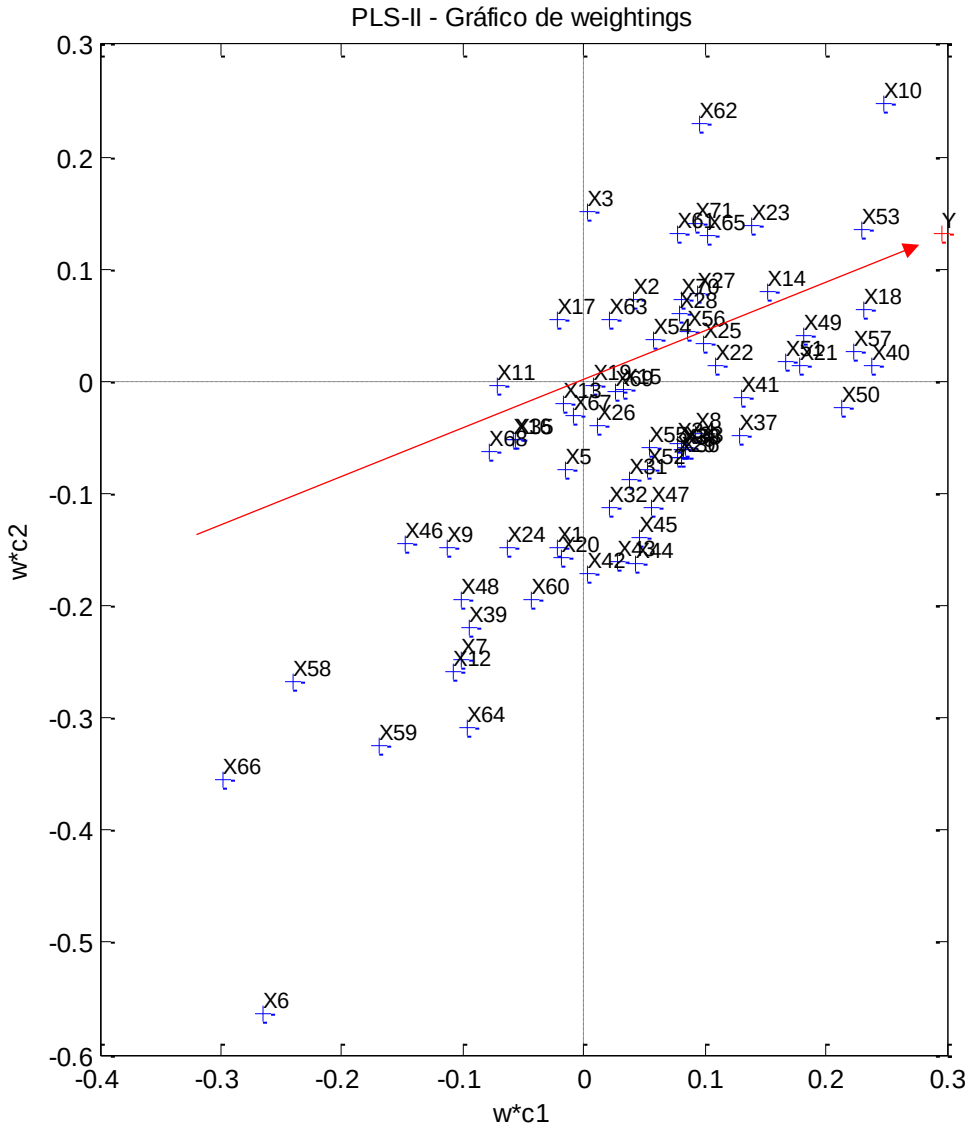

Ilustración 28.- Gráfico de weightings para el PLS-II.

El gráfico de scores de la *Ilustración 29* muestra claramente el gradiente hacia residuos positivos en la dirección Noreste indicada en la *Ilustración 28*, diferenciando periodos de alta y baja eficiencia energética.

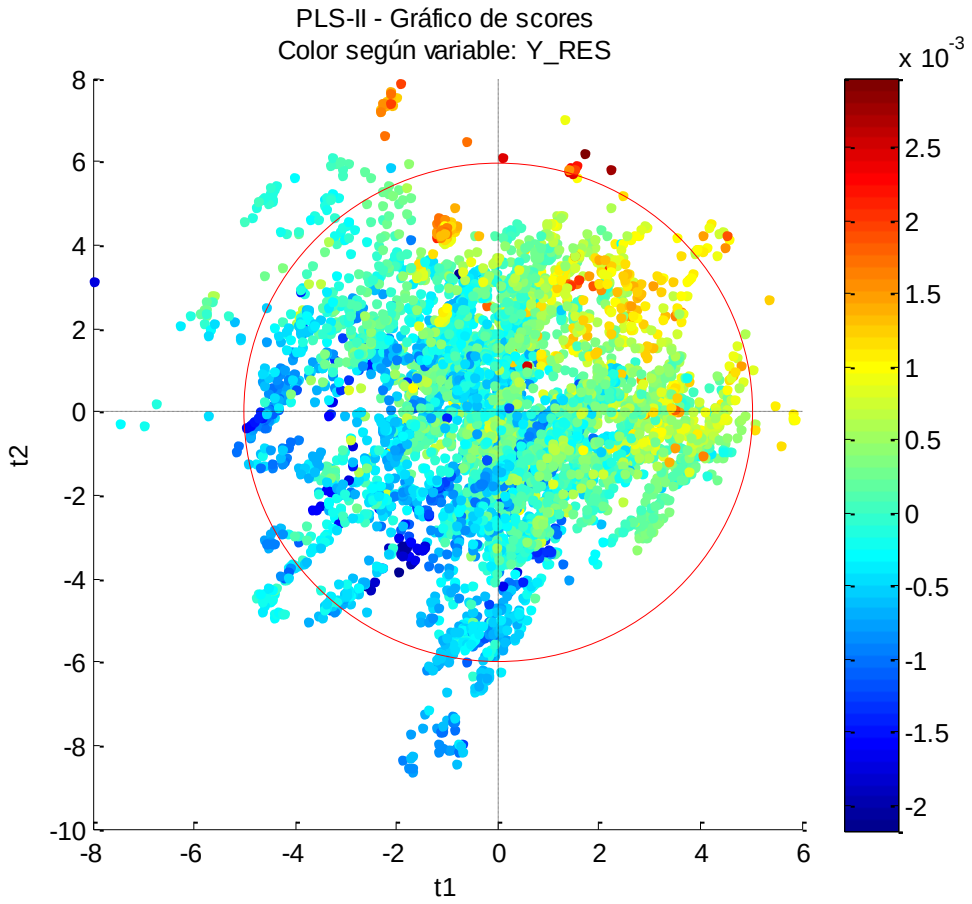

Ilustración 29.- Gráfico de scores para el PLS-II.

En este modelo, el fin es saber qué variables manipulables son más importantes en el proceso a la hora de distinguir los periodos de alta y baja eficiencia energética. Así, en la *Tabla 6* se muestran los coeficientes de las variables con VIP mayor que 1.

Tabla 6.- Coeficientes VIP para el PLS-II.

Variable	VIP	Coeficiente
X1	1,51	6,59E-03
X3	1,5	9,20E-01
X5	1,95	6,83E+00
X6	1,99	-4,68E-01
X7	1,76	-3,09E+00

La interpretación de los coeficientes del modelo es:

- Variables con coeficientes positivos: a menor valor de las variables, los residuos tienden a disminuir (ser más negativos), aumentando la eficiencia.
- Variables con coeficientes negativos: a mayor valor de las variables, los residuos tienden a disminuir (ser más negativos), aumentando la eficiencia.

A continuación, se muestran las variables con coeficientes positivos (*Ilustración 30*), con coeficientes negativos (*Ilustración 32*), así como un diagrama de dispersión entre la variable respuesta (residuos del consumo específico) y una variable con coeficiente positivo (*Ilustración 31*) y negativo (*Ilustración 33*).

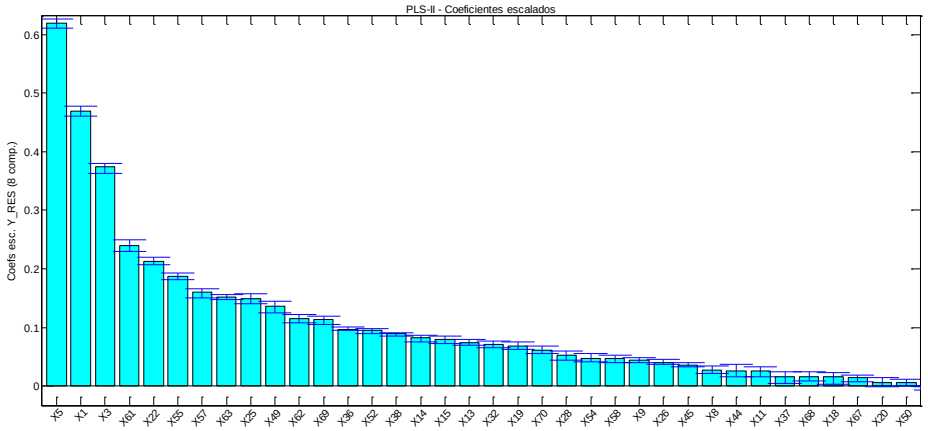

Ilustración 30.- Coeficientes escalados positivos para el PLS-II.

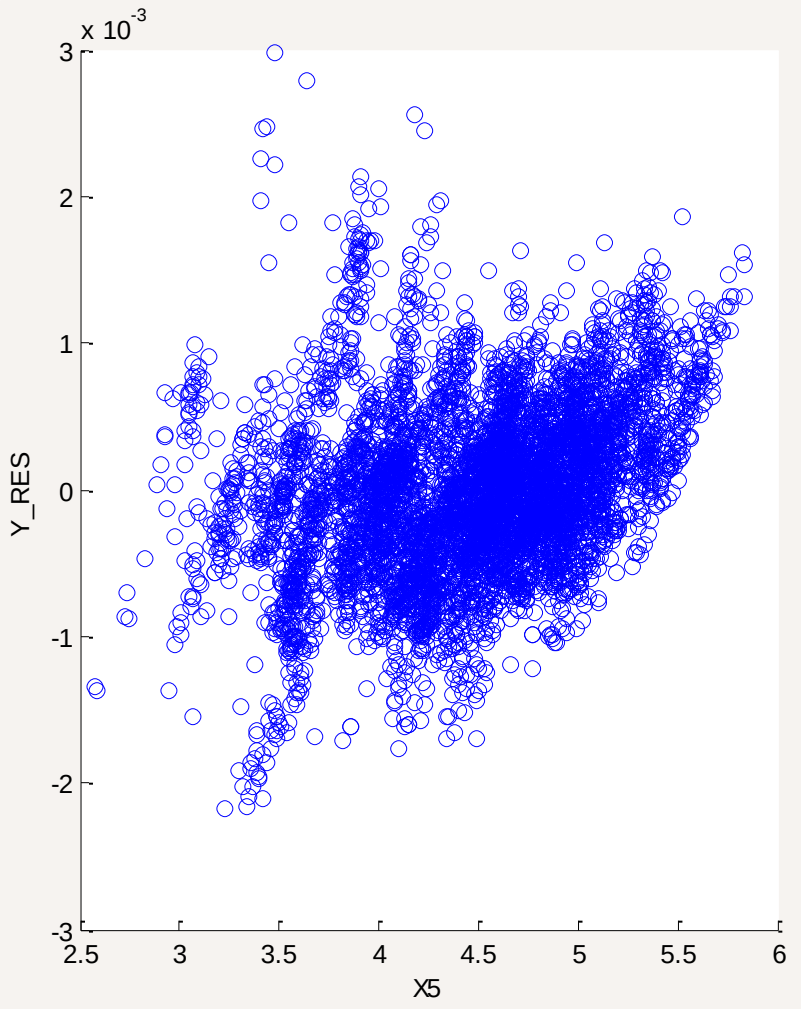

Ilustración 31.- Gráfico de dispersión entre Y_RES y X5.

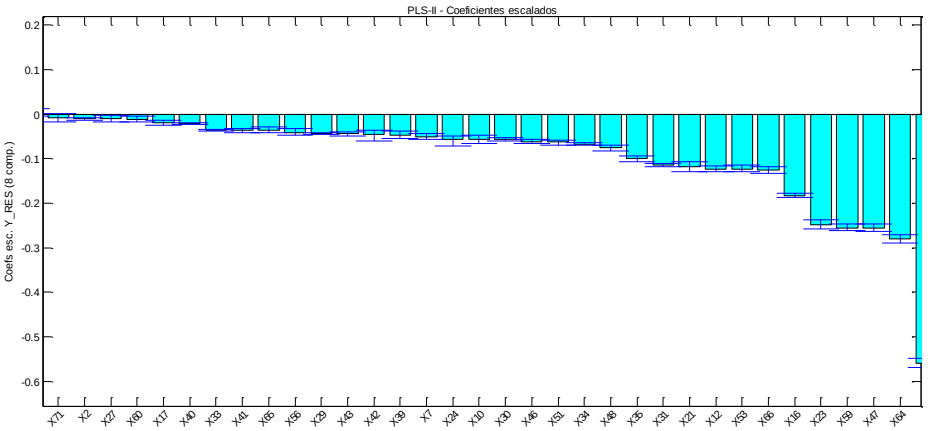

Ilustración 32.- Coeficientes escalados negativos para el PLS-II.

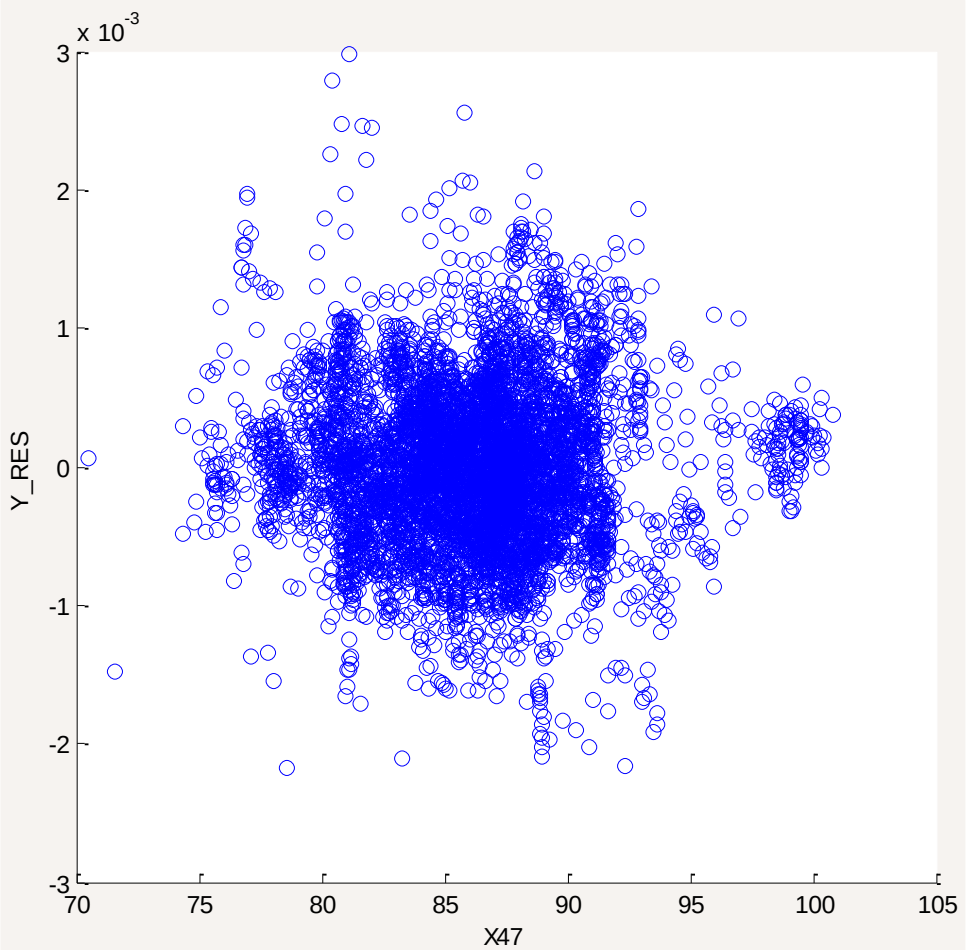

Ilustración 33.- Gráfico de dispersión para Y_RES y X47.

Es necesario destacar el gran valor añadido que aporta esta información. A partir de estos resultados, los operarios pueden saber, dados unos valores de los Drivers, qué configuración de las variables que ellos pueden manipular (No-Drivers) pueden mejorar la eficiencia energética del proceso.

Una vez realizado el PLS, es posible mejorarlo. Observando los intervalos de confianza Jacknife al 95% mostrados en el gráfico de coeficientes, se pueden descubrir las variables que son estadísticamente significativas. Estás serán, aquellas cuyo intervalo de confianza no contenga al cero, y se podrá afirmar con un riesgo de primera especie del 5% que serán influyentes sobre la variable respuesta.

De este modo, haciendo zoom en el gráfico se eliminan las variables destacadas en rojo de la *Ilustración 34* y se repite el análisis:

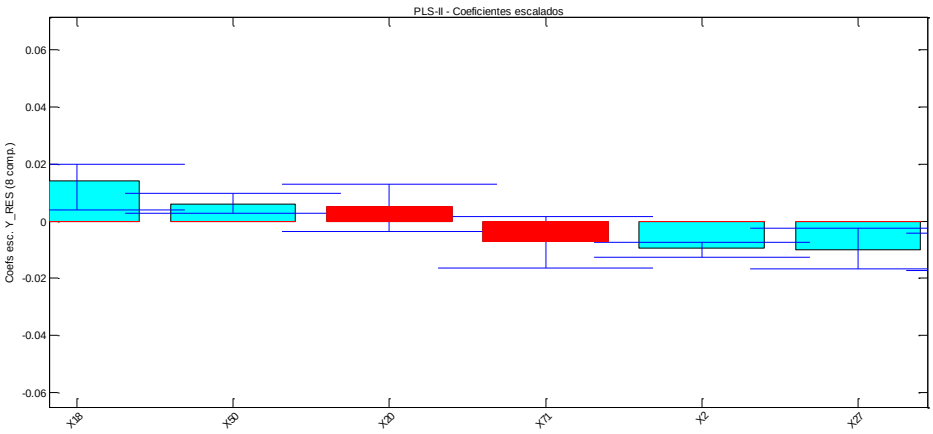

Ilustración 34.- Visión aumentada del gráfico de coeficientes.

Al contrastar los resultados obtenidos en ambos casos, se observa que, si bien la variabilidad de las X se explica mejor, únicamente se produce en este sentido la mejora. Esto se debe a que no vale la pena eliminar dos variables de un total de 88.

3.5 PLS-DA (Discriminant Analysis)

El gráfico de residuos del modelo PLS-I (consumo específico en función de Drivers) es el mostrado en la *Ilustración 35*:

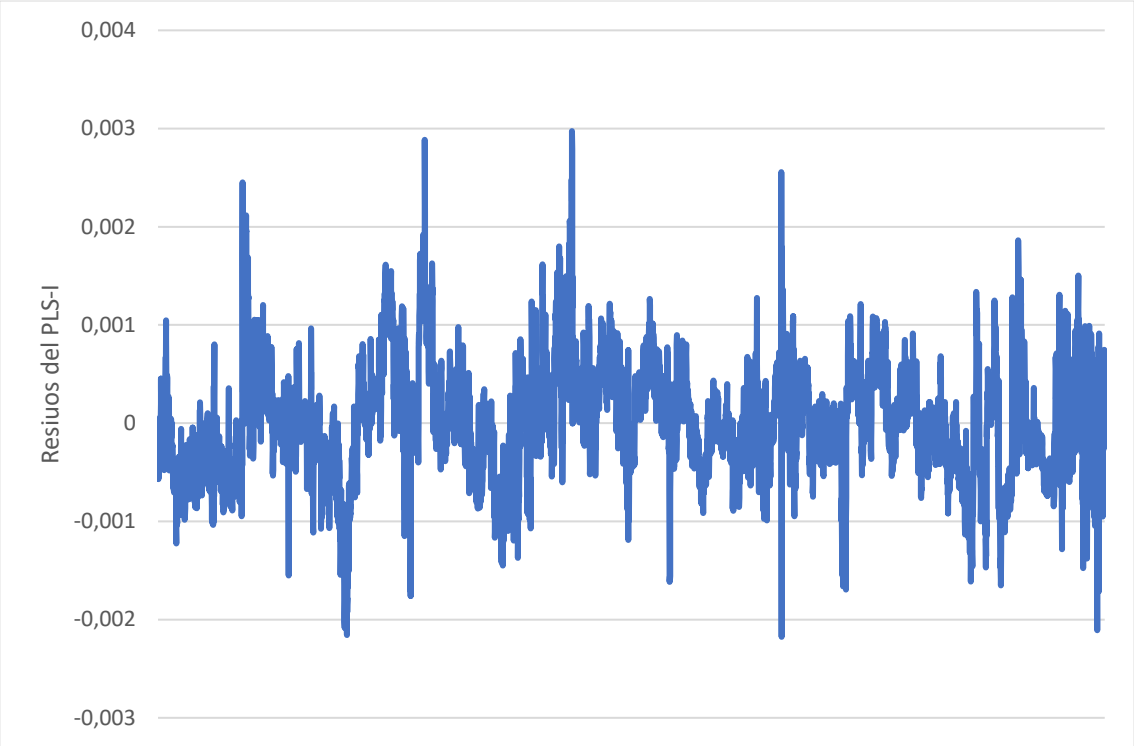

Ilustración 35.- Serie de tiempo de los residuos del PLS-I.

Como ya se ha comentado previamente, residuos positivos indican consumos específicos observados mayores que los predichos, por tanto, periodos de baja eficiencia; mientras que residuos negativos indican consumos específicos observados menores que los predichos, por tanto, periodos de operación eficiente.

Para discriminar qué variables operativas tienen un comportamiento distinto en las situaciones de alta/baja eficiencia (residuos muy negativos/muy positivos, respectivamente), se realizará un PLS-DA de las observaciones con ambos tipos residuos frente a las variables operativas.

Para ello, una vez calculados los residuos de cada observación, estas se clasifican en dos grupos: observaciones con consumos muy superiores a los esperados (residuos muy positivos, baja eficiencia), observaciones con consumos muy inferiores a los esperados (residuos muy negativos, alta eficiencia). Así se categorizarán las observaciones en “Baja eficiencia” y “Alta eficiencia”.

Lo que hay que plantearse es qué significa muy superiores o inferiores. Para ello se proponen dos casos. En el primero, estos límites serán el primer y el tercer cuartil. En el segundo caso se va a trabajar con los percentiles 10% y 90%. Como en el caso anterior, también se dividirá el conjunto de observaciones en un conjunto de entrenamiento (75% de los datos) y otro de validación (25% restante).

- Primer y tercer cuartiles

Teniendo en cuenta que la media tiene un valor de 1.9897e-05, los cuartiles toman estos valores (ver *Tabla 7*):

Tabla 7.- Valores de los cuartiles.

Q1	**-3,4696e-04**
Q3	3,4694e-04

Por tanto, el gráfico resultante es el siguiente (*Ilustración 36*):

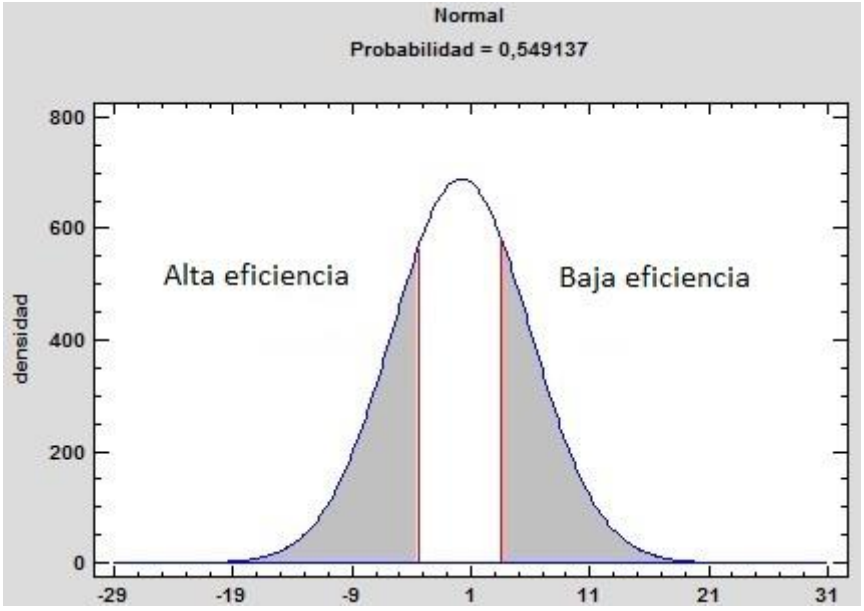

Ilustración 36.- Gráfico de los cuartiles para el PLS-DA.

Escogiendo cinco componentes se consigue un grado de acierto del 91.18% (ver *Tabla 8*), con 91 valores que no se clasifican en ninguna categoría porque son atípicos:

Tabla 8.- Tabla de clasificación para el PLS-DA con los cuartiles.

Tabla de clasificación

Y_Res	Alta eficiencia	Baja eficiencia	<ATIPICO>
Alta eficiencia	941	50	49
Baja eficiencia	53	1065	42

% aciertos: 91.18%

En el gráfico de *weightings* (*Ilustración 37*), se pueden observar las mismas características que en el PLS-II. Es decir, valores altos de variables como X3 provocarán residuos altos. Como lo que se trata es de obtener residuos bajos (consumo por debajo de la predicción), será preferible que variables como estas tomen valores inferiores a su media. Sin embargo, para las variables con pesos negativos, interesará operarlas a valores superiores a su media para generar residuos negativos (consumos menores que los predichos).

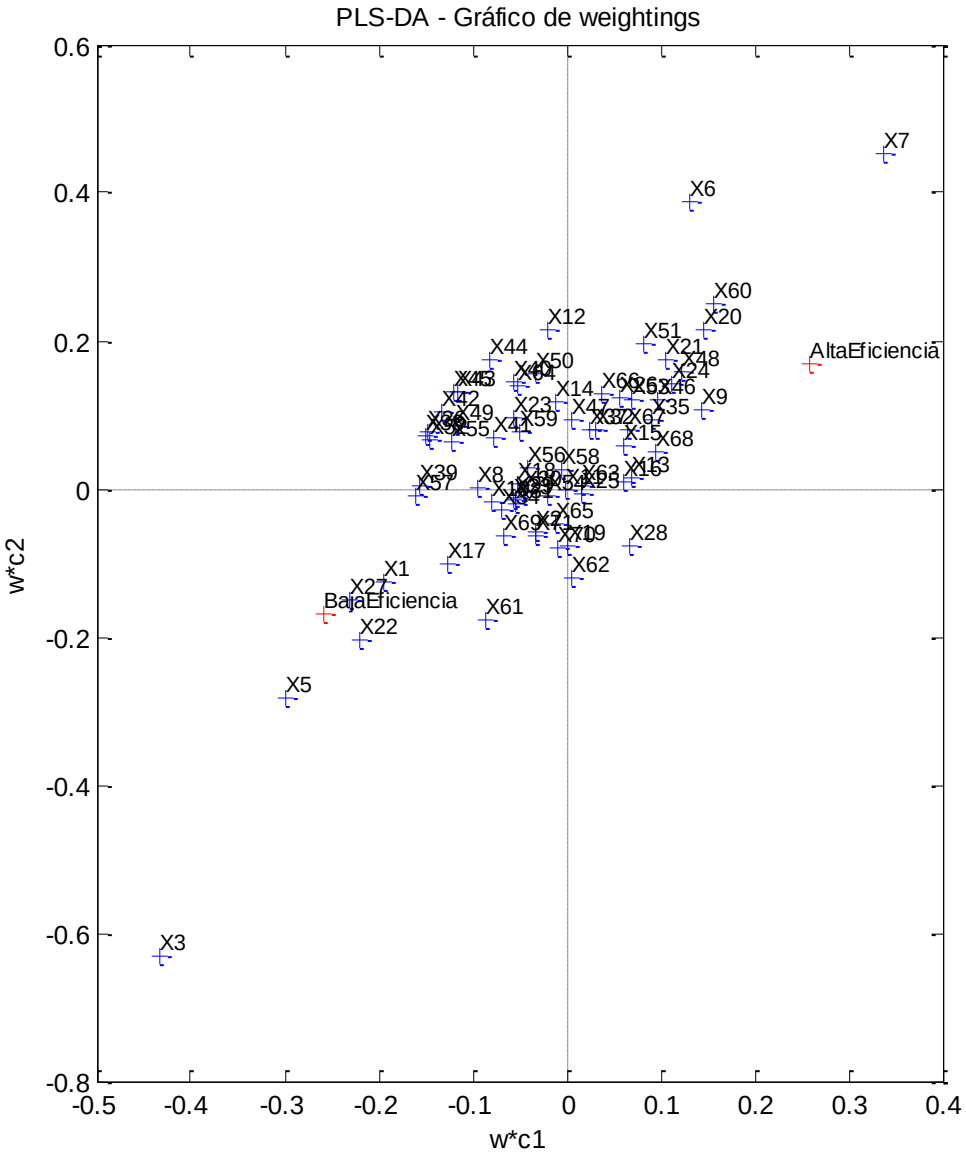

Ilustración 37.- Gráfico de weightings para el PLS-DA con los cuartiles.

Además, en el gráfico de scores de la *Ilustración 38*, se puede apreciar una razonable distinción entre los residuos “Baja eficiencia” y “Alta eficiencia”:

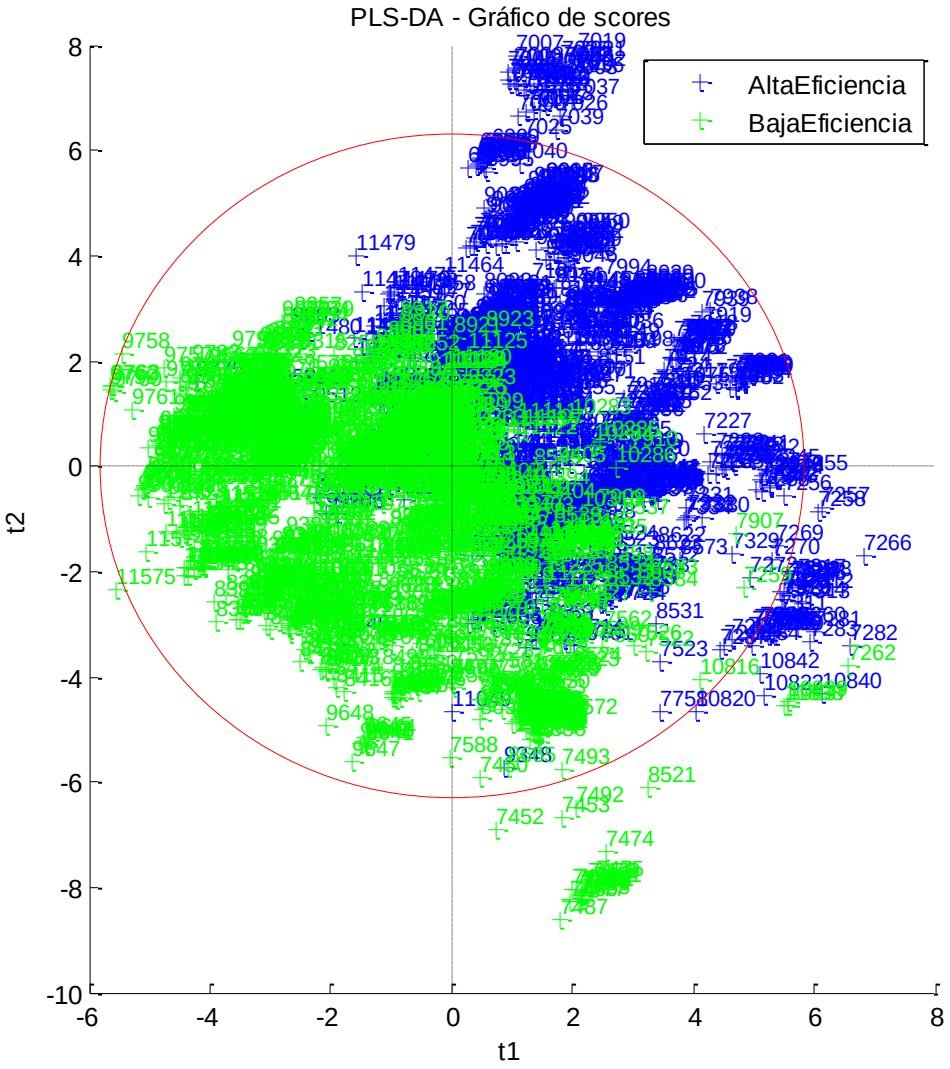

Ilustración 38.- Gráfico de scores para el PLS-DA con los cuartiles.

En la *Ilustración 39* así como en la *42* se muestran los coeficientes ordenados del modelo.

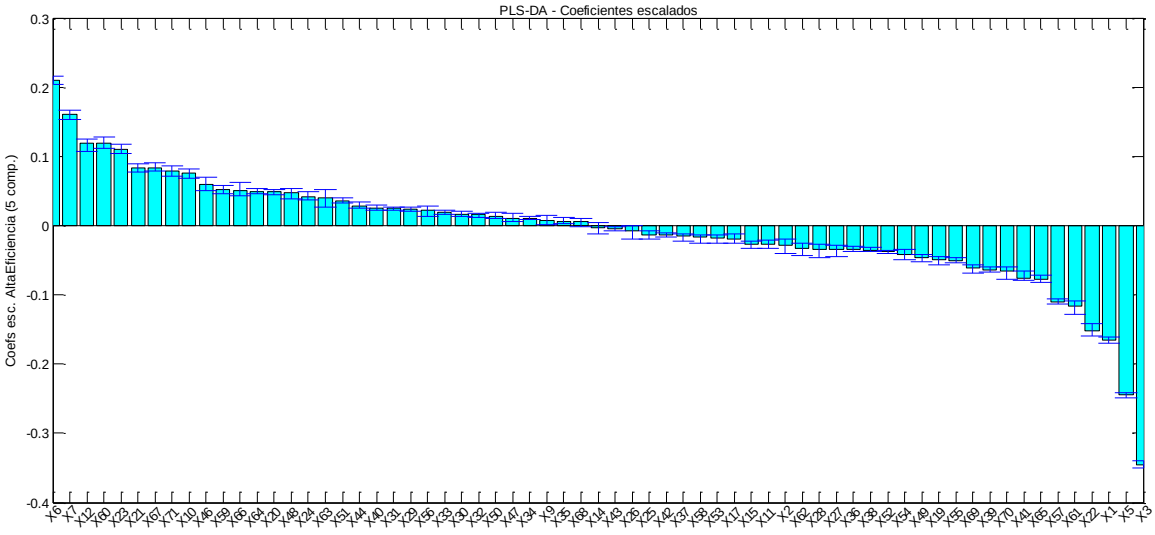

Ilustración 39.- Coeficientes del modelo PLS-DA para predecir la categoría Alta Eficiencia con 5 componentes (Cuartiles).

- Percentiles 10% y 90%

Para este caso, los percentiles toman los valores (*Tabla 9*):

Tabla 9.- Valores de los percentiles.

P10	**-6.5613e-04**
P90	7.3008e-04

Resultando este gráfico de distribución de la *Ilustración 40*:

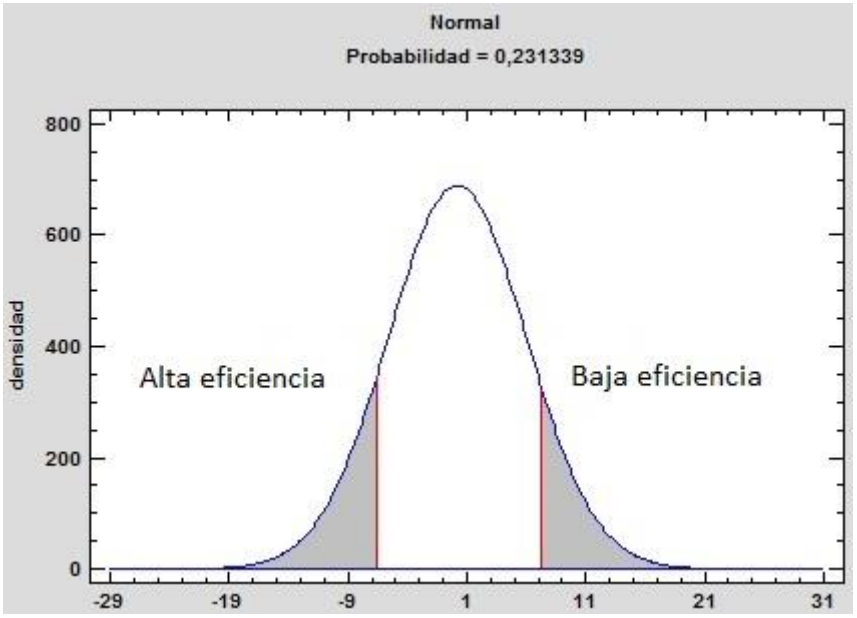

Ilustración 40.- Gráfico de los percentiles para el PLS-DA.

Escogiendo como en el caso anterior, cinco componentes, se obtiene una mejor tasa de acierto, un 95% (ver *Tabla 10*).

Tabla 10.- Tabla de clasificación para el PLS-DA con los percentiles.

Tabla de clasificación

Y_Res	Alta eficiencia	Baja eficiencia	<ATIPICO>
Alta eficiencia	417	4	13
Baja eficiencia	12	438	16

% aciertos: 95.00%

Esto es algo comprensible puesto que ahora las categorías están más separadas y por tanto más fácilmente diferenciables. En cuanto a los resultados, tanto el gráfico de scores como el de *weightings* (no se muestra, pues es similar al de la *Ilustración 37*) dan a entender las mismas conclusiones. Las variables que resultan decisivas son las mismas y habrá que modificarlas en el mismo sentido.

Además, en el gráfico de scores (*Ilustración 41*) se observa una diferenciación incluso mejor que en el caso de los cuartiles.

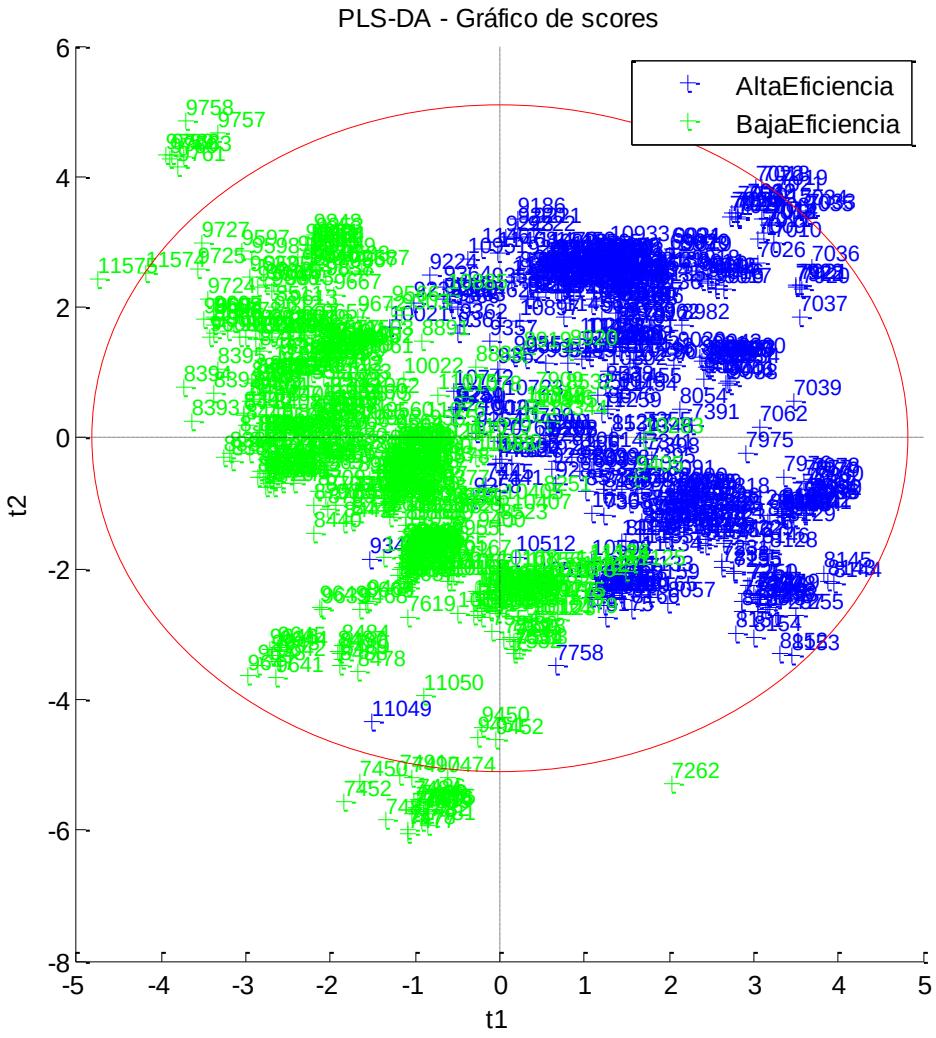

Ilustración 41.- Gráfico de scores para el PLS-DA con los percentiles.

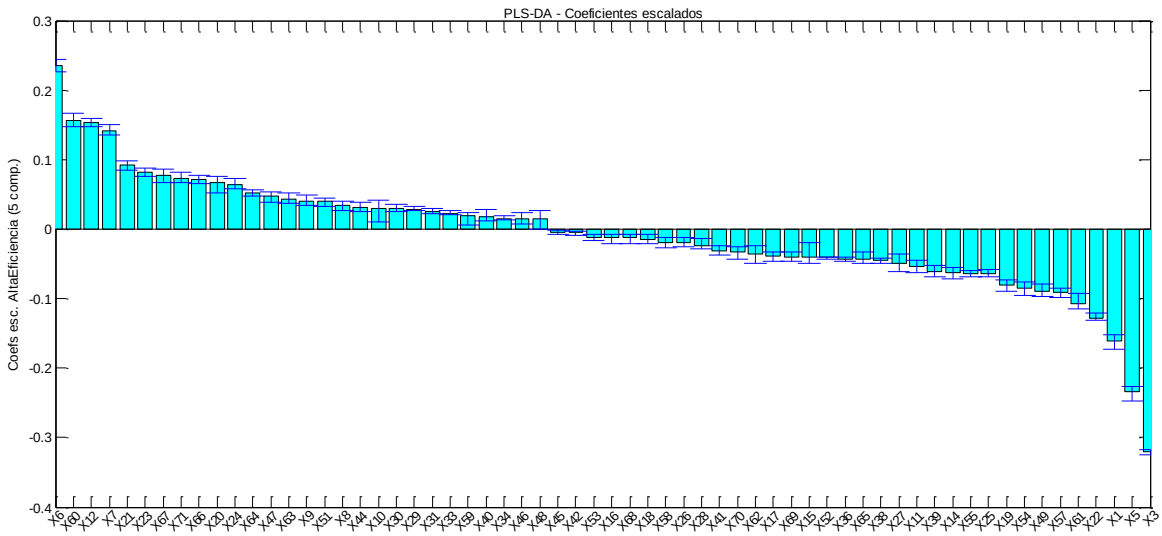

Ilustración 42.- Coeficientes del modelo PLS-DA para predecir la categoría Alta Eficiencia con 5 componentes (Percentiles).

A fin de comparar los modelos PLS-DA con el PLS-II, se han mejorado ambos excluyendo las variables no significativas (aquellas cuyos intervalos Jacknife al 95% asociados a sus coeficientes contienen el cero). Estas han sido las siguientes:

- Modelo con los cuartiles: X8, X13, X14, X16, X18, X45, X68.
- Modelo con los percentiles: X2, X13, X32, X35, X37, X43, X50, X56.

De este modo, además de conseguir una tasa de aciertos superior, también se consigue un gráfico de weightings algo más sencillo de interpretar.

Así, en la *Tablas 11* y *12* se recogen en orden de importancia descendente, las variables que según los tres modelos (PLS-II, PLS-DA *cuartiles* y PLS-DA *percentiles*) favorecen que haya condiciones tanto de “Alta eficiencia” como de “Baja eficiencia”. Es decir, las variables que sean significativas en los tres modelos, tanto con coeficientes positivos como negativos.

Tabla 11.- Variables que favorecen condiciones de Alta eficiencia para los distintos modelos.

PLS-II	PLSDA Cuartiles	PLSDA Percentiles
X6	X6	X6
X64	X7	X60
X47	X12	X12
X23	X60	X7
X59	X23	X21
X16	X21	X23
X66	X67	X67
X53	X71	X71
X12	X10	X66
X31	X46	X20
X21	X59	X24
X35	X66	X64
X48	X64	X47
X34	X20	X63
X24	X48	X9
X10	X24	X51
X30	X63	X8
X39	X51	X44
X46	X44	X10
X51	X40	X30
X42	X31	X29
X7	X29	X31
X65	X56	X33
X43	X33	X59
X29	X30	X40
X56	X32	X34
X33	X50	X46
X17	X47	X48
X41	X34	
X40	X9	
X27	X35	
X60	X68	
X2		

Las celdas resaltadas en azul claro son aquellas variables en común para los distintos modelos. Así, se puede concluir que los tres enfoques llegan a resultados similares.

Tabla 12.- Variables que favorecen condiciones de Baja eficiencia para los distintos modelos.

PLS-II	PLSDA Cuartiles	PLSDA Percentiles
X3	X3	X5
X5	X5	X1
X1	X1	X3
X22	X22	X61
X61	X61	X22
X57	X57	X55
X65	X49	X57
X41	X54	X63
X70	X19	X25
X39	X25	X49
X64	X55	X69
X55	X14	X62
X19	X39	X36
X49	X11	X52
X54	X27	X15
X52	X38	X14
X38	X65	X38
X36	X36	X13
X27	X52	X32
X28	X15	X19
X62	X69	X70
X2	X17	X54
X11	X62	X28
X15	X70	X9
X17	X41	X58
X53	X28	X26
X58	X26	X45
X37	X58	X44
X42	X18	X8
X25	X68	X11
X26	X16	X67
	X68	X68
	X16	X50
	X68	
	X53	
	X42	

Del mismo modo, para las variables cuyos valores altos provocan una baja eficiencia, también se obtienen resultados muy parecidos entre los modelos. Las variables con más capacidad discriminante son las que tienen coeficientes

mayores (en valor absoluto) en el modelo que predice la categoría “Alta eficiencia”. Las de coeficientes positivos deberían operarse a valores mayores que su media, mientras que las de coeficientes negativos deberían operarse a valores inferiores. Por tanto, un valor elevado de la variable *X6* provocará una alta eficiencia mientras que valores elevados de las variables *X1*, *X3* o *X5* provocarán una baja eficiencia.

3.6 Árboles de decisión

A fin de comparar con alguna otra técnica de clasificación/discriminación del ámbito de la inteligencia artificial que pueda proporcionar cierta información discriminante, se ha propuesto usar los árboles de decisión con el objetivo de comprobar cuáles son las variables que tienen poder discriminante sobre la eficiencia energética

De este modo, en el software *MVA-GIEM* se ha empleado el algoritmo de partición *PullLeft* con un ratio de entrenamiento del 0.7.

Prestando atención a los primeros nodos del árbol, se puede corroborar en la *Ilustración 43* que efectivamente, el árbol de decisión ofrece resultados coherentes con los obtenidos mediante PLS:

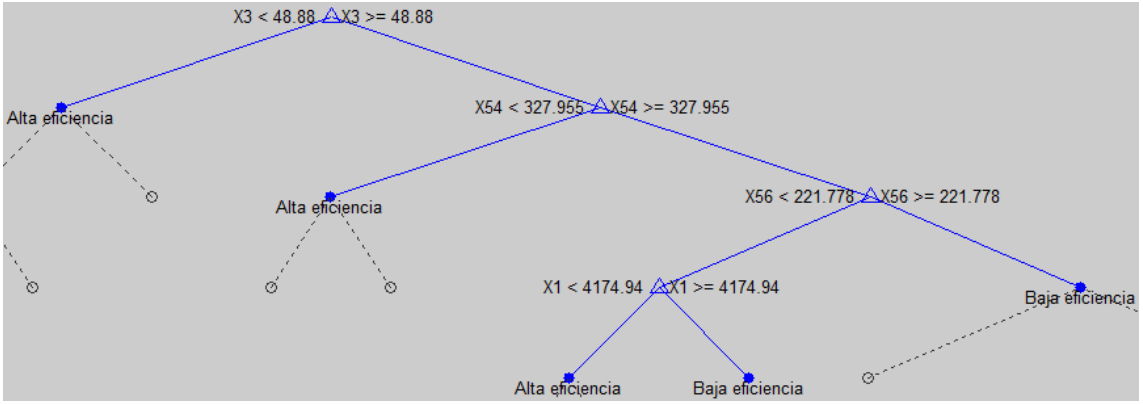

Ilustración 43.- Árbol de decisión.

Así, cuando variables como *X1*, *X3*, *X54*, y *X56* toman valores altos, producen consumos de baja eficiencia. Esto es coherente con los resultados del modelo PLS-II donde estas tres de estas variables (la *X56* no coincide) tienen

coeficientes positivos (*Ilustración 39*). Del mismo modo, en los dos gráficos de coeficientes del modelo PLS-DA (tanto para cuartiles como para percentiles) ocurre lo mismo, estas variables tienen coeficientes negativos para predecir la categoría Alta Eficiencia.

Sin embargo, debido a la propia estructura del árbol (semejante a una selección de variables) los resultados son mucho menos informativos que los obtenidos con el PLS.

3.7 Modelización de alta eficiencia

Para este estudio se intenta obtener un modelo que prediga cuál es el consumo específico mínimo (eficiencia máxima) dadas unas condiciones del proceso prefijadas (es decir, unos *drivers*). Como ya se ha comentado, los periodos de alta eficiencia corresponden a los que tienen residuos más negativos en el modelo PLS construido con los Drivers del proceso (PLS-I).

Para obtener las observaciones correspondientes a estos periodos, se volverá a realizar un filtrado de la base de datos seleccionando las observaciones con residuos más negativos En este caso se analizarán dos escenarios:

- Residuos inferiores a -1 veces la desviación típica: 632 observaciones.
- Residuos inferiores a -1.5 veces la desviación típica: 228 observaciones.

En ambos casos se ajustará un modelo PLS explicando el consumo específico en función de los *drivers*.

Modelo con residuos inferiores a -1s

Tras ajustar el modelo con tres componentes, se consigue explicar más de un 82% de la variabilidad (*Ilustración 44*). En el gráfico SPE se ha eliminado la observación 10842 mientras que en el gráfico T^2 no se han detectado observaciones anómalas (*Ilustraciones 45 y 48*).

R2/Q2 acumulada

A	R2(X)	R2(Y)	Q2
1	0.58426	0.56014	0.55645
2	0.69285	0.7842	0.77737
3	0.74734	0.8285	0.82226

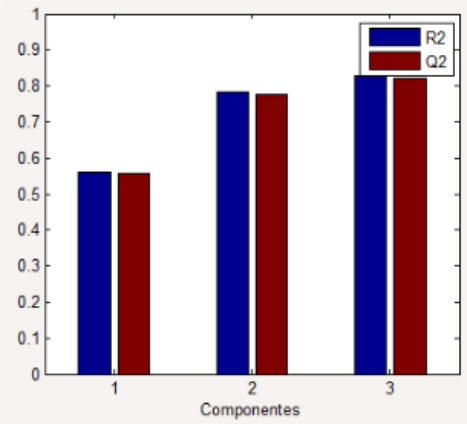

Ilustración 44.- R^2 y Q^2 para el modelo con residuos inferiores a -1s.

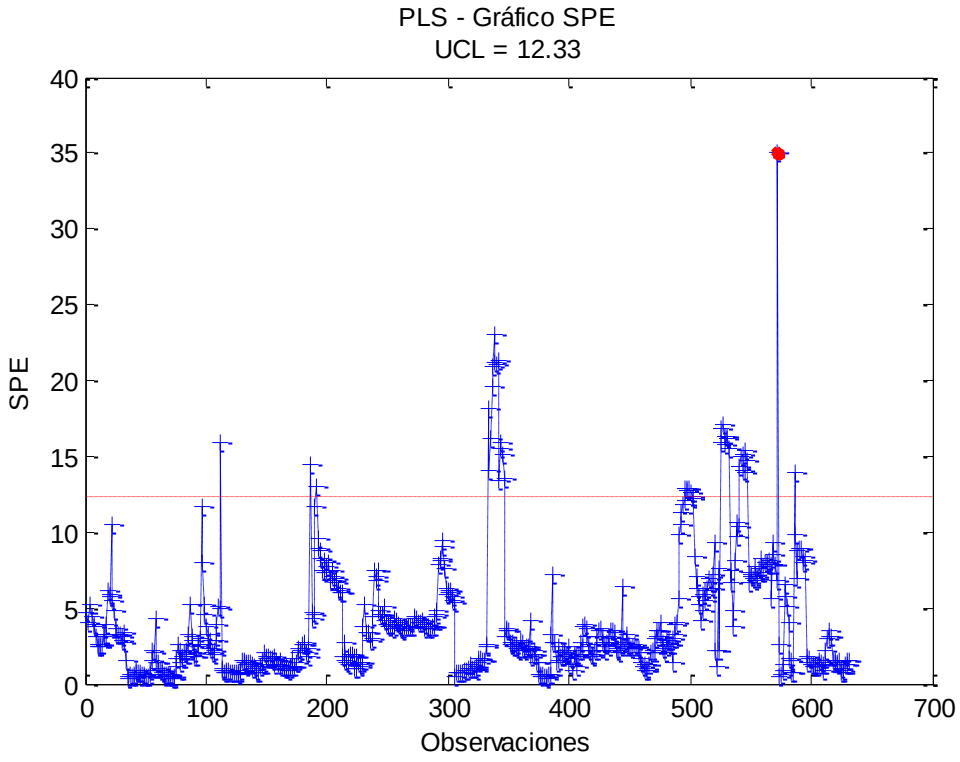

Ilustración 45.- Gráfico SPE para el modelo con residuos inferiores a -1s.

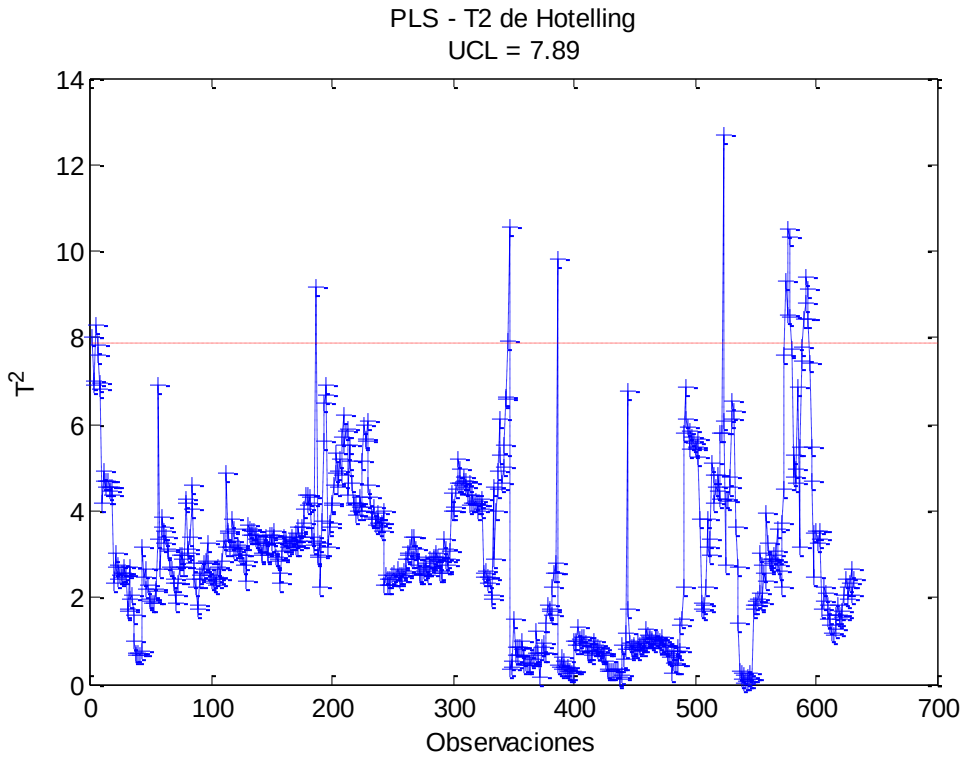

Ilustración 46.- Gráfico T^2 de Hotelling para el modelo con residuos inferiores a -1s.

Con este modelo, el ajuste es bastante bueno, como indica la *Ilustración 47*:

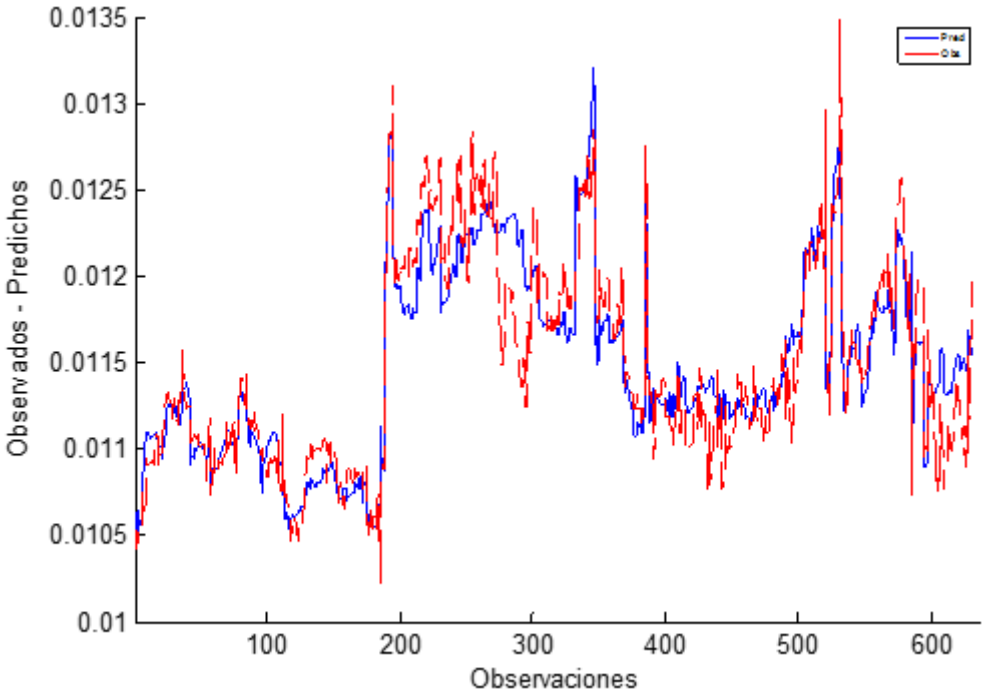

Ilustración 47.- Gráfico observados frente a predichos para el modelo con residuos inferiores a -1s.

A continuación, se utilizará el modelo obtenido para predecir todos los datos, tanto del conjunto de entrenamiento como del de validación. Como era de esperar, en las *Ilustraciones 48 y 49* se observa que, los valores observados tienden a estar por encima de los predichos ya que éstos últimos son los mejores posibles dadas unas condiciones del proceso. Estos gráficos, usados en tiempo real, permitirían a los técnicos del proceso saber cuál sería la eficiencia máxima (consumo específico mínimo) para unas condiciones de operación definidas por los Drivers determinada.

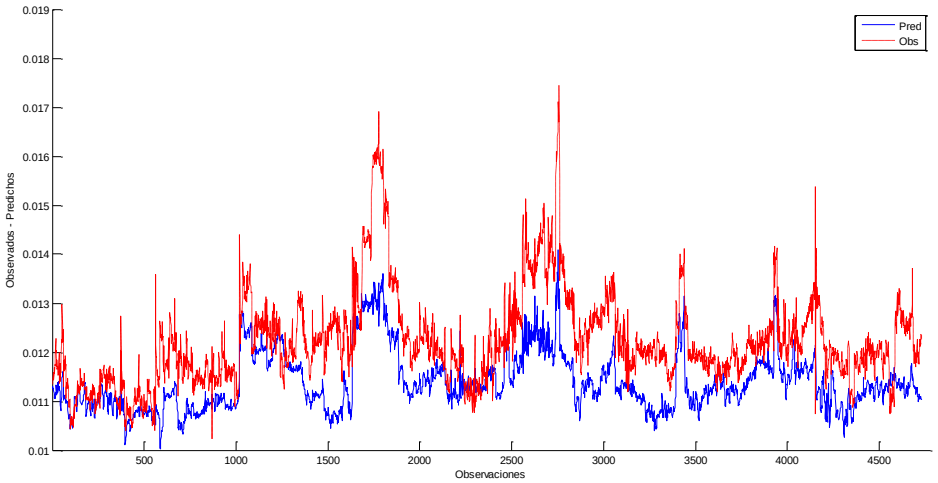

Ilustración 48.- Predicción para el conjunto de datos de entrenamiento.

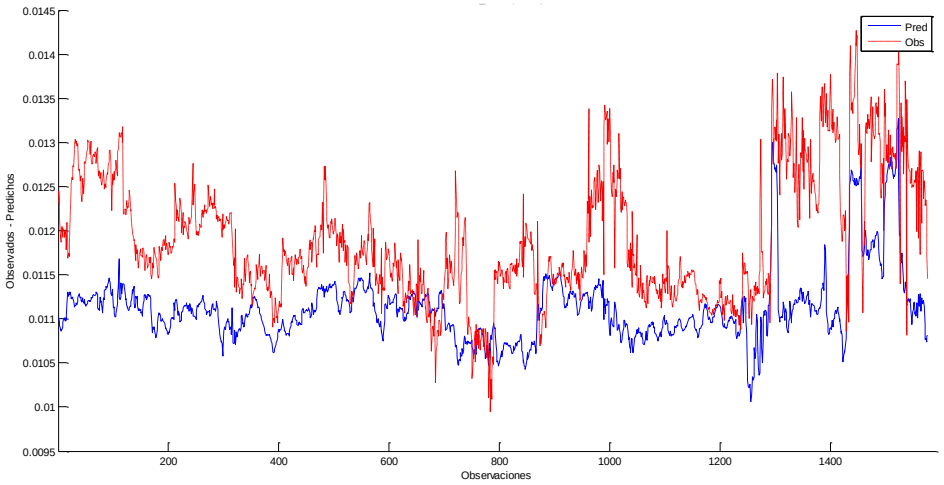

Ilustración 49.- Predicción para el conjunto de datos de validación.

Modelo con residuos inferiores a -1.5s

Para este caso se repite el estudio, pero ahora con los residuos inferiores a -1.5 veces la desviación típica. Se consigue explicar un 83% de la variabilidad con cuatro componentes (*Ilustración 50*).

R2/Q2 acumulada

A	R2(X)	R2(Y)	Q2
1	0.62871	0.64838	0.64287
2	0.72199	0.78803	0.77327
3	0.76427	0.82321	0.80508
4	0.81877	0.83194	0.81795

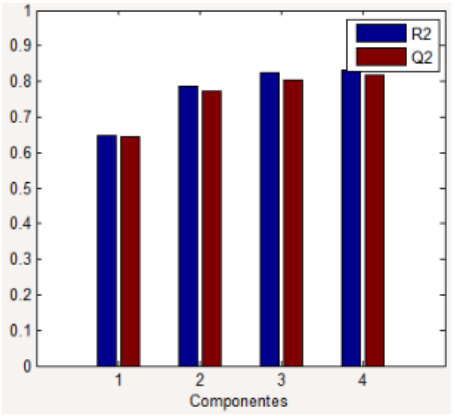

Ilustración 50.- R2 y Q2 para el modelo con residuos inferiores a -1.5s.

Si bien en el gráfico SPE de la *Ilustración 51* no aparece ninguna observación atípica, en el gráfico T^2 de la *Ilustración 52* sí que se ha detectado la observación 7758 como anómala:

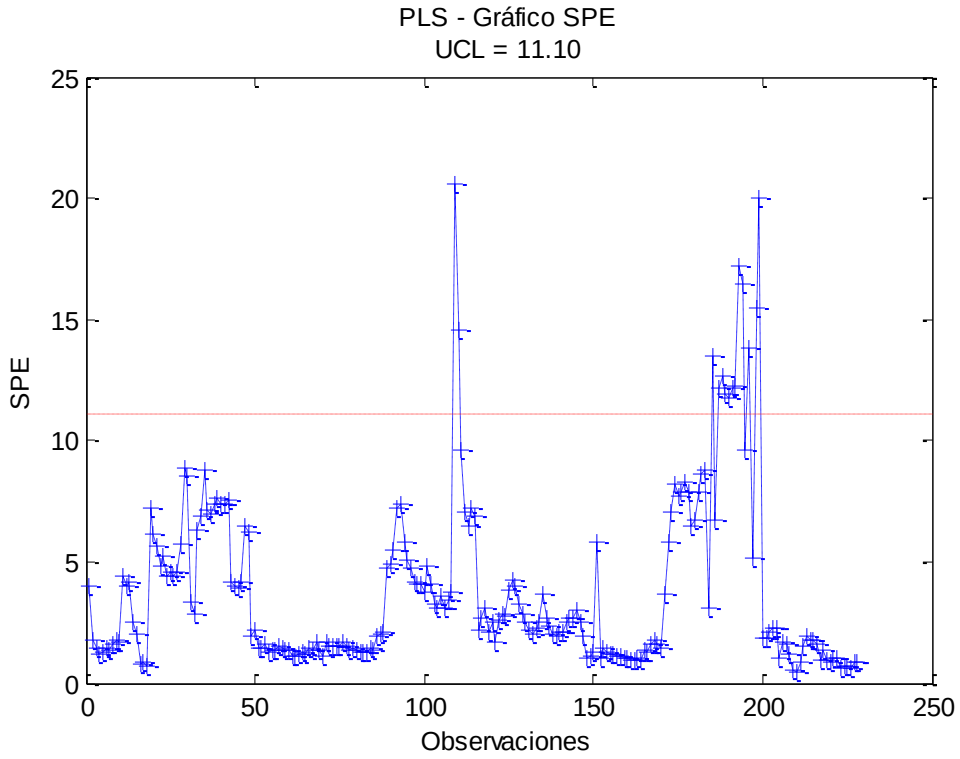

Ilustración 51.- Gráfico SPE para el modelo con residuos inferiores a -1.5s.

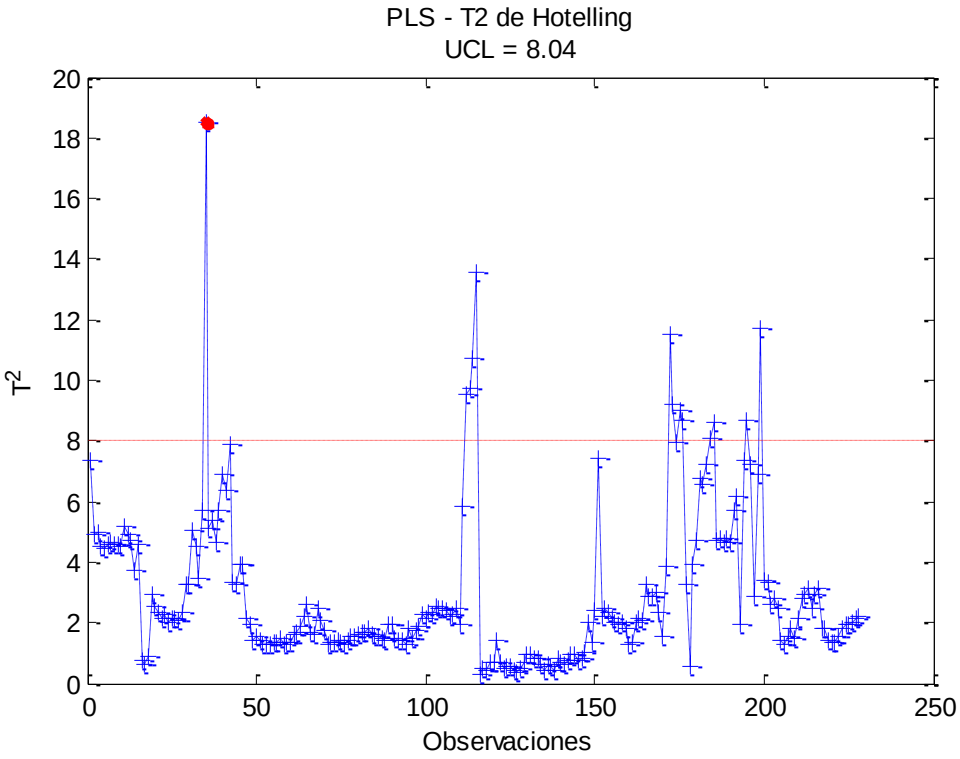

Ilustración 52.- Gráfico T2 para el modelo con residuos inferiores a -1.5s.

Del mismo modo que antes, en las *Ilustraciones 53 y 54* se comprueba que las predicciones, tanto para el conjunto de validación como para el de entrenamiento, son menores que los consumos observados:

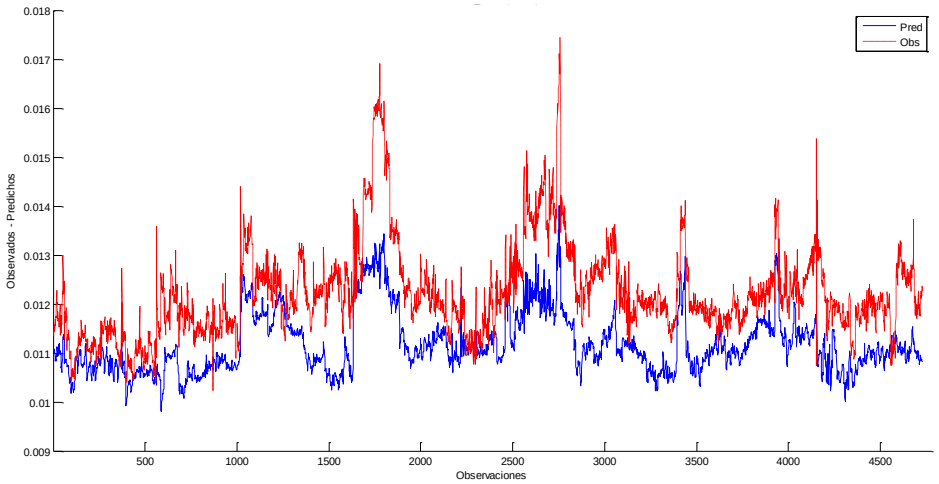

Ilustración 53.- Predicciones para el conjunto de datos de entrenamiento.

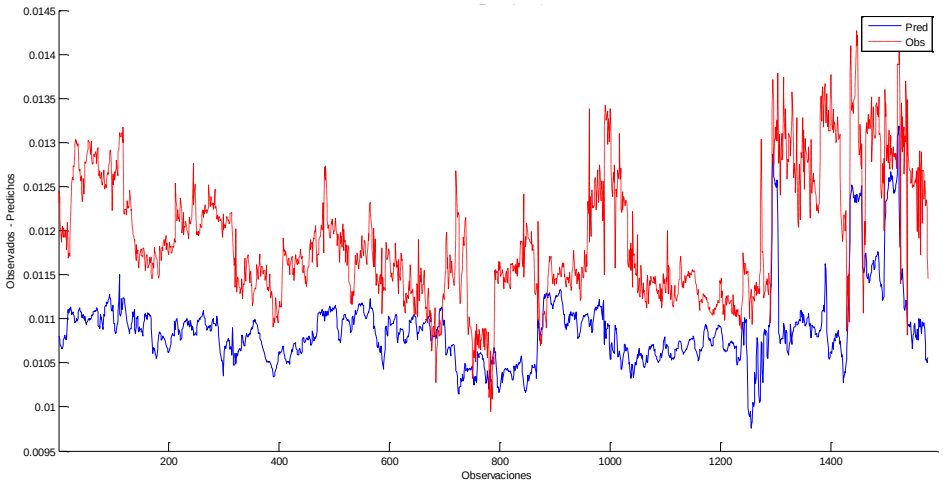

Ilustración 54.- Predicciones para el conjunto de datos de validación.

Por último, se comparan entre sí las predicciones de los modelos (*Ilustraciones 55 y 56*):

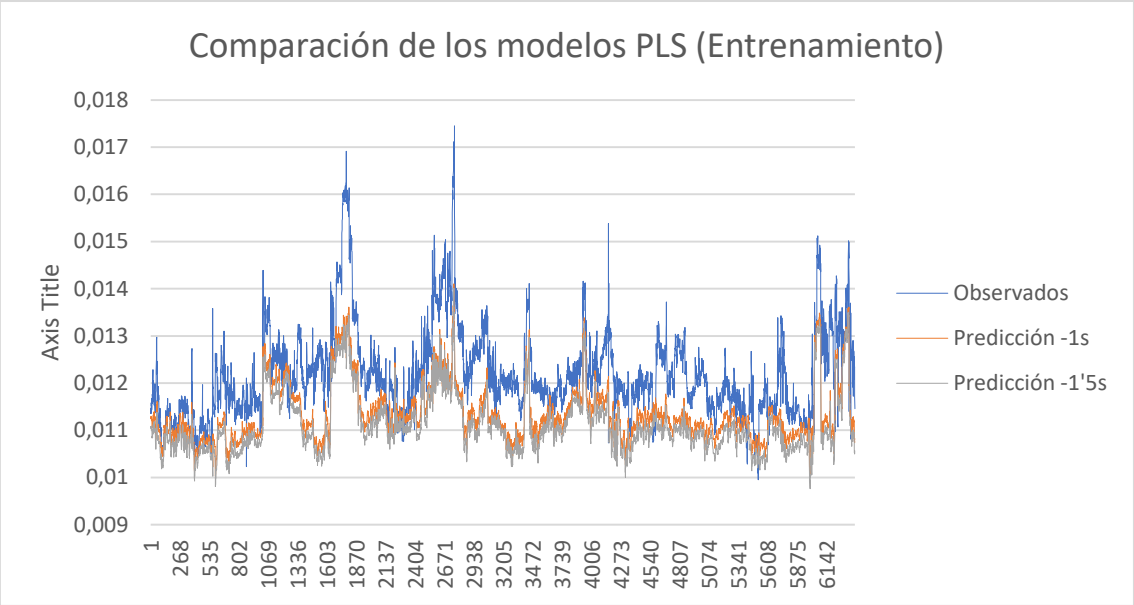

Ilustración 55.- Comparación de los modelos de entrenamiento.

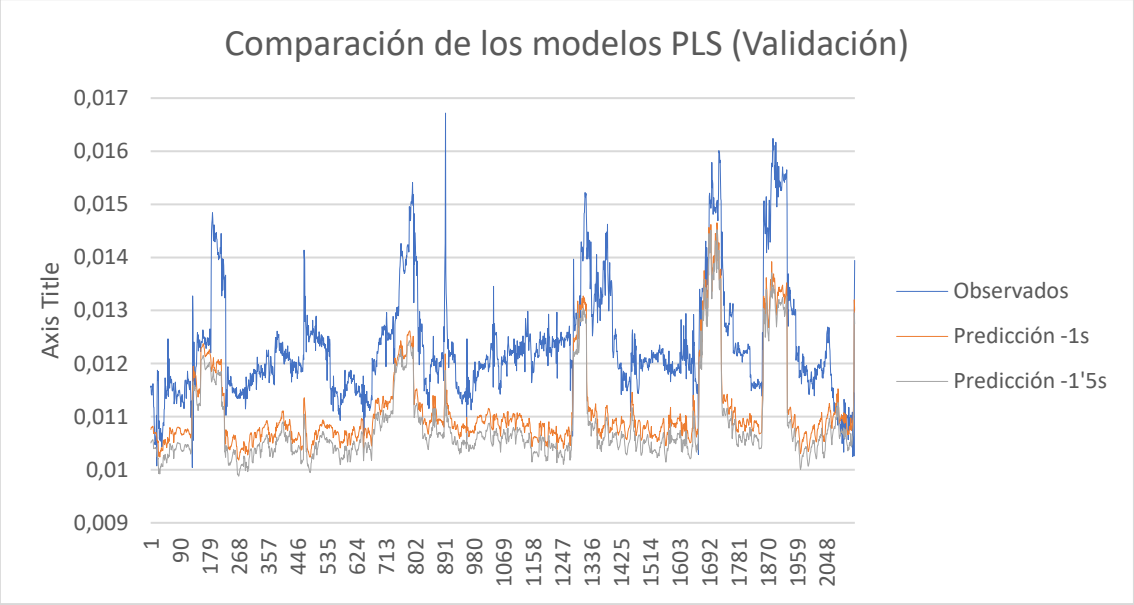

Ilustración 56.- Comparación de los modelos de validación.

3.8 Desarrollo de un modelo predictivo

Análisis de la serie temporal

Para complementar el estudio con otro enfoque, se ha planteado un predictor del consumo específico para la siguiente medida.

Hasta ahora se han conseguido dos de los objetivos: estimar el consumo específico óptimo al que se puede llegar, y saber qué variables manipular para mejorar el proceso. No obstante, es cierto que construir una herramienta para predecir la trayectoria del consumo sería muy interesante.

Teniendo en cuenta que cada media hora se lleva a cabo una medición de cada una de las variables, puede ser una buena idea tener un sistema que permita pronosticar el valor del consumo con antelación. Así, de forma orientativa se puede saber qué comportamiento va a llevar el proceso y quizá, implementar ajustes a tiempo. Hay que tener en cuenta que, en una industria como la petroquímica, tener una herramienta que prediga el consumo incluso en una sola etapa como esta, puede ahorrar grandes cantidades de dinero.

En ese sentido, se ha llevado a cabo un PLS en el cual se ha tomado como variable respuesta el consumo específico en t y como predictores tanto el consumo específico en t-1 como los drivers en t-1.

De este modo, escogiendo cuatro componentes y una vez validado el modelo, se obtiene un $R^2(Y)$ de más de un 90% (ver *Tabla 13*).

Tabla 13.- R2 y Q2 acumulada para el modelo predictor.

R2/Q2 acumulada

A	R2(X)	R2(Y)	Q2
1	0.54896	0.50341	0.50311
2	0.69299	0.80631	0.80591
3	0.7512	0.88865	0.88812
4	0.78016	0.92669	0.92629

Por tanto, se consigue una muy buena predicción que se puede corroborar si se comparan los valores observados con los predichos del 25% de datos que se habían reservado para este fin (*Ilustraciones 57 y 58*):

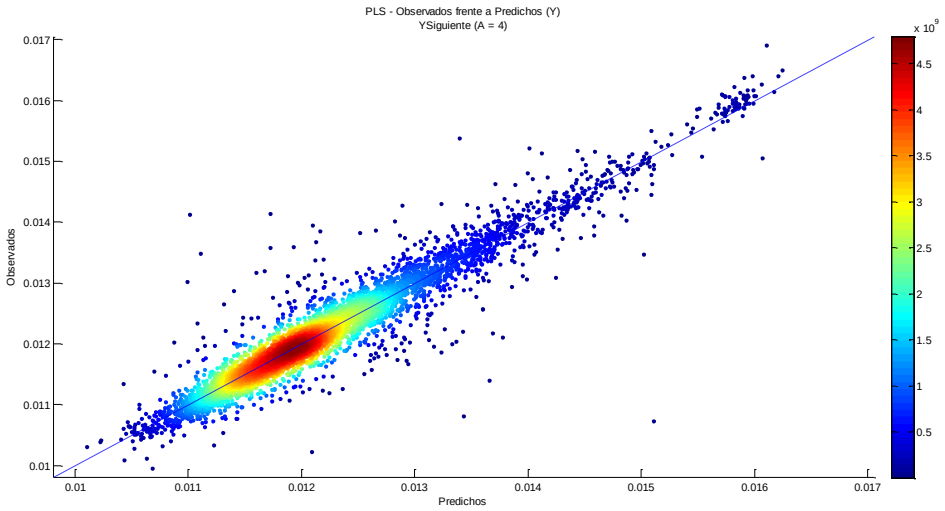

Ilustración 57.- Observados vs. Predichos para el modelo predictor.

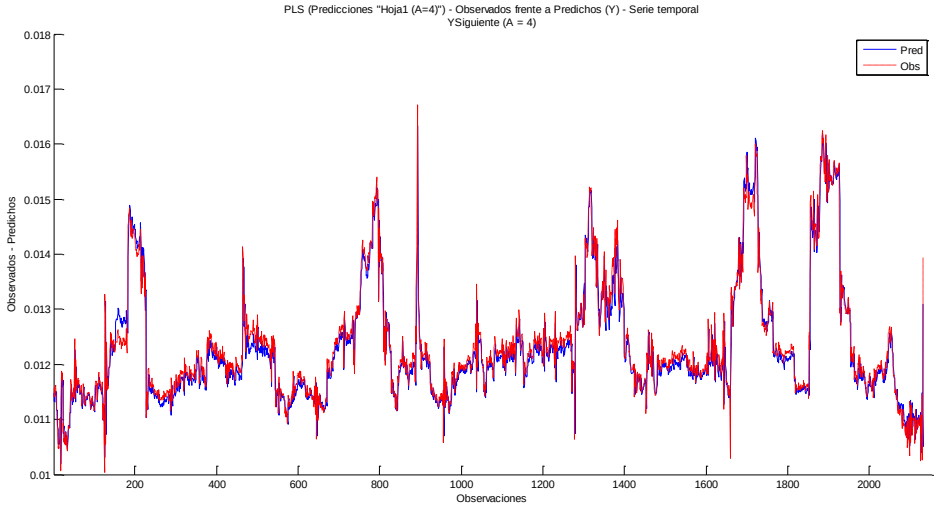

Ilustración 58.- Serie de tiempo de los datos observados vs. predichos.

4 Conclusiones

Al inicio del trabajo, se habían planteado tres objetivos. Por una parte, se pretendía entender el comportamiento de las variables de modo que los operarios pudieran manipularlas convenientemente y así reducir el consumo específico. Por otra parte, se deseaba formular un modelo que predijera la

eficiencia máxima dadas unas condiciones de trabajo. Además, se buscaba crear un modelo predictivo para la siguiente medida del consumo específico.

Para el primer caso, con los modelos PLS y PLS-DA, se ha conseguido entender qué variables son las que favorecen reducir el consumo específico. Esta información, será muy útil para los operarios ya que de este modo sabrán qué magnitud manipular y en qué sentido.

También se ha cumplido el segundo objetivo: se ha propuesto un modelo que haga las veces de “hoja de ruta” a seguir por los operarios siendo ese comportamiento del proceso el camino ideal.

Por otro lado, ha quedado patente la gran importancia que tiene hacer un buen pretratamiento de los datos. A la hora de abarcar un problema de estas características, el primer paso es un tratamiento previo que permita llevar a cabo un buen análisis posteriormente. En realidad, en muchas ocasiones este paso es el que más tiempo requiere y probablemente el más crucial para los resultados futuros.

Además, es igual de indispensable validar cada uno de los modelos que se hagan. De lo contrario, el modelo no tiene ninguna validez puesto que se puede llegar a conclusiones erróneas.

En cuanto a las técnicas de análisis multivariante empleadas, se ha demostrado el gran potencial que tienen, sobre todo si se comparan con otras técnicas más tradicionales como la regresión lineal.

Acerca del modelo predictivo, los resultados han sido muy interesantes y aunque es un enfoque simple, este modelo podría dar lugar a modificaciones como prevenir la medida del siguiente día o tener en cuenta no solo la medida previa sino varias. Se debe tomar esta información de forma orientativa para adelantarse a la trayectoria que está siguiendo el proceso.

Por último, cabe destacar la importancia de los residuos del PLS. No solamente es que no se puedan considerar desechables, sino que realmente pueden llegar a ser determinantes con su significado como se ha demostrado en este trabajo.

5 Referencias

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9):2812.

Folch-Fortuny, A., Arteaga, F., & Ferrer, A. (2015). PCA model building with missing data: New proposals ans a comparative study. *Chemometrics and Intelligent Laboratory Systems*, 146. 10.1016/j.chemolab.2015.05.006. .

Folch-Fortuny, A., Arteaga, F., & Ferrer, A. (2015 y 2017). PLS model building with missing data: new algorithms and a comarative study. *Journal of Chemometrics*, 10.1002/cem.2897.

Kulcsar, T., Koncz, P., Balaton, M., Nagy, L., & Abonyi, J. (2014). Statistical Process Control based Energy Monitoring of Chemical Process. *Computer Aided Chemical Engineering*, 33:397-402.

Prats-Montalbán, J. M., Ferrer, A., Malo, J. L., & Gorbeña, J. (2005). A comparison of different discriminant analysis techniques in a steel industry welding process. *Chemometrics and Intelligent Laboratory Systems* , 80(1):109-119.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37-52.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2-58):109-130.

Apéndice

Regresión Stepwise

La variante más utilizada para la Regresión Stepwise es la *forward*; que parte de un modelo sin ninguna variable, al que se le va incorporando, en cada paso y según un criterio establecido, la variable que mejor explica la parte de la variable dependiente que todavía no está explicada por las variables ya introducidas en el modelo

Por tanto, se ha empleado un modelo lineal puro de tipo *forward* explicando el consumo específico a partir de los *drivers*.

En las siguientes tablas (*Tablas 14 y 15*) se comparan los coeficientes de la Regresión lineal múltiple y de la Regresión Stepwise.

Tabla 14.- Coeficientes de la Regresión Stepwise y de la Regresión Lineal Múltiple.

Driver	Reg. Step.	Reg. Lin. Mul.
D6	-5,16E-01	-5,16E-01
D2	-0.0023625	-0.0023625
D1	-0.016477	-0.016477
D3	4,90E-01	4,90E-01
D9	-7,53E-02	1,95E-01
D12	5,42E-01	2,72E-01
D11	4,30E-01	1,60E-01
D8	2,70E-01	-2,70E-01
D10	5,43E-02	0
D7	3,85E-01	5,43E-02
D16	-1,40E-01	3,85E-01
D4	-3,22E-01	-1,40E-01
D14	-0.04086	-3,22E-01
D13	-0.0021257	-0.04086
D5	0.00013418	0.0021257
D15	-5,16E-01	0.00013418
Indep.	-0.0023625	-5,16E-01

Exceptuando cuatro Drivers (marcados en rojo en la *Tabla 6*), el resto coinciden en signo en ambos modelos.

Cabe destacar que al estudiar los coeficientes del PLS, se aprecia que algunos de ellos no coinciden con los valores de las regresiones. Esto puede entenderse si se observa la *Ilustración 14* en la cual gracias al gráfico de loadings se puede ver (dentro del círculo azul) que muchas de ellas están fuertemente relacionadas (variables *D3*, *D9*, *D8*, *D10* Y *D7*)

Tabla 15.- Coeficientes del PLS-I

Driver	PLS
D6	-9,69E-03
D2	-0.0080
D1	-0.0024
D3	-2,44E-02
D9	-6,66E-04
D12	1,05E-02
D11	8,06E-03
D8	-2,33E-02
D10	-5,27E-03
D7	-2,24E-03
D16	4,08E-01
D4	-2,59E-02
D14	-4,77E-01
D13	-0.0457
D5	-0.0066
D15	2,88E+00
Indep.	0.0621