

Análisis de la relación precio marginal y demanda de electricidad mediante conglomerados

Andrés Ramos¹, Gonzalo Cortés¹, Jesús M^a Latorre¹, Santiago Cerisola¹

¹ Instituto de Investigación Tecnológica. Escuela Técnica Superior de Ingeniería - ICAI. Universidad Pontificia Comillas. Alberto Aguilera 23, 28015 Madrid. andres.ramos@iit.ica.upcomillas.es

Resumen

El presente artículo analiza la relación entre el precio marginal del mercado de electricidad y la demanda mediante el uso de conglomerados. Se obtienen conglomerados para las muestras conjuntas de precios y demanda mediante los algoritmos de gas neuronal y k-medias. Posteriormente se determina el número óptimo de los mismos de acuerdo con diferentes criterios. Para dicho número óptimo de conglomerados se analiza la relación lineal entre el precio y la demanda para cada hora del día.

Palabras clave: métodos de clasificación, análisis univariante, mercado diario de electricidad.

1. Introducción

La desregulación del sector eléctrico en España en 1998 dio origen a un mercado de producción de electricidad donde se casan las ofertas de generación y de demanda. Éste a su vez está organizado en diversos mercados secuenciales para diferentes productos o servicios: mercado diario, mercados intradiarios y mercados de servicios complementarios. El mercado diario es el más relevante tanto desde el punto de vista de la energía casada (un 97 % aproximadamente) como de su impacto en el precio final a los consumidores (un 90 %).

La predicción de precios de la electricidad es una tarea difícil y no existen técnicas de predicción claramente preestablecidas. Las razones estriban, por una parte, en que los precios de la electricidad difieren claramente entre diferentes mercados nacionales y, por otra parte, los mercados suelen estar sometidos a reformas que cambian la dinámica de los precios y la volatilidad de los mismos. El precio de la electricidad depende fundamentalmente de la demanda y de la potencia disponible. Sin embargo, la potencia disponible a futuro no es conocida, no sólo por su posible fallo fortuito sino por la aparición de nuevas instalaciones. Independientemente, está el comportamiento propio de los agentes, cambiante a lo largo del tiempo, y los precios de los combustibles.

En este artículo se realiza un análisis del comportamiento histórico del precio marginal de la electricidad en el mercado diario en función de la demanda casada. Este es un primer paso para establecer el comportamiento del precio en función de las variables del sistema. En el apartado de conclusiones se sugieren futuras líneas de extensión de este trabajo para explicar el precio en función de otras variables. Este análisis está dividido en dos etapas:

- I. Obtención de conglomerados (*clusters*) que agrupan los días similares en precio y demanda de electricidad para las 24 horas. El objetivo de esta etapa es segmentar el conjunto total de datos disponibles en subconjuntos homogéneos desde el punto de vista de demanda y precio de la electricidad.

II. Cálculo de las curvas de regresión que determinan el comportamiento histórico de la relación entre precio y demanda para cada hora del día en cada conglomerado. Estas rectas permiten estimar precios medios horarios en cada conglomerado.

El resto del artículo está organizado así. La sección 2 muestra parámetros estadísticos generales de las variables de precio y demanda horarios. La siguiente sección presenta diferentes técnicas de obtención de conglomerados y su aplicación a los datos del mercado. La sección 4 analiza los conglomerados obtenidos e interpreta físicamente los mismos. En la sección 5 se obtienen las curvas de regresión que permiten estimar precios en función de la demanda. Finalmente, se presentan las conclusiones del artículo y futuras líneas de trabajo.

2. Análisis estadísticos generales

El análisis de la relación entre precio y demanda se basa en la información histórica disponible de los precios del mercado y de la demanda casada para el mercado español de electricidad de los últimos años. Se dispone de la demanda y precios horarios desde el 1-jul-2001 al 30-jun-2005 (1461 días), ver OMEL. Los días de cambio de la hora en el último fin de semana de marzo y de octubre (con 23 ó 25 horas, respectivamente) han sido corregidos para tener la duración habitual de 24 horas.

El crecimiento de la demanda ha sido significativo en estos años, con valor aproximado de un 5 % anual, que para una demanda media diaria de 22500 MW supone 1125 MW entre un día y el semejante del año siguiente. Se ha optado por no descontar esta tendencia ya que la relación precio-demanda está además muy influida por el precio de los combustibles.

La tabla 1 presenta algunos estadísticos preliminares sobre los datos. Un primer análisis de la variabilidad horaria de la demanda y del precio lo obtenemos mediante el coeficiente de variación horario (ratio entre la desviación típica y la media en cada hora). En el gráfico de la tabla se puede observar que este coeficiente es más de 3 veces superior para el precio que para la demanda en cualquier hora del día. Esto indica que la variación del precio es mucho mayor y, por consiguiente, sólo podrá ser explicada parcialmente por la variación de la demanda. La volatilidad del precio puede deberse a varios factores: la necesidad de recuperación de costes fijos o de arranque de los generadores, la estocasticidad de los precios de combustible o las ineficiencias en el mercado. Por otra parte, se observa que la mayor variación se da en las horas de rampas crecientes de la mañana (9 a 10 h) y de la tarde (18 a 20 h).

Tabla 1. Estadísticos del precio y de la demanda y coeficiente de variación horario.

	Precio [c€/kWh]	Demanda [MW]
Media	3.43	22333
Mediana	3.28	22431
Mínimo	0.00	12608
Máximo	15.84	35689
Desviación típica	1.54	3475
Coeficiente de variación total	0.45	0.16
Curtosis	6.63	2.60
Asimetría	1.36	0.14

En la figura 1 se muestran los histogramas horarios de la variable precio (a la izquierda) y de la variable demanda (a la derecha), así como la nube de puntos de una variable con respecto a la otra y viceversa (en el centro). Como se puede apreciar en los histogramas y en la asimetría, los precios son claramente asimétricos a la derecha, mientras que la demanda presenta una distribución prácticamente simétrica. El coeficiente de correlación entre ambas variables es de 0.61, lo que implica que la demanda explica una parte importante de la variación del precio.

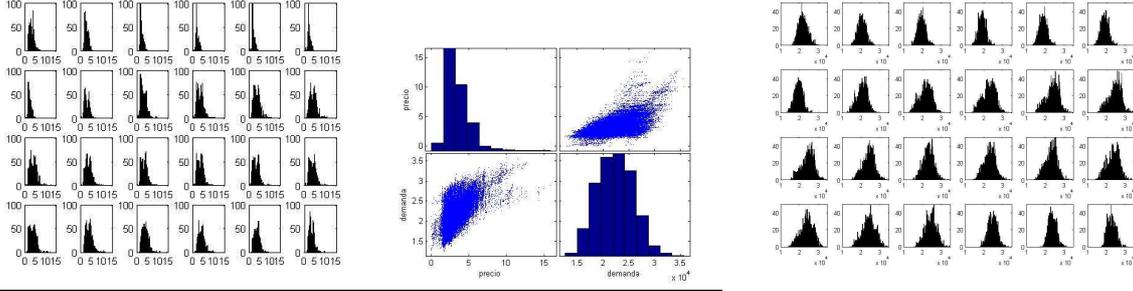


Figura 1. Histogramas horarios (precio a la izquierda) y nubes de puntos respectivas del precio y de la demanda.

3. Obtención de conglomerados

La clasificación en conglomerados agrupa los datos originales en subconjuntos de datos similares entre sí. Adicionalmente la media de los valores de los datos de cada conglomerado determina un representante del grupo que en la literatura es denominado centroide. De este modo, la determinación de un conjunto de conglomerados adecuado viene motivada por la selección de unos centroides que minimicen la función de representación, que se define a continuación.

Sea $\{X_i\}_{i \in I}$ el conjunto de muestras. Se dispone de I días totales. Cada muestra X_i está compuesta por 24 valores de precios y demandas tipificadas, correspondientes a las 24 horas del día $X_i = (X_{ih}^p, X_{ih}^d)_{h=1}^{24}$. Sea c cada uno de los centroides y C en número total de ellos.

Consideremos $\{X_c\}_{c \in C}$ el conjunto de centroides, que al igual que las muestras, están formados por 24 parejas de precios y demandas $X_c = (X_{ch}^p, X_{ch}^d)_{h=1}^{24}$. Sea

$D_{i'}$ = $\sqrt{\sum_{h=1}^{24} [(X_{ih}^p - X_{i'h}^p)^2 + (X_{ih}^d - X_{i'h}^d)^2]}$ la *distancia entre las muestras* X_i y $X_{i'}$ de los días i e i' en todas las horas, ver Gordon (1999). Se ha utilizado la distancia euclídea, la más habitual, aunque se podrían haber utilizado otras distancias: la distancia al cuadrado, la de valor absoluto, etc. La *distancia día-centroide* de una muestra i a un centroide c se denota como D_{ic} y se define como $D_{ic} = \sqrt{\sum_{h=1}^{24} [(X_{ih}^p - X_{ch}^p)^2 + (X_{ih}^d - X_{ch}^d)^2]}$, Gordon (1999).

El error de representación $E(C)$ cuantifica, para cada muestra i , la distancia al centroide más cercano, que es su representante. El conjunto de muestras que son representadas por un centroide c se denomina región de Voronoi V_c . Una vez identificado el representante adecuado para cada muestra, $c(i)$, el error de representación (o error de cuantización vectorial) viene definido como la suma de las distancias de las muestras a sus representantes

$$E(C) = D_C = \sum_{c=1}^C \sum_{i \in V_c} D_{ic} = \sum_{i=1}^I D_{ic(i)} \quad (1)$$

En este contexto, los centroides óptimos son aquellos que minimizan la anterior función de error. Los algoritmos para su elección se clasifican en *jerárquicos* y de *partición*, Gordon (1999). Los primeros son métodos en los que el número de conglomerados no está fijado de antemano, sino que va aumentando hasta alcanzar un criterio que determina la cantidad de representantes escogidos. Por el contrario, en los métodos de *partición*, el número de conglomerados está especificado de antemano y el número adecuado debe obtenerse posteriormente mediante repeticiones de este método. Dentro de esta última familia, métodos de *partición*, existen gran variedad de algoritmos. En los denominados de aprendizaje rígido (*hard competitive learning*) cada muestra modifica un único centroide o representante en cada

iteración [método LBG, ver Lloyd (1957)], [método k-medias, MacQueen (1967)]. Por el contrario, en los denominados de aprendizaje flexible (*soft competitive learning*) (método gas neuronal, ver Martinetz (1991) y Latorre *et al.* (2006)), cada muestra puede modificar uno o varios representantes en cada iteración. Este último tipo de métodos presenta ventajas desde el punto de vista numérico. En los algoritmos de aprendizaje rígido, es común la existencia de centroides que evolucionan a valores que no representan a ninguna muestra de los datos originales. En los métodos de aprendizaje flexible, si bien la imposibilidad de este fenómeno no está garantizada, su ocurrencia es mucho menor. En cualquier caso, los métodos que minimizan la función de error solamente garantizan la convergencia hacia un mínimo local de dicha función.

En la aplicación numérica presentada en este artículo el método empleado ha sido el *k-medias*. Las diferencias observadas entre los resultados obtenidos con este método y otros como el *gas neuronal* son poco significativas para los datos que se han analizado. La asignación de muestras a los conglomerados es similar al igual que el error de representación. En la siguiente figura podemos observar como dicho error disminuye a medida que aumenta el número de conglomerados en los que se dividen los datos de entrada. Como se observa en la figura 2 la diferencia en el error de representación entre el método de gas neuronal y el de k-medias para diferente número de conglomerados es poco significativa. Aunque el método gas neuronal da un menor mínimo local, a partir de ahora, por simplicidad y por ser el más utilizado se van a presentar sólo los resultados para el algoritmo k-medias.

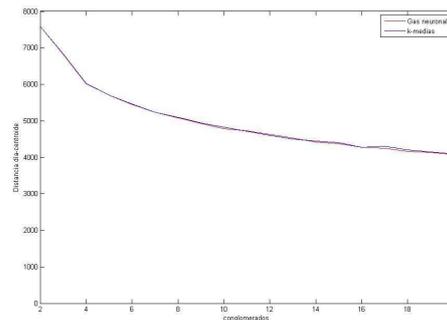


Figura 2. Error de representación con el algoritmo gas neuronal y con el k-medias.

En los métodos de clasificación se pueden distinguir dos etapas:

- I. Detección de la existencia de conglomerados naturales en los datos y
- II. Determinación del número óptimo de conglomerados.

Para este tipo de análisis se utilizan diversos índices basados en las distancias anteriormente definidas. La *distancia intraconglomerado* es un valor de la concentración de las muestras en una partición. La *distancia interconglomerado*, por el contrario, representa la separación entre conglomerados de una partición. El *valor de la silueta* es una medida de la calidad de la asignación de puntos a conglomerados. La existencia de agrupaciones naturales en los datos puede identificarse mediante el *intervalo estadístico*, el *coeficiente de emparejamiento* S_M y *coeficiente de Jaccard* S_J . Finalmente, para la determinación del número óptimo de conglomerados destacamos el *índice Calinski and Harabasz*, la maximización de la silueta media y la detección del codo de la curva del error de representación.

- *Distancia intraconglomerado*

Para cada muestra i , se define D_{Intra_i} como el valor medio de las distancias de la muestra al conjunto de valores del conglomerado al que pertenece, Gordon (1999).

$$D_{Intra_i} = \sum_{i' \in c} D_{ii'} / N_c \quad (2)$$

siendo N_c el número de puntos del conglomerado c . De modo global, se define $DIntra_c$ como $DIntra_c = \sum_i DIntra_i$. Obsérvese que dicho índice es nulo si todos los puntos pertenecientes al conglomerado son iguales.

- *Distancia interconglomerado*

Para cada punto i de un conglomerado c a todos los puntos i' de otro conglomerado c' , se define $DInter_i$ como la *distancia media interconglomerado*

$$DInter_i = \sum_{i' \in c'} D_{ii'} / N_{c'} \quad (3)$$

- *Silueta*

Una medida de la calidad de la asignación de puntos a conglomerados es el valor de la *silueta*, Kaufman and Rousseeuw (1990). Es una medida de la similitud de dicho punto a los puntos de su propio conglomerado comparado con puntos en otros conglomerados. Su rango está entre -1 y $+1$. Un valor negativo indica mala asignación de un punto a su conglomerado. La silueta media es la media de la silueta de los puntos. Valores de la silueta media cercanos a uno indican gran separación entre los conglomerados. Valores bajos indican cercanía entre los conglomerados. Se definen como

$$S_i = \left(\min_{c'} DInter_i - DIntra_i \right) / \max \left(\min_{c'} DInter_i, DIntra_i \right) \quad (4)$$

siendo la silueta media $\bar{S} = \sum_i S_i / I$.

Para la detección de la existencia o no de conglomerados existen dos métodos en la literatura:

- *Estadístico intervalo GAP*

El cálculo de este índice proporciona un valor para la detección de la existencia o no de conglomerados naturales en los datos. Dicho índice se define como, Hastie *et al.* (2000),

$$GAP_C = E_n \left[\ln(DIntra_c) \right] - \ln(DIntra_c) \quad (5)$$

siendo un número n de muestras suficiente y $E_n(\cdot)$ la esperanza matemática. Este índice compara $\ln(DIntra_c)$ para los datos originales con su esperanza obtenida para una distribución uniforme de los datos de entrada $E_n \left[\ln(DIntra_c) \right]$. La esperanza se calcula considerando distribuciones uniformes entre los valores mínimos y máximos de los datos originales para cada hora. Para el cálculo del estadístico intervalo se siguen estos pasos: i) Obtener muestras de valores con una distribución uniforme entre los valores extremos en cada dimensión; ii) Agrupar en un número dado de conglomerados y calcular $DIntra_c$; iii) Repetir los pasos 1 y 2; iv) Hacer una media de los logaritmos de estas distancias; v) Calcular el estadístico intervalo con la expresión (5). El número de conglomerados óptimo será el que dé una mayor diferencia positiva entre ambos y que cumpla $GAP_C \geq GAP_{C+1} - s_{C+1}$ siendo s_{C+1} la desviación típica muestral de las n muestras de $\ln(DIntra_c)$. Para este estudio se han tomado 20 muestras.

- *Coficiente de emparejamiento S_M y coeficiente de Jaccard S_J*

Estos índices aparecen al detectar la estabilidad de los conglomerados con respecto a perturbaciones en los datos de entrada, ver Ben-Hur *et al.* (2002). Una partición de los datos en C conglomerados se dice que es estable, que captura su estructura inherente, si las particiones en C conglomerados de dos subconjuntos de muestras¹ de los datos originales (con un porcentaje de un 80 % por ejemplo) son similares. Los índices que a

¹ Se han tomado 20 muestras.

continuación se definen miden la similitud entre dos particiones. Se denominan coeficiente de emparejamiento S_M y coeficiente de Jaccard S_J y se calculan como

$$S_M = (N_{00} + N_{11}) / (N_{00} + N_{01} + N_{10} + N_{11}) \quad (6)$$

$$S_J = N_{11} / (N_{01} + N_{10} + N_{11}) \quad (7)$$

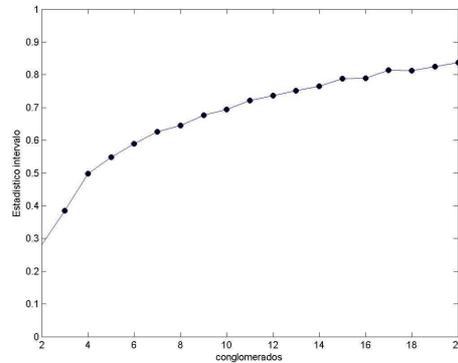


Figura 3. Estadístico intervalo GAP_C .

Donde N_{11} es el número de parejas de datos que pertenecen al mismo conglomerado tanto en la primera como en la segunda partición; N_{00} es el número de parejas de datos que no pertenecen al mismo conglomerado ni en la primera ni en la segunda partición; N_{10} es el número de parejas de datos que pertenecen al mismo conglomerado en la primera partición pero que pertenecen a conglomerados distintos en la segunda partición y N_{01} es el número de parejas de datos que no pertenecen al mismo conglomerado en la primera partición pero que pertenecen al mismo en la segunda partición.

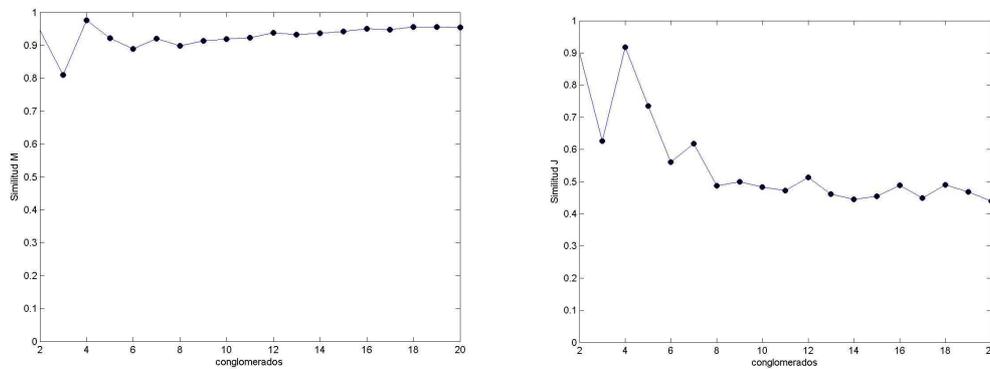


Figura 4. Índices de similitud de emparejamiento y de Jaccard.

Los dos métodos anteriores indican si hay o no conglomerados naturales. Los siguientes métodos determinan su número óptimo mediante unas reglas de parada que han funcionado bien, Gordon (1999) y Milligan and Cooper (1985). Se basan en índices globales basados en criterios heurísticos y se recomienda utilizar varios de ellos para tomar la decisión:

- **Índice Calinski and Harabasz CH**

El cálculo de este índice está orientado a determinar el número óptimo de conglomerados, ver Gordon (1999) y Milligan and Cooper (1985). Dicho número óptimo puede determinarse con el primer máximo del índice CH , Calinski and Harabasz (1974):

$$CH_C = \left[\bar{D}_C / (C - 1) \right] / \left[D_C / (I - C) \right] \quad (8)$$

siendo $\bar{D}_C = \sum_{i=1}^I \sum_{c \neq c(i)} D_{ic}$ la distancia total a los centroides de otros conglomerados.

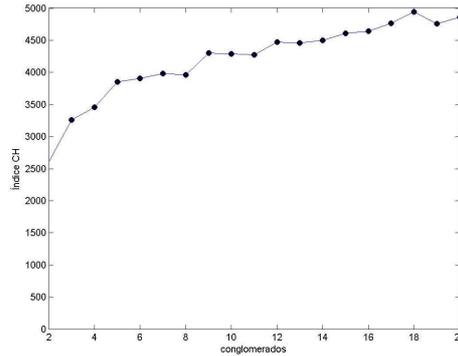


Figura 5. Índice CH_C .

- *Maximización de la silueta media*
 Alternativamente al índice CH , Kaufman and Rousseeuw (1990) propusieron maximizar la silueta media \bar{S} como medio para determinar el número óptimo de conglomerados.

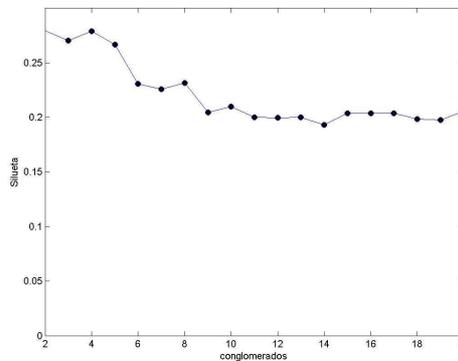


Figura 6. Índice silueta media \bar{S} .

- *Detección del codo de la curva del error de representación*
 La detección del codo de la curva de error tiene también el propósito de determinar el número óptimo de conglomerados. Para tal fin, Salvador *et al.* (2003), proponen el empleo del método L . Dicho método calcula el error que se produce cuando se aproxima la función de error $E(C)$ mediante dos rectas. La primera de estas rectas aproxima dicha función entre valores de 2 a c conglomerados. La segunda recta aproxima la función de error para valores entre $c+1$ y C conglomerados, siendo C un número elevado de conglomerados. En la figura 7 se puede observar la evolución de $RMSE_c$, raíz del error cuadrático medio de la recta que aproxima la curva en función del valor de c escogido para la dividir la función $E(C)$.

$$RMSE_c = RMSE(L_c)(c-1)/(C-1) + RMSE(R_c)(C-c)/(C-1) \quad (9)$$

Una vez calculado el error de aproximación, el número de conglomerados óptimo se determina como el valor de c para el que el error de aproximación es menor.

4. Análisis de conglomerados

Veamos a continuación en la tabla 2 los resultados de los criterios previos. Parecen determinar que el número óptimo de conglomerados es de aproximadamente cuatro y se toma este número de conglomerados como el adecuado para los siguientes análisis de las secciones posteriores del artículo.

Los centroides obtenidos se pueden observar en las siguientes figuras. En la demanda se pueden apreciar más claramente perfiles de demanda de días laborables, de sábado y de domingo. En los precios se aprecian perfiles similares, aunque menos claros, y un centroide en particular que representa los valores más elevados.

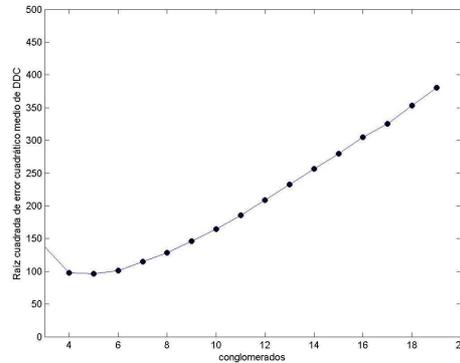


Figura 7. Índice $RMSE_c$.

Tabla 2. Número óptimo de conglomerados según el índice.

Índice	Estadístico GAP_C	Similitud	Índice CH_C	Silüeta \bar{S}	Índice $RMSE_c$
Número conglomerados	17 ó 18	4	9 a 11	4	4 ó 5

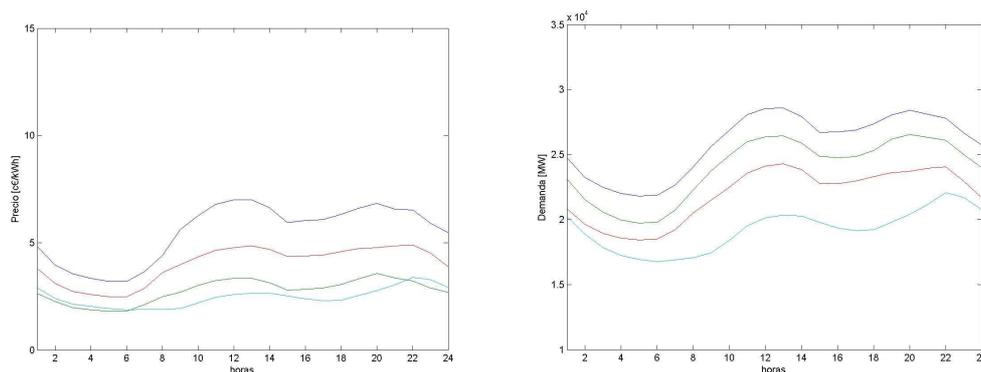


Figura 8. Centroides de precio y demanda.

En la figura 9 se representan los conglomerados de precio y demanda y se observa su variabilidad, siempre mayor proporcionalmente en los precios. Hay conglomerados como el primero que presenta gran variabilidad en precios, aunque en demanda ésta es menor. Para mejorar la aproximación se podrían haber eliminado los valores extremos.

Una vez obtenidos los conglomerados se hace un análisis de los mismos para identificarlos y etiquetarlos, es decir, tratar de encontrar su denominación característica. Se definen como laborables los días de lunes a viernes y festivos sábado y domingo. Se definen como invierno los tres primeros meses del año y así sucesivamente. Se presentan unos mapas que asignan colores a cada uno de los días de los años analizados.

Nótese que la separación laborable-festivo sí segmenta los conglomerados. Por ejemplo, los conglomerados 1 a 3 representan a días laborables, mientras que el conglomerado 4 agrupa a los días festivos. Sin embargo, la división estacional no es muy significativa ya que en cada conglomerado hay días de todas las estaciones. Se observa también que los conglomerados corresponden aproximadamente con la secuencia de los años. Los días del conglomerado 3 corresponden principalmente a días laborables de 2001 y 2002, los del 2 a días laborables del invierno del 2003 y al 2004 y los del 1 a días laborables del invierno y primavera del 2005. Es decir, el crecimiento de la demanda es lo que más influencia tiene en el proceso de creación

de los conglomerados. Los días fríos de finales de 2001 y de comienzos de 2002 están en el conglomerado 1 correspondiente a días de 2005. Este hecho sugiere la posibilidad de aumentar el número de conglomerados para capturar mejor otras diferencias.

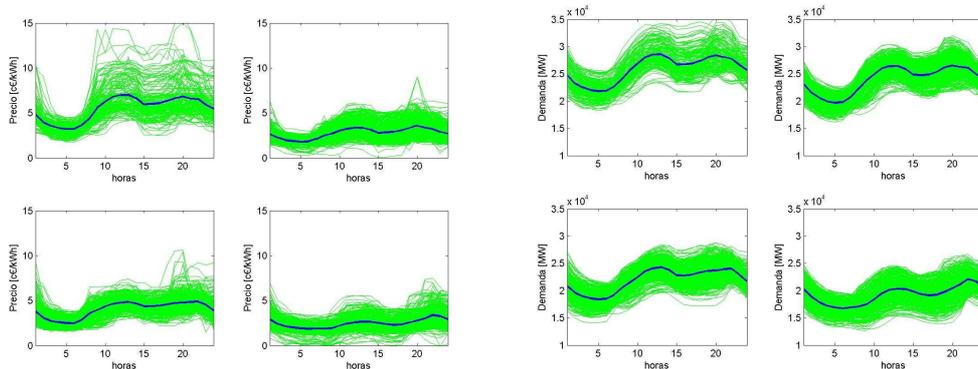


Figura 9. Conglomerados de precio y demanda.

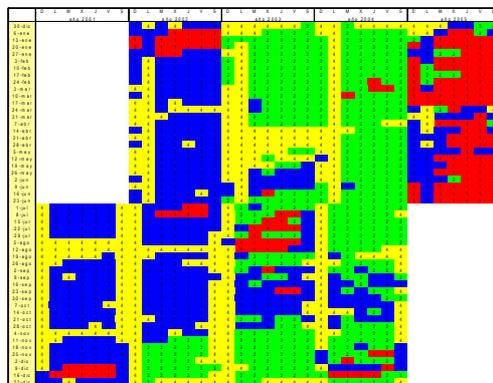


Figura 10. Asignación de días a conglomerados.

Tabla 3. Atributos de cada conglomerado.

conglomerado	población	laborables	festivos	invierno	primavera	verano	otoño
1	182	166	16	82	56	27	17
2	438	389	49	147	94	80	117
3	467	389	78	69	119	151	128
4	374	100	274	63	95	110	106

5. Obtención de curvas de regresión

Para el conjunto de puntos de un conglomerado y para cada hora se obtiene la recta de regresión de precio-demanda. Se ha utilizado también regresión no lineal (cuadrática) y mejora la aproximación. La aproximación de los precios estimados mediante regresión lineal a los originales de las muestras pertenecientes al conglomerado 3 se puede ver en la figura 11. Obsérvese que las pendientes, expresadas en c€/kWhMW, son más elevadas en las horas de rampa de subida (8 a 10 h, 18 a 20 h) y de bajada (24 h). En las horas de madrugada (6 a 7 h) son muy bajas. Las ordenadas en el origen están expresadas en c€/kWh. Esta aproximación lineal es válida para el rango de variación de los valores de demanda.

Para estimar precios a largo plazo a partir de esta información se asignan días a conglomerados y dentro de cada conglomerado los precios se obtienen con estas rectas de regresión.

6. Conclusiones

Este artículo ha presentado un procedimiento para segmentar la información de precio y demanda del mercado de electricidad para poder hacer análisis de la información disponible y estimaciones a largo plazo del precio en función de la demanda.

Como continuación natural de este trabajo está la extensión a regresión multivariante introduciendo nuevas variables como producciones o precios de los combustibles como elementos destacados. Por otra parte, el análisis de conglomerados se puede hacer no sobre las variables originales sino sobre otras obtenidas mediante transformaciones en componentes principales que maximizan la variabilidad y reducidas tras un análisis factorial.

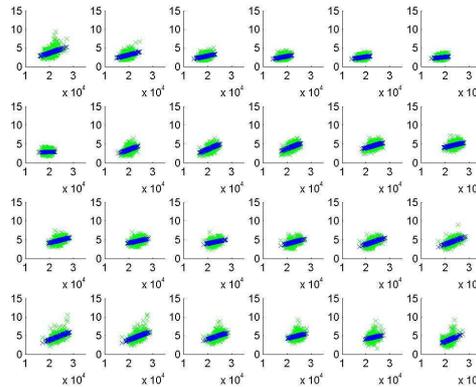


Figura 11. Precios originales y estimados para las 24 horas de un conglomerado.

Tabla 4. Rectas de regresión de precios en función de la demanda.

Hora	1	2	3	4	5	6	7	8	9	10	11	12
Pendiente	0.000231	0.000177	0.000144	0.000133	0.000109	0.000077	0.000031	0.000247	0.000278	0.000254	0.000195	0.000159
Ordenada	-1.01	-0.38	0.00	0.12	0.48	1.06	2.28	-1.46	-2.00	-1.37	0.05	0.94
Hora	13	14	15	16	17	18	19	20	21	22	23	24
Pendiente	0.000173	0.000147	0.000139	0.000184	0.000225	0.000250	0.000257	0.000259	0.000224	0.000175	0.000155	0.000279
Ordenada	0.66	1.18	1.20	0.19	-0.74	-1.25	-1.35	-1.38	-0.51	0.67	0.96	-2.18

Referencias

- Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002) *A stability based method for discovering structure in clustered data* Pacific Symposium on Biocomputing 7:6-17.
- Calinski, T., Harabasz, J. (1974) *A Dendrite Method for Cluster Analysis* Communications in Statistics, 3(1), 1–27.
- Gordon, A.D. (1999) *Classification. 2nd Edition* Chapman and Hall/CRC.
- Hastie, T., Tibshirani, R. and Walther, G. (2000) *Estimating the number of data clusters via the Gap statistic*. Department of Statistics Division of Biostatistics. Stanford University.
- Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data. An Introduction to Cluster Analysis* John Wiley & Sons, Inc., New York, NY.
- Latorre, J.M., Cerisola, S., Ramos, A. (2006) *Clustering Algorithms for Scenario Tree Generation. Application to Natural Hydro Inflows* European Journal of Operational Research (in press).
- Lloyd, S. P. (1957) *Least squares quantization in pcm*. Technical note, Bell Laboratories. published in 1982 in *IEEE Transactions on Information Theory*.
- Martinetz, T.M. and Schulten, K.J. (1991) *A neural-gas network learns topologies* in T. Kohonen, K. Mäkisara, O. Simula, J. Kangas, editors, *Artificial Neural Networks*, pp. 397-402. North-Holland, Amsterdam.
- MacQueen, J.B. (1967) *Some Methods for classification and Analysis of Multivariate Observations* Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.
- Milligan, G.W. and Cooper, M.C. (1985) *An examination of procedures for determining the number of clusters in a data set* Psychometrika, 50, 159-179.
- OMEL Operador del Mercado Ibérico de Energía <http://www.omel.es/frames/es/index.jsp>
- Salvador, S. and Chan, Ph. (2003) *Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms* Technical Report CS-2003-18 Department of Computer Sciences. Florida Institute of Technology.