Análisis Big Data para la Respuesta a la Demanda en el Mercado Eléctrico

José Antonio Fábregas,
José María Luna-Romera
and José C. Riquelme Santos
Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Sevilla, España

Ángel Arcos-Vargas
Organización Industrial y Gestión de Empresas
Universidad de Sevilla
Sevilla, España

Javier Tejedor Aguilera
Endesa S.A.
Madrid, España

***Abstract*—El modelo de negocio tradicional de las compañías energéticas está cambiando los últimos años. La introducción de los contadores inteligentes ha conllevado un aumento exponencial del volumen de datos disponibles, y su análisis puede ayudar a encontrar patrones de consumo entre los clientes eléctricos para reducir costes y proteger el medioambiente. Las centrales generan energía eléctrica para poder cubrir los picos de consumo en momentos puntuales. Un conjunto de técnicas denominadas "demand response" intenta dar solución a este problema usando propuestas de inteligencia artificial. En este documento se propone una metodología para el procesado de los grandes volúmenes de datos como los que generan los contadores inteligentes. Tanto para el preprocesado como para la optimización y realización de este análisis se utilizan técnicas big data. En concreto, una versión distribuida del algoritmo k-means y de varios índices de validación interna de clustering para big data en Spark. Los datos de origen corresponden los consumos de clientes eléctricos de Cataluña durante el año 2016. El análisis de estos consumidores realizado en este trabajo ayuda a su caracterización. Este mayor conocimiento sobre los hábitos de consumos y tipos de clientes, puede facilitar a las compañías eléctricas la labor.**

***Index Terms*—Big data, respuesta a la demanda, clustering, contadores inteligentes, consumo eléctrico**

I. Introducción

Durante la mayor parte del siglo XX, la relación entre los usuarios de electricidad y las empresas de distribución se mantuvo sin cambios. No se eligieron proveedores y, por lo tanto, no había necesidad de tratar a los consumidores como clientes. Sin embargo, la desregulación, la agenda verde y los continuos saltos tecnológicos han cambiado esta relación. Nuevas limitaciones como la seguridad del suministro, la competitividad y la sostenibilidad son los tres ejes prioritarios para cambiar el modelo energético actual, que pueden lograrse mediante objetivos como la reducción de las emisiones y la mejora de la generación de energía renovable y la eficiencia energética.

Una herramienta esencial en este nuevo modelo son los llamados "contadores inteligentes", que no deben entenderse sólo como dispositivos que miden el consumo sino como verdaderos sensores para una red eléctrica. Estos sensores facilitan una red altamente flexible y adaptable que integra de forma inteligente las acciones de los usuarios que se conectan a ella para conseguir un suministro eficiente, seguro y sostenible.

Uno de los principales problemas del sector eléctrico es la necesidad de disponer de capacidad de generación y de una red sobredimensionada para cubrir los picos de alto consumo de los clientes en determinados momentos. Sin embargo, existen soluciones basadas en la adaptación de la demanda a la energía disponible en lugar de aumentar la oferta para satisfacer la demanda. Esto se denomina "Respuesta a la Demanda" y su objetivo es cambiar los hábitos de consumo de electricidad de los clientes en respuesta a los cambios en los precios del suministro. El principal inconveniente es el gran volumen de información disponible en estas redes, ya que sólo puede manejarse con técnicas big data.

Nuestra propuesta se basa en el procesado de estos grandes conjuntos de datos de manera distribuida y paralela. En particular, la aplicación de técnicas de minería de datos para comprender mejor los patrones de consumo de los clientes. Por un lado, haremos uso de HDFS [1] para el almacenamiento de datos distribuido. Mientras que el procesado lo realizaremos con Spark [2], una plataforma de computación distribuida y paralela. En concreto, utilizaremos la implementación del algoritmo k-means de la librería MLlib [3] de Spark, así como cuatro índices de validación de clustering para big data [4]. Este estudio podría ayudar a la planificación de las conexiones de las fuentes de energía renovables a la red, con un doble objetivo: la reducción de precios y la sostenibilidad medioambiental.

La estructura de este artículo es la siguiente. En la sección II se describen los trabajos relacionados. En la sección III se detallan las características del dataset. En la sección IV se muestran los resultados de los experimentos realizados para preprocesar los datos y aplicar técnicas de clustering. Para finalizar, en la sección V se presentan las conclusiones de la investigación realizada.

II. Trabajo relacionados

La irregularidad de la demanda de electricidad es uno de los principales problemas del sector. Esto se debe a que las compañías eléctricas deben tener tanto una capacidad de generación sobredimensionada, como redundancia de la red para hacer frente a grandes cantidades de demanda que sólo se requieren unas pocas horas al año. Normalmente, se establece un umbral del 20% para la generación de electricidad latente, que debe cubrir aproximadamente el 5% del tiempo de servicio de la red (pico de demanda) [5]. Algunos de los recursos para resolver este problema necesitan la implicación de los usuarios. Estas soluciones se estudian bajo el nombre de 'respuesta a la demanda' (Demand Response DR) [6]. En contraste con las ideas convencionales de aumentar la oferta para satisfacer la demanda, las soluciones apuntan a satisfacerla con la energía disponible.

El objetivo es cambiar las pautas de consumo de energía de los clientes en respuesta a los cambios en los precios ofrecidos. Esto permitirá a las compañías eléctricas gestionar mejor la demanda con un mejor ajuste de las predicciones y una reducción del coste de la energía para los clientes. Existen múltiples iniciativas de posibles esquemas de fijación de precios, que en algunos casos incluso mantienen los beneficios para las empresas proveedoras [7]. Una de las principales ventajas de la respuesta a la demanda es ofrecer una opción sostenible con una generación de energía más volátil. Sobre todo en España, donde existe una alta presencia de fuentes de generación renovables. Para implementar los mecanismos de respuesta a la demanda, las redes eléctricas deben evolucionar a una infraestructura que permita el flujo de información entre los diferentes participantes del sistema eléctrico. En este campo es donde los grandes datos se convierten en una tecnología esencial para analizar este flujo de información y convertirlo en conocimiento útil.

Estos datos de consumo de los clientes, obtenidos mediante contadores inteligentes, no son sino múltiples series temporales. El análisis de series temporales puede entenderse como una secuencia de valores observados a lo largo del tiempo y ordenados cronológicamente [8]. Como el tiempo es una variable continua, las muestras se registran en puntos sucesivos igualmente espaciados. Por lo tanto, las series temporales son una secuencia de datos de tiempo discreto.

En el contexto de la minería de datos de series temporales, el principal desafío es cómo representar los datos. El enfoque más común es transformar las series temporales en otro ámbito para la reducción de la dimensionalidad y desarrollar un mecanismo de indexación. La medida de la similitud entre las sub-secuencias de series temporales y la segmentación son las dos tareas principales en la minería de series temporales que corresponden con las tareas clásicas de la minería de datos. El uso cada vez mayor de datos de series cronológicas ha dado lugar a una gran cantidad de intentos de investigación y desarrollo en el campo de la minería de datos [9].

En este trabajo nos centraremos en el clustering, un método de minería de datos para agrupar instancias no etiquetadas de conjuntos de datos. La idea es que las instancias recogidas en un mismo grupo tendrán un comportamiento similar [10]. El clustering de series temporales surge como un enfoque útil para minar patrones comunes a partir de datos dependientes del tiempo [11] que se caracterizan por tener una alta dimensionalidad y un gran tamaño.

Centrándonos en el clustering a partir de datos de consumo de energía, son muchas las propuestas enmarcadas en este campo: En [12] se presenta el efecto de las medidas de similitud en la aplicación de la agrupación para descubrir los patrones energéticos de los edificios. Para obtener perfiles de carga típicos de los clientes, en [13] se propone un índice de estabilidad para elegir el algoritmo de agrupamiento que mejor se adapte a este problema de reconocimiento de patrones. Además, se propone otro índice de prioridad (basado en el índice de estabilidad) para determinar el rango de prioridad de los agrupamientos. En [14] desarrolla una técnica de clustering de particiones para extraer información útil de los precios de la electricidad. Mientras que en [15] se usan técnicas de clustering con el objetivo de agrupar y etiquetar las muestras de un conjunto de datos para pronosticar el comportamiento de las series temporales basadas en la similitud de las secuencias de patrones.

En relación a la gestión inteligente de la demanda de electricidad, en [16] los autores proponen un Virtual Power Player como gestor para satisfacer la demanda y reserva de energía eléctrica requerida. En [17] se presenta un análisis de los datos de los contadores inteligentes de los clientes para comprender mejor la demanda máxima y las principales fuentes de variabilidad en su comportamiento.

Además de los métodos clásicos de gestión de datos, el enfoque de big data ha surgido recientemente debido a la disponibilidad de una gran cantidad de datos, sistemas de ficheros distribuidos, y potentes motores de procesamiento distribuido. Esto ha propiciado que muchos de los algoritmos de minería de datos se hayan adaptado al entorno de big data, como por ejemplo los algoritmos de clustering. En cuanto al campo del consumo de energía, en la actualidad han surgido varias grandes soluciones de datos, como las optimizaciones de redes inteligentes en [18] y los patrones de consumo energético en [19].

III. Características del dataset

Los datos originales se encuentran almacenados en 34 tablas, divididas en cientos de archivos CSV. Estas tablas contienen una amplia información sobre consumos, tarifas, contadores, datos geográficos o personales de consumidores eléctricos de Endesa en Cataluña entre 2010 y 2016 para un tamaño total de 1,8 TB.

Los clientes en los que centramos este estudio son aquellos que poseen tarifas 2.0A y 20DHA. Ambas del mercado libre de Endesa para potencias contratadas menores a 10kW, donde se encuentran inmensa mayoría de hogares y pequeños locales. La tarifa 2.0A mantiene un precio fijo durante todo el año, mientras que la 20DHA tiene discriminación horaria de dos periodos. En esta última, los periodos punta y valle marcan dos precios distintos según la hora en la que se consuma la energía: punta de 12h a 22h en invierno y de 13h a 23h en verano; valle de 22h a 12h en invierno y de 23h a 13h en verano.

En la Tabla Ia se observa la distribución de los clientes según su tarifa. Debido a la gran variedad de potencias contratadas, La Tabla Ib muestra una distribución de los clientes por cada rango de potencia. Estos rangos están basados en los valores estándar de potencia que actualmente pueden contratarse al tener contadores inteligentes.

TABLE I: Distribución de clientes

(a) Por tarifa contratada

Tarifa	Clientes
2.0A	102,123
2DHA	6,614

(b) Por potencia contratada

Potencia(kW)	Clientes
[0.1-2.3)	8,972
[2.3-3.45)	21,969
[3.45-4.60)	38,578
[4.60-5.75)	22,328
[5.75-6.90)	8,772
[6.90-8.05)	3,134
[8.05-9.20)	4,946
[9.20-10.0)	82

IV. Análisis realizados

En esta sección se presentan los resultados de los análisis realizados. En el apartado IV-A se detalla el procesado de los datos En el apartado IV-B se analizan cuatro índices de validación de clustering obtener el k óptimo. En el apartado IV-C se muestran los clusters obtenidos con el algortimo k-means. Por último, en el apartado IV-D se realiza una evaluación de los resultados.

Para la realización de los experimentos se han utilizado los siguientes entornos:

- Cluster propio: 72 procesadores Intel Xeon E7-4820, 128 GB RAM y 8 TB de almacenamiento.
- Cluster de EMR (Elastic Map Reduce) de AWS: cinco instancias de m3.2xlarge con 16 procesadores Intel Xeon E5-2670 v2 (Ivy Bridge), 30 GB RAM y 2 SSDs DE 80 GB cada una.

A. Procesado de datos

El primer objetivo es conseguir un conjunto de datos que tenga sentido minar. Primero, almacenaremos la gran cantidad de datos de los que disponemos en un sistema de ficheros distribuido (HDFS) configurado en nuestro cluster. Para reducir el tamaño de los datos, los ficheros CSV se pasan a Parquet, un formato de datos orientado a columnas que los comprime y codifica. Una vez almacenados y formateados, todo el procesamiento de estos datos se realiza de forma distribuida y paralela con Spark. Debido a que estos datos también se procesarán con herramientas online de Amazon Web Service, se almacenan además en el S3, el sistema de almacenamiento en línea de Amazon. Posteriormente se estudian y seleccionan los atributos necesarios para construir un primer dataset. Las tablas procesadas se muestran a continuación (Tabla II):

TABLE II: Tablas procesadas

Tabla	Elementos (millones)	Tamaño (MB)
Clientes	20.6	716
Contratos	40.6	560
Maestro Contratos	33.1	666
Curvas de carga	2,094.44	340,992

Una vez construido este primer dataset, se seleccionan los usuarios con: una potencia contratada igual o inferior a 10kW, una tarifa 2.0A o 20DHA, y que tengan con todas las lecturas de consumo del año 2016. Por último, se descartan las instancias con valores nulos. Este dataset está compuesto por 47,829,235 instancias correspondientes a las 365 curvas de carga de 2016 de 131,039 clientes. De cada consumidor, tenemos 24 lecturas de consumo horario para cada una de sus 365 instancias.

Para construir las series temporales de 2016 es necesario transformar este dataset de instancias diarias en uno de instancias anuales .De esta forma, el nuevo dataset constaría de 131,039 instancias, una por cliente, con: 8760 (365x24) lecturas horarias de 2016, la cups (Código Universal del Punto de Suministro), la tarifa y la potencia contratadas.

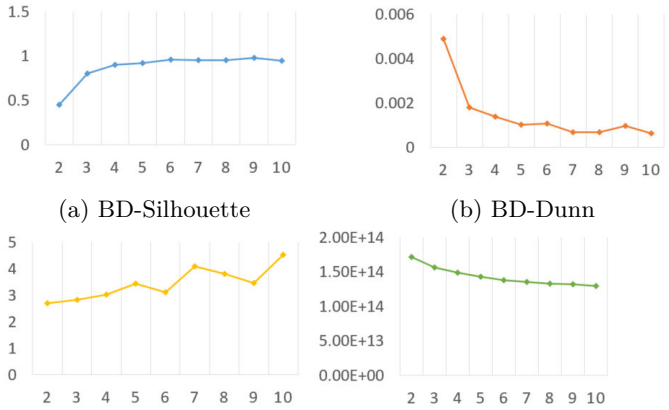
A continuación generamos un nuevo atributo con el que categorizamos a los consumidores en función a la potencia contratada. Por último, construimos un dataset alternativo con diferencias de consumo normalizadas: hallamos la diferencia entre cada par de valores de consumo consecutivo y la dividimos por la media de consumo de ese día. De esta forma tendremos un dataset con los consumos horarios de 2016 y otro con las diferencias normalizadas de consumo del mismo periodo. Estos datasets se utilizarán tanto de forma conjunta como individual en los siguientes apartados con el objetivo de encontrar patrones en los hábitos de consumo eléctrico de los clientes.

B. Determinación del número óptimo de clusters

Antes de aplicar algoritmos de clustering a nuestros datasets es necesario determinar cuál es el número óptimo de clusters (k) a obtener. Para ello, aplicamos a cada uno de los datasets cuatro índices de validación de clustering para big data (BD-CVIs) [4]: BD-Silhouette [4], BD-Dunn [4], Davies-Bouldin [20] y Within Set Sum of Square Errors (WSSSE) [21].

En la Figura 1a se muestra la representación gráfica resultados índice BD-Silhouette. Para este índice los valores óptimos de k son sus máximos, 6 y 9. Estos valores coincides con los máximos de la gráfica correspondiente al índice BD-Dunn (Figura 1b). En el caso del índice Davies-Bouldin los valores óptimos se encuentran en los mínimos,

que vuelven a coincidir en 6 y 9, tal y como se observa en la Figura 1c. Por último, los resultados del índice WSSSE representados en la Figura 1d no arrojan un valor claro. En este índice buscamos una estabilización de valores y, como podemos ver, no hay un valor concreto en lo que esto ocurra. Tras analizar estos resultados, hemos obtenido los valores 6 y 9 como óptimos para la realización del clustering.



(a) BD-Silhouette (b) BD-Dunn

(c) Davies-Bouldin (d) WSSSE

Fig. 1: Índices big data de validación de clustering

Al igual que para el dataset de comsumos, hemos vuelto a aplicar estos índices al dataset de diferencias normalizadas comentado al final del apartado IV-A. En este caso, los resultados para el valor óptimo de k fueron 5 y 7.

C. Clustering

Una vez calculado el número óptimo de clusters, aplicamos a cada dataset la versión implementada en Spark del algoritmo k-means. Esta implementación fue desarrollada para poder extraer patrones en sistemas paralelos y distribuidos. A la hora de ejecutar el algoritmo, le daremos como entrada el objeto RDD (Resilient Distributed Dataset) y el k obtenido anteriormente. Como resultado obtendremos una serie de clusters con elementos de cada uno de los datasets.

Para el dataset de consumos, dos de los clusters obtenidos con k=9 tenían menos de 5 elementos, por lo que se decidió trabajar con el otro valor óptimo obtenido de k=6. En el caso del dataset de diferencias normalizadas, dos de los clusters para k=5 contenían un único elemento. Por este motivo, se optó por tomar el valor k=7.

La distribución de los elementos en los 6 clusters del dataset de consumos se muestra en la Tabla IIIa. De la misma forma, en la Tabla IIIb podemos ver cómo se distribuyen los elementos correspondientes a los 7 clusters del dataset de diferencias normalizadas.

En la Figura 2 se representan las curvas de consumo horario formadas por los centroides de cada uno de los clusters durante una semana de enero de 2016. En ella destaca que la mayoría de los consumidores, agrupados

TABLE III: Clusters de los datasets para k óptimo

(a) Dataset de consumos para k=6

Cluster	Elementos
0	12,029
1	50,643
2	1002
3	138
4	1116
5	43,789

(b) Dataset de diferencias normalizadas para k=7

Cluster	Elementos
0	42,276
1	8241
2	2544
3	18,994
4	28,462
5	7191
6	1029

en los clusters 1 y 5, tienen consumos inferior a 1 kWh. También puede observarse como un grupo reducido de clientes, que conforma el cluster 3, tiene un consumo muy alto por la noche (de 19 a 8 horas) y prácticamente nulo durante el día.

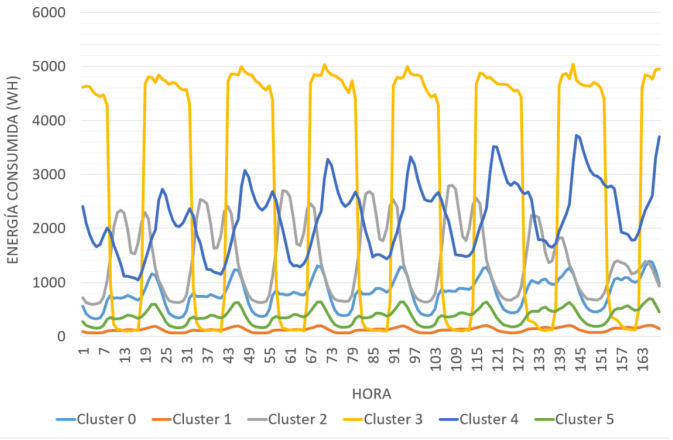


Fig. 2: Curvas de consumo horario de los centroides durante una semana de enero

En la Figura 3 se representan las las curvas de los mismos centroides durante una semana de julio. Se observa que los consumidores de los clusters 0, 1 y 5 mantienen un consumo prácticamente igual al de enero, aunque los picos máximo se acercan más a la media noche. Sin embargo, los del cluster descienden de forma radical hasta equipararse a los del 0. También destacan las curvas de consumo del cluster 3, donde las horas de alto consumo se reducen a 6 (00 a 6 horas).

Por su parte, la Figura 4 muestra las curvas de las diferencias horarias normalizadas en el mismo periodo de tiempo que la Figura 2. En ella podemos observar como los mayores picos de diferencias de consumo entre horas pertenecen a los elementos de los clusters 1,2,5 y 6. Estos clusters resultan ser los menos numerosos, representando un 17% del total de los consumidores. Lo que nos indica que el consumo de la mayoría de los usuarios a lo largo del día mantiene cierta uniformidad.

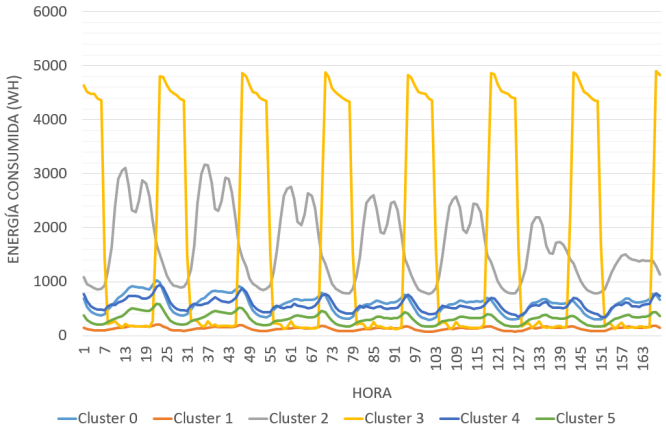


Fig. 3: Curvas de consumo horario de los centroides durante una semana de julio

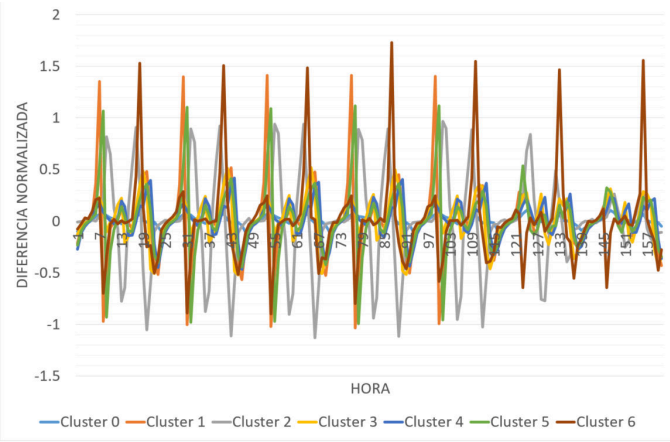


Fig. 4: Curvas de diferencias horarias normalizadas de los centroides

D. Evaluación de resultados

1) Evaluación no supervisada: En esta última fase evaluaremos los resultados obtenidos tras aplicar el algoritmo k-means a los datasets. Para ello, se generan las tablas de contingencia entre los clusters obtenidos para los distintos datasets, cruzando los resultados entre los mismos. Esto nos permite encontrar patrones que definan a los consumidores en relación a las características de los diferentes conjuntos de datos.

La tabla de la Figura 5 muestra por filas los 6 clusters del dataset de consumos (C0 a C5), y por columnas los 7 obtenidos del dataset de diferencias normalizadas (D0 a D6). Estos valores representan los porcentajes relativos al total de cada fila. Es decir, el porcentaje de consumidores de cada cluster de consumos presente en cada uno de los clusters de diferencias normalizadas.

En la Figura 5 destaca que casi la totalidad del cluster C3 está formado por consumidores del cluster D6. Si atendemos a la Figura 2 esto caracterizaría a un grupo de clientes con un consumo nocturno muy alto. Si además

		CLUSTER DIFERENCIAS NORMALIZADAS							
		D0	D1	D2	D3	D4	D5	D6	
CLUSTER CONSUMOS	C0	29.11%	10.20%	3.85%	19.03%	27.53%	9.85%	0.43%	*100.00%*
	C1	53.06%	5.25%	1.87%	13.03%	20.61%	4.74%	1.44%	*100.00%*
	C2	52.20%	1.60%	27.64%	7.98%	7.29%	3.19%	0.10%	*100.00%*
	C3	5.07%	0.00%	0.00%	0.00%	0.00%	0.72%	94.20%	*100.00%*
	C4	56.16%	7.83%	0.09%	7.13%	16.99%	8.36%	3.43%	*100.00%*
	C5	24.51%	9.70%	1.96%	22.71%	32.99%	7.94%	0.18%	*100.00%*

Fig. 5: Tabla de contingencia de valores relativos al total de cada fila

nos fijamos en la Figura 4, observamos que tienen un mínimo y un máximo en las diferencias horarias que se mantiene constante los 7 días de la semana. Esto podría indicar un perfil de consumidor muy concreto, con un consumo energético nocturno alto y uniforme. Por otro lado, podemos observar que la mitad cluster C1, el más numeroso, está formado por consumidores del cluster D0, también el más numeroso. Aquí se identifica un amplio grupo de usuarios de consumo bajo y con pocos cambios, cuyas mayores variaciones en el consumo se producen a las 8h.y 18h.

A continuación vamos a analizar el cluster C5, de un tamaño similar al C1 y con un consumo un poco más alto. En este caso, sólo se compone en un 24.51% de consumidores del cluster D0. Sin embargo, aumenta considerablemente el número de elementos pertenecientes a los clusters D3 y D4. En ellos, las diferencias de consumo horario (algo más altas que en el anterior) se concentran en las 19h. en los consumidores del cluster D3 y las 21 h. en los de D4. Por lo que podemos interpretar que los clientes con un consumo algo mayor que los del cluster C1 son más heterogéneos en cuanto a su comportamiento.

Ahora analizaremos los resultados desde la otra perspectiva. En la Figura 6 se representan los porcentajes relativos al total de cada columna. Es este caso, el porcentaje de consumidores de cada cluster de diferencias normalizadas presente en cada uno de los clusters de consumos.

		CLUSTER DIFERENCIAS NORMALIZADAS						
		D0	D1	D2	D3	D4	D5	D6
CLUSTER CONSUMOS	C0	8.28%	14.89%	18.20%	12.05%	11.63%	16.48%	5.05%
	C1	63.56%	32.29%	37.15%	34.74%	36.68%	33.39%	70.65%
	C2	1.24%	0.19%	10.89%	0.42%	0.26%	0.45%	0.10%
	C3	0.02%	0.00%	0.00%	0.00%	0.00%	0.01%	12.63%
	C4	1.51%	1.08%	0.04%	0.43%	0.68%	1.32%	3.79%
	C5	25.39%	51.55%	33.73%	52.36%	50.76%	48.35%	7.77%
		100.00%	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*

Fig. 6: Tabla de contingencia de valores relativos al total de cada columna

Si nos fijamos en la Figura 6, lo primero que podemos destacar es que todos los clusters de diferencias normalizadas están compuestos mayoritariamente por los clusters C1 y C5. Algo normal debido entre los dos aglutinan el 86% del total de consumidores. Los clusters D0 y D1 tienen una mayoría de elementos de C1, mientras que los clusters D1, D3, D4 y D5 tienen una mayor presencia de elementos del C5. Por otro lado, en el D2 los clusters C1

y C5 no tienen una representación tan alta debido al 10% de elementos del cluster C2.

Centrándonos en el cluster D6, vemos que el 70.65% de sus clientes pertenecen al cluster C1. Mientras que observando la Figura 5 se denota que sólo el 1.44% de los del cluster C1 de consumos pertenecen al cluster D6.

2) Evaluación semi-supervisada: En este apartado tomaremos como referencia el rango de potencia contratada al que pertenece cada consumidor. Por lo que cruzamos elementos de los clusters del dataset de comsumo (Ver Tabla IIIa) con los de cada rango de potencia (ver Tabla Ib). El objetivo es encontrar relaciones entre los distintos tipos de consumidores y sus potencias contratadas. En la tabla de contingencia de la Figura 7 se representan los clusters del dataset de consumos por filas (C0 a C5) y los rangos de potencia por columnas (P1 a P8). Estos valores representan los porcentajes relativos al total de cada fila. Es decir, el porcentaje de consumidores de cada cluster de consumos presente en cada uno de los rangos de potencia.

		RANGO DE POTENCIA								
		P1	P2	P3	P4	P5	P6	P7	P8	
CLUSTER CONSUMOS	C0	2.65%	9.51%	27.46%	23.09%	17.67%	6.66%	12.80%	0.17%	*100.00%*
	C1	13.76%	26.37%	34.08%	17.74%	4.43%	1.50%	2.08%	0.03%	*100.00%*
	C2	2.69%	5.29%	12.28%	16.87%	14.87%	19.56%	26.85%	1.60%	*100.00%*
	C3	3.62%	0.72%	10.87%	21.74%	10.14%	47.10%	4.35%	1.45%	*100.00%*
	C4	4.15%	9.01%	12.54%	15.19%	26.50%	19.08%	12.90%	0.62%	*100.00%*
	C5	3.67%	16.70%	40.50%	23.28%	8.89%	2.50%	4.41%	0.05%	*100.00%*

Fig. 7: Tabla de contingencia de valores relativos al total de cada fila

Si observamos los clusters C1 y C5, los de menor consumo (ver Figura 2), comprobamos que la mayoría de sus elementos pertenece a los rangos P1 a P4, las potencias bajas. Mientras, en los clusters C2 y C4 aumentan de forma considerable los consumidores con potencias más altas (P5, P6 y P7). Como es lógico, la mayoría de los clientes con consumo bajo tienen contratadas potencias medias y bajas. Además, más de la mitad de los clientes con un consumo alto contrató una potencia alta. Un caso particular es el C3, donde casi la mitad de sus elementos pertenecen a P6. Esto indica que cerca del 50% de los clientes con alto consumo nocturno tienen contratados entre 6.90 y 8.05kW.

Ahora vamos a analizarlo desde el punto de vista de la potencia contratada. Si observamos la Figura 8, se representan los porcentajes relativos al total de cada columna. Es decir, el porcentaje de consumidores de cada rango de potencia presente en cada uno de los clusters de consumos.

Podemos ver como más del 85% de los consumidores de los rangos P1 a P4 pertenecen a los clusters C1 y C5. Por lo que la relación entre estos grupos de clusters y rangos de potencia existe en ambos sentidos: Clientes con bajo consumo contrató un nivel potencia baja o media y viceversa. En el análisis de la Figura 7 vimos como los clientes con alto consumo tenían potencias altas contratadas. Pero, si nos fijamos en los rangos P5 a P8,

		RANGO DE POTENCIA							
		P1	P2	P3	P4	P5	P6	P7	P8
CLUSTER CONSUMOS	C0	3.54%	5.20%	8.55%	12.42%	24.34%	25.53%	31.10%	24.39%
	C1	77.66%	60.80%	44.75%	40.25%	25.72%	24.31%	21.33%	15.85%
	C2	0.30%	0.24%	0.32%	0.76%	1.71%	6.25%	5.44%	19.51%
	C3	0.06%	0.00%	0.04%	0.13%	0.16%	2.07%	0.12%	2.44%
	C4	0.52%	0.46%	0.37%	0.77%	3.44%	6.89%	2.95%	8.54%
	C5	17.92%	33.28%	45.98%	45.66%	44.63%	34.94%	39.06%	29.27%
		100.00%	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*

Fig. 8: Tabla de contingencia de valores relativos al total de cada columna

observamos que éstos también están compuestos entre un 45% y un 70% por clientes con bajo consumo. Por lo que, aunque los clientes con alto consumo tienen potencias altas contratadas, la mayoría de consumidores con estas potencias tienen un bajo consumo. Al igual que antes, nos vamos a centrar el caso de P6 y C3. Mientras que el 47.1% de los consumidores de C3 pertenecían a P6, sólo el 2.07% de los de P6 se encuentran en C3. Esto refuerza la conclusión anterior, ya que más del 85% de los consumidores con potencia entre 6.9 y 8.05kW tienen consumos que rondan entre los 0.5 y 1.5kWh.

V. Conclusiones y trabajo futuro

En este trabajo se ha presentado la aplicación de técnicas big data para el análisis de datos de consumidores eléctricos. La caracterización de estos clientes obtenida como resultado da lugar a las siguientes conclusiones:

- Más del 85% de los clientes presentan curvas de carga de consumo donde los valores máximos no superan el kWh.
- Existe grupo de 138 clientes con un consumo nocturno muy alto. Y, aunque la cantidad de horas de consumo durante el verano es muy inferior al invierno, siempre se producen en periodos valle. El 97.8% de estos usuarios tienen contratada una tarifa adaptada a sus consumos, con discriminación horaria.
- En su inmensa mayoría, los clientes con un bajo consumo tenían contratada una baja potencia y viceversa.
- Los usuarios con un consumo medio-alto contrataron una potencia media-alta. Sin embargo, el 77.32% de los clientes que contrataron estos niveles de potencia, consumió valores de energía que no llegaron a 1kWh. Por lo que más de 3/4 partes de estos clientes tienen potencias contratadas muy por encima de lo que necesitan.
- Durante todo el año, los picos de consumo que se alcanzan por las mañanas y al mediodía se producen en horarios valle. Lo mismo ocurre con los picos nocturnos en verano, ya que estos aparecen entre las 23 y la 1. Esto indica que las tarifas de discriminación horaria podrían ser beneficiosas para los clientes con estos hábitos de consumo. Sin embargo, sólo el 5% (4,784 de 95,569 clientes) tienen contratada este tipo de tarifa.

- En trabajos futuros se caracterizará a los consumidores en función de sus consumos y tarifas. Además, se analizarán y recomendarán las tarifas y potencias óptimas a contratar de forma personalizada para cada tipo de cliente.

References

[1] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, May 2010, pp. 1–10.

[2] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," vol. 10, pp. 10–10, 07 2010.

[3] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar, "Mllib: Machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, no. 34, pp. 1–7, 2016.

[4] J. M. Luna-Romera, J. García-Gutiérrez, M. Martínez-Ballesteros, and J. C. Riquelme Santos, "An approach to validity indices for clustering techniques in big data," *Progress in Artificial Intelligence*, vol. 7, no. 2, pp. 81–94, Jun 2018.

[5] H. T. Haider, O. H. See, and W. Elmenreich, "A review of residential demand response of smart grid," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 166 – 178, 2016.

[6] I. Hussain, S. Mohsin, A. Basit, Z. A. Khan, U. Qasim, and N. Javaid, "A review on demand response: Pricing, optimization, and appliance scheduling," *Procedia Computer Science*, vol. 52, pp. 843 – 850, 2015.

[7] H. T. Haider, O. H. See, and W. Elmenreich, "Residential demand response scheme based on adaptive consumption level pricing," *Energy*, vol. 113, pp. 301 – 308, 2016.

[8] F. Martínez-Álvarez, A. Troncoso, G. Asencio-Cortés, and J. C. Riquelme, "A survey on data mining techniques applied to electricity-related time series forecasting," *Energies*, vol. 8, no. 11, pp. 13 162–13 193, 2015.

[9] T. chung Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164 – 181, 2011.

[10] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering – a decade review," *Information Systems*, vol. 53, pp. 16 – 38, 2015.

[11] H. Wang, W. Wang, J. Yang, and P. Yu, "Clustering by pattern similarity in large data sets," vol. 3, 10 2002.

[12] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, 2013.

[13] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 153–160, Feb 2012.

[14] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme, "Partitioning-clustering techniques applied to the electricity price time series," in *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, ser. IDEAL'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 990–999.

[15] F. M. Alvarez, A. Troncoso, J. C. Riquelme, and J. S. A. Ruiz, "Energy time series forecasting based on pattern sequence similarity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1230–1243, Aug 2011.

[16] P. Faria, Z. Vale, and J. Baptista, "Demand response programs design and use considering intensive penetration of distributed generation," *Energies*, vol. 8, no. 6, pp. 6230–6246, 2015.

[17] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136–144, Jan 2016.

[18] J.-S. Chou and N.-T. Ngo, "Smart grid data analytics framework for increasing energy savings in residential buildings," *Automation in Construction*, vol. 72, pp. 247 – 257, 2016.

[19] R. Perez-Chacon, R. L. Talavera-Llames, F. Martinez-Alvarez, and A. Troncoso, "Finding electric energy consumption patterns in big time series data," in *Distributed Computing and Artificial Intelligence, 13th International Conference*, S. Omatu, A. Semalat, G. Bocewicz, P. Sitek, I. E. Nielsen, J. A. García García, and J. Bajo, Eds. Cham: Springer International Publishing, 2016, pp. 231–238.

[20] D. L. Davies and D. Bouldin, "A cluster separation measure," vol. PAMI-1, pp. 224 – 227, 05 1979.

[21] "Spark clustering rdd based api documentation for spark 2.3.0. 2017." https://spark.apache.org/docs/2.3.0/mllib-clustering.html, accessed: 2018-06-11.