Automatic data clustering to identify changes in consumption patterns

Patricio G. Donato, Marcos A. Funes y Carlos M. Orallo

***Abstract*—The global energy scenario has increasingly posed challenges for the management of electricity distribution systems. In recent decades, machine learning tools for processing electrical variables and managing networks have been used more and more frequently. To identify the potential of these algorithms in specific cases, it is necessary to evaluate them using real data. This paper presents the first results obtained from the analysis of historical power demand data using unsupervised clustering algorithms. The results show that it is possible not only to identify common routines for working and non-working days, or periods with and without activity, but also to detect anomalous behavior. This opens the door to future developments and studies that will be useful for energy management in public buildings and other similar consumption scenarios.**

***Index Terms*—Clustering algorithms, machine learning, energy demand, demand patterns.**

***Abstract*—El contexto energético mundial plantea cada vez más desafíos para la gestión de los sistemas de distribución de electricidad. En las últimas décadas se han comenzado a emplear, en forma cada vez más habitual, herramientas aprendizaje automático para el procesamiento de variables eléctricas y la gestión de las redes. Para identificar las potencialidades de estos algoritmos en casos concretos, es necesario evaluarlos con datos reales. En este trabajo se presentan los primeros resultados obtenidos a través del análisis de datos históricos de demanda de potencia mediante algoritmos de agrupamiento no supervisados. Los resultados obtenidos muestran que es posible no solo identificar rutinas habituales de días laborales y no laborables, u horarios con y sin actividad, si no que se pueden identificar comportamientos anómalos. Esto abre las puertas a futuros desarrollos y estudios que sean útiles para la gestión de la energía en edificios públicos y otros consumos similares.**

***Index Terms*—Algoritmos de agrupamiento, aprendizaje automático, demanda de energía, perfiles de consumo.**

I. Introducción

El incremento constante en la demanda de energía eléctrica, impulsado por el crecimiento demográfico y los patrones de consumo de la sociedad moderna, genera incertidumbre en el panorama energético futuro. La solución a esta problemática no radica únicamente en la implementación de dispositivos de generación de energía eléctrica a partir de fuentes alternativas, sino también en la reducción de la demanda mediante un consumo más racional y eficiente. La eficiencia energética se integra en el concepto más amplio de las Redes Eléctricas Inteligentes (REI) [1]. Estas redes implican la fusión de la infraestructura eléctrica tradicional con tecnologías avanzadas de información y comunicación (TIC), sistemas de generación distribuida y almacenamiento energético. Las REI deben ser capaces de recopilar datos en tiempo real desde diversos puntos de la red utilizando medidores inteligentes y otros dispositivos de medición, además de intervenir en la red cuando sea necesario mediante actuadores adecuados como interruptores y convertidores de potencia. Este proceso es complejo, ya que requiere un soporte robusto de comunicaciones para gestionar grandes volúmenes de datos procedentes de múltiples puntos de la red, además de hardware y software avanzados para procesarlos y extraer información relevante.

Entre los temas comprendidos dentro del procesamiento de datos relativos a parámetros eléctricos, hay una notoria tendencia en los últimos años a emplear herramientas de inteligencia computacional para la extracción de información útil a partir de los datos registrados de la red. Una de las técnicas más destacadas en este ámbito es el uso de algoritmos de agrupamiento (también conocidos por su expresión en idioma inglés como '*clustering*'), los cuales permiten agrupar grandes volúmenes de datos en conjuntos homogéneos basados en características similares. Estos algoritmos son especialmente útiles para identificar patrones ocultos y segmentar datos, facilitando así la detección de anomalías, la predicción de comportamientos y la optimización de operaciones en la red eléctrica [2][3].

En este trabajo se presentan los primeros resultados de un estudio realizado sobre series de datos históricos de potencia consumida en la Facultad de Ingeniería de la UNMDP. El objetivo es comparar los resultados obtenidos con diferentes algoritmos de agrupamiento y determinar si éstos son capaces de detectar comportamientos anómalos y patrones que puedan ser relevantes. Este es un primer paso dentro de un estudio mayor, que se extenderá a diferentes bases de datos de consumo de la Universidad. El trabajo se organiza de la siguiente manera: en la sección II se describe la base de datos utilizada, mientras que las técnicas consideradas en el estudio se resumen en la sección III. La sección IV resume los resultados más importantes del estudio y en la sección V se comentan las conclusiones generales.

Este trabajo ha sido realizado gracias al financiamiento de CONICET (PIP 2643) y la Universidad Nacional de Mar del Plata

P. G. Donato es investigador del CONICET y profesor de la UNMDP, Mar del Plata, Argentina (e-mail: donatopg@fi.mdp.edu.ar).

M. A. Funes es investigador del CONICET y profesor de la UNMDP, Mar del Plata, Argentina (e-mail: mafunes@fi.mdp.edu.ar).

C. M. Orallo es profesor de la UNMDP, Mar del Plata, Argentina (e-mail: orallo@fi.mdp.edu.ar).

.

II. Base de datos analizada

La base de datos empleada para este análisis corresponde a la serie temporal de datos de demanda de potencia eléctrica de la Facultad de Ingeniería de la Universidad Nacional de Mar del Plata (UNMDP). Esta base de datos comprende un conjunto de mediciones de potencia activa, reactiva y aparente que abarca desde enero de 2021 a agosto de 2024, con muestras tomadas cada 15 minutos. En la Tabla I se muestran las principales características del conjunto de datos empleados en este estudio.

TABLA I
Características de la serie de muestras de potencia empleadas en este estudio

	Facultad de Ingeniería
Total de muestras	126074
Potencia activa máxima (kW)	104,98
Potencia activa media (kW)	21,51
Desviación estándar potencia activa (kW)	13,85
Potencia reactiva máxima (kVAR)	21,99
Potencia reactiva media (kVAR)	9,61
Desviación estándar potencia reactiva (kVAR)	3,22
Potencia aparente máxima (kVA)	105,01
Potencia aparente media (kVA)	24,37
Desviación estándar potencia aparente (kVA)	12,78

III. Técnicas de agrupamiento consideradas

Los algoritmos de agrupamiento intentan agrupar grandes volúmenes de datos en conjuntos homogéneos basados en características similares. Son muy útiles para la identificación de patrones ocultos, la detección de anomalías y la segmentación de datos. Existen numerosos enfoques de agrupamiento en la bibliografía específica [4] [5]. En este trabajo se emplearon tres técnicas de agrupamiento diferentes, las cuales se describen a continuación.

A. K-Means

Es un algoritmo de agrupamiento que divide un conjunto de datos en torno a K grupos definidos por la distancia euclidiana entre cada dato individual y los centroides de cada grupo. Este algoritmo se inicializa con K centroides elegidos aleatoriamente, y se le asigna a cada dato del conjunto al centroide que está más cercano (ver ejemplo de la Fig. 1 para un caso con dos variables independientes). Luego, se recalcula la posición de los centroides como la media de los puntos asignados a cada grupo y repite este proceso hasta que los centroides converjan o hasta que se alcance un número máximo de iteraciones. El algoritmo K-Means está diseñado para minimizar la varianza intra-grupo, es decir, busca asegurar que los puntos dentro de un mismo grupo sean lo más similares posible entre sí y que cada grupo esté representado por su centroide.

B. Fuzzy C-Means (FCM)

Este algoritmo de agrupamiento puede verse como una variante de K-Means pero con la introducción del concepto de lógica difusa (fuzzy), que permite que un punto de los datos pertenezca a múltiples grupos simultáneamente, asignando una pertenencia difusa a cada uno de ellos en lugar de una asignación binaria. Se basa en la minimización de una función objetivo que considera la distancia euclidiana intragrupo ponderada entre cada punto y los centroides de los grupos, utilizando para ello una función de membresía o pertenencia que indica el grado de pertenencia de cada punto a cada grupo (ver Fig. 2). Durante cada iteración, se recalculan los centroides y la matriz de pertenencia hasta que se alcance una convergencia. Fuzzy C-Means es útil cuando los puntos de datos tienen grados de pertenencia ambiguos a múltiples grupos.

Aunque el algoritmo Fuzzy C-Means proporciona una asignación difusa de los datos a los grupos, en muchas aplicaciones prácticas es necesario tener una clasificación definitiva de los datos. Esto implica convertir las asignaciones difusas en asignaciones binarias, donde cada punto de datos se asocia de manera exclusiva con un solo agrupamiento. Este proceso se realiza transformando la matriz de pertenencias difusas en una matriz binaria, donde cada punto está asociado únicamente con el agrupamiento que tiene el mayor grado de pertenencia.

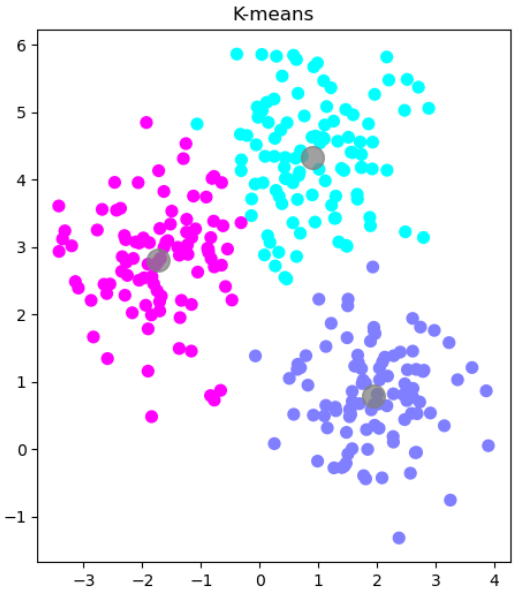

Fig. 1. Ejemplo de agrupamiento de datos aleatorios obtenido con el algoritmo K-Means. Se pueden ver los centroides (puntos grises) y los puntos de colores asignados a cada centroide.

C. Modelo de Mezclas Gaussianas (Gaussian Mixture Models, GMM)

El modelo de mezclas Gaussianas es un algoritmo de agrupamiento probabilístico que asume que los datos están generados por una mezcla de varias distribuciones gaussianas. Busca ajustar estas distribuciones para maximizar la probabilidad de observar los datos. Durante el proceso de entrenamiento, se estima la probabilidad de que cada punto de datos pertenezca a cada una de las distribuciones gaussianas, utilizando el algoritmo de esperanza-maximización

(Expectation-Maximization, EM). GMM se caracteriza por su flexibilidad para crear agrupamientos de formas y tamaños variados, además de ser capaz de asignar probabilidades a la pertenencia de cada punto a cada agrupamiento (ver Fig. 3). Suele tener un buen desempeño en conjuntos de datos donde los grupos tienen formas elipsoidales y no necesariamente están bien separados. Por otro lado, es muy sensible a los valores de inicialización, lo cual puede llevar a soluciones subóptimas, y su costo computacional puede ser restrictivo en el caso de conjuntos de datos muy grandes y con muchas dimensiones.

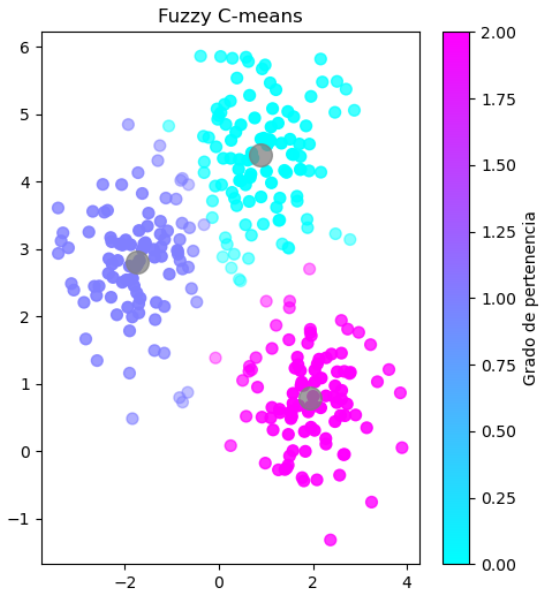

Fig. 2. Ejemplo de agrupamiento de datos alatorios obtenido con el algoritmo Fuzzy C-Means. Se pueden ver los centroides (puntos grises) y los puntos de colores asignados a cada centroide con un grado de pertenencia difuso.

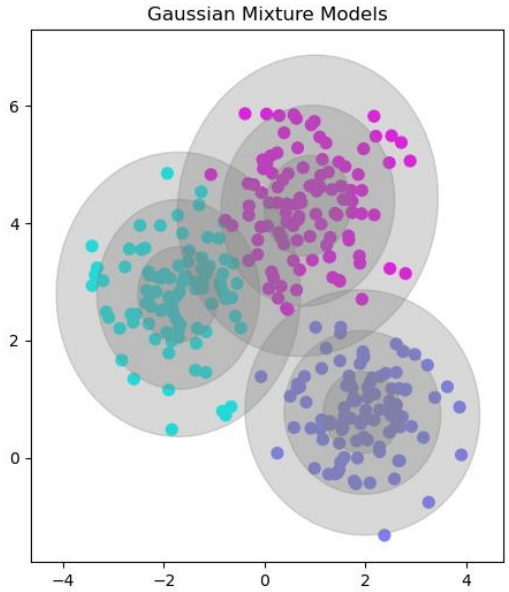

Fig. 3. Ejemplo de agrupamiento de datos aleatorios obtenido con el algoritmo de mezcla Gaussiana. Se puede ver que los agrupamientos se organizan en torno a centroides que tienen una distribución Gaussiana, identificada mediante las elipses de color gris.

D. Tratamiento de los datos

Los datos fueron estandarizados antes de aplicar el algoritmo de ordenamiento correspondiente. El intervalo de tiempo comprendido por las diferentes bases de datos incluye años con diferente comportamiento, como es el caso del año 2021, donde todavía estaban vigentes muchas de las medidas de restricción de movilidad y funcionamiento virtual debido a la pandemia de Covid-19. Por el contrario, en los años 2022, 2023 y el primer semestre de 2024, hubo funcionamiento normal, con presencialidad total en las aulas y los laboratorios. Por ello se procedió a estandarizar los datos por año, para no alterar las mediciones de años de diferente naturaleza. La estandarización de los datos antes del *clustering* garantiza que todas las características contribuyan de manera uniforme al proceso de agrupamiento, evitando que aquellas con rangos de valores más amplios dominen el proceso de agrupamiento y distorsionen los resultados finales.

Los datos de entrada para cada algoritmo fueron vectores compuestos por las mediciones de potencias activa (P) y reactiva (Q). Se hicieron algunos ensayos agregando un tercer vector con los datos de potencia aparente (S), pero los resultados obtenidos no difieren sustancialmente del caso donde se usa solo P y Q, donde la complejidad es menor. Esto puede explicarse por la relación directa que hay entre la potencia S y las potencias P y Q. Los algoritmos de agrupamiento se aplicaron a intervalos de datos que se correspondían con un mes de calendario. En un principio se planteó hacer el análisis por día, aplicando el agrupamiento al conjunto de todos los días lunes del año, al de los días martes, y así sucesivamente. Sin embargo, este criterio de análisis está condicionado por la estacionalidad, ya que es muy diferente la característica del consumo de potencia un día hábil de enero que uno de junio o septiembre. Eso lleva a que los agrupamientos obtenidos no siempre son intuitivos y su utilidad es menos tangible. Por el contrario, limitando el uso del algoritmo a un intervalo temporal de un mes se evitan los problemas de la estacionalidad, ya que los datos de consumo dentro del mes son del mismo tipo. De esta forma, los algoritmos de agrupamiento logran separar los datos de una forma más intuitiva.

E. Métricas

Para la comparativa se emplearon dos criterios diferentes. Por un lado, se calcularon las métricas Silhouette y Davies-Bouldin, para ponderar la calidad del agrupamiento. Se adoptaron estas dos métricas por ser las que presentan el mejor desempeño en la mayoría de los casos, aunque en futuros trabajos se podría ampliar a otras menos usuales [6]. Por otro lado, se analizaron las series temporales con las etiquetas temporales correspondientes, a fin de identificar las diferencias y similitudes entre los agrupamientos obtenidos con cada técnica.

La métrica Silhouette se utiliza para evaluar la cohesión y la separación de los grupos, es decir, qué tan bien los datos se agrupan naturalmente. Esta métrica asigna un valor a cada dato basado en qué tan similar es con su propio agrupamiento en comparación con otros agrupamientos. Un valor cercano a 1 indica que el agrupamiento es bueno y que cada dato está

asignado al clúster más adecuado. Un valor cercano a 0 indica que la asignación de los datos a cada grupo está cerca del límite de decisión entre dos agrupamientos. Por último, un valor cercano a -1 indica que la asignación a ese grupo es incorrecta.

Por otro lado, la métrica de Davies-Bouldin se utiliza como medida de la similitud promedio entre cada grupo y su grupo más similar, teniendo en cuenta también la diferencia promedio dentro de cada grupo. Se basa en la idea de que un buen agrupamiento tiene grupos compactos y bien separados entre sí. Para calcular la métrica Davies-Bouldin, se considera la dispersión intragrupo (la distancia promedio entre los puntos dentro de cada grupo) y la dispersión intergrupo (la distancia entre los centroides de los diferentes grupos). Un valor reducido de este índice indica una mejor separación entre los agrupamientos y una mayor cohesión dentro de cada uno de ellos. El mejor valor posible es cero, lo que indica que los agrupamientos son bien definidos y están claramente separados.

IV. Resultados Obtenidos

Se usaron K=2 agrupamientos en todos los algoritmos, porque configuraciones con mayor número de grupos dieron lugar a peores resultados en lo que respecta a las métricas de calidad, como se verá más adelante. En el caso particular de Fuzzy C-Means se usó m=10 como índice o grado de difusión. Este parámetro es el que controla el nivel de difusión de los agrupamientos, es decir, el grado de pertenencia de un determinado dato a más de un agrupamiento. El uso de valores menores producían resultados casi indistinguibles de K-Means, y valores más elevados no muestran cambios en los resultados.

A. Ubicación de los centroides

El resultado de la aplicación de cualquiera de los algoritmos de agrupamiento es, por un lado, la delimitación de la pertenencia de cada dato a cada grupo, y por otro, la posición de los centroides, que son los puntos del espacio de datos en torno a los que se conforma cada agrupamiento. La ubicación de los centroides da información acerca de los tipos de comportamiento observados. En las figuras 4 y 5 se resume la evolución de la posición de los centroides de cada una de las potencias consideradas, a lo largo del tiempo en el intervalo que va de enero de 2021 a agosto de 2024. Del análisis de estas figuras se pueden resaltar las siguientes cuestiones:

- En los meses de enero de todos los años los centroides de potencia activa de todos los algoritmos son casi iguales.
- Durante 2021 los centroides de potencia activa muestran poca diferencia, lo cual es lógico, ya que la demanda de potencia en ese año fue bastante plana y de magnitud reducida.
- Durante los meses de mayor consumo, entre febrero y diciembre, se observa la mayor diferenciación entre los centroides, lo cual se corresponde con la notable diferencia que hay entre los horarios con actividad en el edificio universitario y las horas donde no hay ocupación.
- Salvando algunas diferencias relativamente menores, la ubicación de los centroides de potencia activa en los tres algoritmos es muy similar.
- Los centroides de potencia reactiva muestran un comportamiento errático, no pudiéndose determinar, a priori, un patrón de comportamiento claro. Esto, sin embargo, se podrá apreciar mejor más adelante, cuando se muestren los resultados en forma temporal.

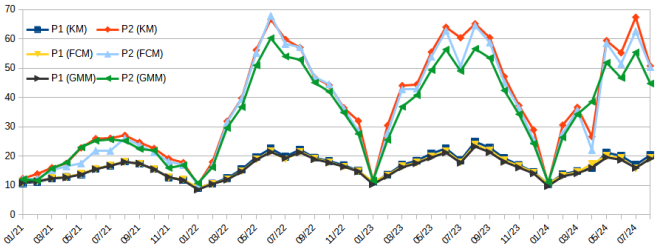

Fig. 4. Ubicación de los centroides de potencia activa a lo largo del período 2021-2024.

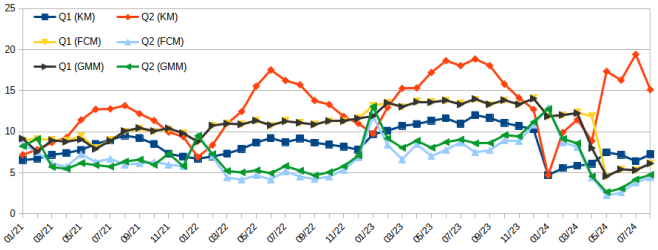

Fig. 5. Ubicación de los centroides de potencia reactiva a lo largo del período 2021-2024.

B. Métricas de calidad del agrupamiento

Tal como se mencionó en la sección anterior, se hizo una evaluación numérica a través de las métricas Silhouette y Davies-Bouldin. La métrica de Silhouette es más adecuada cuando se busca una medida más general de la calidad del agrupamiento que no dependa de la forma o densidad de los grupos en sí mismo. Por otro lado, el índice de Davies-Bouldin es más útil para evaluar la separación entre agrupamientos y la cohesión dentro de los mismos. En lo que respecta a los valores, un valor más grande de Silhouette y un valor más pequeño de Davies-Bouldin son indicativos de una mejor calidad del agrupamiento. Sin embargo, los rangos específicos de estos valores pueden variar dependiendo del conjunto de datos y del algoritmo de agrupamiento utilizado. En general, un valor de Silhouette cercano a 1 y un valor de Davies-Bouldin cercano a 0 son considerados buenos.

En las figuras 6 y 7 se observa la evolución en el tiempo de las métricas Silhouette y Davies-Bouldin, desde enero de 2021 hasta agosto de 2024. Llama la atención en ambas figuras que ambos índices experimentan un deterioro notable en el mes de enero, donde la magnitud de Silhouette decae bruscamente por debajo de 0,45, llegando a 0,29 en el caso del algoritmo GMM en enero de 2024. Por su parte, en esos mismos meses el índice Davies-Bouldin aumenta notoriamente respecto de los

meses anteriores y posteriores. En ambos casos esto indica que la calidad del agrupamiento no es buena. Un análisis más detallado, nos muestra que durante el mes de enero de 2024 (donde se obtuvieron las peores métricas), la potencia reactiva consumida fue, en proporción, mayor a la consumida en el mismo mes en años anteriores. Como se verá en algunos de los ejemplos de la subsección IV. C. , las variaciones de potencia reactiva afectan la forma en que se agrupan las mediciones, y en particular el algoritmo GMM parece ser el más sensible en ese sentido. Otra posible explicación de este fenómeno es que en enero el consumo de potencia total baja notoriamente respecto del consumo habitual en los meses de actividad normal (de febrero a diciembre). Al ser el consumo más reducido y a su vez al no haber variaciones notorias según las bandas horarias (como sí ocurre en los demás meses, donde se aprecia una importante diferencia entre el consumo en la banda horaria laboral y la que no lo es), los conjuntos de datos están muy juntos y es difícil separarlos, tal como se ve en los centroides mostrados en la subsección previa (figuras 4).

Por otro lado, se observa en que en los meses de mayor consumo, asociados con los periodos de tiempo donde hay actividad académica, concurrencia de alumnos y está la mayor cantidad de personal universitario en el edificio, las métricas muestran los mejores resultados. Esto guarda relación con lo visto en la figura 4, donde se ve que en esos meses es donde la ubicación de los centroides se diferencia más, permitiendo separar mejor los datos que se agrupan en torno a cada centroide.

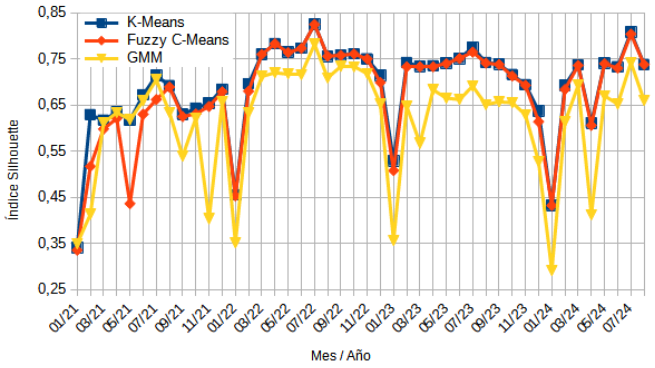

Fig. 6. Evolución de la métrica Silhouette a lo largo del período 2021-2024. Los valores más cercanos a 1 indican una buena calidad del agrupamiento, mientras que los que tienden a 0 corresponden a agrupamientos que están en el límite del umbral de decisión.

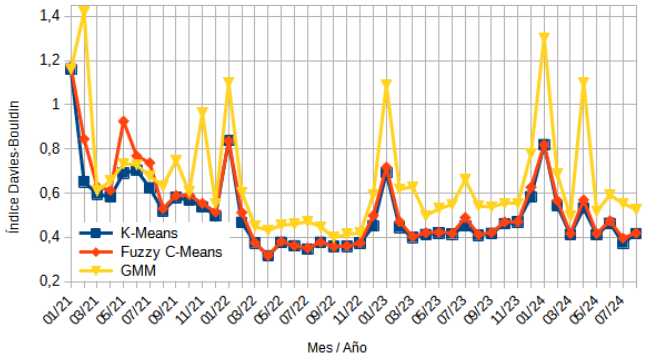

Fig. 7. Evolución de la métrica Davies-Bouldin a lo largo del período 2021-2024.

C. Comportamiento en función del tiempo

En cuanto al análisis temporal, se observaron algunas diferencias entre los agrupamientos obtenidos con cada algoritmo, en función del mes y año. En general los agrupamientos son similares, identificando los intervalos horarios en los que hay presencia / ausencia de personas en el edificio y diferenciando los días hábiles de los días correspondientes a los fines de semana y los feriados. Sin embargo, se detectaron pequeñas diferencias entre los métodos, principalmente entre los agrupamientos obtenidos con K-Means y Fuzzy C-Means, respecto de GMM. La más importante se observó en el mes de abril de 2024, donde el agrupamiento obtenido con GMM es completamente diferente al obtenido con los otros dos métodos (Fig. 8). El criterio de clasificación para K-Means y Fuzzy C-Means parece haberse determinado en función del comportamiento de la potencia reactiva, que, como se ve en la gráfica temporal, experimenta un descenso notorio a partir del día 11 de abril. Este cambio en la potencia reactiva se debió a la desconexión de un banco de capacitores obsoleto que afectaba el factor de potencia de la Facultad de Ingeniería.

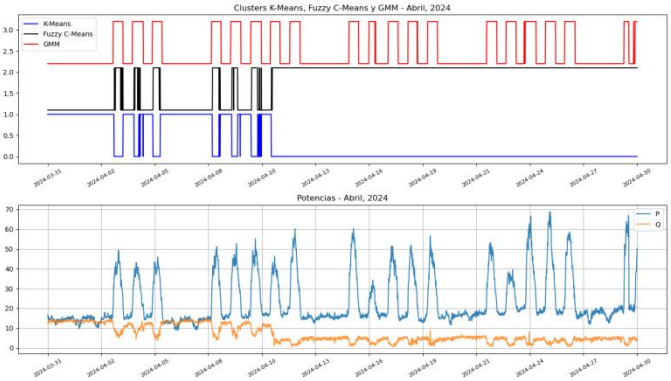

Fig. 8. En la parte inferior de la figura se observan las curvas de P y Q para el mes de abril de 2024, mientras que en la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) se muestra con dos niveles, alto y bajo, que corresponde a cada uno de los dos centroides (los cuales fueron graficados en las figuras 4 y 5).

En el mes de mayo de 2022 se obtienen agrupamientos idénticos para todos los algoritmos, excepto para el día miércoles 18 de mayo (Fig. 9). En ese día se realizó el Censo Nacional 2022, por lo cual no hubo actividad en el ámbito de la universidad. Eso se ve claramente en la curva de potencia activa, aunque casi no se aprecia en la potencia reactiva. Aquí los tres algoritmos muestran comportamientos diferentes. Mientras K-Means interpreta las mediciones de potencia como correspondientes a un periodo de inactividad, el algoritmo Fuzzy C-Means detecta alguna anomalía, mientras que GMM lo clasifica como si fuera un día laborable. Se podría deducir que en este caso, a diferencia del visto en la figura 8, GMM agrupa los datos tomando como referencia la variación en la potencia reactiva.

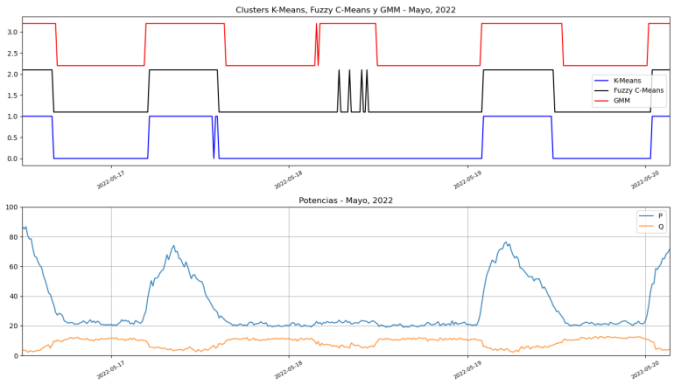

Fig. 9. En la parte inferior de la figura se observa un detalle de las curvas de P y Q para la tercera semana del mes de mayo de 2022, en donde se observa el cambio de comportamiento del día 18 de mayo, que corresponde con el Censo Nacional 2022. En la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) se muestra con dos niveles, alto y bajo, que corresponde a cada uno de los dos centroides (los cuales fueron graficados en las figuras 4 y 5).

En mayo de 2024 se observa una situación similar a la del Censo de 2022, aunque en este caso debido a otro evento no programado, como lo fue el paro nacional del 9 de mayo. En este caso se observa un marcado descenso en la potencia activa, al punto tal que los algoritmos K-Means y Fuzzy C-Means agrupan las mediciones de todo ese día como correspondientes a un periodo de inactividad. Sin embargo, GMM identifica a esos consumos como correspondientes a un día de actividad (Fig. 10).

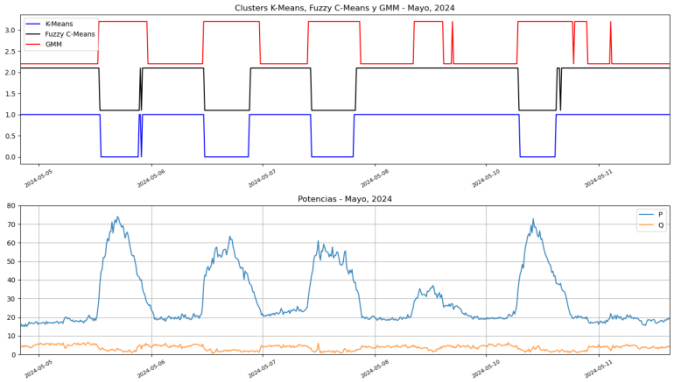

Fig. 10. En la parte inferior de la figura se observa un detalle de las curvas de P y Q para el mes de mayo de 2024, mientras que en la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) se muestra con dos niveles, alto y bajo, que corresponde a cada uno de los dos centroides (los cuales fueron graficados en las figuras 4 y 5).

En febrero de 2023 se observa otra diferencia entre los algoritmos, que parece estar asociada a variaciones en la potencia activa (Fig. 11).

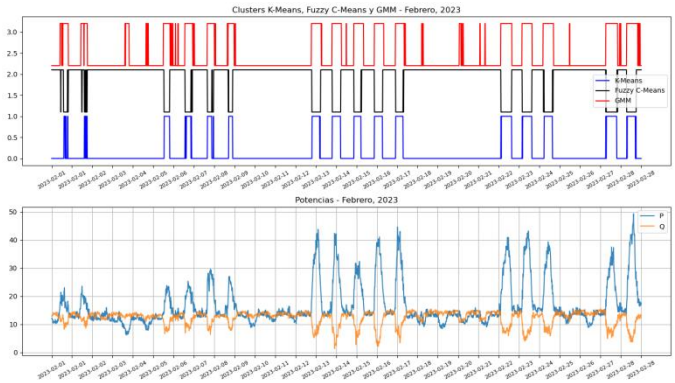

Fig. 11. En la parte inferior de la figura se observan las curvas de P y Q para el mes de febrero de 2023, mientras que en la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) se muestra con dos niveles, alto y bajo, que corresponde a cada uno de los dos centroides (los cuales fueron graficados en las figuras 4 y 5).

D. Agrupamientos con K>2

Agrupando los datos en dos grupos (K=2) permite identificar dos tipos de comportamientos bien diferentes, que pueden asociarse a los intervalos de tiempo de actividad y falta de actividad, que generalmente van de la mano de la cantidad de personas que ocupan el edificio. Adicionalmente, este planteo permite identificar algunos comportamientos anómalos, tal como se vio en los ejemplos previamente mostrados. El agrupamiento de datos en tres o más grupos (K>2), si bien identifica mayor variedad de comportamientos, no tiene una correspondencia tan clara con respecto a eventos o condiciones de actividad en el edificio universitario. El análisis de todo el periodo 2021-2024 empleando K=3 muestra que las métricas de calidad del agrupamiento se deterioran considerablemente, tal como se aprecia en las figuras 12 y 13. En particular se nota un gran deterioro en los agrupamientos obtenidos con Fuzzy C-Means respecto de los obtenidos con K=2.

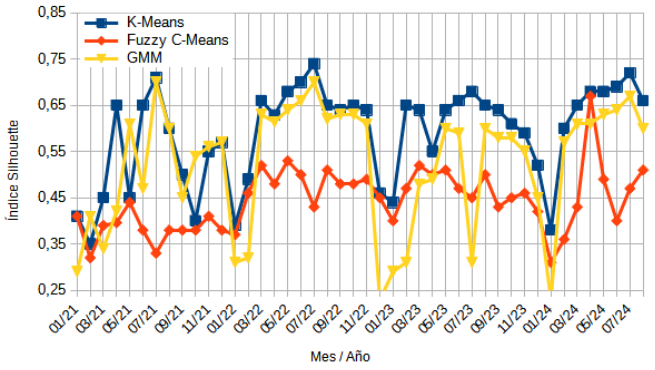

Fig. 12. Evolución de la métrica Silhouette a lo largo del período 2021-2024, empleando K=3.

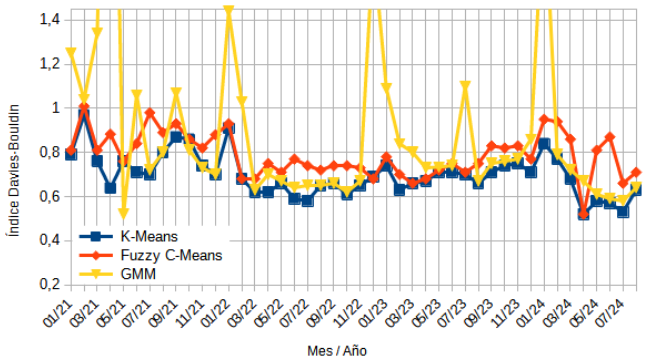

Fig. 13. Evolución de la métrica Davies-Bouldin a lo largo del período 2021-2024, empleando K=3.

El análisis de los agrupamientos respecto del tiempo muestra resultados diferentes respecto del caso con K=2, lo cual es lógico, teniendo en cuenta que ahora los agrupamientos se organizan en torno a tres centroides en lugar de dos. A modo de ejemplo se presenta el caso de abril de 2024, el cual fue analizado previamente (ver Fig. 8). Al evaluarse los mismos datos con K=3, se obtienen los agrupamientos que se muestran en la figura 14. A diferencia del caso con K=2, donde la reducción de la potencia reactiva era advertida solo por los algoritmos K-Means y Fuzzy C-Means, en el caso de K=3 se ve como GMM ahora identifica el cambio en la potencia reactiva.

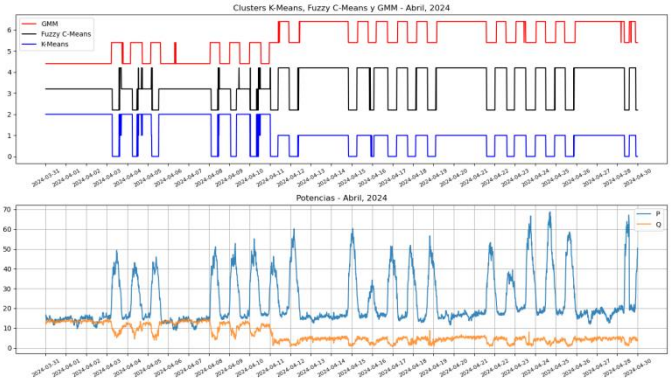

Fig. 14. En la parte inferior de la figura se observan las curvas de P y Q para el mes de abril de 2024, mientras que en la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) tiene tres niveles, que corresponden a cada uno de los tres centroides.

Otro ejemplo donde se aprecia que el uso de un número mayor de grupos puede no ser conveniente es cuando se analiza el comportamiento de mayo de 2024, en donde hubo un día laborable sin actividad debido a un paro nacional (Fig. 10). En la figura 15 se puede ver el resultado del agrupamiento en este caso. Nótese que solo GMM identifica los periodos de actividad y ausencia de la misma, mientras que los otros dos algoritmos agrupan los primeros días de la semana de una forma que no tiene un claro sentido técnico. Además, el día del paro, donde se aprecia un descenso de la actividad a través de la brusca caída del consumo, es interpretado por GMM como un día de actividad normal, mientras que los otros dos algoritmos parecieran no identificarlo apropiadamente.

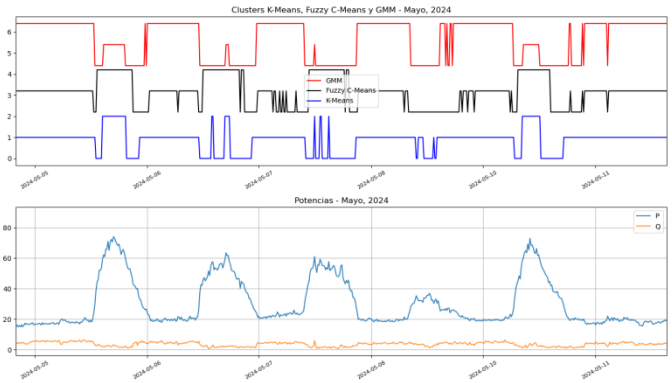

Fig. 15. En la parte inferior de la figura se observa un detalle de las curvas de P y Q para el mes de mayo de 2024, mientras que en la parte superior se observa cómo se asocian los datos temporales a cada grupo y con cada algoritmo. Cada una de las trazas correspondiente a cada algoritmo (KM, FCM y GMM) se muestra con tres niveles, que corresponden a cada uno de los centroides.

V. Conclusiones

En este trabajo se han presentado los primeros resultados obtenidos de la aplicación de algoritmos de agrupamiento automático, no supervisados, a la base de datos de demanda de potencia activa, reactiva y aparente correspondiente al período 2021 a 2024. Se observa que los tres algoritmos evaluados presentan desempeños muy similares, aunque con algunas diferencias puntuales que corresponden a eventos anómalos en las características del consumo, especialmente algunos relacionados con variaciones de la demanda de potencia reactiva. También se ha comprobado que la forma en que se considera la potencia reactiva en el proceso de agrupamiento es bastante diferente entre los algoritmos más tradicionales, como K-Means y Fuzzy C-Means, y otro muy diferente como el de mezclas Gaussianas. Se requiere profundizar el estudio de estos algoritmos, tanto por medio de una evaluación exhaustiva de diferentes configuraciones de parámetros de los mismos como mediante la evaluación de desempeño con otras bases de datos de demanda. También, como posibles trabajos futuros, habría que considerar otros algoritmos de agrupamiento basados en la densidad, como OPTICS y DBSCAN, aunque algunos ensayos preliminares realizados al respecto han arrojado resultados poco satisfactorios.

VI. Referencias

[1] P.G. Donato, “Las redes eléctricas inteligentes en Argentina: una cuestión estratégica para la presente década”, *Revista Electrónica de Divulgación de Metodologías Emergentes en el Desarrollo de las STEM*, vol. 3(2), pp. 3-19, 2021. https://www.revistas.unp.edu.ar/index.php/rediunp/article/view/317

[2] L. Sun, Y. Chen, Q. Du, R. Ding, Z. Liu, Q. Cheng, “Topology Identification of Low-Voltage Power Lines Based on IEC 61850 and the Clustering Method”, *Energies*, vol. 16, pp. 1-20, enero 2023. https://doi.org/10.3390/en16031126

[3] L. Sun, Y. Chen, Q. Du, H. Xu, W. Wang, “Identification of low-voltage phase lines using IEC 61850 and K-means clustering”, Electric Power Systems Research, vol. 234, 2024. https://doi.org/10.1016/j.epsr.2024.110597

[4] A.E. Ezugwu, A.M. Ikotun, O.O. Oyelade, L. Abualigah, J.O. Agushaka, C.I. Eke, A.A. Akinyelu, “A comprehensive survey of clustering

algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects", *Engineering Applications of Artificial Intelligence*, vol. 110, pp. 1-43, 2022, https://doi.org/10.1016/j.engappai.2022.104743.

[5] Scikit-learn Project. Clustering. https://scikit-learn.org/stable/modules/clustering.html [Accedido el 16/08/24].

[6] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez, I. Perona, "An extensive comparative study of cluster validity índices", *Pattern Recognition*, vol. 46, Issue 1, pp. 243-256, 2013. https://doi.org/10.1016/j.patcog.2012.07.021

VII. BIOGRAFÍAS

Patricio G. Donato nació en Puerto Madryn, Argentina, en 1975. Recibió el grado de Ingeniero Electrónico por parte de la Universidad Nacional de la Patagonia San Juan Bosco (Comodoro Rivadavia, Argentina), en 2000, y el grado de Doctor en Electrónica por parte de la Universidad de Alcalá, (Alcalá de Henares, España), en 2005. Es investigador independiente del CONICET y profesor asociado en la UNMDP. Actualmente trabaja en el ICYTE, instituto de doble dependencia CONICET-UNMDP. Su trabajo de investigación está dedicado a las Redes Eléctricas Inteligentes, incluyendo temas específicos como procesamiento de señales, inteligencia computacional y calidad de la energía eléctrica.

Marcos A. Funes nació en Mar del Plata, Argentina, en 1974. Recibió el grado de Ingeniero Electrónico por parte de la Universidad Nacional de Mar del Plata en 1999, y el grado de Doctor en Electrónica por parte de la misma universidad en 2007. Es investigador independiente del CONICET y profesor asociado en la UNMDP. Actualmente trabaja en el ICYTE, instituto de doble dependencia CONICET-UNMDP y tiene el cargo de Director del Laboratorio de Instrumentación y Control (LIC). Su trabajo de investigación está dedicado a procesamiento de señales, calidad de la energía eléctrica, microrredes de corriente continua y control de convertidores electrónicos de potencia.

Carlos M. Orallo nació en Mar del Plata, Argentina, en 1982. Recibió el grado de Ingeniero Electrónico por parte de la Universidad Nacional de Mar del Plata en 2011, y el grado de Doctor en Electrónica por parte de la misma universidad en 2015. Es profesor adjunto en la UNMDP y actualmente trabaja en el ICYTE, instituto de doble dependencia CONICET-UNMDP. Su trabajo de investigación está dedicado al procesamiento de señales y la calidad de la energía eléctrica en el marco de las redes eléctricas inteligentes.