



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Pronóstico de la demanda de energía en Colombia a corto plazo basado en un modelo híbrido adaptativo

José Fernando Montoya Cardona

Universidad Nacional de Colombia
Facultad de Minas, Área curricular de Sistemas de Informática
Medellín, Colombia
2021

Pronóstico de la demanda de energía en Colombia a corto plazo basado en un modelo híbrido adaptativo

José Fernando Montoya Cardona

Trabajo final de investigación presentado como requisito parcial para optar al título de:
Magister en Ingeniería Analítica

Director:

Juan David Velásquez Henao, MSc, PhD

Línea de Investigación:

Analítica de Mercados de Energía

Universidad Nacional de Colombia
Facultad de Minas, Área curricular de Sistemas de Informática
Medellín, Colombia
2021

La mayor motivación es la pasión

Agradecimientos

El autor desea agradecer a las siguientes personas que fueron importantes para el desarrollo exitoso de esta tesis:

- Al profesor Juan David Velásquez Henao, director de esta investigación, por sus enseñanzas, lineamientos y recomendaciones
- A mi familia por todo el apoyo emocional y consejos para la realización de este trabajo
- A todas aquellas personas que apoyaron mi proceso de formación académica con sus recomendaciones y sugerencias.

Resumen

Pronóstico de la demanda de energía en Colombia a corto plazo basado en un modelo híbrido adaptativo

La precisión del pronóstico de la demanda horaria de energía eléctrica es fundamental para realizar una planificación adecuada de los recursos de generación, ya que las desviaciones altas en el pronóstico generan sobrecostos en la operación del sistema. En este trabajo se propone una metodología novedosa de pronóstico basada en el agrupamiento de series de tiempo y los modelos ARIMA; específicamente, el modelo realiza el agrupamiento por tipos de día; seguidamente, agrega las series pertenecientes a un mismo grupo; luego, descompone las series agregadas usando descomposición aditiva; después, se pronostican las series con modelos ARIMA donde se utilizan como variables exógenas las componentes espectrales de Fourier para considerar la estacionalidad y finalmente, se combinan los pronósticos. El modelo propuesto fue utilizado para pronosticar la demanda horaria desde el 13 de enero de 2020 hasta el 15 de marzo de 2020. Los pronósticos del modelo propuesto fueron comparados con los pronósticos del modelo del Centro Nacional de Despacho (Colombia), encontrándose mejoras de hasta un 50% en la precisión con el modelo propuesto.

Palabras clave: demanda de energía, precisión del pronóstico, agrupamiento, series de tiempo, k-means, ARIMA, Fourier.

Abstract

The forecasting energy demand in Colombia in the short term based on an adaptive hybrid model

The forecasting accuracy of the hourly electricity demand is essential for planning the resources of generation, since high deviations in the forecast generate cost overruns in the system's operation. In this research, a novel forecasting methodology based in clustering time series and ARIMA models is proposed; specifically, the model performs the clustering by types of day, then adds the time series belonging to the same cluster; it later decomposes the aggregate series using additive decomposition, then time series are forecasted with ARIMA models where the Fourier spectral components are used as exogenous variables to consider seasonality and finally, the results of the forecast are combined. The proposed model was used to forecast hourly demand from January 13, 2020 to March 15, 2020. The results of the proposed model were compared with the model of the National Dispatch Center (Colombia), getting improvements of up to 50% of accuracy with the proposed model.

Keywords: energy demand, forecasting accuracy, clustering, time series, k-means, ARIMA, Fourier

Contenido

| | Pág. |
|--|-------------|
| Resumen | IX |
| Lista de figuras..... | XIII |
| Lista de tablas | XIV |
| 1 Introducción | 15 |
| 1.1 Definición del problema de negocio | 16 |
| 1.2 Definición del problema de analítica | 17 |
| 1.3 Marco teórico y antecedentes..... | 17 |
| 1.4 Revisión de literatura..... | 20 |
| 1.4.1 Metodología..... | 21 |
| 1.4.2 Resultados obtenidos | 23 |
| 1.4.3 Discusión | 29 |
| 1.4.4 Conclusiones..... | 32 |
| 1.5 Discusión y justificación de la propuesta del trabajo final..... | 33 |
| 1.6 Información disponible | 34 |
| 1.7 Hipótesis | 34 |
| 1.8 Objetivos de la investigación..... | 34 |
| 1.8.1 Objetivo general..... | 34 |
| 1.8.2 Objetivos específicos..... | 34 |
| 2 Metodología propuesta | 35 |
| 2.1 Esquema general de la metodología propuesta | 35 |
| 2.1 Definición del modelo matemático..... | 36 |
| 2.1.1 Representación mediante coeficientes espectrales de Fourier | 36 |
| 2.1.2 Análisis de Componentes Principales..... | 36 |
| 2.1.3 k-means..... | 37 |
| 2.1.4 Descomposición Estacional y de Tendencia por Loess | 37 |
| 2.1.5 Modelo ARIMA..... | 38 |
| 2.1.6 Combinación de pronósticos..... | 38 |
| 2.2 Estimación de parámetros..... | 38 |
| 2.2.1 Determinación de cantidad óptima de grupos..... | 38 |
| 2.2.2 Criterio de Información Akaike | 40 |
| 3 Calibración y evaluación del modelo propuesto para el mercado colombiano..... | 41 |
| 3.1 Información utilizada | 41 |
| 3.2 Determinación de los perfiles típicos de los días | 41 |
| 3.3 Agrupamiento para la identificación de los patrones de consumo | 41 |
| 3.4 Pronóstico del modelo propuesto | 44 |
| 3.5 Evaluación de la precisión | 46 |
| 4 Evaluación desde el punto de vista del negocio | 49 |

| | | |
|----------|--|-----------|
| 5 | Conclusiones | 51 |
| 5.1 | Respuesta a la pregunta de investigación, ¿Los modelos de pronóstico basados en técnicas de agrupamiento son más precisos que los modelos de pronóstico utilizados en el CND? | 51 |
| 5.2 | Cumplimiento del objetivo general..... | 51 |
| 5.3 | Cumplimiento de los objetivos específicos..... | 52 |
| 5.3.1 | Modelar la dinámica de la demanda de energía a corto plazo usando un modelo híbrido adaptativo que represente adecuadamente el comportamiento de los consumos de los usuarios, manteniendo los modelos de pronóstico del CND | 52 |
| 5.3.2 | Evaluar el desempeño del modelo híbrido adaptativo del objetivo anterior respecto a la metodología actual que se tiene para el pronóstico de la demanda de energía a corto plazo | 52 |
| | Bibliografía | 55 |

Lista de figuras

| | Pág. |
|--|------|
| Figura 1-1 Visión general del algoritmo de agrupamiento adaptativo. Tomado de [10] | 19 |
| Figura 1-2 Puntajes de calidad de artículos en relación con el año de publicación | 24 |
| Figura 3-1 Análisis para el tipo de día ORD (ordinarios). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2. | 42 |
| Figura 3-2 Análisis para el tipo de día SAB (sábados). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2. | 43 |
| Figura 3-3 Análisis para el tipo de día DOM (domingos). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2. | 43 |
| Figura 3-4 Análisis para el tipo de día FESTIVO (festivos). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Consumo de los usuarios para el grupo 3; (d) Agregación de consumos de los usuarios del grupo 1; (e) Agregación de consumos de los usuarios del grupo 2; (f) Agregación de consumos de los usuarios del grupo 3. | 44 |
| Figura 3-5 Pronóstico de la demanda del modelo propuesto, del modelo del CND versus la demanda real para nueve (9) semanas del año 2020 | 46 |
| Figura 3-6 Dispersión del error porcentual absoluto para las nueve (9) semanas de pronóstico | 47 |
| Figura 4-1 Distribución de los errores porcentuales del modelo propuesto y del CND | 49 |

Lista de tablas

| | Pág. |
|---|-------------|
| Tabla 1-1 Estudios seleccionados | 25 |
| Tabla 1-2 Evaluación de calidad..... | 28 |
| Tabla 3-1 Resultados de estimación parámetros..... | 42 |
| Tabla 3-2 Parámetros (p, d, q) del modelo ARIMA de las series descompuestas en primer orden. Parámetro k = 75 | 45 |
| Tabla 3-3 Parámetros (p, d, q) del modelo ARIMA de las series descompuestas en segundo orden. Parámetro k = 65 | 45 |
| Tabla 3-4 Métricas de error calculadas para las nueve (9) semanas de pronóstico..... | 46 |
| Tabla 4-1 Desviaciones en el pronóstico con valores mayores al 5%..... | 50 |

1 Introducción

La energía eléctrica es un bien intangible muy importante para el desarrollo sostenible de los sectores económicos de un país y por esto la relevancia de mantener la confiabilidad del sistema eléctrico. Una de las principales características del sistema eléctrico es mantener la frecuencia dentro de unos límites estrictos para que el suministro eléctrico se realice en condiciones de calidad aceptables, y por esto, se debe garantizar el equilibrio entre la generación y la demanda [1]. Por lo anterior, la precisión de los modelos de pronóstico de la demanda es un factor relevante para asegurar la estabilidad del sistema eléctrico [2] mediante la programación de los recursos de generación.

La predicción de la demanda de energía es un problema multivariado de series de tiempo [2]. Históricamente, dentro de la literatura, los modelos más usados para pronosticar la demanda son: regresión multivariada, cointegración, funciones de transferencia y modelos ARIMA [3]. Sin embargo, se ha demostrado que el uso de modelos no lineales como lo son las redes neuronales convolucionales profundas [2] o modelos híbridos de métodos grupales de manejo de datos con máquinas de soporte vectorial de mínimos cuadrados (GLSSVM) [4] tienen un mejor desempeño que los algoritmos de pronóstico clásico [4], [5].

En el caso de Colombia, el pronóstico oficial de la demanda del país lo realiza la Unidad de Planeación Minero-Energética -UPME- con una resolución anual usando modelos econométricos que relacionan variables como el Producto Interno Bruto -PIB-, las tarifas de energía y el crecimiento de la población, y que posteriormente son desagregadas a nivel mensual mediante modelos ARIMA y pronóstico condicional [3]. Estas proyecciones de largo plazo se utilizan como señales del mercado que son estratégicas para la planeación adecuada de los recursos del país. Específicamente para el corto plazo, el pronóstico de la demanda en Colombia, lo realiza el Centro Nacional de Despacho (CND) con la información reportada por los agentes participantes del Mercado de Energía Mayorista (MEM) a través de sus fronteras comerciales registradas ante el Administrador del Sistema de Intercambios Comerciales -ASIC- y teniendo en cuenta el procedimiento y la metodología establecida en el Acuerdo CNO 1303 (2020) [6]. Esta metodología se basa principalmente en realizar el pronóstico de la demanda por Operador de Red (OR), teniendo en cuenta la Ecuación de Demanda (ED):

$$ED = \text{Importaciones} + \text{Generación Interna} - \text{Exportaciones} \quad (1-1)$$

donde las Importaciones y Exportaciones corresponde a las medidas de las fronteras entre agentes, y la Generación Interna a las medidas de las fronteras de generación conectadas a un mismo OR y que no estén conectadas directamente al Sistema de Transmisión Nacional (STN).

La metodología planteada en el Acuerdo CNO 1303 (2020) [6], toma como insumo principal las series de tiempo de la demanda por OR, las cuales se ven afectadas por las variables descriptoras de la ED. Éstas corresponden a los consumos de las fronteras comerciales registradas ante el ASIC y a los consumos de los usuarios que pertenecen al Mercado de Comercialización (MC) de cada OR que no se encuentran registrados ante el ASIC. Esto implica que las variables explicativas de las series de tiempo están definidas por el comportamiento del consumo de los usuarios que a su vez, se caracterizan principalmente por atributos como el tipo de la actividad económica, la ubicación y la carga instalada.

Dado que la demanda de un MC asociado a un OR está compuesta por los consumos de los usuarios, este estudio propone un modelo híbrido, el cual está basado en la aplicación de técnicas de agrupamiento de las

series de tiempo de los usuarios considerando variables exógenas y metodologías para el pronóstico de la demanda de la demanda del Sistema Interconectado Nacional (SIN), que son aplicadas por el CND.

1.1 Definición del problema de negocio

En Colombia, las actividades de planeación y coordinación de la operación de los recursos de generación del Sistema Interconectado Nacional (SIN) están a cargo de XM S.A. E.S.P. en su calidad de Centro Nacional de Despacho (CND) y en cumplimiento de lo establecido por las Leyes 142 y 143 de 1994. Adicionalmente, estas actividades se encuentran con estricta sujeción a lo dispuesto en el Reglamento de Operación expedido por la Comisión de Regulación de Energía y Gas (CREG), los acuerdos expedidos por el Consejo Nacional de Operación (CNO) y la normatividad vigente aplicable en Colombia.

En la actividad de planeación de la operación se encuentran establecidos, conforme a la Resolución CREG 025 de 1995, horizontes de planeamiento tales como: Largo, Mediano, Corto y muy Corto Plazo. El horizonte del Largo Plazo es de cinco años, el Mediano Plazo de cinco semanas, el Corto Plazo de 24 horas y el Muy Corto Plazo desde la hora actual hasta el final del día. En el caso del horizonte de Corto Plazo, se realiza el despacho económico que corresponde al proceso mediante el cual se obtiene para un período de 24 horas, el programa horario de generación de los recursos del SIN despachados centralmente y se efectúa con el criterio de minimizar el costo de atender la demanda. Uno de los procedimientos que se encuentran establecidos en el despacho económico, es la predicción horaria de la demanda de energía, la cual está definida en el Numeral 3.1 de la Resolución CREG 025 de 1995 o aquellas que la modifiquen o sustituyan y en el Acuerdo CNO 1303 de 2020.

Los resultados de la predicción de demanda se publican de manera preliminar todos los miércoles con el pronóstico de la siguiente semana (se considera la semana de lunes a domingo) teniendo en cuenta la información hasta el viernes anterior a ese miércoles y a la Ecuación de Demanda (ED) (ver Ecuación (1-1)). Luego, el viernes el CND publica los pronósticos oficiales los cuales corresponden a los validados por los agentes que participan en el Mercado de Energía Mayorista (MEM) y que, adicionalmente, tienen la opción de realizar ajustes a los valores pronosticados durante la siguiente semana. Se resalta que, a partir del 23 de marzo de 2020, los modelos de pronósticos del CND tuvieron modificaciones debido al inicio de la pandemia producto del COVID-19, donde se tuvo en cuenta más el conocimiento de un experto para el pronóstico; por lo cual, en esta investigación, no es posible comparar el desempeño de la metodología propuesta con respecto a la del CND durante la pandemia porque no tienen la misma base de comparación.

El CNO, mediante acuerdos establece diferentes indicadores de seguimiento de la operación del SIN, dentro de los cuales se encuentra el indicador de calidad del pronóstico oficial. Este indicador, mide para el mes y lo corrido del año, el porcentaje de periodos en que la desviación del pronóstico oficial de demanda para el SIN con respecto a la demanda real, es mayor al valor absoluto del 5%. Estos indicadores permiten un seguimiento continuo de los pronósticos de demanda para asegurar la operación segura y confiable del SIN; y de allí, la importancia de la precisión en las predicciones realizadas por el CND.

Es una necesidad constante la búsqueda de métodos que permitan mejorar los indicadores de seguimiento ya que esto permite la reducción de costos del sistema. En este trabajo final, el problema de negocio considerado es evaluar si los métodos de pronóstico basados en técnicas de agrupamiento (que son descritos más adelante) permiten una mejor precisión del pronóstico, y, por ende, reducir el costo de operación del sistema.

1.2 Definición del problema de analítica

Actualmente, el CND tiene diferentes modelos para el pronóstico de la demanda de energía, dentro de las cuales se encuentran metodologías tales como SARIMA, LSTM, LR, entre otros. Dichos modelos se encuentran entrenados con la información de las series de tiempo de los Mercados de Comercialización (MC) que pertenecen a un Operador de Red, o en el caso de los usuarios que se encuentran conectados directamente al Sistema de Transmisión Nacional (STN) se consideran MC independientes. Estos modelos se actualizan con la mejor información disponible basados en la Ecuación de Demanda (ED) (ver Ecuación (1-1)).

Las series de tiempo de los Mercados de Comercialización (MC) están compuestas por agrupaciones de usuarios pertenecientes a cada MC, por lo cual la demanda de los MC corresponde a la agregación la demanda de cada usuario. En este sentido, el comportamiento de la demanda de las series de cada MC se ven afectadas por los patrones de consumo individuales, ocasionando una posible pérdida de información en la serie de los MC debido a que se ignoran los cambios temporales en el comportamiento del consumo de cada usuario, y por tanto, causando una mayor incertidumbre en los resultados de los modelos de pronóstico, lo que se traduce a una menor precisión de los mismos.

La precisión de los modelos es de gran relevancia, dado que de los resultados del pronóstico de la demanda depende la programación de los recursos de generación para los días futuros, es decir para la adecuada operación del SIN manteniendo y asegurando los criterios de confiabilidad, seguridad y economía. También es importante resaltar, que la CREG le ha dado relevancia a este tema, como se evidencia en el proyecto de Resolución CREG 100 de 2019, donde se establecen una serie de penalizaciones en los casos donde el pronóstico se desvíe ciertos umbrales con respecto a la demanda real. Actualmente, existe un seguimiento de los indicadores para la operación del SIN definidos en el Acuerdo CNO 1406 de 2021, donde se evalúan el número de desviaciones que se han presentado en cada periodo por encima del 5% respecto a la demanda real.

Recientemente se han incrementado el número de publicaciones relacionadas con el pronóstico de la demanda de energía eléctrica donde se presentan diferentes metodologías que demuestran mejores resultados en la precisión de los modelos. De esta manera, surge la necesidad de la búsqueda de nuevas metodologías que permitan mejorar la precisión del pronóstico, en particular, las metodologías basadas en técnicas de agrupamiento las cuales permiten capturar los cambios en el comportamiento temporal de la demanda de los usuarios.

En conclusión, se desea responder la siguiente pregunta de investigación: ¿Los modelos de pronóstico basados en técnicas de agrupamiento son más precisos que los modelos de pronóstico utilizados en el CND?

1.3 Marco teórico y antecedentes

Las nuevas condiciones del mercado enmarcado en la diversificación de la matriz energética para asegurar el crecimiento sostenible motivan el mejoramiento de los modelos de pronósticos que permitan predecir la demanda de energía en el corto plazo. Las principales componentes de la variable demanda, según la UPME, son: el crecimiento poblacional, tarifa de la energía y el PIB [3]. Sin embargo, el cálculo de la demanda del SIN se realiza con agregación de las medidas de las fronteras comerciales registradas ante el ASIC [6] lo cual implica que la demanda del SIN depende de los comportamientos de consumo de los usuarios finales. En consecuencia, se podría pensar en un modelo que permita caracterizar la demanda obteniendo perfiles de los consumos de los usuarios finales para así, mejorar la precisión de las metodologías de pronóstico aplicadas a dichos perfiles.

En diferentes investigaciones se resalta la importancia del agrupamiento de las series de tiempo teniendo en cuenta variables exógenas seguido del pronóstico sobre cada una de las agrupaciones, es decir, del desarrollo de un modelo híbrido para mejorar la precisión del pronóstico de las series de tiempo.

En [7] se desarrolla un modelo adaptativo que se compone de tres etapas: agrupamiento, rastreo y pronóstico. En la primera etapa, los autores construyeron una red para representar interrelaciones entre entidades, donde las interrelaciones están representadas por aristas y las entidades por nodos. Esta estructura en red permite relacionar virtualmente a los clientes que tengan un comportamiento homogéneo. Para su construcción, los autores se basaron en [8] y establecieron métricas de similitud tanto de forma como de proximidad como lo son las distancias Dynamic Time Warping y Kullback-Leibler (forma) y las distancias euclidianas, cuadrática media y Minkowski (proximidad). Finalmente, en esta etapa, utilizaron el algoritmo Fastgreedy [9] para dividir la red en subredes donde los nodos de cada subred se encuentran estrechamente conectados entre sí y escasamente conectados con los nodos de otras subredes, con el objetivo de determinar diferentes grupos que, a su vez, poseen un patrón representativo que refleja los patrones de consumo de los clientes dentro de dicho grupo. En la etapa de rastreo, los autores aplicaron el principio de seguimiento de grupos (subredes) de redes sociales, el cual consiste en alinear las subredes que tienen nodos idénticos en secuencias. En este caso, rastrearon los grupos en función de los perfiles de carga representativos. En la etapa de pronóstico, los autores utilizaron modelos como el SARIMA, MLP, SVR y regresiones LASSO, Bayesiana y KNN, los cuales son aplicados al conjunto de series de perfiles de carga representativos que fueron determinados en la segunda etapa. Los autores realizaron dos experimentos con información de los consumos de energía extraídos de Smart Data Set for Sustainability (SDSS) y el sitio web EnerNOC3 (EnerNOC GreenButton Data). Para el caso de los datos SDSS usaron información de 114 apartamentos residenciales durante 6 meses (enero a junio de 2016) con intervalos de lectura de 15 minutos y para el caso de EnerNOC3 información de 100 establecimientos comerciales e industriales durante el año 2012 con intervalos de lectura de 5 minutos. Los resultados obtenidos para cada uno de los diferentes modelos de pronóstico tuvieron un valor de promedio de error absoluto de 0.000284, lo cual demostró para todos los casos, al realizar la etapa de agrupamiento y rastreo anterior al pronóstico disminuye considerablemente el promedio de error absoluto, teniendo en cuenta que para los otros casos en promedio se tiene un valor de 0.0216. Con esto, los autores propusieron a la comunidad de investigación una metodología para el mejoramiento de los pronósticos de la demanda de energía donde se plantea la incorporación de etapas anteriores al pronóstico de la serie de tiempo de manera adaptativa.

Por otra parte, en [10] también se plantea un modelo adaptivo y recursivo de agrupamiento para las series de tiempo de la demanda de energía con un enfoque en K-medias. El algoritmo de agrupamiento adaptativo inicialmente requiere de tres parámetros: el primer conjunto de perfiles de carga típico, el número de centros de los grupos y la matriz de distancia. Para esto, utilizaron la agrupación de consenso (o conjunto) para determinar dichos parámetros y generar una partición robusta. La agrupación de consenso consiste en ejecutar en paralelo varias instancias de algoritmos de agrupación (posiblemente diferentes) con diferentes valores de K_q que devuelven q particiones Π_q de datos históricos. Posteriormente, los autores generan una matriz de coocurrencia la cual especifica que si dos puntos están en un mismo grupo en cada instancia se crea un enlace entre cada par de cargas. Los beneficios de la agrupación por consenso son múltiples; todas las instancias se usan para construir la matriz de distancia de probabilidad; y la partición final es un consenso entre las diferentes instancias, por lo que reduce el sesgo de cada instancia; se libera de un sesgo potencial debido a la elección del conjunto inicial de centros de clúster. En resumen, en la Figura 1-2, se puede observar de manera general el funcionamiento del modelo adaptativo incluyendo el desarrollo de la etapa del consenso. Luego, los autores plantean un agrupamiento ascendente jerárquico implementada sobre la matriz de distancia de probabilidad con el criterio Ward, para que el usuario pueda determinar el número óptimo de agrupaciones de inicialización.

Con esto, se calcula el promedio de los consumos en cada grupo, los cuales pertenecen a los centroides de las series de tiempo. Finalmente, se actualizan los centroides como lo indica la Figura 1-1.

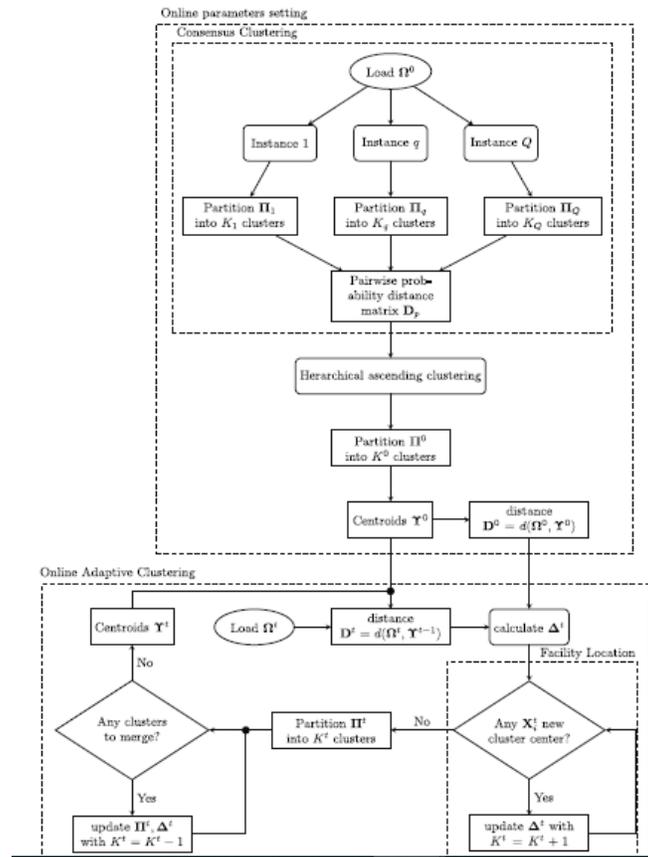


Figura 1-1 Visión general del algoritmo de agrupamiento adaptativo. Tomado de [10]

Los autores aplicaron esta metodología con dos conjuntos de datos reales: (i) cargas de calefacción del distrito central de 97 edificios en Copenhague con resolución por hora durante un mes, (ii) 13,241 cargas eléctricas de industrias, negocios y hogares con PV (es decir, para fines de facturación) con resolución por hora durante todo un año.

Los resultados obtenidos fue un puntaje general RMSE del algoritmo de agrupación adaptativa en línea de 7.2% contra 7.5% y 7.9% para K-medias y SOM respectivamente. Por lo anterior, los autores concluyeron que el algoritmo de agrupación adaptativa es más ágil que el perfil de carga basado en agrupamiento tradicional, ya que es recursivo y utiliza los datos más recientes para actualizar los perfiles de carga típicos diarios y es computacionalmente eficiente, ya que procesa solo series de tiempo cortas en cada paso del agrupamiento en línea.

Otro campo de estudio como las finanzas también resaltan la importancia de tener modelos híbridos que contengan inicialmente algoritmos de agrupamiento para mejorar la precisión de los modelos de pronóstico de precios [11].

Por último, con base en lo expuesto anteriormente, se concluye que la incorporación de técnicas de agrupamiento para la identificación de los perfiles de demanda de los usuarios utilizada como insumo para realizar el pronóstico de la demanda, mejora la precisión de los modelos de pronóstico.

1.4 Revisión de literatura

En esta sección se desarrolla la revisión de literatura, que tiene como fin, la identificación los aportes teóricos en el desarrollo y aplicación de modelos de pronóstico de la demanda de energía eléctrica basados en técnicas de agrupamiento.

La precisión del pronóstico de la demanda de energía eléctrica tiene un gran impacto sobre la confiabilidad y seguridad de un sistema eléctrico, debido a que los errores en los pronósticos generan desviaciones significativas entre la generación programada y la demanda real, y, por ende, inestabilidades en el sistema [1]. Esto ha llevado a una exploración constante de metodologías más precisas y a ser un campo ampliamente investigado en la literatura.

Los modelos de pronóstico de la demanda de energía eléctrica pueden ser aplicados, principalmente, en tres casos: el primero, a series de tiempo de tiempo individuales, es decir, el modelo se aplica a los consumos de cada usuario final (hogares, fábricas, establecimientos comerciales, etc). El segundo, a series de tiempo que son obtenidas al utilizar alguna técnica de agrupamiento; esto es, las series que representan los patrones de consumo para cada grupo. Y, por último, a una serie de tiempo agregada, la cual corresponde a los consumos de todos los usuarios finales. Con respecto al primer caso, al aplicar un modelo de pronóstico para cada usuario final se necesitaría, en la mayoría de los casos, tantos modelos como usuarios finales se tengan para lograr un buen ajuste del modelo, debido a que los patrones de consumo de cada usuario no dependen de otros, y, por ende, sería ineficiente utilizar este enfoque. El segundo caso, corresponde a los modelos de pronóstico basados en técnicas de agrupamiento, los cuales permiten identificar patrones de consumo de los usuarios finales con el fin de asociarlos en grupos similares. Estas técnicas logran capturar la temporalidad de las series de los consumos, es decir los cambios en los patrones de consumo de un intervalo a otro, además de que se reduce la cantidad de series de consumo a pronosticar con relación al primer caso, asociado a los modelos de las series individuales. El tercer caso, correspondiente al pronóstico de una única serie de tiempo determinada por la agregación de todos los consumos de los usuarios finales, permite reducir significativamente el volumen de datos, y, por ende, el tiempo de respuesta del modelo. No obstante, al agregar las series de consumo de los usuarios finales se está desconociendo que cada uno de ellos es independiente, y, por tanto, el patrón de consumo de un usuario puede ser diferente de los demás usuarios. Esto conllevaría a una posible pérdida de información puesto que el comportamiento de la serie agregada ignora los hechos estilizados asociados a los consumos de los usuarios [7]. Por ejemplo, si un usuario cambia su patrón de consumo de un intervalo a otro, debido a cambios en su comportamiento de consumo, esto no sería capturado por la serie agregada. Por lo anterior, si el comportamiento de los consumos de los usuarios no es homogéneo, es decir sus patrones de consumo cambian de un intervalo a otro, los modelos de pronóstico aplicados a las series agregadas tendrían menor precisión en relación con los modelos que son capaces de capturar dichos cambios como los modelos de pronóstico basados en técnicas de agrupamiento.

Por lo tanto, las técnicas de agrupamiento se consideran pasos importantes en el preprocesamiento de los datos proporcionando información sobre los atributos del comportamiento de los consumos de los usuarios y capturando la temporalidad de sus patrones de consumo con tiempos de respuesta y precisión aceptables. Como ejemplo de las técnicas de agrupamiento se tienen: la agrupación aglomerativa, mapas autoorganizados, la

agrupación de colonias de hormigas, el vecino más cercano basado en lógica difusa, la deformación dinámica del tiempo, la detección comunitaria en redes, entre otros [7], [12]–[28].

Consecuentemente con lo anterior, las técnicas de agrupamiento son metodologías prometedoras para mejorar la calidad del pronóstico sin caer en altos costos computacionales. Sin embargo, desde la práctica es necesario que el modelador seleccione las técnicas más apropiadas para realizar la agrupación de las series de tiempo, las cuales son dependientes de los datos particulares. De esta forma, surge la necesidad de realizar una revisión de literatura que permita establecer cómo han sido aplicadas las técnicas de agrupamiento para el pronóstico de las series de consumo de energía; esto con el fin de poder desarrollar un esquema experimental que permita llegar a los mejores modelos para el pronóstico de la demanda en el caso colombiano. En efecto, se plantearon las siguientes preguntas de investigación:

- RQ1. ¿Cuáles son los métodos de agrupamiento de series de tiempo que han sido propuestos en la literatura?
- RQ2. ¿Qué técnicas se utilizan para preprocesar las series de tiempo antes de realizar la detección de grupos?
- RQ3. ¿Cómo se mide la similitud de las series de tiempo para realizar su agrupamiento?
- RQ4. ¿Qué técnicas se utilizan para determinar la cantidad óptima de grupos de series de tiempo?
- RQ5. ¿Qué métodos se han utilizado para pronosticar los grupos de serie de tiempo?
- RQ6. ¿Qué métodos se han utilizado para la agregación de pronósticos de los grupos?
- RQ7. ¿Qué beneficios se han reportado en la literatura sobre la agrupación de series de tiempo orientada al pronóstico?

La relevancia de esta revisión se basa principalmente en tres aspectos: primero, en lograr identificar las diferentes técnicas que se han utilizado para el agrupamiento de series de tiempo desde el punto de vista procedimental y de su formulación matemática; segundo, en determinar los algoritmos de pronóstico utilizados en las series de tiempo y las metodologías para agregarlos; y tercero, en conocer el impacto que tiene la aplicación de las técnicas de agrupamiento en los modelos de pronóstico.

El resto de este capítulo está organizado de la siguiente manera: en la Sección 1.4.1 se presenta la metodología de investigación, en la Sección 1.4.2 se describen los resultados obtenidos y por último, en las Secciones 1.4.3 y 1.4.4 se presenta la discusión y las conclusiones respectivamente.

1.4.1 Metodología

En este estudio se aplicó la metodología de Revisión Sistemática de Literatura -RSL- propuesta por Kitchenham [29] en el contexto de la ingeniería de software. Esta metodología permite buscar, identificar y sintetizar toda la evidencia documental con el fin de responder a una determinada pregunta de investigación [30]. La RSL, además permite solucionar una de las principales falencias en las revisiones de literatura, la cual corresponde a la mala o falta de formulación de las preguntas de investigación.

1.4.1.1 Proceso de búsqueda

El proceso de búsqueda consistió en una consulta manual utilizando el sistema bibliográfico SCOPUS, el cual es una de las bases de datos más grande de referencias y resúmenes de literatura que existen en el mundo. Se utilizó el siguiente criterio de búsqueda:

```
TITLE-ABS-KEY ( cluster* AND "time series" ) AND ( LIMIT-
TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-
```

```
TO ( SUBJAREA , "MATH" ) ) AND ( LIMIT-TO ( EXACTKEYWORD , "Energy
Utilization" ) OR LIMIT-TO ( EXACTKEYWORD , "Smart Power Grids" ) OR LIMIT-
TO ( EXACTKEYWORD , "Electric Power Utilization" ) OR LIMIT-
TO ( EXACTKEYWORD , "Electricity-consumption" ) OR LIMIT-TO ( EXACTKEYWORD , "Wind
Power" ) OR LIMIT-TO ( EXACTKEYWORD , "Efficiency" ) OR LIMIT-
TO ( EXACTKEYWORD , "Electric Power Plant Loads" ) OR LIMIT-
TO ( EXACTKEYWORD , "Electric Utilities" ) OR LIMIT-TO ( EXACTKEYWORD , "Energy
Efficiency" ) OR LIMIT-TO ( EXACTKEYWORD , "Wind" ) )
```

1.4.1.2 Criterios de inclusión y exclusión

En esta revisión se definieron diferentes criterios de inclusión y exclusión para la selección de los artículos a analizar, los cuales se presentan a continuación:

Criterios de inclusión:

- Se propone un modelo de pronóstico de grupos de series de tiempo.

Criterios de exclusión:

- El artículo aplica una metodología previamente utilizada en otro artículo con modificaciones mínimas.
- Presenta únicamente aspectos metodológicos sin un caso de aplicación.

1.4.1.3 Evaluación de la calidad

Para realizar el proceso de evaluación de calidad de los documentos seleccionados se propusieron las siguientes preguntas:

- QA1. ¿Es explícita la formulación matemática del método de agrupamiento?
- QA2. ¿Es explícita la técnica de transformación de datos (preprocesamiento) de series de tiempo?
- QA3. ¿Son explícitas las métricas de similitud de los grupos?
- QA4. ¿Es explícito el proceso o técnica de estimación óptima de grupos?
- QA5. ¿Es explícita la formulación matemática del modelo de pronóstico?
- QA6. ¿Es explícita la metodología aplicada para la agregación de pronósticos?
- QA7. ¿Se demuestran mejores resultados en los pronósticos basados en agrupamiento?

Las preguntas de calidad fueron calificadas mediante una valoración cualitativa y cuantitativa representadas por las letras S, P y N que corresponden a valores de 1, 0.5 y 0 respectivamente. A continuación, se define para cada pregunta el significado de cada valoración:

- QA1: S (sí), la formulación matemática del método de agrupamiento es presentada de manera explícita en el artículo; P (parcialmente), la formulación matemática es presentada de manera implícita y; N (no), no se define explícitamente la formulación matemática ni puede ser inferida del contenido del artículo.
- QA2: S (sí), es explícita la formulación matemática de la técnica de transformación de datos; P (parcialmente), se menciona la técnica de transformación de datos, pero no presenta de manera explícita su formulación matemática y; N (no), no se aborda en el artículo el tema de transformación de datos en el preprocesamiento.

- QA3: S (sí), es explícita la formulación matemática de la(s) métrica(s) de similitud de los grupos; P (parcialmente), se menciona(n) la(s) métrica(s) de similitud, sin embargo, no está(n) definida(s) matemáticamente y; N (no), no se aborda el tema de métricas de similitud entre grupos.
- QA4: S (sí), es explícita la formulación matemática y el procedimiento para la estimación óptima de grupos; P (parcialmente), el procedimiento de estimación óptima de grupos se realiza implícitamente o hace referencia a otros artículos y; N (no), no se aborda el tema de estimación óptima de grupos.
- QA5: S (sí), la formulación matemática del modelo de pronóstico es presentada de manera explícita en el artículo; P (parcialmente), la formulación matemática es presentada de manera implícita y; N (no), no se define explícitamente la formulación matemática ni puede ser inferida del contenido del artículo.
- QA6: S (sí), se presenta de manera explícita la formulación matemática y el método aplicado para la agregación de pronósticos; P (parcialmente), es mencionado o se puede inferir el método de agregación de pronósticos, pero no es explícita la formulación matemática y; N (no), no se aborda el tema de agregación de pronósticos ni es posible inferirlo del artículo.
- QA7: S (sí), se demuestra de manera explícita mejores resultados en el pronóstico basado en agrupamiento con un caso de aplicación; P (parcialmente), los resultados de pronóstico poseen un desempeño similar a otros modelos no basados en agrupamiento y; N (no), no se demuestran mejores resultados de los modelos de pronóstico basados en agrupamiento o no se realiza una comparación con otros modelos.

Adicionalmente, se consideraron los siguientes criterios:

- QA8: Valor del indicador Scimago Journal Rank (SJR), el cual refleja el prestigio de la fuente, esto es, el valor de las citas ponderadas por documento de acuerdo con el año de publicación del artículo.
- QA9: Impacto normalizado por fuente del artículo (SNIP), que define la razón del conteo de citas de la revista por artículo y el potencial de citas en el campo de interés.

1.4.1.4 Recolección de datos

La información recolectada de cada documento de estudio fue la siguiente: nombre del artículo publicado, referencia completa, resumen del estudio, modelos propuestos, la calidad de la investigación, y los indicadores Source-Normalized Impact per Paper (SNIP) y Scimago Journal Rank (SJR) en el año de publicación del artículo.

1.4.2 Resultados obtenidos

A continuación, se presenta los resultados de la revisión sistemática de literatura.

1.4.2.1 Resultados de la búsqueda

Al utilizar la cadena de búsqueda en el sistema bibliográfico SCOPUS, se obtuvieron 298 resultados, de los cuales, al aplicar de forma manual los criterios de inclusión y exclusión, establecidos en la Sección 1.4.1.2, se seleccionaron 28 artículos que corresponden a un 9% y cuyas referencias se pueden observar en la Tabla 1-1.

De este proceso, se resalta que, a pesar de los 298 resultados, muy pocos artículos presentan metodologías novedosas para el pronóstico de series de tiempo basados en agrupamiento y que contengan casos de aplicación con datos reales. Adicionalmente se encontró que, de los artículos seleccionados, la mayor cantidad se presentaron en el año 2019 con un total de 6 artículos.

1.4.2.2 Evaluación de la calidad

Para los 28 estudios analizados, los cuales se muestran en la Tabla 1-1 con su respectiva referencia, año de publicación, modelo propuesto (agrupamiento y pronóstico) y conclusiones, se les aplicaron los criterios de calidad definidos en el numeral 1.4.2.2.

La información de la evaluación de la calidad se presenta en la Tabla 1-2 que se compone de los siguientes campos: identificación del artículo, el cual está relacionado con los señalados en la Tabla 1-1, la referencia, en las columnas desde la QA1 hasta la QA7 se muestran los puntajes de calidad dados a cada artículo, la columna de puntaje total que corresponde a la suma de los valores obtenidos de las columnas QA1 a QA7 y las columnas QA8 y QA9 se observan los datos asociados a los indicadores de SJR y SNIP respectivamente.

Los resultados obtenidos muestran que el puntaje promedio fue de 3.9 ± 1 , es decir los puntajes de calidad varían entre 2.9 y 4.9. Se resalta que para los años 2012 al 2015, 2017 y 2020 los puntajes obtenidos son homogéneos, es decir están concentrados sobre la media, contrario a los puntajes de los años 2016, 2018 y 2019, los cuales son más heterogéneos (Véase la Figura 1-3). Adicionalmente, se puede observar que ningún artículo cumple con los 7 criterios de calidad y que el mayor puntaje se presentó para el año 2018, con un valor de 6 puntos y corresponde al artículo [18].

Por otra parte, se encontró que no existe relación entre los indicadores SJR y SNIP (columnas QA8 y QA9 de la Tabla 1-2) y el puntaje total obtenido según los valores de los coeficientes de correlación Spearman, los cuales fueron de -0.1468 para el caso del indicador SJR y de -0.1805 para el SNIP.

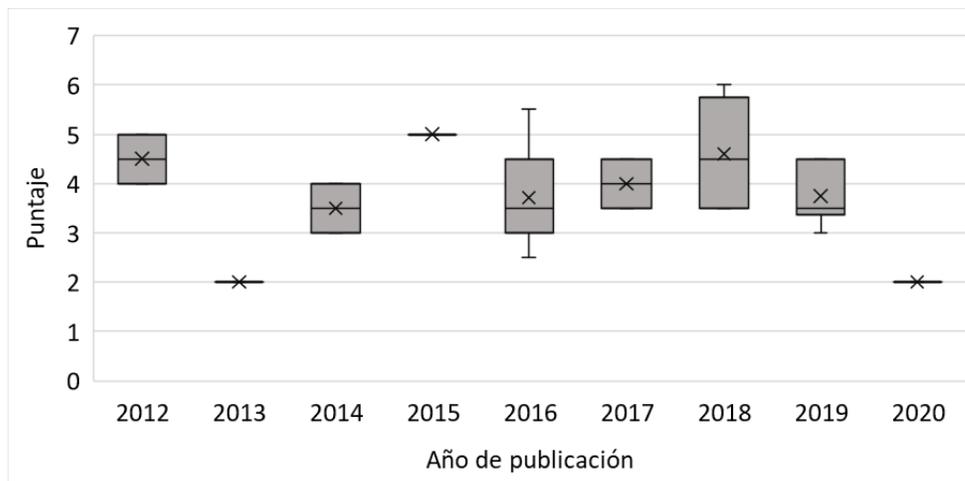


Figura 1-2 Puntajes de calidad de artículos en relación con el año de publicación

Tabla 1-1 Estudios seleccionados

| ID | Autores [Referencia] | Año | Modelo propuesto | Conclusiones |
|----|---|------|---|--|
| S1 | Zhang, W., Mu, G., Song, C., Yan, G., Heuveline, V. [31] | 2020 | One-class SVM, DPMM | Se propone una metodología de pronóstico basada en características espacio-temporales. Se reconocen estados anormales de carga espaciales y se realiza el agrupamiento con el método de DPMM (Dirichlet Process Mixture Model). Se evidencia una alta correlación de las distribuciones espaciales de las cargas y son representadas mediante grafos espacio-estado. Finalmente, se demuestra una mayor precisión en el pronóstico. |
| S2 | Laurinec, P., Lóderer, M., Lucká, M., Rozinajová, V. [13] | 2019 | Bootstrapping (K-means, K-means.N), DBSCAN-OPTICS | Se propone una metodología de bootstrapping para generar los grupos utilizando K-means y K-means.N (Generación de datos aleatorios con distribución gaussiana normal), determinando el k-óptimo de grupos con la métrica del índice DB (Davies-Boulding). Se aplica el método de ensamble DBSCAN-OPTICS (Density-Based Spatial Clustering of Applications with Noise - Ordering Points To Identify the Clustering Structure) utilizando coeficientes de regresión estimados, k-means. Finalmente, se demuestra una mayor precisión en el pronóstico. |
| S3 | Sun, S., Fu, J., Li, A. [14] | 2019 | DBSCAN-TSD-BSA-LSSVM, DBSCAN-TSD-BSA-ELM, DBSCAN-TSD-BSA-WNN, and DBSCAN-TSD-ARIMA | Se propone un método de ensamble mediante cuatro técnicas de pronóstico que son integrados por medio de la optimización de pesos con BSA (Backtracking Search Algorithm). Cada técnica de pronóstico utiliza DBSCAN como técnica de agrupamiento y luego TSD (Two-Stage Decomposition) para descomponer la serie en componentes más estables. Finalmente, se utiliza modelos de pronóstico como LSSVM (Least Square Support Vector Machine), ELM (Extreme Learning Machine), WNN (Wavelet Neural Networks), ARIMA (Autoregressive Integrated Moving Average) donde los parámetros de cada modelo son optimizados mediante BSA (Backtracking Search Algorithm). |
| S4 | Wang, Y., Shen, Y., Mao, S., Chen, X., Zou, H. [15] | 2019 | k-means++, LLIFA (LASSO-LSTM Integrate Forecasting Algorithm) | Se propone un modelo de pronóstico LASSO (Least Absolute Shrinkage and Selection Operator) y LSTM (Long Short Term Memory) basado en series agrupadas mediante k-means++. Se demuestra una mejora en el pronóstico frente a modelos como ARIMA (Autoregressive Integrated Moving Average) y SVM (Support Vector Machine). |
| S5 | Laurinec, P., Lucká, M. [16] | 2019 | k-medoids (basado en PAM), ClippedStream, Forecasting (EXP, CART, CTREE, ARIMA, RF) | Se propone una metodología denominada ClippedStream que consiste en el acortamiento de las series de tiempo mediante la representación en FeaClips (características de la serie de tiempo transformada en binaria). Luego se aplica una técnica de k-medoids (PAM-Partition Around Medoids) para el agrupamiento de las series. Finalmente, se aplican diferentes modelos de pronóstico como EXP (Triple Exponential Smoothing), CART (Recursive Partitioning Regression Tree), CTREE (Conditional Inference Trees), ARIMA (Autoregressive Integrated Moving Average) y RF (Random Forests). |

| ID | Autores [Referencia] | Año | Modelo propuesto | Conclusiones |
|-----|---|------|---|--|
| S6 | Tajeuna, E.G., Bouguessa, M., Wang, S. [7] | 2019 | Detección comunitaria en redes. SARIMA, regresión LASSO y Ridge para el pronóstico | Se propone un modelo de agrupamiento, rastreo y pronóstico. Para el agrupamiento, se utiliza un método adaptativo basado en la detección de redes comunitarias. Luego se aplica una técnica de rastreo donde se dividen los grupos en subredes similares para seleccionar el nodo que representa a la red. Finalmente, se puede utilizar cualquier modelo de pronóstico de series de tiempo, en este caso se realizó con SARIMA, regresión LASSO y Ridge para el pronóstico. Se evidencia una gran mejoría al utilizar un modelo híbrido (clustering, tracking y forecasting) para la predicción de la demanda. |
| S7 | Jain, R., Jain, N., Gupta, Y., (...), Chugh, T., Hemanth, D.J. [17] | 2019 | kmFuzzy (k-means based Fuzzy logic), Fuzzy based logic forecasting algorithm | Se propone una metodología de lógica difusa para el pronóstico. Se aplica k-means como técnica para determinar el número de relaciones de los conjuntos difusos. Esto con el fin de que las reglas estén relacionadas con los patrones de consumo. |
| S8 | Gajowniczek, K., Zabkowski, T. [12] | 2018 | DTW, Lp, Fourier coefficients, and periodogram-based distances, RF, NNET, SVR | Se propone una serie de métodos para realizar agrupamiento y pronóstico. Para el caso de agrupamiento, se presentan mejores resultados con técnicas como DTW, Lp, Fourier coefficients y periodogram-based distances. Para el caso de pronóstico, se propone modelos de RF, NNET y SVR. |
| S9 | Singh S., Yassine A. [18] | 2018 | Incremental clustering based on appliance associations and k-means, BN | Se propone una metodología incremental de agrupamiento basado en asociaciones de distintos dispositivos (PC, teteras, lavadora, TV, etc) y k-means. Para esto se aplica diferentes técnicas de medición como Kulczynski e IR (Imbalance Ratio) para medir la correlación y el balance de los datos. Además, se aplica k-means como técnica adicional para realizar las agrupaciones. Por último, se aplica BN (Bayesian Network) para el modelo de pronóstico. |
| S10 | Shchetinin, E.Y. [19] | 2018 | k-means, STL+ARIMA, GBM, SVR, RF, Bagging | Se propone una metodología de pronóstico basada en agrupamiento mediante k-means. Se evalúan modelos como STL+ARIMA (Seasonal decomposition of time series based on loses regression + Autoregressive Integrated Moving Average), GBM (Gradient Boosting Machine), SVR (Support Vector Machine), RF (Random Forest) y Bagging. Para agregar los pronósticos, se utilizan diferentes técnicas como Median, HW (Holt-Winters exponential smoothing), GAM (Generalized Additive Model), RLM (Robust Linear Model), SUM. |
| S11 | Sun, W., Zhang, C. [32] | 2018 | FA-ACC-BA-ELM | Se propone un modelo FA-ACC-BA-ELM (Factor Analysis-Ant Colony Clustering-Bat Algorithm-Extreme Learning Machine). Inicialmente, se realiza un análisis de correlación de diferentes variables exógenas, luego se aplica ACC como técnica de agrupamiento y finalmente, para el pronóstico, se utiliza BA-ELM (Bat Algorithm-Extreme Learning Machine). BA es utilizado para optimización de los parámetros. |

| ID | Autores [Referencia] | Año | Modelo propuesto | Conclusiones |
|-----|--|------|--|--|
| S12 | Auder, B., Cugliari, J., Goude, Y., Poggi, J.-M. [20] | 2018 | DWT, HAC, PAM, KWF | Se propone un modelo de arriba hacia abajo (bottom-up), el cual consiste, inicialmente, en realizar extracción de características de las series de tiempo mediante DWT (Discrete Wavelet Transform). Luego, se aplica HAC (Hierarchical Agglomerative Clustering) para realizar agrupamiento y se utiliza la técnica de PAM (Partitioning Around Medoids) para fijar el número óptimo de particiones. Finalmente, se utiliza un modelo KWF (Kernel-Wavelet-Functional) para realizar el pronóstico, es decir el principio general del modelo de pronóstico es encontrar en el pasado situaciones similares al presente y combinar linealmente sus futuros para construir los pronósticos. Se demuestra mejores resultados en el pronóstico con agrupamiento respecto a pronosticar la serie total. |
| S13 | Li, L., Ota, K., Dong, M. [21] | 2017 | SOM, Deep Learning based Short-term Forecasting (DLSF) | Se propone un método de pronóstico a corto plazo basado en aprendizaje profundo (DLSF). Se transforma la tarea de pronóstico en un problema de procesamiento de imágenes y se diseña una red convolucional profunda de doble rama para extraer las características de los datos de entrada para el proceso de agrupamiento. Para el agrupamiento se utiliza la técnica SOM (Self-Organized Map). Para la predicción, se propone una red neuronal multicapa para pronosticar futuras variaciones de carga. En comparación con otros modelos de predicción simples y métodos SVM, el método DLSF propuesto tiene en cuenta todos los factores de influencia externos, p. temperatura, humedad, velocidad del viento, etc., y supera los métodos existentes en la precisión del pronóstico. |
| S14 | Ghofrani, M., Azimi, R., Najafabadi, F.M., Myers, N. [22] | 2017 | GTSOM, BNN | Se propone una metodología de agrupamiento denominada GTSOM (Game Theory Self-Organized Map), la cual es una modificación a la técnica de SOM. Para el pronóstico, se aplica BNN (Bayesian Neural Network). |
| S15 | Yang, J., Ning, C., Deb, C., (...), Sekhar, C., Tham, K.W. [23] | 2017 | K-Shape, SVR | Se propone un modelo de pronóstico SVR (epsilon-Support Vector Regression) donde se embebe un modelo de agrupamiento k-Shape que se compara con DTW (Dynamic Time Warping) con 7 índices de validación de clasificación (Silhouette index, Dunn index, Context-independent Optimality and Partiality index, Davies-Bouldin index, Modified Davies-Bouldin index, Calinski-Harabasz index, and Score Function). Se demuestran mejores resultados en el pronóstico con la clasificación (prueba T-test). |
| S16 | Ghofrani, M., Carson, D., Ghayekhloo, M. [33] | 2016 | k-means, BNN | Se propone un modelo de agrupamiento mediante k-means y un modelo de pronóstico BNN (Bayesian Neural Network). |
| S17 | Panapakidis I.P. [24] | 2016 | minCEntropy, ANNs | Se propone un modelo de agrupamiento mediante minCEntropy y un modelo de pronóstico ANN (Artificial Neural Network). |
| S18 | Laurinec, P., Loderer, M., Vrablecova, P., (...), Rozinajova, V., Ezzeddine, A.B. [34] | 2016 | GAM, k-means++, Bagging | Se propone una metodología de agrupamiento mediante k-means++ (se diferencia de k-means principalmente por el inicio aleatorio de los centroides) partiendo de una representación de las series de tiempo con GAM (Generalized Additive Model). Finalmente, se utiliza Bagging como modelo de pronóstico. |
| S19 | Liang, X., Hong, T., Shen, G.Q. [25] | 2016 | k-means, DTL | Se propone un modelo de agrupamiento utilizando k-means. Se utiliza DTL (Decision Tree Learning) para el pronóstico el cual es basado en árboles de decisiones. |

| ID | Autores [Referencia] | Año | Modelo propuesto | Conclusiones |
|-----|--|------|---------------------------------------|--|
| S20 | Laouafi, A., Mordjaoui, M., Laouafi, F., Boukelia, T.E. [35] | 2016 | FKW-ANFIS | Se propone una metodología denominada FKW-ANFIS la cual combina los beneficios de agrupamiento de FCM (Fuzzy c-means), el algoritmo k-NN (k-Nearest neighbors), WPD (Wavelet Packet Decomposition) y Adaptive Neuro-Fuzzy Inference System (ANFIS). |
| S21 | Talavera-Llames, R.L., Pérez-Chacón, R., Martínez-Ballesteros, M., Troncoso, A., Martínez-Álvarez, F. [36] | 2016 | kNN | Se propone una metodología de big data para el pronóstico. Se utiliza kNN (k-Nearest Neighbours) como técnica de agrupamiento y un método WNN (weighted nearest neighbours) como pesos para el pronóstico el cual es el resultado de una combinación lineal. |
| S22 | Chen, Y.H., Hong, W.-C., Shen, W., Huang, N.N. [26] | 2016 | GHSA-FTS-LSSVM | Se propone un modelo GHSA-FTS-LSSVM (Global Harmony Search Algorithm-Fuzzy Time Series-Least Squares Support Vector Machines), el cual utiliza FTS como técnica de agrupamiento y LSSVM como modelo para el pronóstico el cual es optimizado mediante GHSA. |
| S23 | Wang, W., Pedrycz, W., Liu, X. [37] | 2015 | DTW-FCM | Se propone un modelo de agrupamiento de FCM (Fuzzy c-means) basado en DTW (Dynamic Time Warping) y se determina el peso de cada regla difusa con respecto a la observación de entrada y se usan dichos pesos para determinar la salida predicha en función del esquema de interpolación de reglas difusas múltiples. |
| S24 | Bennett, C.J., Stewart, R.A., Lu, J.W. [38] | 2014 | Correlation, ARIMAX | Se propone una metodología de agrupamiento mediante análisis estadístico (correlación). Finalmente, se aplica un modelo ARIMAX para el pronóstico. |
| S25 | Chaouch, M. [27] | 2014 | CFWK | Se propone un modelo CFWK (Clustering Functional Wavelet Kernel) el cual consiste en aplicar un algoritmo jerárquico para el agrupamiento y se pronostica aplicando un modelo funcional FWK. |
| S26 | Humeau, S., Wijaya, T.K., Vasirani, M., Aberer, K. [39] | 2013 | k-means, SVR | Se propone un modelo k-means para el agrupamiento. Luego se aplica un modelo de pronóstico SVR a cada grupo y finalmente se agregan los resultados mediante la operación de la suma. |
| S27 | Mori, H., Takahashi, M. [28] | 2012 | DA, GRBFN | Se propone un modelo de agrupamiento DA (Deterministic Annealing), el cual es una mejora al k-means debido a que no tiene dependencia de dónde se inicia (parámetros iniciales) y posteriormente se aplica un modelo GRBFN (Generalized Radial Basis Function Network) para el pronóstico de los grupos. |
| S28 | Ding, Z., Yang, P., Yang, X., Zhang, Z. [40] | 2012 | Hierarchical Tree, Mathematical model | Se propone una metodología de agrupamiento mediante un árbol jerárquico, luego se modela una ecuación (polinomio de grado n) ajustada a cada centroide de cada grupo y finalmente se pronostica de acuerdo con esa función ajustada. |

Tabla 1-2 Evaluación de calidad

| ID | Referencia | QA1 | QA2 | QA3 | QA4 | QA5 | QA6 | QA7 | Puntaje total | QA8 | QA9 |
|----|------------|-----|-----|-----|-----|-----|-----|-----|---------------|------|------|
| S1 | [31] | S | P | P | N | N | N | N | 2 | 0.61 | 1.72 |
| S2 | [13] | P | N | P | P | N | P | S | 3 | 0.42 | 1.09 |
| S3 | [14] | N | P | N | P | P | S | S | 3.5 | 0.42 | 1.09 |
| S4 | [15] | P | N | N | P | S | P | S | 3.5 | 1.40 | 3.87 |

| ID | Referencia | QA1 | QA2 | QA3 | QA4 | QA5 | QA6 | QA7 | Puntaje total | QA8 | QA9 |
|-----|------------|-----|-----|-----|-----|-----|-----|-----|---------------|------|------|
| S5 | [16] | P | S | P | P | P | P | S | 4.5 | 0.91 | 2.67 |
| S6 | [7] | S | N | S | P | P | P | S | 4.5 | - | - |
| S7 | [17] | P | P | N | P | P | P | S | 3.5 | 0.51 | 0.99 |
| S8 | [12] | P | N | S | S | S | S | S | 5.5 | 0.54 | 0.86 |
| S9 | [18] | S | P | S | S | S | P | S | 6 | 0.61 | 1.16 |
| S10 | [19] | P | N | P | P | P | P | S | 3.5 | 0.17 | 0.39 |
| S11 | [32] | P | P | N | P | S | P | P | 3.5 | 0.61 | 1.16 |
| S12 | [20] | P | S | N | P | S | P | S | 4.5 | 0.61 | 1.16 |
| S13 | [21] | P | S | S | N | P | P | S | 4.5 | - | - |
| S14 | [22] | S | N | P | P | P | P | S | 4 | - | - |
| S15 | [23] | P | P | P | P | P | N | S | 3.5 | 1.93 | 2.12 |
| S16 | [33] | S | N | S | N | S | P | P | 4 | - | - |
| S17 | [24] | N | N | P | N | P | P | S | 2.5 | 1.26 | 1.99 |
| S18 | [34] | P | P | P | P | P | P | P | 3.5 | - | - |
| S19 | [25] | S | P | P | P | P | P | S | 4.5 | 1.88 | 2.20 |
| S20 | [35] | P | P | P | N | P | P | P | 3 | 1.26 | 1.99 |
| S21 | [36] | S | N | P | N | P | P | P | 3 | 0.28 | 0.71 |
| S22 | [26] | S | S | P | S | P | P | S | 5.5 | 0.61 | 1.16 |
| S23 | [37] | S | N | S | P | S | S | P | 5 | 0.88 | 1.81 |
| S24 | [38] | S | N | S | P | P | P | P | 4 | 2.05 | 1.82 |
| S25 | [27] | P | N | N | N | S | P | S | 3 | 3.36 | 3.51 |
| S26 | [39] | N | N | P | N | N | S | P | 2 | - | - |
| S27 | [28] | S | N | S | P | S | P | S | 5 | - | - |
| S28 | [40] | S | P | S | N | S | P | N | 4 | 1.26 | 1.99 |

1.4.3 Discusión

En este capítulo se presentan las respuestas a las preguntas de investigación

1.4.3.1 RQ1. ¿Cuáles son los métodos de agrupamiento de series de tiempo que han sido propuestos?

Los métodos o técnicas de agrupamiento encontrados en los estudios relacionados en la Tabla 1-1 fueron los siguientes:

- K-means [13], [19], [25], [33], [39]
- K-means++ [15], [34]
- K-means based Fuzzy Logic (kmFuzzy) [17]
- K-Nearest Neighbour (kNN) [36]
- K-medoids [16]
- K-Shape [23]
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [14]

- Dynamic Time Warping (DTW) [12]
- Dynamic Time Warping – Fuzzy Cmeans (DTW-FCM) [37]
- Ant Colony Clustering (ACC) [32]
- Hierarchical Tree (HT) [40]
- Hierarchical Agglomerative Clustering (HAC) [20]
- Self-Organized Map (SOM) [21]
- Game Theory Self-Organized Map (GTSOM) [22]
- minCEntropy [24]
- Fuzzy kNN Wavelet (FKW) [35]
- Fuzzy Time Series (FTS) [26]
- Clustering Functional Wavelet Kernel (CFWK) [27]
- Deterministic Annealing (DA) [28]
- Community Detection in Networks [7]
- A priori [18]
- Dirichlet Process Mixture Model (DPMM) [31]

1.4.3.2 RQ2. ¿Qué técnicas se utilizan para preprocesar las series de tiempo antes de realizar la detección de grupos?

De los estudios relacionados en la Tabla 1-1, se han utilizado las siguientes técnicas para el preprocesamiento de las series de tiempo:

- Normalización z-score [12], [13], [23], [31], [34]
- Extracción de coeficientes de una regresión lineal múltiple [13]
- Escalamiento Max-Min
- Transformaciones en representaciones binarias FeaClips [16]
- Extracción de características mediante Convolutional Neural Network (CNN) [21]
- Extracción de características basado en conocimiento experto [24], [32]
- Transformaciones de la serie de tiempo con GAM (Generalized Additive Models) y HW (Holt-Winters exponential smoothing) [19]
- Extracción de características mediante DWT (Discrete Wavelet Transform) [20]

1.4.3.3 RQ3. ¿Cómo se mide la similitud de las series de tiempo para realizar su agrupamiento?

La similitud de las series de tiempo, según los estudios relacionados en la Tabla 1-1, se miden con las siguientes métricas:

- Media [31]
- Desviación estándar [31]
- Distancia Euclidiana [13], [15], [16], [18], [19], [33], [34]
- Distancia Manhattan
- Distancia Minkowshi [7], [21]
- Distancia basada en coeficientes de Fourier [12]
- Distancia basada en periodograma [12]
- Distancia basada en correlación cruzada [12]

- Distancia TQuest [12]
- DTW [7], [12], [17]
- DBSCAN [14]
- Kullback-Leiber [7]
- ACC [32]
- Similitud de forma de series de tiempo (SBD) [7], [23]
- Enlace Ward para las técnicas HAC [20]
- Estrategias de teoría de juegos con diferentes funciones de costos [22]
- Criterio mínimo condicional de entropía (kernel Gaussiano) [24]

1.4.3.4 RQ4. ¿Qué técnicas se utilizan para determinar la cantidad óptima de grupos de series de tiempo?

De los estudios relacionados en la Tabla 1-1, se han utilizado las siguientes técnicas para determinar la cantidad óptima de grupos de series de tiempo:

- Davies-Bouldin Index (DBI) [13], [16], [19], [21], [22], [34]
- Partitioning Around Medoids (PAM) [20]
- Método del codo (Thorndike) [17], [23]
- Criterio experto de parada de un algoritmo iterativo (K-means) [33]
- Coeficiente de Silhouette [18]
- Voto mayoritario de los índices [12]:
 - Calinski and Harabasz index (ch)
 - Duda index (duda)
 - C index (cindex)
 - Ptbiserial index (ptbiserial)
 - DB index (db)
 - Frey index (frey)
 - Hartigan index (hartigan)
 - Ratkowsky index (ratkowsky)
 - Ball index (ball)
 - McClain index (mcclain)
 - KL index (kl)
 - Silhouette index (silhouette)
 - Dunn index (dunn)
 - SD index (sdindex)

1.4.3.5 RQ5. ¿Qué métodos se han utilizado para pronosticar los grupos de serie de tiempo?

Los métodos que se han utilizado para pronosticar los grupos de serie de tiempo, según los estudios relacionados en la Tabla 1-1, son los siguientes:

- Conditional Inference Trees (CTREE) [16]
- Random Forest (RF) [16]
- Autoregressive Integrated Moving Average (ARIMA) [13], [14], [38]

- Support Vector Machine (SVM) [23], [31], [40]
- Seasonal Autoregressive Integrated Moving Average (SARIMA) [7]
- Least Square Support Vector Machine (LSSVM) [14], [26]
- Extreme Learning Machine (ELM) [14], [32]
- Wavelet Neural Network (WNN) [14], [36]
- Long-Short Term Memory (LSTM) [15]
- Regresión LASSO [7]
- Reglas de lógica difusa [17], [37]
- Bayesian Neural Network (BNN) [18], [22], [33]
- Gradient Boosting Machine (GBM) [19]
- Kernel Wavelet Functional (KWF) [20]
- Multi Layer Neural Network (MLNN) [21]
- Artificial Neural Network (ANN) [24]
- Generalized Radial Basis Function Network (GRBFN) [28]
- Bagging [13], [34]

1.4.3.6 RQ6. ¿Qué métodos se han utilizado para la agregación de pronósticos de los grupos?

De los estudios relacionados en la Tabla 1-1, se han utilizado los siguientes métodos para la agregación de pronósticos de los grupos:

- Mediana [13]
- Media [13]
- Backtracking Search Algorithm (BSA) [14]
- Suma [12], [37], [39]

1.4.3.7 RQ7. ¿Qué beneficios se han reportado en la literatura sobre la agrupación de series de tiempo orientada al pronóstico?

En general, de los estudios relacionados en la Tabla 1-1, los beneficios que se han reportado en la literatura sobre la agrupación de series de tiempo orientada al pronóstico corresponden a mayor precisión que las metodologías tradicionales ya que la modelación se realiza sobre series de tiempo que tienen un patrón de comportamiento, con lo cual los modelos se ajustan en mayor medida a dicho patrón y, en consecuencia, se evidencia una reducción en tiempos de entrenamiento de los modelos.

1.4.4 Conclusiones

En esta sección se llevó a cabo la Revisión Sistemática de Literatura (RSL) cuyo objetivo fue la identificación los aportes teóricos en el desarrollo y aplicación de modelos de pronóstico de la demanda de energía eléctrica basados en técnicas de agrupamiento. Esta revisión fue desarrollada en tres secciones: una de metodología, otra de resultados y, por último, la discusión. En la metodología fueron definidos los criterios de búsqueda, de inclusión y de exclusión para la selección de los artículos y las preguntas con las cuales se evalúan la calidad de cada uno de los estudios. En la Sección de resultados, se evidencian algunas estadísticas obtenidas como respuesta a las preguntas de evaluación de calidad de los 28 estudios seleccionados, los cuales fueron elegidos al aplicar los criterios de inclusión y exclusión de la sección metodológica. En la discusión, fueron resueltas

las preguntas de investigación planteadas al inicio de esta Sección, las cuales estaban orientadas a identificar cada una de las técnicas o metodologías propuestas para el desarrollo de modelos de pronóstico de demanda de energía eléctrica basada en técnicas de agrupamiento, empezando desde la etapa de preprocesamiento de los datos hasta la etapa de agregación de los pronósticos de cada grupo. En general, se identificaron: 8 métodos para el preprocesamiento de los datos, 22 tipos de técnicas de agrupamiento dentro de las cuales existen 17 tipos de métricas para medir la similitud de los grupos y 6 técnicas para determinar la cantidad óptima de grupos; 18 métodos para el pronóstico de los grupos y 4 metodologías para la agregación de dichos pronósticos. Finalmente, en la mayoría de los estudios analizados los modelos de pronóstico basados en técnicas de agrupamiento demuestran una mayor precisión y reducción de tiempos de entrenamientos de los modelos con relación a las metodologías tradicionales. En conclusión, se evidencia que las metodologías de pronóstico basadas en técnicas de agrupamiento permiten llegar a los mejores modelos para el pronóstico de la demanda, los cuales pueden ser aplicados en el caso colombiano.

1.5 Discusión y justificación de la propuesta del trabajo final

Hasta aquí se ha demostrado que las metodologías de pronóstico basadas en técnicas de agrupamiento son más precisas que los modelos tradicionales. No obstante, para obtener un resultado de alto impacto, esta investigación se debe enfocar en un único problema. Se hace necesario modificar la metodología actual del pronóstico de la demanda a corto plazo mediante un modelo basado en técnicas de agrupamiento que permita mejorar la precisión de la predicción, con el objetivo de disminuir la incertidumbre en la programación de los recursos de generación, además de la que se tiene con la incorporación de las fuentes de energía renovables no convencionales en el sistema eléctrico.

La metodología actual del pronóstico se define en el Acuerdo CNO 1303 (2020) [6], el cual establece que el pronóstico debe realizarse a las series de tiempo agregadas de la demanda del MC de cada OR, lo que conlleva a que se ignoren los hechos estilizados asociados a los consumos de los usuarios; además, se ignora la incapacidad de capturar la temporalidad de las series de los consumos, es decir los cambios en los patrones de consumo que se den de un intervalo a otro. En ese sentido, los modelos de pronóstico basados en técnicas de agrupamiento permiten que la estimación de los valores futuros de la demanda tenga menor incertidumbre respecto con la metodología que actualmente realiza el CND.

Las metodologías de pronóstico basadas en técnicas de agrupamiento se basan principalmente en tres etapas: agrupamiento, pronóstico de grupos y agregación de los pronósticos. En la etapa de agrupamiento, se identifican los patrones de consumo de los usuarios finales con el fin de asociarlos en grupos similares, lo cual se puede lograr mediante la aplicación de algunas de las técnicas mencionadas en la Sección 1.4.3.1. En el caso del pronóstico de grupos, el objetivo es estimar los valores futuros de cada uno de los grupos determinados en la etapa anterior y, por último, la etapa de agregación de los pronósticos, la cual tiene como fin calcular mediante alguna técnica de agregación, que pueden ser de las mencionadas en la Sección 1.4.3.6, el valor total de la estimación del modelo basados en los valores estimados del pronóstico de cada grupo.

Finalmente, queda la inquietud si para el caso de Colombia, la metodología actual para el pronóstico de la demanda puede ser modificada mediante la aplicación de una metodología de pronóstico basada en técnicas de agrupamiento que considere la temporalidad de las series de consumo de los usuarios finales con el objetivo de mejorar la precisión de los modelos actuales.

1.6 Información disponible

Los datos para poder realizar el tema desarrollado en el presente documento de acuerdo con los objetivos definidos en la Sección 1.8, son tomados desde la fuente de XM S.A. E.S.P. donde la información de la demanda por usuario es de carácter comercial, y, por lo tanto, de conformidad con lo establecido en el Código de Comercio se encuentra protegida con la reserva legal. En este sentido, se procede a modificar la información de tal manera que no fuese posible la identificación de los usuarios.

Se cuenta con la información desde el año 2017 hasta el año 2020 de manera horaria y expresada en unidades de kWh para cada uno de los usuarios de los diferentes Operadores de Red a los cuales se encuentren conectados. Esta información corresponde a la demanda real de cada usuario, es decir se encuentra referida al nivel de tensión del Sistema de Transmisión Nacional (STN), que corresponde con niveles mayores o iguales 220kV conforme a como lo establece la regulación vigente para realizar los pronósticos de demanda.

1.7 Hipótesis

Es posible implementar un modelo híbrido adaptativo para el pronóstico a corto plazo de la demanda de energía en Colombia que sea más preciso que los métodos usados actualmente por el CND.

1.8 Objetivos de la investigación

A continuación, se presenta el objetivo general y los objetivos específicos que tendrá esta investigación.

1.8.1 Objetivo general

Proponer un modelo híbrido adaptativo para el pronóstico a corto plazo de la demanda de energía en Colombia modificando únicamente los insumos que tienen los modelos de pronóstico del CND.

1.8.2 Objetivos específicos

1. Modelar la dinámica de la demanda de energía a corto plazo usando un modelo híbrido adaptativo que represente adecuadamente el comportamiento de los consumos de los usuarios, manteniendo los modelos de pronóstico del CND.
2. Evaluar el desempeño del modelo híbrido adaptativo del objetivo anterior respecto a la metodología actual que se tiene para el pronóstico de la demanda de energía a corto plazo.

2 Metodología propuesta

La metodología con la cual se desarrolla el modelo implementado en esta investigación se compone de múltiples pasos, donde se pueden resumir principalmente en dos etapas: agrupamiento y pronóstico. En la primera etapa, se desarrolla el preprocesamiento de los datos y la aplicación de la técnica de agrupamiento, generando así, múltiples series de tiempo que representan diferentes patrones de consumo de los usuarios. En la segunda etapa, se implementa el modelo de pronóstico para cada serie de tiempo resultante del agrupamiento y finalmente se aplica una metodología para agrupar dichos pronósticos.

2.1 Esquema general de la metodología propuesta

Los pasos ejecutados en la metodología planteada se explican a continuación:

1. Se toman todas las series de tiempo asociadas a los consumos de cada usuario pertenecientes a un mismo Mercado de Comercialización -MC- y se segmentan por los 4 tipos de días que, para el caso de Colombia, son principalmente los días ordinarios (lunes, martes, ..., viernes), sábados, domingos y festivos. Dicha segmentación será explicada en mayor detalle en la Sección 3.1.
2. Se realiza una representación de cada serie de tiempo con los primeros sesenta (60) coeficientes espectrales de la transformada de Fourier, definida en la Sección 2.1.1 en la Ecuación 2-1, lo cual dicha representación equivale al número de características de cada serie.
3. Se aplica una normalización z-score para cada una de las series transformadas y una técnica de reducción de características denominada Análisis de Componentes Principales (PCA, por sus siglas en inglés). Ver Ecuación 2-2.
4. Se aplica la técnica de agrupamiento k-means la cual es explicada en la Sección 2.1.3 y se determina la cantidad óptima de grupos de acuerdo con lo indicado en la Sección 2.2.1.
5. Se agregan los consumos de las series agrupadas por grupo, obteniendo así, k series de tiempo que representan los diferentes patrones de consumo para cada tipo de día.
6. Se realizan dos descomposiciones aditivas de las series de tiempo (STL, por sus siglas en inglés) de acuerdo con mencionado en la Sección 2.1.4, en la Ecuación 2-3, donde cada serie es representada como la agregación de su parte estacional, de tendencia y residual. La primera descomposición, es de primer orden y la segunda, de segundo orden.
7. Se aplica el modelo de pronóstico ARIMA, explicado en la Sección 2.1.5, donde son entrenados dos modelos. El primero, para el caso de la serie de tiempo descompuesta de primer orden y el otro para el caso de la descomposición de segundo orden. Esta metodología es aplicada para cada grupo.
8. Se aplica la metodología de ensamble de los resultados de ambos modelos mediante la media y se obtiene la serie pronosticada.

9. Se realiza la comparación de los resultados del pronóstico de esta metodología con los resultados del pronóstico realizado por el CND y con la serie de demanda real.

El proceso de pronóstico es repetido incrementalmente de manera que cada semana que se obtienen nuevos datos reales, es decir que ya se cuenta con la información de lo que pasó, y el modelo es ajustado con esta nueva información para pronosticar los nuevos valores futuros.

2.1 Definición del modelo matemático

En esta sección se presentan las definiciones matemáticas de las técnicas y modelos desarrollados para la implementación de la metodología planteada, así como la explicación de los diferentes métodos para la estimación de los parámetros.

2.1.1 Representación mediante coeficientes espectrales de Fourier

Las series de Fourier permiten representar una señal periódica mediante la superposición de diferentes series de senos y cosenos. Dentro de una de sus representaciones, se encuentra la representación compleja de la serie de Fourier, la cual está definida como [41]:

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{j\omega_n t}$$

Donde:

$$c_0 = \frac{1}{T} \int_{-T/2}^{T/2} f(t) dt = \frac{a_0}{2}$$

$$c_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-j\omega_n t} dt = \frac{1}{2} (a_n - jb_n)$$

$$c_{-n} = \frac{1}{T} \int_{-T/2}^{-T/2} f(t) e^{j\omega_n t} dt = \frac{1}{2} (a_n + jb_n)$$

$$|c_n| = \frac{1}{2} \sqrt{a_n^2 + b_n^2} \quad (2-1)$$

Así las cosas, en esta investigación, se tomaron los primeros sesenta (60) coeficientes espectrales de Fourier $|c_n|$ para la representación de las series de tiempo, lo que es equivalente a decir los primeros sesenta (60) armónicos de la serie transformada en el espacio de la frecuencia.

2.1.2 Análisis de Componentes Principales

El Análisis de Componentes Principales (PCA) busca reducir la dimensionalidad de un conjunto de datos X , compuesto por n variables manteniendo la variabilidad. Esto se logra transformando los datos originales en un nuevo conjunto de variables denominadas componentes principales, las cuales no son correlacionadas entre sí y son ordenadas de tal manera que las primeras variables representan la mayor variabilidad del conjunto de datos originales.

Sea z un vector de J variables aleatorias, PCA se concentra en las varianzas de las J variables aleatorias. Primero, se halla una función lineal $\omega_1^T z$ de los elementos de z , con máxima varianza, donde ω_1 es un vector de J constantes $[\omega_{11}, \omega_{12}, \dots, \omega_{1J}]^T$, se tiene entonces:

$$\omega_1^T z = \omega_{11}z_1 + \omega_{12}z_2 + \dots + \omega_{1J}z_J = \sum_{m=1}^J \omega_{1m}z_m$$

Luego, se busca una función lineal $\omega_2^T z$, no correlacionada con $\omega_1^T z$, con la varianza máxima, y así sucesivamente, hasta que en la m -ésima etapa se encuentre con la función lineal $\omega_m^T z$ con la máxima varianza, sujeto a que no sea correlacionada con $\omega_1^T z, \omega_2^T z, \dots, \omega_{m-1}^T z$. La m -ésima variable derivada, $\omega_m^T z$ es el m -ésimo componente principal. Se pueden encontrar hasta M componentes principales, donde la mayor parte de la varianza de z se encuentre en los primeros q componentes principales ($q \ll M$).

El cálculo de las componentes principales se basa en la matriz de covarianza σ_X cuyo elemento (i, j) corresponde a la covarianza entre los elementos i y j de X cuando $i \neq j$. Para $m = 1, 2, \dots, M$, el k -ésimo componente principal está dado por $\gamma_m = \omega_m^T z$, donde ω_m es el vector propio de σ_X correspondiente al m -ésimo mayor valor propio λ_m .

El conjunto de variables ω_m conforman una matriz de transformación $W \in \mathbb{R}^{n \times q}$ que mapea el conjunto original $X \in \mathbb{R}^{n \times m}$ en un subconjunto, tal que:

$$Y = WX \quad (2-2)$$

Donde $Y \in \mathbb{R}^{n \times q}$ y corresponde al conjunto de características relevantes de X_n .

2.1.3 k-means

La técnica de agrupamiento k-means es utilizada para separar n muestras de datos en k grupos de igual varianza. Esta técnica requiere como parámetro de entrada la cantidad k de grupos que se desean formar. En esta investigación, fue implementada la técnica k-means++ la cual permite disminuir la probabilidad de que los centroides seleccionados en la inicialización de la técnica converjan a un mínimo local.

La técnica k-means divide el conjunto de datos en k grupos C disjuntos, cada uno descritos por la media μ_j de las muestras en el grupo. Las medias son usualmente denominadas centroides. Esta técnica tiene como función objetivo elegir centroides que minimicen la inercia o suma de cuadrados dentro de cada grupo:

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

2.1.4 Descomposición Estacional y de Tendencia por Loess

La descomposición de las series de tiempo (STL) es un método el cual descompone una serie estacional en tres componentes: la estacional, tendencia y residual (ruido). La componente estacional es encontrada por Loess (regresión local) suavizada, la cual puede ser reemplazado por la media. Los valores de la componente estacional son removidos de la serie, quedando así los valores de la componente residual y de tendencia [19]. Para el caso de esta investigación fue utilizada la descomposición aditiva, es decir:

$$s(t) = E(t) + T(t) + \varepsilon(t) \quad (2-3)$$

Donde $s(t)$ representa la serie de tiempo, y $E(t)$, $T(t)$ y $\varepsilon(t)$ representan las diferentes componentes: estacional, tendencia y residual respectivamente.

2.1.5 Modelo ARIMA

El modelo Autorregresivo Integrado de media móvil (ARIMA por sus siglas en inglés) proporciona una descripción de un proceso estocástico estacionario en términos de dos polinomios, uno para la parte autorregresiva y otro para la media móvil que es aplicado en algunos casos donde los datos muestran evidencia de no estacionariedad donde al realizar un proceso de diferenciación, la cual corresponde a la parte integrada del modelo, se podría eliminar la no estacionariedad. Dada una serie de tiempo L_i , un modelo ARIMA (p, d, q) está definido como:

$$\left(1 - \sum_{j=1}^p \varphi_j B^j\right) (1 - B)^d L_i = \delta + \left(1 + \sum_{j=1}^q \theta_j B^j\right) \varepsilon_j$$

Donde p es el orden del modelo autorregresivo (número de rezagos), d es el grado de diferenciación (el número de veces que se aplica diferenciación en los datos), q es el orden de la media móvil del modelo, B^j es el operador de rezago, φ_j son los parámetros de la parte autorregresiva del modelo, θ_j son los parámetros de la parte de media móvil, ε_j son los errores y δ es un factor de derivación $\mu(1 - \varphi_1 - \dots - \varphi_p)$ donde μ es la media de $(1 - B)^d L_i$. Los términos de error ε_j son generalmente asumidos como independientes, idénticamente distribuidos de una distribución normal con media cero [12].

2.1.6 Combinación de pronósticos

Para la combinación de los pronósticos, los cuales corresponden a dos modelos (explicados en la Sección 3.4) para cada grupo y tipo de día (definidos en la Sección 3.2), se promedian los resultados de los modelos, obteniendo así, el pronóstico para cada grupo. Adicionalmente, para determinar el valor final del pronóstico para cada tipo de día, es decir para cada día, se agregan los resultados de los pronósticos de cada grupo [13].

2.2 Estimación de parámetros

En esta sección se presentan las diferentes metodologías aplicadas en esta investigación para determinar los parámetros utilizados en los modelos desarrollados.

2.2.1 Determinación de cantidad óptima de grupos

Para la determinación de la cantidad óptima de grupos se utilizaron las siguientes métricas de evaluación o índices:

- Davies Bouldin Index (DBI): Este índice es una función entre la relación de dispersión dentro de los grupos con separación que se encuentran los grupos. Este índice es calculado utilizando la Ecuación 2-4.

$$DBI(q) = \frac{1}{q} \sum_{k=1}^q \max_{k \neq l} \left(\frac{\delta_k + \delta_l}{d_{kl}} \right) \quad (2-4)$$

Donde,

- $k, l = 1, \dots, q$: cantidad de grupos
- $d_{kl} = \sqrt[p]{\sum_{j=1}^p |c_{kj} - c_{lj}|^p}$: distancia entre los centroides de los grupos C_k y C_l (para $v = 2$, d_{kl} es la distancia Euclidiana)
- p : número de variables
- $\delta_k = \sqrt[u]{\frac{1}{n_k} \sum_{i \in C_k} \sum_{j=1}^p |x_{ij} - c_{kj}|^u}$: medida de dispersión dentro de un grupo C_k (para $u = 2$, δ_k es la desviación estándar de la distancia de las observaciones de un grupo C_k al centroide de este grupo)
- n_k : número de observaciones en el grupo C_k

El valor de q que minimiza el índice es considerado como la cantidad óptima de grupos del conjunto de datos. [12], [42]

- Silhouette Index (SI): Este índice es calculado de acuerdo con la Ecuación 2-5.

$$SI = \frac{\sum_{i=1}^n S(i)}{n}, SI \in [-1, 1] \quad (2-5)$$

Donde,

- $S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}$
- $a(i) = \frac{\sum_{j \in \{C_r \setminus i\}} d_{ij}}{n_r - 1}$: es el promedio de la disimilitud de la observación i -ésima a todas las observaciones del grupo C_r
- $b(i) = \min_{s \neq r} \{d_{iC_s}\}$
- $d_{iC_s} = \frac{\sum_{j \in C_s} d_{ij}}{n_s}$: es el promedio de la disimilitud de la observación i -ésima a todas las observaciones del grupo C_s

El máximo valor de este índice es considerado como la cantidad óptima de grupos del conjunto de datos. Es importante mencionar que $S(i)$ no está definido para $k = 1$ (un solo grupo). [12], [42]

- Calinski y Harabasz Index (CHI): Este índice está definido de acuerdo con la Ecuación 2-6.

$$CHI(q) = \frac{tr(B_q)/(q-1)}{tr(W_q)/(n-q)} \quad (2-6)$$

Donde,

- $B_q = \sum_{k=1}^q n_k (c_k - \bar{x})(c_k - \bar{x})^T$: es la matriz de dispersión entre grupos para los datos agrupados en q grupos
- n_k : número de observaciones en el grupo C_k
- \bar{x} : centroide de los datos de la matriz X
- $X = \{x_{ij}\}, i = 1, 2, \dots, n; j = 1, 2, \dots, p$: matriz $n \times p$ de p variables y n observaciones

- $W_q = \sum_{k=1}^q \sum_{i \in C_k} (x_i - c_k)(x_i - c_k)^T$: es la matriz de dispersión entre las observaciones de un mismo grupo perteneciente a una agrupación de q grupos.
- x_i : es un vector de dimensión p de la observación i -ésima del grupo C_k

El valor de q que maximiza el índice es considerado como la cantidad óptima de grupos para el conjunto de datos.

Basado en el resultado de cada índice, se define la cantidad óptima de grupos por mayoría, es decir, se toma el número de grupo con mayor moda. En caso de no existir consenso entre los índices, se toma como cantidad óptima de grupos el valor del índice Silhouette.

2.2.2 Criterio de Información Akaike

Para determinar el orden de un modelo ARIMA, usualmente es utilizado el criterio denominado Criterio de Información Akaike (AIC por sus siglas en inglés) definido en la Ecuación 2-7:

$$AIC = -2 \log(L) + 2(p + q + k + 1) \quad (2-7)$$

Donde L es el máximo valor de la función de verosimilitud y k es el número de parámetros en el modelo que se ajusta a los datos. En este caso, la función para determinar el orden del modelo está establecida por la minimización del criterio de información Akaike. [12]

3 Calibración y evaluación del modelo propuesto para el mercado colombiano

En esta sección se presentan los resultados obtenidos con la metodología propuesta en esta investigación para el mercado eléctrico colombiano.

3.1 Información utilizada

La información utilizada corresponde a las series de consumos de los usuarios de un Mercado de Comercialización (MC) perteneciente al sistema eléctrico colombiano. El conjunto de datos de entrenamiento considerado es de tres (3) años de información, es decir, el periodo comprendido entre el año 2017 hasta el año 2019. La información del año 2020 fue utilizada para evaluar el desempeño de la metodología.

3.2 Determinación de los perfiles típicos de los días

Para el desarrollo de la metodología implementada en esta investigación se toma como base la tipificación principal de los días que actualmente tienen los modelos de pronóstico del CND. Estos tipos de días corresponden a: ordinarios (ORD – corresponden a los lunes, martes, ..., viernes que no sean festivos), sábados (SAB), domingos (DOM) y festivos (FESTIVO). Lo anterior, implica que el modelo de pronóstico explicado en la Sección 3.3, considera que no existe dependencia de los valores pasados entre los diferentes tipos de días, por ejemplo, los valores de un día tipo sábado no dependen de los valores del viernes anterior que corresponde a un día tipo ordinario. Este tipo de metodología se plantea en el artículo [7].

3.3 Agrupamiento para la identificación de los patrones de consumo

La técnica de agrupamiento utilizada en esta investigación es la definida en la Sección 2.1.3, la cual corresponde a k-means. Esta técnica es aplicada al conjunto de datos de los consumos de los usuarios a nivel horario pertenecientes a uno de los MC que tiene sistema eléctrico colombiano para cada tipo de día definido en la Sección 3.2. Este conjunto de datos, antes de aplicar la técnica de agrupamiento, es transformado mediante la representación, para cada serie de consumo, de los primeros sesenta (60) coeficientes espectrales de Fourier calculados mediante la Ecuación 2-1. Cada uno de estos coeficientes representan las diferentes características para cada serie tiempo. Luego, es reducida la cantidad de características (dimensionalidad) con el objetivo de mejorar la eficiencia en el cómputo y tener la mayor representación de la varianza (información) del conjunto de características. Dicha reducción se realizó según lo definido en la Sección 2.1.2

Como resultado de la aplicación de la técnica de agrupamiento, se obtiene diferentes cantidades de grupos para cada tipo día. La cantidad óptima de grupos para cada tipo de día fue determinada de acuerdo con lo indicado en la Sección 2.2.1. Los resultados obtenidos de la cantidad de características utilizadas para cada tipo de día

luego de aplicar la reducción de dimensionalidad mediante PCA y de la cantidad óptima de grupos se presentan en la Tabla 3-1.

Tabla 3-1 Resultados de estimación parámetros

| Tipo de día | Cantidad de características (PCA) | Cantidad óptima de grupos |
|-------------|-----------------------------------|---------------------------|
| ORD | 29 | 2 |
| SAB | 40 | 2 |
| DOM | 39 | 2 |
| FEST | 39 | 3 |

A continuación, se presentan los diferentes grupos generados para cada tipo de día y la serie representativa de cada grupo la cual es calculada como la sumatoria de todos los consumos de cada usuario en un mismo grupo. Dichas series representan el patrón de consumos de los usuarios para cada tipo de día. Estas series, indicadas en las Figura 3-1 (c, d), Figura 3-2 (c, d), Figura 3-3 (c, d) y Figura 3-4 (d, e, f), son utilizadas en la modelación para el pronóstico explicado en la Sección 3.3.

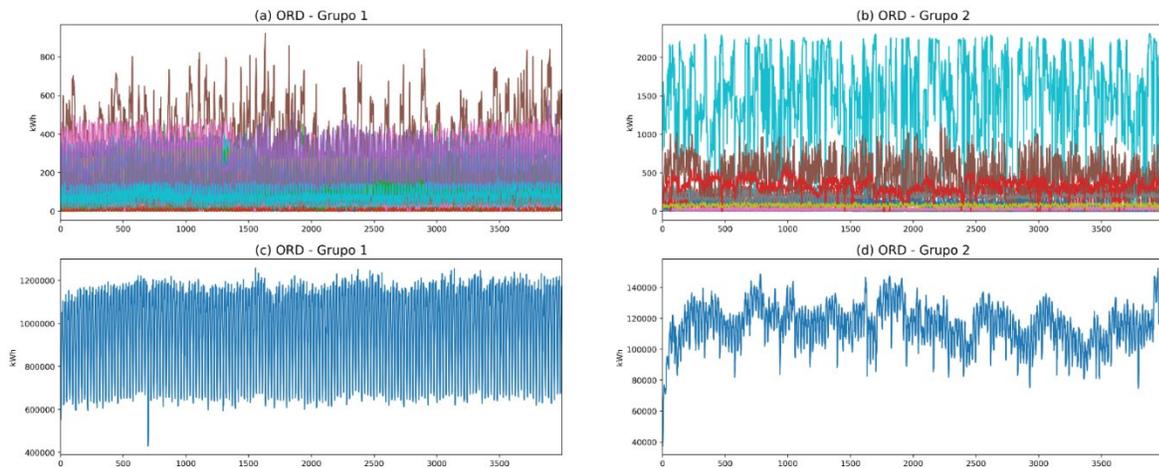


Figura 3-1 Análisis para el tipo de día ORD (ordinarios). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2.

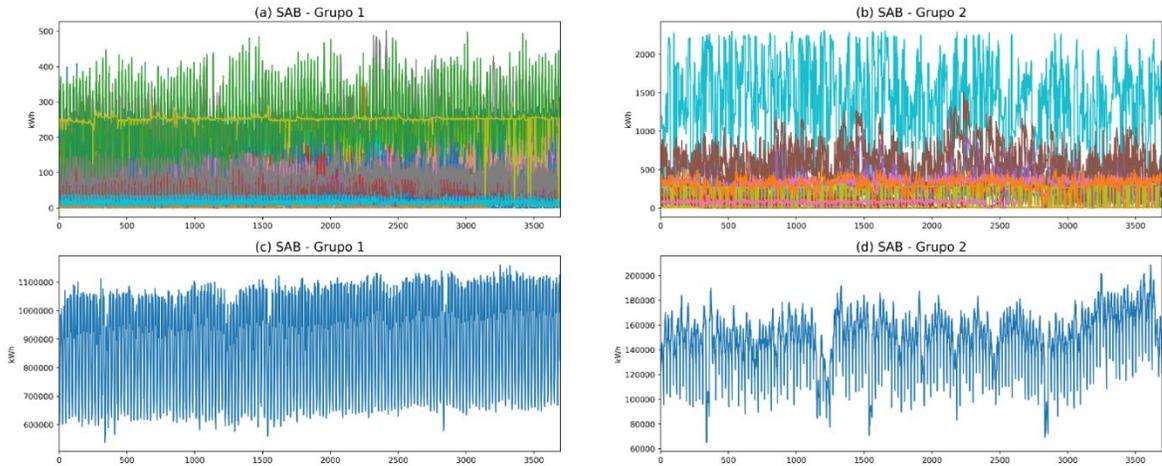


Figura 3-2 Análisis para el tipo de día SAB (sábados). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2.

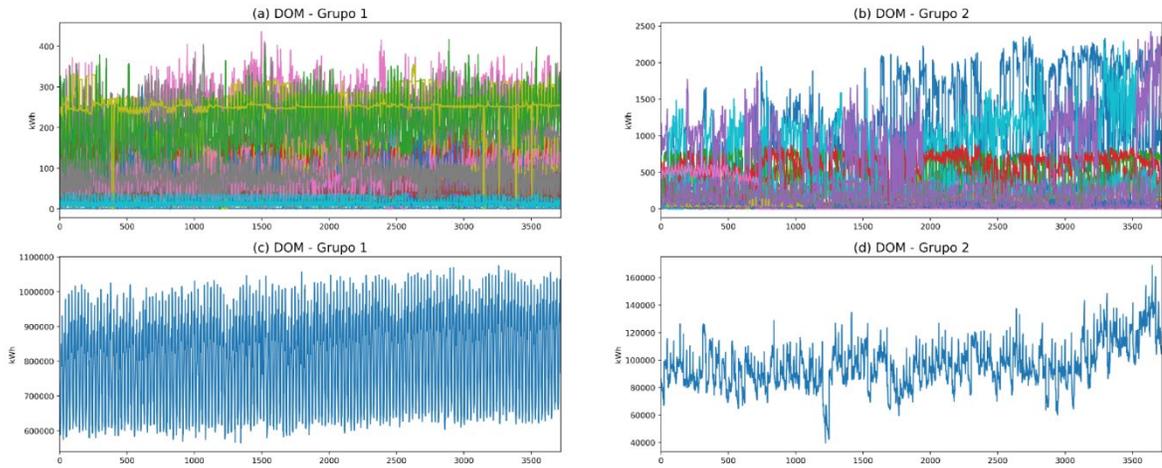


Figura 3-3 Análisis para el tipo de día DOM (domingos). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Agregación de consumos de usuarios del grupo 1; (d) Agregación de consumos de usuarios del grupo 2.

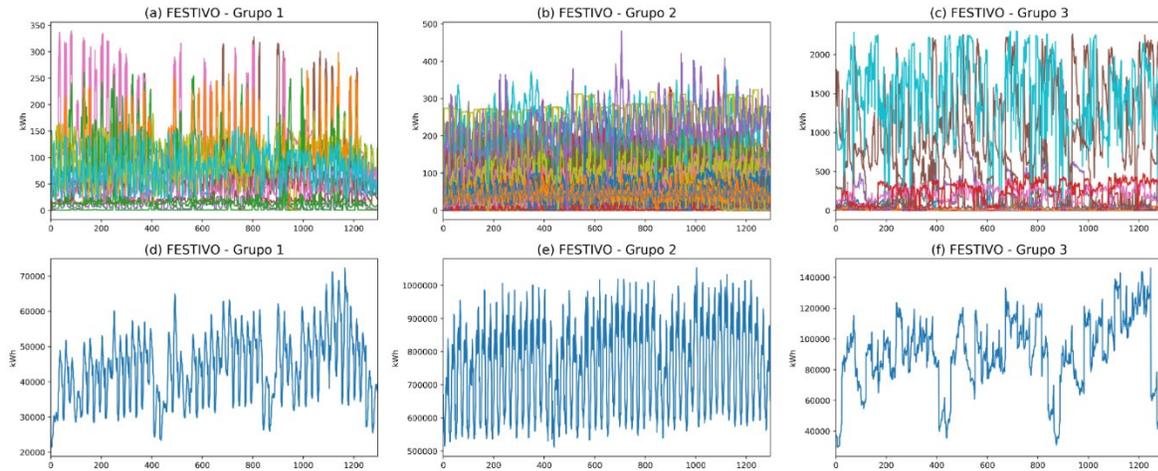


Figura 3-4 Análisis para el tipo de día FESTIVO (festivos). (a) Consumo de los usuarios para el grupo 1; (b) Consumo de los usuarios para el grupo 2; (c) Consumo de los usuarios para el grupo 3; (d) Agregación de consumos de los usuarios del grupo 1; (e) Agregación de consumos de los usuarios del grupo 2; (f) Agregación de consumos de los usuarios del grupo 3.

3.4 Pronóstico del modelo propuesto

El modelo de pronóstico propuesto en esta investigación es el modelo ARIMA definido en la Sección 2.1.5. Sin embargo, podrían haberse utilizado otros modelos de serie de tiempo, pero esto está fuera del alcance de este trabajo. Este modelo es entrenado con las series de tiempo resultantes de la técnica de agrupamiento explicada en la Sección 3.3. El proceso para realizar el entrenamiento y el pronóstico consta principalmente de tres etapas: descomposición de las series de tiempo, estimación y actualización de los parámetros del modelo con la nueva información de cada semana.

La descomposición de las series de tiempo se realiza con el fin de remover la parte estacional y entrenar un modelo ARIMA con las otras dos componentes. El tipo de descomposición aplicado en esta investigación se calcula mediante la Ecuación 2-3. La componente estacional se considera constante dado que la estacionalidad de la serie tiene cambios muy mínimos en el tiempo, al menos en la descomposición de primer y segundo orden que son los utilizados para entrenar los dos modelos ARIMA.

En esta investigación se utiliza la implementación del modelo ARIMA disponible en la librería `pmdarima` [43] para el lenguaje Python. Los parámetros del modelo son seleccionados mediante la minimización del criterio de información Akaike de acuerdo con la Ecuación 2-7. Adicionalmente, se consideraron variables exógenas al modelo, las cuales corresponden a las características de las series de Fourier, lo cual permite realizar un aplicar un ARIMA y no un SARIMA, es decir si aún persiste estacionalidad en la serie esta es modelada mediante las series de Fourier con las variables exógenas agregadas al modelo ARIMA [43]. La adición de las características de las series de Fourier al modelo implica un parámetro k adicional que se debe estimar en la modelación. La estimación del parámetro k es calculado con la minimización el error medio absoluto (MAE) para cada modelo ARIMA. Los valores de los parámetros de los dos modelos ARIMA aplicados a cada serie de tiempo resultante del proceso de agrupamiento de la Sección 3.3 para cada tipo de día se muestran en la Tabla 3-2 y Tabla 3-3.

Tabla 3-2 Parámetros (p, d, q) del modelo ARIMA de las series descompuestas en primer orden. Parámetro $k = 75$

| Tipo de día | Grupo 1 | Grupo 2 | Grupo 3 |
|-------------|---------|---------|---------|
| ORD | (0,1,4) | (1,1,3) | |
| SAB | (1,1,2) | (2,1,2) | |
| DOM | (3,1,3) | (2,1,4) | |
| FESTIVO | (2,1,2) | (1,1,4) | (2,1,3) |

Tabla 3-3 Parámetros (p, d, q) del modelo ARIMA de las series descompuestas en segundo orden. Parámetro $k = 65$

| Tipo de día | Grupo 1 | Grupo 2 | Grupo 3 |
|-------------|---------|---------|---------|
| ORD | (1,1,4) | (2,2,3) | |
| SAB | | | |
| DOM | (3,1,1) | (0,1,3) | |
| FESTIVO | (2,1,1) | (1,1,1) | (1,0,0) |

En la Tabla 3-3 se puede observar que para el caso del tipo de día SAB (sábados) no se cuenta con valores (p, d, q) del modelo ARIMA de las series descompuestas en segundo orden. Esto se debe a que en la evaluación de los resultados del modelo mediante las métricas de error mencionadas en la Sección 3.5 no mejora el desempeño del modelo para este tipo de día, en consecuencia, y basado en el principio de Parquedad (Parsimonia), únicamente se utiliza el modelo ARIMA de la serie descompuesta en primer orden para el tipo de día SAB (sábados).

Para la actualización de los parámetros del modelo, el pipeline de la librería pmdarima [43] es utilizado para la implementación de la metodología propuesta en esta investigación; esta librería permite realizar esta actualización partiendo de los parámetros estimados para el conjunto de entrenamiento, es decir los indicados en la Tabla 3-2 y Tabla 3-3. La selección del modelo óptimo se realiza mediante la minimización del criterio de información Akaike de acuerdo con la Ecuación 2-7. Los parámetros son actualizados con la información de cada semana s . El día de la semana s hasta donde se consideran los datos para el modelo es el viernes, y con esto se realiza el pronóstico de las siguientes dos semanas, es decir hasta la semana $s + 2$ considerando como día de finalización de la semana el domingo. Este proceso es realizado todos los miércoles de cada semana por el CND y se evalúan los resultados del pronóstico de la semana siguiente.

Una vez realizado el pronóstico para cada tipo de día con los dos modelos, es utilizada la metodología de la media mencionada en la Sección 2.1.6 para combinar los pronósticos de ambos modelos, obteniendo así, un único valor para cada periodo correspondiente a un tipo de día.

De acuerdo con lo anterior, en la Figura 3-5 se muestran los resultados de nueve (9) semanas del año 2020 de los pronósticos del modelo propuesto, los pronósticos realizados por el CND versus la demanda real. El desempeño del modelo propuesto es explicado en la Sección 3.5, sin embargo, se puede observar que el modelo propuesto se ajusta mucho mejor al comportamiento real de la serie de la demanda que el modelo utilizado en el CND.

Por último, se resalta que, a partir del 23 de marzo de 2020, los modelos de pronósticos del CND tuvieron modificaciones debido al inicio de la pandemia producto del COVID-19, donde se tuvo en cuenta más el conocimiento de un experto para el pronóstico; por lo cual, en esta investigación, no es posible comparar el desempeño de la metodología propuesta con respecto a la del CND durante la pandemia porque no tienen la misma base de comparación.

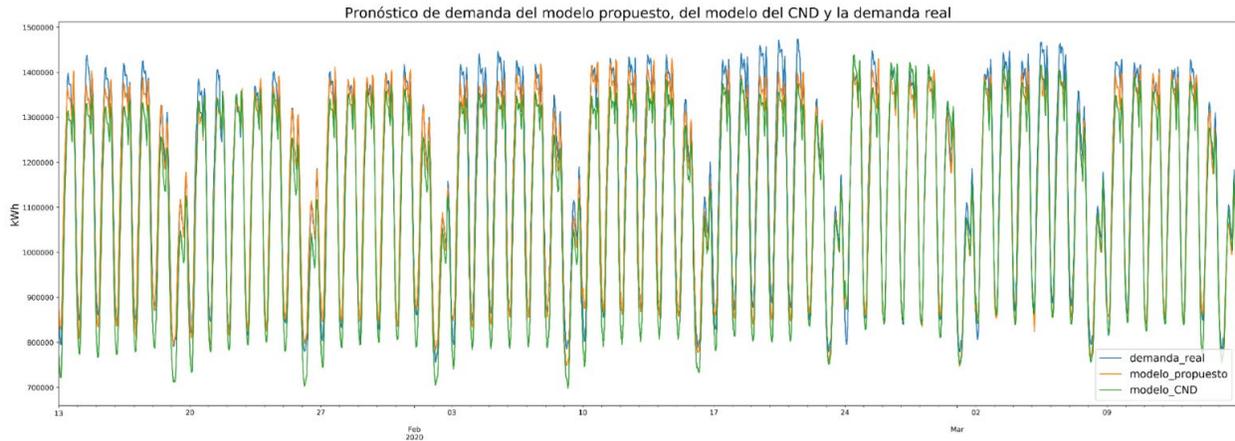


Figura 3-5 Pronóstico de la demanda del modelo propuesto, del modelo del CND versus la demanda real para nueve (9) semanas del año 2020

3.5 Evaluación de la precisión

Para la evaluación de la precisión del modelo propuesto en esta investigación, se utilizaron cuatro (4) métricas que calculan el error del pronóstico, las cuales corresponden a: error cuadrático medio (RMSE, por sus siglas en inglés), error medio absoluto (MAE, por sus siglas en inglés), error medio absoluto porcentual (MAPE, por sus siglas en inglés), r^2 y el error máximo absoluto. En la Tabla 3-4 se muestran los resultados de las métricas de error del modelo propuesto y del modelo del CND donde se puede inferir que en todas las métricas calculadas el modelo propuesto en esta investigación tiene mejor desempeño que el modelo del CND.

Tabla 3-4 Métricas de error calculadas para las nueve (9) semanas de pronóstico

| Modelo\Métrica de error | RMSE | MAE | MAPE | r^2 | Error máximo absoluto |
|-------------------------|-----------|-----------|-------|-------|-----------------------|
| Propuesto | 31,822.49 | 25,566.53 | 2.22% | 0.98 | 102,576.57 |
| CND | 61,170.61 | 53,921.21 | 4.77% | 0.91 | 135,136.27 |

Por otra parte, en la Figura 3-6 se muestra la dispersión de los errores porcentuales calculados para cada periodo en la ventana de las nueve (9) semanas de pronóstico, donde se observa que el 75% de los datos para el caso del modelo propuesto tienen un error por debajo del 4% y para el caso del modelo del CND para la misma proporción de población, tienen un error por debajo del 7%. Adicionalmente, el error máximo del modelo propuesto se encuentra aproximadamente 2 puntos porcentuales por debajo del error máximo del modelo del CND.

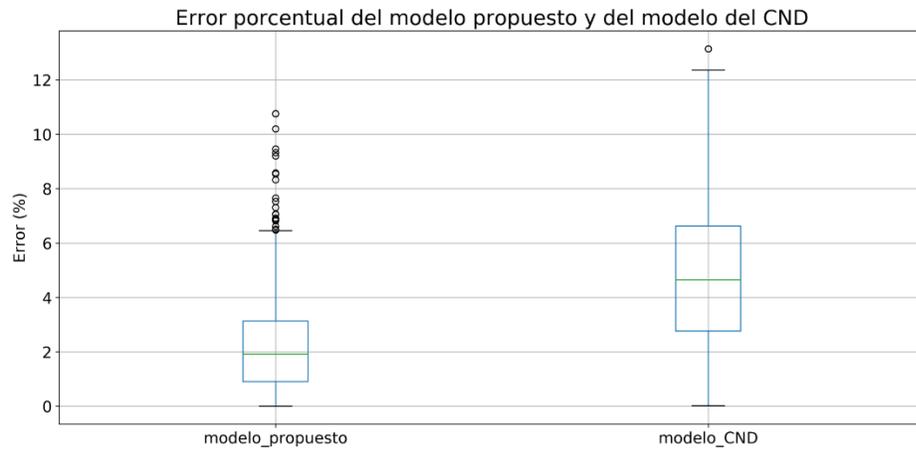


Figura 3-6 Dispersión del error porcentual absoluto para las nueve (9) semanas de pronóstico

4 Evaluación desde el punto de vista del negocio

Actualmente, existe un seguimiento de los indicadores para la operación del SIN definidos en el Acuerdo CNO 1406 de 2021, donde se evalúan el número de desviaciones que se han presentado en cada periodo por encima del 5% respecto a la demanda real. Las desviaciones en el pronóstico de la demanda real implican mayores costos para el sistema eléctrico debido a que impacta directamente en la programación de los recursos de generación que serán tenidos en cuenta en el día de operación, lo que puede ocasionar que algunos recursos de generación entren a reconciliar por encontrarse fuera de mérito. El valor de las reconciliaciones es recaudado a través de los generadores y otra parte del valor es trasladado a las restricciones del sistema, y estas últimas, son cobradas a través de los comercializadores a los usuarios finales. Por lo anterior, la importancia de tener un modelo de pronóstico de la demanda de energía más preciso.

La metodología propuesta en esta investigación demuestra un mejor desempeño del modelo de pronóstico que el utilizado por el CND de acuerdo con lo mencionado en la Sección 3.5, y, en consecuencia, aportaría con la reducción en los costos de operación del sistema eléctrico colombiano. En la Figura 4-1 se presenta la cantidad de veces que se tuvo un determinado valor de error porcentual; es decir, en qué porcentaje se ha desviado el pronóstico y cuántas veces a ocurrido esa desviación. En la Figura 4-1 se puede observar que el modelo propuesto tiene la mayor concentración de datos en valores de error porcentual por debajo del 4% a diferencia del modelo del CND que está por encima de este valor. Adicionalmente, en la Tabla 4-1 se muestra la cantidad de veces que el modelo propuesto y el modelo del CND tuvieron desviaciones por encima del 5%, donde se resalta que el modelo propuesto tuvo desviaciones mayores al 5% en el 6.08% de los periodos comprendidos en las nueve (9) semanas de pronóstico, lo cual es significativamente menor al 45.3% de los periodos que superaron este porcentaje para el caso del modelo del CND.

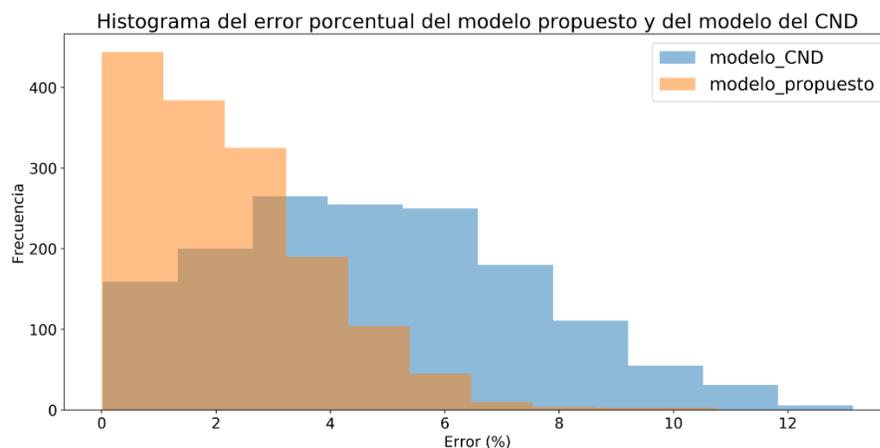


Figura 4-1 Distribución de los errores porcentuales del modelo propuesto y del CND

Tabla 4-1 Desviaciones en el pronóstico con valores mayores al 5%

| Modelo | Cantidad de desviaciones mayores al 5% | Porcentaje de representación de la muestra |
|-----------|--|--|
| Propuesto | 92 | 6.08% |
| CND | 685 | 45.30% |

5 Conclusiones

A continuación, se presentan las conclusiones de esta investigación asociadas con los objetivos planteados en la Sección 1.8 y con la pregunta de investigación mencionada en la Sección 1.2.

5.1 Respuesta a la pregunta de investigación, ¿Los modelos de pronóstico basados en técnicas de agrupamiento son más precisos que los modelos de pronóstico utilizados en el CND?

Para responder a la pregunta de investigación planteada en la Sección 1.2, se consideran los resultados obtenidos en la Sección 3.5, donde se evidencia que la metodología de pronóstico propuesta en esta investigación, la cual está basada en técnicas de agrupamiento, tiene una mayor precisión que la utilizada en el CND en las nueve (9) semanas evaluadas. Lo anterior, se demuestra en las métricas de error RMSE, MAE y MAPE que tienen aproximadamente un 50% de reducción de su valor respecto a los valores de dichas métricas en la evaluación del modelo del CND.

La precisión de la metodología propuesta se logra gracias a que se capturan los cambios temporales de las series de tiempo que tienen los diferentes patrones de consumo de los usuarios con el fin de tener un modelo más ajustado y generalizado para cada patrón; permitiendo que el modelo de pronóstico tenga un mejor desempeño que el empleado en el CND, el cual utiliza únicamente la serie de tiempo que agrega los consumos de todos los usuarios.

De esta manera, se concluye que las metodologías de pronóstico basadas en técnicas de agrupamiento es una buena alternativa para mejorar la precisión de los modelos de pronóstico utilizados por el CND, de acuerdo con los resultados obtenidos en esta investigación; tanto del modelo propuesto como de los planteados en los artículos de la revisión de literatura de la Sección 1.4. Adicionalmente, se resalta que, a partir del 23 de marzo de 2020, los modelos de pronósticos del CND tuvieron modificaciones debido al inicio de la pandemia producto del COVID-19, donde se tuvo en cuenta más el conocimiento de un experto para el pronóstico; por lo cual, en esta investigación, no es posible comparar el desempeño de la metodología propuesta con respecto a la del CND durante la pandemia porque no tienen la misma base de comparación.

5.2 Cumplimiento del objetivo general

Actualmente, en Colombia, el CND para el pronóstico de la demanda de energía eléctrica a corto plazo utiliza las series de consumo que representan cada Mercado de Comercialización (MC) que tiene el Sistema Interconectado Nacional. Cada Mercado Comercialización está compuesto por las series de consumo de los usuarios pertenecientes al mismo, por lo que la serie representativa de cada MC es afectada por todos los patrones de consumo que tienen los usuarios y en consecuencia, ocasione una posible pérdida de información puesto que el comportamiento de la serie agregada ignora los hechos estilizados asociados a los consumos de los usuarios [7].

Por lo anterior, en esta investigación se implementó una metodología de pronóstico basada en técnicas de agrupamiento que toma como insumo las series de los consumos de cada usuario pertenecientes a un MC con el fin de lograr capturar los cambios en los patrones de consumo de un usuario dados por los cambios en el comportamiento de consumo. La metodología implementada, utiliza el modelo autorregresivo ARIMA,

también usado por el CND, para el pronóstico de la demanda de energía eléctrica en el corto plazo (en este caso, corresponde a dos semanas). El modelo de pronóstico se realiza para cada una de las series representativas asociadas a los diferentes patrones de consumo de los usuarios determinados por la técnica de agrupamiento descrita en la Sección 3.3, que luego, son combinados mediante el método de la media, explicado en la Sección 2.1.6. Este modelo utiliza la información hasta el viernes de la semana s para pronosticar los valores hasta la semana $s + 2$, considerando una semana de lunes a domingo, tal cual como se realiza el CND conforme a lo establecido en la Resolución CREG 025 de 1995 o aquellas que la modifiquen o sustituyan. Con los nuevos datos de cada semana, los parámetros de los modelos son actualizados permitiendo que los modelos consideren la nueva información de demanda de cada semana.

En ese sentido, y en cumplimiento al objetivo general de esta investigación planteado en la Sección 1.8.1, se implementa una metodología híbrida de pronóstico, la cual considera una técnica de agrupamiento para agrupar patrones similares de consumo y determinar series representativas para cada grupo; y, un modelo autorregresivo ARIMA, también usado por el CND, para pronosticar la demanda para cada una de esas series representativas. Adicionalmente, es adaptativo, dado que los parámetros de los modelos son actualizados en la medida que se cuenta con nueva información cada semana.

5.3 Cumplimiento de los objetivos específicos

5.3.1 Modelar la dinámica de la demanda de energía a corto plazo usando un modelo híbrido adaptativo que represente adecuadamente el comportamiento de los consumos de los usuarios, manteniendo los modelos de pronóstico del CND

En esta investigación, se implementa una metodología de pronóstico basada en técnicas de agrupamiento la cual permite determinar diferentes patrones de consumo de los usuarios que se ajustan a los comportamientos de las series de tiempo de consumo capturando los cambios que ocurren en el comportamiento de un intervalo de tiempo a otro. La técnica de agrupamiento para determinar dichos patrones es k-means y realiza el agrupamiento con las series de consumo representadas por los primeros sesenta (60) coeficientes espectrales de Fourier, como se explica en la Sección 2.1.1. Luego, se aplica el modelo autorregresivo ARIMA, también usado por el CND, para pronosticar las diferentes series representativas de cada grupo y luego, se combinan los pronósticos mediante la media, explicada en la Sección 2.1.6.

De acuerdo con lo anterior, y en cumplimiento con este objetivo específico, se logra modelar la dinámica de la demanda de energía mediante la utilización de técnicas de agrupamiento que permiten representar los diferentes patrones de consumo de los usuarios y, además, pronosticar las diferentes series de tiempo de consumo con el modelo autorregresivo ARIMA que también es utilizado por el CND.

5.3.2 Evaluar el desempeño del modelo híbrido adaptativo del objetivo anterior respecto a la metodología actual que se tiene para el pronóstico de la demanda de energía a corto plazo

En la Sección 3.5, se evalúa el desempeño de la metodología propuesta en esta investigación, donde se midieron las métricas de error RMSE, MAE y MAPE las cuales también fueron calculadas para el modelo que actualmente tiene el CND obteniendo mejoras de hasta de un 50% en la precisión con el modelo propuesto. Adicionalmente, se evalúa la cantidad de veces que los valores pronosticados tienen una desviación mayor al

5% respecto a la demanda real; este porcentaje de desviación está establecido por el Acuerdo del CNO 1406 de 2021 para realizar seguimiento a los modelos de pronóstico de demanda. Con el modelo propuesto, se obtuvo que sólo el 6.08% de las nueve (9) semanas pronosticadas a nivel horario tuvieron desviaciones mayores al 5% en comparación con el modelo del CND que obtuvo un 45.3% para el mismo periodo de pronóstico.

De acuerdo con lo anterior, y en cumplimiento con este objetivo específico, se midieron diferentes métricas que permitieron evaluar el desempeño de la metodología propuesta en esta investigación respecto con la que es utilizada por el CND donde se evidencian mejoras significativas en la precisión del pronóstico de la demanda de energía a corto plazo.

Bibliografía

- [1] P. Ledesma, «Regulación de frecuencia y potencia», p. 33.
- [2] T.-Y. Kim y S.-B. Cho, «Predicting residential energy consumption using CNN-LSTM neural networks», *Energy*, vol. 182, pp. 72-81, 2019, doi: 10.1016/j.energy.2019.05.230.
- [3] J. Davidvelásquez, C. J. Franco, y H. A. García, «A non-linear model for forecasting the monthly demand for electricity in Colombia», *Estud. Gerenciales*, vol. 25, n.o 112, pp. 37-54, 2009, doi: 10.1016/S0123-5923(09)70079-8.
- [4] A. S. Ahmad et al., «A review on applications of ANN and SVM for building electrical energy consumption forecasting», *Renew. Sustain. Energy Rev.*, vol. 33, pp. 102-109, 2014, doi: 10.1016/j.rser.2014.01.069.
- [5] R. Y. M. Li, S. Fong, y K. W. S. Chong, «Forecasting the REITs and stock indices: Group method of data handling neural network approach», *Pac. Rim Prop. Res. J.*, vol. 23, n.o 2, pp. 123-160, 2017, doi: 10.1080/14445921.2016.1225149.
- [6] «Acuerdo 1303 Por el cual se actualizan los procedimientos para la gestión integral de la demanda | C.N.O». <https://www.cno.org.co/content/acuerdo-1303-por-el-cual-se-actualizan-los-procedimientos-para-la-gestion-integral-de-la> (accedido jun. 29, 2020).
- [7] E. G. Tajeuna, M. Bouguessa, y S. Wang, «A network-based approach to enhance electricity load forecasting», 2019, vol. 2018-November, pp. 266-275. doi: 10.1109/ICDMW.2018.00046.
- [8] L. N. Ferreira y L. Zhao, «A time series clustering technique based on community detection in networks», 2015, vol. 53, n.o 1, pp. 183-190. doi: 10.1016/j.procs.2015.07.293.
- [9] M. E. J. Newman, «Fast algorithm for detecting community structure in networks», *Phys. Rev. E - Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, n.o 6, p. 5, 2004, doi: 10.1103/PhysRevE.69.066133.
- [10] G. Le Ray y P. Pinson, «Online adaptive clustering algorithm for load profiling», *Sustain. Energy Grids Netw.*, vol. 17, 2019, doi: 10.1016/j.segan.2018.100181.
- [11] L. F. S. Vilela, R. C. Leme, C. A. M. Pinheiro, y O. A. S. Carpinteiro, «Forecasting financial series using clustering methods and support vector regression», *Artif. Intell. Rev.*, vol. 52, n.o 2, pp. 743-773, 2019, doi: 10.1007/s10462-018-9663-x.
- [12] K. Gajowniczek y T. Zabkowski, «Simulation Study on Clustering Approaches for Short-Term Electricity Forecasting», *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/3683969.
- [13] P. Laurinec, M. Lóderer, M. Lucká, y V. Rozinajová, «Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption», *J. Intell. Inf. Syst.*, vol. 53, n.o 2, pp. 219-239, 2019, doi: 10.1007/s10844-019-00550-3.
- [14] S. Sun, J. Fu, y A. Li, «A compound wind power forecasting strategy based on clustering, two-stage decomposition, parameter optimization, and optimal combination of multiple machine learning approaches», *Energies*, vol. 12, n.o 18, 2019, doi: 10.3390/en12183586.
- [15] Y. Wang, Y. Shen, S. Mao, X. Chen, y H. Zou, «LASSO and LSTM integrated temporal model for short-term solar intensity forecasting», *IEEE Internet Things J.*, vol. 6, n.o 2, pp. 2933-2944, 2019, doi: 10.1109/JIOT.2018.2877510.
- [16] P. Laurinec y M. Lucká, «Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting», *Data Min. Knowl. Discov.*, vol. 33, n.o 2, pp. 413-445, 2019, doi: 10.1007/s10618-018-0598-2.
- [17] R. Jain, N. Jain, Y. Gupta, T. Chugh, T. Chugh, y D. J. Hemanth, «A Modified Fuzzy Logic Relation-Based Approach for Electricity Consumption Forecasting in India», *Int. J. Fuzzy Syst.*, vol. 22, n.o 2, pp. 461-475, 2020, doi: 10.1007/s40815-019-00704-z.

- [18] S. Singh y A. Yassine, «Big data mining of energy time series for behavioral analytics and energy consumption forecasting», *Energies*, vol. 11, n.o 2, 2018, doi: 10.3390/en11020452.
- [19] E. Y. Shchetinin, «Cluster-based energy consumption forecasting in smart grids», *Commun. Comput. Inf. Sci.*, vol. 919, pp. 445-456, 2018, doi: 10.1007/978-3-319-99447-5_38.
- [20] B. Auder, J. Cugliari, Y. Goude, y J.-M. Poggi, «Scalable clustering of individual electrical curves for profiling and bottom-up forecasting», *Energies*, vol. 11, n.o 7, 2018, doi: 10.3390/en11071893.
- [21] L. Li, K. Ota, y M. Dong, «Everything is image: CNN-based short-term electrical load forecasting for smart grid», 2017, vol. 2017-November, pp. 344-351. doi: 10.1109/ISPAN-FCST-ISCC.2017.78.
- [22] M. Ghofrani, R. Azimi, F. M. Najafabadi, y N. Myers, «A new day-ahead hourly electricity price forecasting framework», presentado en 2017 North American Power Symposium, NAPS 2017, 2017. doi: 10.1109/NAPS.2017.8107269.
- [23] J. Yang et al., «k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement», *Energy Build.*, vol. 146, pp. 27-37, 2017, doi: 10.1016/j.enbuild.2017.03.071.
- [24] I. P. Panapakidis, «Clustering based day-ahead and hour-ahead bus load forecasting models», *Int. J. Electr. Power Energy Syst.*, vol. 80, pp. 171-178, 2016, doi: 10.1016/j.ijepes.2016.01.035.
- [25] X. Liang, T. Hong, y G. Q. Shen, «Occupancy data analytics and prediction: A case study», *Build. Environ.*, vol. 102, pp. 179-192, 2016, doi: 10.1016/j.buildenv.2016.03.027.
- [26] Y. H. Chen, W.-C. Hong, W. Shen, y N. N. Huang, «Electric load forecasting based on a least squares support vector machine with fuzzy time series and global harmony search algorithm», *Energies*, vol. 9, n.o 2, pp. 1-13, 2016, doi: 10.3390/en9020070.
- [27] M. Chaouch, «Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves», *IEEE Trans. Smart Grid*, vol. 5, n.o 1, pp. 411-419, 2014, doi: 10.1109/TSG.2013.2277171.
- [28] H. Mori y M. Takahashi, «Application of preconditioned Generalized radial Basis Function Network to prediction of photovoltaic power generation», presentado en IEEE PES Innovative Smart Grid Technologies Conference Europe, 2012. doi: 10.1109/ISGTEurope.2012.6465877.
- [29] B. Kitchenham, «Procedures for Performing Systematic Reviews», p. 33.
- [30] M. R. Cogollo y J. D. Velásquez, «Methodological advances in artificial neural networks for time series forecasting», *IEEE Lat. Am. Trans.*, vol. 12, n.o 4, pp. 764-771, 2014, doi: 10.1109/TLA.2014.6868881.
- [31] W. Zhang, G. Mu, C. Song, G. Yan, y V. Heuveline, «Extraction of Spatial Features of Bus Loads in Electric Grids Through Clustering in a Dynamic Model Space», *IEEE Access*, vol. 8, pp. 5852-5861, 2020, doi: 10.1109/ACCESS.2019.2963071.
- [32] W. Sun y C. Zhang, «A hybrid BA-ELM model based on factor analysis and similar-day approach for short-term load forecasting», *Energies*, vol. 11, n.o 5, 2018, doi: 10.3390/en11051282.
- [33] M. Ghofrani, D. Carson, y M. Ghayekhloo, «Hybrid clustering-time series-Bayesian neural network short-term load forecasting method», presentado en NAPS 2016 - 48th North American Power Symposium, Proceedings, 2016. doi: 10.1109/NAPS.2016.7747865.
- [34] P. Laurinec, M. Loderer, P. Vrabecova, M. Lucka, V. Rozinajova, y A. B. Ezzeddine, «Adaptive Time Series Forecasting of Energy Consumption Using Optimized Cluster Analysis», 2016, vol. 0, pp. 398-405. doi: 10.1109/ICDMW.2016.0063.
- [35] A. Laouafi, M. Mordjaoui, F. Laouafi, y T. E. Boukelia, «Daily peak electricity demand forecasting based on an adaptive hybrid two-stage methodology», *Int. J. Electr. Power Energy Syst.*, vol. 77, pp. 136-144, 2016, doi: 10.1016/j.ijepes.2015.11.046.
- [36] R. L. Talavera-Llames, R. Pérez-Chacón, M. Martínez-Ballesteros, A. Troncoso, y F. Martínez-Álvarez, «A nearest neighbours-based algorithm for big time series data forecasting», *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 9648, pp. 174-185, 2016, doi: 10.1007/978-3-319-32034-2_15.
- [37] W. Wang, W. Pedrycz, y X. Liu, «Time series long-term forecasting model based on information granules and fuzzy clustering», *Eng. Appl. Artif. Intell.*, vol. 41, pp. 17-24, 2015, doi: 10.1016/j.engappai.2015.01.006.
- [38] C. J. Bennett, R. A. Stewart, y J. W. Lu, «Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system», *Energy*, vol. 67, pp. 200-212, 2014, doi: 10.1016/j.energy.2014.01.032.

- [39] S. Humeau, T. K. Wijaya, M. Vasirani, y K. Aberer, «Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households», presentado en 2013 Sustainable Internet and ICT for Sustainability, SustainIT 2013, 2013. doi: 10.1109/SustainIT.2013.6685208.
- [40] Z. Ding, P. Yang, X. Yang, y Z. Zhang, «Wind power prediction method based on sequential time clustering support vector machine», Dianli Xitong ZidonghuaAutomation Electr. Power Syst., vol. 36, n.o 14, pp. 131-135+149, 2012, doi: 10.3969/j.issn.1000-1026.2012.14.025.
- [41] «Complex Fourier Series». 2020. [En línea]. Disponible en: [https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Book%3A_Electrical_Engineering_\(Johnson\)/04%3A_Frequency_Domain/4.02%3A_Complex_Fourier_Series](https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Book%3A_Electrical_Engineering_(Johnson)/04%3A_Frequency_Domain/4.02%3A_Complex_Fourier_Series)
- [42] M. Charrad, N. Ghazzali, V. Boiteau, y A. Niknafs, «NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set», J. Stat. Softw., vol. 61, n.o 1, Art. n.o 1, nov. 2014, doi: 10.18637/jss.v061.i06.
- [43] «pmdarima», API Reference. <https://alkaline-ml.com/pmdarima/modules/classes.html>