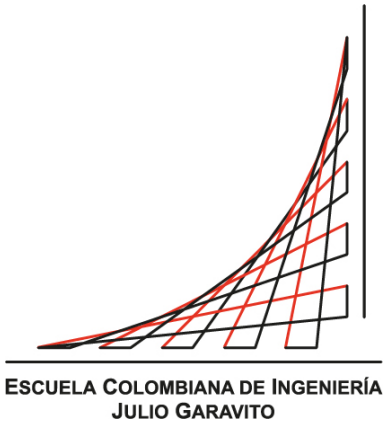


ESCUELA COLOMBINA DE INGENIERÍA JULIO GARAVITO

Programación para el análisis de datos

Análisis del mercado de energía eléctrica Colombiano basado en clusterización y algoritmos de predicción de demanda

Entrega final

Presentado por :
Lilian Daniela Suárez Riveros
Laura Melisa Patarroyo Godoy
Santiago Jejen Salinas

Índice

1. Entendimiento del negocio

En Colombia, la mayor proporción de la demanda de energía eléctrica está conectada al Sistema Interconectado Nacional (SIN), que es un conjunto de centrales de generación eléctrica y sistemas de distribución conectadas a través del sistema de transmisión nacional y que permiten atender la demanda de energía eléctrica para Colombia. El SIN es uninodal, es decir, todos los generadores y toda la demanda están conectados en un mismo punto, lo que implica que existe un precio de bolsa con el cual se vende y se compra energía.

Aunque existen otras modalidades para la compra y venta de energía, para el desarrollo del trabajo se abarcará solo este mercado, conocido también como mercado Spot. El precio del mercado spot se fija con relación a la demanda y a la oferta, en otras palabras, los generadores publican las cantidades disponibles y los precios a los cuales ofertan estas, y con la curva de la demanda se determina cuántos generadores se necesitan para satisfacerla.

Igualmente, el precio de bolsa es el precio del último generador necesario para satisfacer la demanda, al cual se le conoce como precio marginal. Así, los generadores que oferten a menor precio son los que participan para atender la demanda. Este es variable para todas las horas de cada día de la semana.

De esta forma, el mercado spot de Colombia funciona bajo el esquema de competencia perfecta y este es uno de los temas que se desea analizar en el desarrollo del trabajo.

Por otra parte, en Colombia existe un organismo encargado de coordinar el mercado energético, Xm Colombia, quien realiza la gestión del sistema de energía eléctrica de Colombia en tiempo real y por lo cual disponen de bases de datos disponibles al público. En este trabajo se emplean los datos de demanda, generación y precio encontrados allí.

1.1. Objetivo

Establecer la relación entre demanda, precio y generación del mercado de energía a nivel nacional de la compañía Interconexión Eléctrica S.A. E.S.P. - ISA – con base en los datos históricos del año 2019 y 2020, para así predecir la demanda a nivel horario-diario de la energía eléctrica en Colombia.

1.2. Alcance

Desarrollo del perfilamiento para los valores de la demanda a nivel horario y según el día de la semana, para detectar y establecer patrones y rangos o franjas en los que se puedan agrupar los datos para un mejor análisis y comprensión del problema.

La predicción de la demanda de energía eléctrica de Colombia se realizará para una hora sobresaliente de cada uno de los clústers resultantes del análisis descriptivo que se realice, bajo la aplicación de 4 modelos de minería de datos

: Regresión lineal, Random Forest, Máquina de vector de soporte con kernel polinómico de grado 2 y Máquina de vector de soporte con kernel polinómico de grado 3 y estableciendo como el mejor, aquel modelo que genere el menor error absoluto medio porcentual.

2. Clustering

El agrupamiento de series de tiempo busca resaltar la estructura inherente en el conjunto de datos por medio de un agrupamiento homogéneo y coherente para encontrar patrones o perfiles desconocidos que puedan ayudar al entendimiento de las mismas. Esta tarea tiene dificultades adicionales debido a la estructura única de las series de tiempo, haciendo que los métodos tradicionales de agrupamiento no se puedan aplicar directamente. Los componentes a considerar son la selección adecuada de una medida de similitud o distancia y el algoritmo de agrupamiento, ya que sin importar el algoritmo, se requiere una medida de similitud o distancia para la comparación de las series de tiempo [1].

De acuerdo con lo anterior, se aplico un agrupamiento jerárquico con la medida de similitud del coseno y el método de Ward, ya que el caso de estudio examina patrones de consumos y captura los comportamientos similares caracterizados por picos a la misma hora del día como se puede observar en trabajos previos [2, 3, 4].

Como fase inicial, se realizó un análisis exploratorio previo para identificar la relación entre *demanda*,*precio* y *generación* de la energía eléctrica. Los resultados de la matriz de correlación y el gráfico de dispersión, confirmaron el alto grado de relación lineal entre la *demanda* y la *generación* (siendo siempre un poco mayor en proporción la *generación*), y nos empezó a mostrar señales de la ausencia de relación lineal entre la *demanda* y el *precio*. Por lo anterior, y como última medida de verificación, se optó por realizar clusterización para las 3 variables y así comprobar lo anteriormente planteado.

Luego, se aplico una clusterización por días de la semana y por horas (Hora0 a Hora23), con el conjunto de datos de los años 2019 y 2020, para las 3 variables.

Para la clusterización de la demanda horaria se utilizo k=7, pero debido a la proximidad de uno de ellos, se agrupó la hora 4 con la hora 0,1,2 y 3, como se observa en la tabla 1b. De manera análoga, se realizo la clusterización de la generación horaria.

Al analizar los resultados, se pudo concluir que la *demanda* y el *precio* de la energía eléctrica en Colombia no poseen una relación, como se puede observar en la tabla 1, pues los clusters de precio y demanda difieren para ciertas horas, dando a entender que el mercado de energía eléctrica en Colombia no es un mercado de competencia perfecta, sino por el contrario, puede ser un oligopolio. Si el mercado se comportara en competencia perfecta, todas las horas clusterizadas para el precio, coincidirían con todas las horas clusterizadas para la demanda,

pero se observan horas donde el consumo es bajo pero la energía es costosa, y en sentido contrario, horas en que la demanda es alta pero el precio es bajo.

Cluster	Horas	Centroides
1	0 a 4	210,7034
2	5 y 6	220,3497
3	7,8 y 23	239,4320
6	20 a 22	270,1575
4	9 a 13 y 15 a 17	270,6961
5	14,18 y 19	286,2048

(a) Clusters de precio

Cluster	Horas	Centroides
1	0 a 4	6.170.005
2	5 y 6	6.429.960
6	22 y 23	7.245.343
3	7 a 13	7.847.692
4	14 a 18	8.196.142
5	19,20 y 21	8.524.556

(b) Clusters de demanda

Tabla 1: Clusters

Por otra parte, para la clusterización de demanda horaria por variable, las horas en cada clúster fueron muy similares para ambos año; sin embargo, a partir de la clusterización diaria para ambos años no se pudo llegar a ninguna conclusión relevante, pues los diferentes días aparecían en todos los clusters y no se observaba ningún patrón contundente de agrupación. Por lo anterior, se realizó nuevamente la clusterización diaria, pero esta vez particionando el conjunto de datos por año.

Al comparar los resultados del clustering por día de la semana, particionada por año, para el año 2020 se observó que no se podía establecer ningún patrón claro, pues el comportamiento de la demanda fue muy similar sin importar el día. De lo anterior, se concluyó que el comportamiento especial del año 2020 puede tener su causa debido a la situación de pandemia y emergencia sanitaria que atraviesa el país actualmente y por lo tanto, el conjunto de datos del 2020, al producir ruido, no sería útil al momento de realizar las predicciones.

Teniendo en cuenta lo anterior, se optó por realizar los análisis y pasos siguientes solamente para la variable de la *demanda* y con los datos históricos del año 2019.

Finalmente, se realizó una clusterización diaria con k=3 para el año 2019 y su distribución fue la siguiente:

Cluster	**Rango**
1	sábado,domingo, días festivos
2	lunes, martes
3	miércoles, jueves, viernes

Tabla 2: Clusters de demanda

3. Predicciones

Para realizar las predicciones, se tomó solo el conjunto de datos correspondiente al año 2019 y se seleccionó 1 hora característica de cada uno de los clusters de demanda horaria obtenida para predecirla. Por lo tanto, las horas a predecir fueron: hora 5, hora 10, hora 18, hora 21 y hora 23.

Como conjunto de entrenamiento/test se tomaron en cuenta las variables día, mes, día de la semana (con valores del 1 al 7), clúster diario obtenido previamente (tabla 2), y la demanda de la hora 0 a la hora 3.

Se usaron 4 algoritmos de predicción: Regresión Lineal, Random Forest, Máquina de vector de soporte con kernel polinómico de grado 2 y grado 3, y se ejecutaron para cada una de las 5 horas a predecir. En la figura 1, se encuentra la demanda real vs la predicción de cada hora con cada uno de los algoritmos utilizados y se obtuvo:

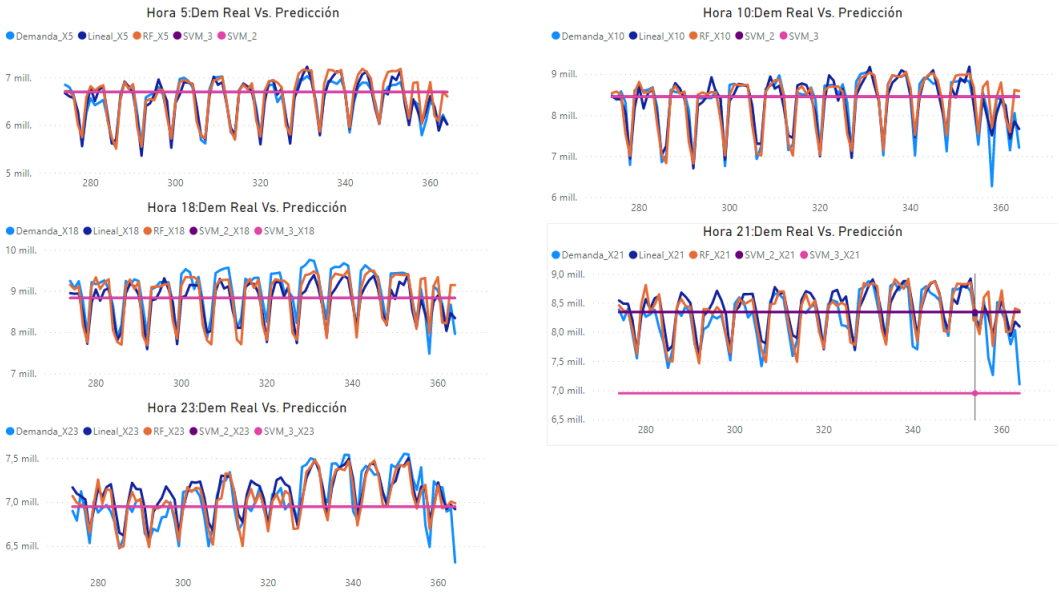


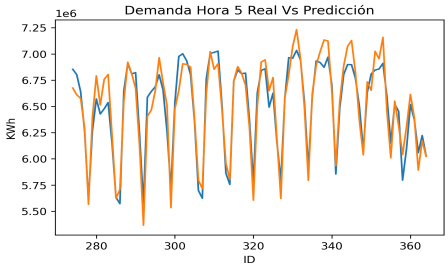
Figura 1: Resultado Algoritmos aplicados

Para realizar la comparación y selección del mejor, se calculó el error porcentual absoluto medio (MAPE), obteniendo como resultado los siguientes valores:

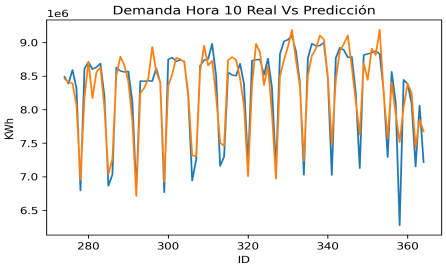
Algoritmo	Hora 5	Hora10	Hora18	Hora21	Hora23
Lineal	1,721036	2,470657	3,094869	2,648409	2,142871
Random Forest	7,897552	8,722274	7,103597	5,688358	4,418050
SVM -2	5,591749	6,372205	5,586693	4,217401	3,444969
SVM -3	5,591791	6,372184	5,586649	4,217417	3,444926

Tabla 3: MAPE

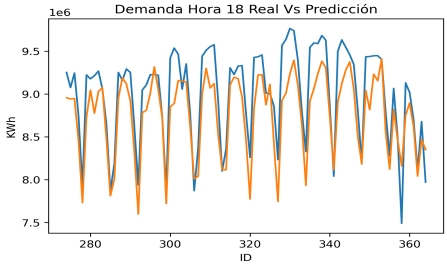
Como se puede observar en la tabla 3 el menor MAPE fue para el algoritmo de Regresión Lineal, por lo que este sería el seleccionado y bajo el cual se presentan las siguientes comparaciones de demanda real vs predicción para las 5 horas respectivas:



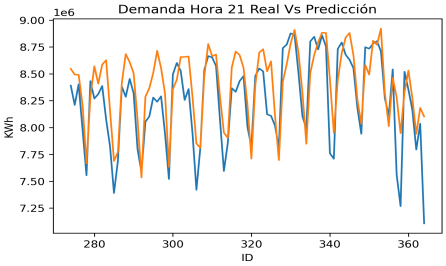
(a) Demanda Hora 5



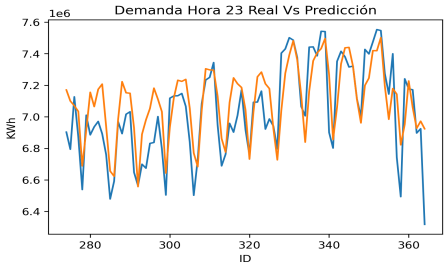
(b) Demanda Hora 10



(c) Demanda Hora 18



(d) Demanda Hora 21



(e) Demanda Hora 23

Figura 2: Demanda Real Vs Predeción

Intentando mejorar los modelos de predicción, se aplicaron los algoritmos mencionados anteriormente para cada hora pero esta vez particionando la data por el clúster diario obtenido en los análisis previos. En términos generales el MAPE que se obtuvo fue muy cercano al MAPE resultante para la data completa observado anteriormente en la tabla 3, y para algunas horas con la data particionada, el MAPE era más alto, por lo tanto, se decidió usar los algoritmos sin discriminarlos por clúster diario.

4. Conclusiones

- Al relacionar los clusters de la demanda horaria de energía eléctrica con los precios de bolsa por hora de la energía eléctrica, se observa que el mercado de energía eléctrica no es completamente dinámico, ya que para algunas horas con demanda baja los precios son altos, y para algunas horas con demanda alta, los precios son bajos.
- Para los días festivos de octubre y diciembre todos los algoritmos de predicción para todas las horas de la demanda, presentaron un error significativo (no pudieron reconocer el pico del conjunto de datos real), por lo tanto se sugiere para trabajos futuros realizar un modelo único para estos días.
- Para trabajos futuros, se recomienda realizar algoritmos de predicción para el precio de bolsa, y para esto se recomienda tener variables adicionales que ayuden a entender y establecer mejor su comportamiento, tales como la hidrología y variables climáticas de Colombia.
- Si el MAPE resultante para la data completa del año 2019 fuese muy alto, existiría la necesidad de discriminar y particionar la misma por clúster diario obtenido, y realizar las predicciones de igual manera.
- Para realizar buenos modelos no siempre es necesario recurrir a algoritmos 'caja negra' o complicarse con modelos muy complejos, un ejemplo de lo anterior es que nuestro modelo ganador fue el más clásico:la regresión lineal, por encima de los algoritmos más evolucionados.
- Para trabajos futuros, se necesitaría esperar que los comportamientos de la demanda se reajustara por la situación especial (pandemia) que se se vive en el 2020, o en dado caso, esperar un tiempo determinado para que un modelo pudiese aprender a detectar estas situaciones anómalas.

Referencias

[1] Rodr, J. E. (2011) *Agrupamiento De Datos De Series De Tiempo. Estado Del Arte.* Revista Vínculos, 8(1), 210–231. https://doi.org/10.14483/2322939X.4191.

[2] Candelieri, A., & Archetti, F. (2014) *Identifying typical urban water demand patterns for a reliable short-term forecasting - The icewater project approach. Procedia Engineering.* 89, 1004–1012. https://doi.org/10.1016/j.proeng.2014.11.218

[3] Conti, D., Gibert, K., & Rosa, D. D. La. (2018) *Characterization of electric energy consumption with clustering techniques: a case study in Northern Mexico Characterization of electric energy consumption with clustering techniques: a case study in Northern Mexico.*

[4] Servidone, G., Conti, D. (2016) *Discovering and labelling of temporal granularity patterns in electric power demand with a Brazilian case study.* Pesquisa Operacional, 36(3), 575–595. https://doi.org/10.1590/0101-7438.2016.036.03.0575