Estructura básica

Introducción

* Contexto y justificación
* Objetivos

Análisis exploratorio de la demanda eléctrica

* Introducción
* Evolución y estacionariedad de la demanda eléctrica en el período de estudio
  + análisis de la evolución de la demanda eléctrica
  + estacionariedad de la demanda eléctrica
  + funciones de autocorrelación (ACF) y autocorrelación parcial (PACF)
* Evolución de la demanda eléctrica
  + evolución de la demanda eléctrica diaria y semanal
  + evolución de la demanda eléctrica mensual
* Influencia de la temperatura en la demanda eléctrica
* Resumen de variables exógenas

Extracción y preparación de los datos

* Introducción
* Conexión con la API de E-SIOS
* Extracción de datos de temperatura
* Incorporación de variables exógenas

Modelos de series temporales

* Introducción
* Métricas escogidas
* Modelo de Holt- Winters como modelo de partida
* SARIMAX
* XGBoost
* Prophet
* Comparación entre modelos

Conclusiones

Pasos a mejorar

Bibliografía

**Resumen**

El objetivo del presente trabajo es el análisis de la demanda eléctrica en el Estado Español durante el período comprendido entre los años 2014-2021 mediante modelos de series temporales, poniendo énfasis en la adición de variables exógenas a los modelos para ver si de verdad ayudan o no a mejorar las predicciones.

A través del análisis de la serie temporal que conforma la demanda eléctrica para el período de tiempo comentado, se han podido extraer las principales variables exógenas que después se han utilizado en los diferentes modelos.

Mediante la comparativa de métricas se han determinado cómo afectan las variables exógenas, así como el mejor modelo descriptivo de la serie.

**1.- Introducción**

**1.1.- Contexto y justificación**

Los mercados de electricidad se diferencian de otros mercados por la necesidad de producir la energía en el mismo instante que es consumida, manteniendo un balance casi perfecto.

El análisis de la demanda de energía eléctrica es importante para comprender no sólo los patrones de consumo de los usuarios y su relación con factores como el económico y climático, sino también para identificar uno de los elementos que afecta la formación de precio

En el momento en el que se escribe esta memoria (año 2022) todo lo relacionado con el Sistema Eléctrico español lleva siendo tendencia en los debates a nivel nacional, tanto social como político. Principalmente por el alza de los costes de generación y cómo afecta a nivel económico y social, pero también por cómo se tiene que generar la energía en la necesidad de una transición ecológica, sumándose ahora la incertidumbre por los suministros de energía en el nuevo contexto internacional que marcan la guerra en Ucrania y la tensión de las relaciones con Argelia.

Es en este contexto que surge el deseo de conocer mediante el análisis de series temporales cómo se comporta la demanda energética en el Estado Español para el período 2014-2021: qué tendencia global sigue, comportamiento estacional, variables que afectan a dicha demanda, etc. Para ello se utilizarán diferentes técnicas para generar modelos y cómo afecta la adición de variables exógenas a la bondad de dichos modelos.

**1.2.- Objetivos**

Los objetivos del presente trabajo son los siguientes:

* Caracterización de la demanda eléctrica en el Estado Español.
* Identificación de las principales variables que afectan a la demanda eléctrica.
* Modelización
* Identificación del peso de las variables exógenas en la precisión de los modelos.

**2.- Análisis exploratorio de la demanda eléctrica**

**2.1.-Introducción**

(pendiente)

**2.2.-Evolución y estacionariedad de la demanda eléctrica en el período de estudio**

**2.2.1.- Análisis de la evolución de la demanda eléctrica**

Dado que nuestro período de análisis abarca los años 2014-2021, el perfil de la demanda eléctrica durante ese período es el siguiente:

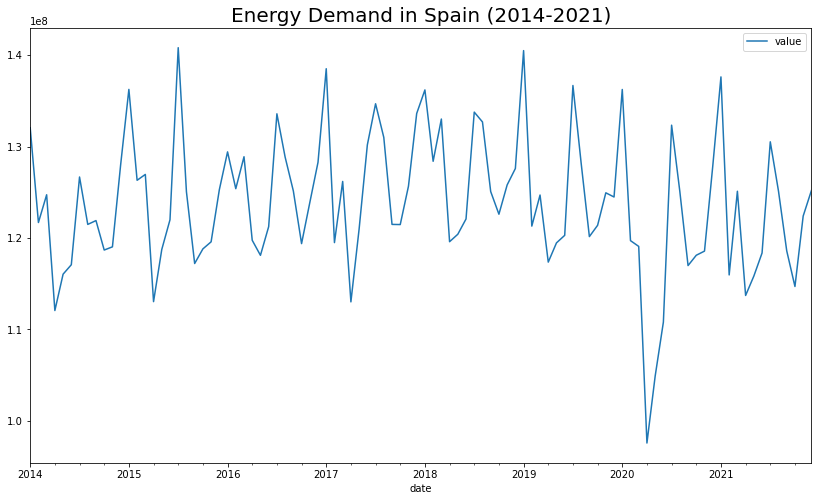


Fig. 1: Demanda energética en España (2014-2021)

Podemos ver en esta representación cómo la demanda de energía oscila de forma periódica a lo largo del tiempo con unos máximos y mínimos dentro de un margen bien definido entre (1.1-1.4)\*108 Mwh, excepto en 2020 donde vemos una caída abrupta que coincide con la irrupción del COVID-19 en España y el inicio del confinamiento estricto que duró casi 3 meses. Después podemos ver cómo remonta la demanda de energía coincidiendo con la vuelta a cierta normalidad.

Podemos intuir que el pico máximo cada año se produce a principios de año y en los meses de verano debido a la mayor necesidad de energía en los hogares para calefacción o aire acondicionado, respectivamente. Por otra parte, los meses de menor consumo son los correspondientes a los meses de primavera y otoño. Esto sería un primer indicativo de cómo afecta la temperatura a la demanda eléctrica: los valores extremos implican un mayor consumo de energía. Más adelante veremos qué rangos de temperaturas provocan estas mayores demandas.

Veamos cómo se distribuye la demanda de energía por año y mes en el intervalo de tiempo que tratamos. La demanda anual nos permite ver una tendencia a lo largo del tiempo, mientras que la demanda mensual nos permite ver si hay estacionalidad.

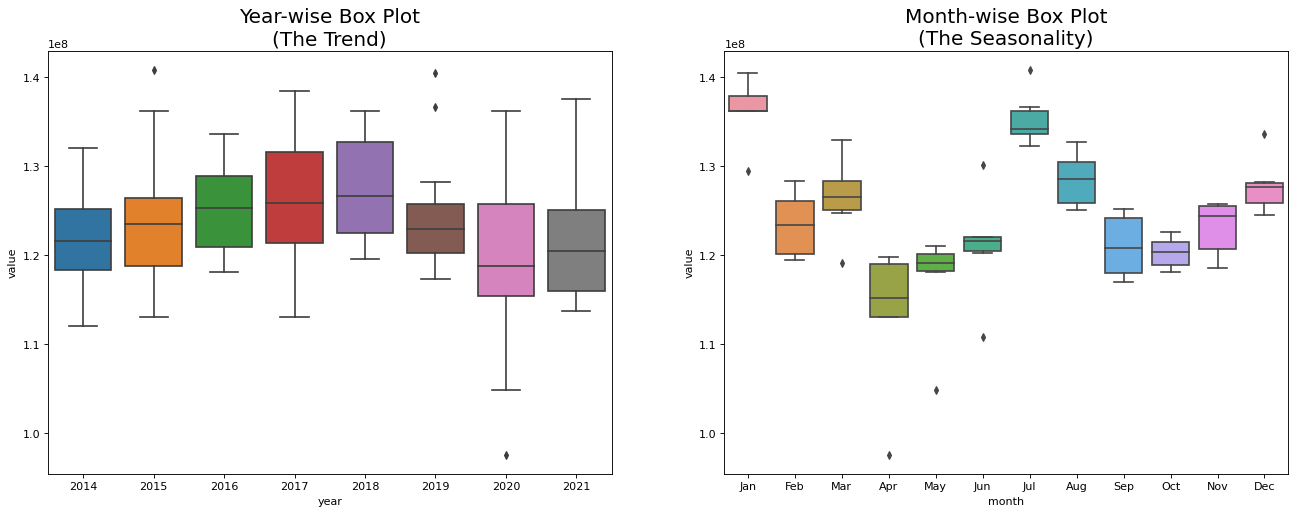


Fig.2: Distribución promedio de la demanda energética por año (izda.). Distribución promedio de la demanda energética por mes.

En el primer gráfico, podemos ver cómo la media de la demanda de energía cada año sube hasta llegar al año 2018, a partir del cual se produce un descenso bastante notable en 2019 para caer aún más en 2020 y volver a remontar en 2021, aunque todavía bastante por debajo de 2018 y los años anteriores.

Ya explicamos que la caída en la demanda de energía en 2020 se debe a la aparición del COVID y el inicio del confinamiento, pero lo indicado en el 2018 como el año de mayor consumo podría ser indicativo de que este año marca el inicio de un período de contracción económica.

En el segundo gráfico confirmamos lo que habíamos dicho antes: los mayores consumos se producen en los meses de invierno (diciembre-marzo) y verano (julio-agosto), mientras que los de menor consumo se corresponden con los meses de primavera (abril-junio) y otoño (septiembre-noviembre).

Vamos a ver una comparativa de la demanda energética por año de forma superpuesta para ver mejor los patrones de demanda anuales.

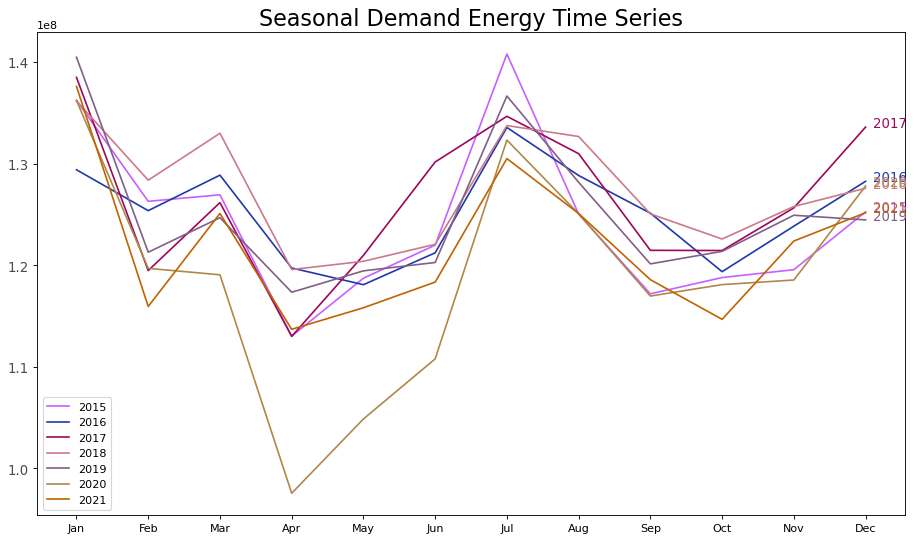


Fig. 3: Distribución estacional de la demanda energética (2014-2021)

Podemos ver a nivel general cómo cada año se repiten los patrones de demanda de energía en los mismos meses.

Destaca 2015 con el máximo valor registrado en la serie en el mes de julio, mientras que el valor mínimo aparece en abril de 2020.

En 2017 podemos ver cómo una vez alcanzado el mínimo en abril, la demanda de energía crece de forma rápida manteniendo valores muy cercanos en los meses de junio, julio y agosto, lo que nos puede indicar que fue un año bastante caluroso y no solo en julio, como viene siendo habitual en el resto de años. En ese aspecto, también destaca 2018 por tener un consumo muy parecido entre julio y agosto.

De la tendencia y estacionalidad que hemos analizado antes podemos obtener conclusiones parecidas descomponiendo la señal en sus 3 componentes: tendencia, estacionalidad y ruido.

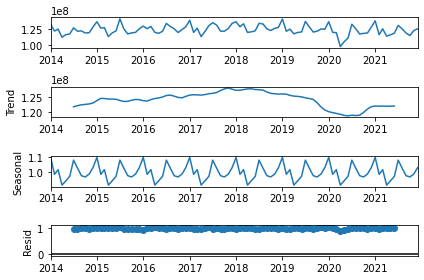


Fig. 4: Descomposición de la demanda energética (2014-2021) en sus componentes (tendencia, estacionalidad y residual)

**2.2.2.- Estacionariedad de la demanda eléctrica**

Antes de realizar cualquier modelo de series temporales, es importante saber si es estacionaria o no, es decir, si su media y varianza no están en función del tiempo. La estacionariedad es importante porque la mayoría de métodos estadísticos para realizar análisis y predicciones se basan en la suposición de que las propiedades estadísticas de la serie (media, varianza, correlación, etc.) son constantes en el tiempo. Para comprobarlo vamos a usar el test de Dickey-Fuller que establece como hipótesis nula (H0) que la serie es estacionaria.

(DUDA: se pondría código aquí?)

Obtenemos el siguiente resultado:

**ADF Test Statistic:** -1.34347257117945

**p-value:** 0.6090564874005809

**Lags Used:** 12

**Number of Observations Used:** 83

**Conclusion:** weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary

Por lo tanto, podemos ver cómo nuestra serie no es estacionaria.

**2.2.3.- Funciones de autocorrelación (ACF) y autocorrelación parcial (PACF)**

Antes de construir los modelos necesitamos saber cómo se correlacionan los datos entre sí mediante las funciones de autocorrelación (ACF) y las funciones de autocorrelación parcial (PACF). Los análisis de autocorrelación ayudan a detectar patrones ocultos y estacionalidad, así como buscar aleatoriedad.

ACF mide y dibuja la correlación promedio entre los datos de una serie temporal y sus valores previos para diferentes valores de retrasos (lags). También mide el grado de asociación lineal que existe entre dos variables del mismo proceso estocástico.

Una PACF es similar a una ACF midiendo la correlación entre dos variables separadas por k periodos cuando no se considera la dependencia creada por los retardos intermedios existentes entre ambas. Es decir, mide la autocorrelación que existe entre dos variables separadas k períodos descontando los posibles efectos debidos a variables intermedias.

Gracias a ACF podemos responder las siguientes preguntas:

* ¿Es nuestra serie temporal ruido blanco o random?
* ¿Una observación está relacionada con su adyacente?...¿y la siguiente? ¿y la siguiente?
* ¿La serie temporal se puede describir con un modelo MA? ¿de qué orden?

Mientras que PACF:

* ¿La serie temporal se puede modelar con un modelo AR? ¿de qué orden?

Por lo tanto, el valor de un ACF y PACF en el primer retraso es igual porque ambos miden la correlación entre datos en un tiempo t con puntos atrasados un tiempo t-1. Sin embargo, en el segundo retraso, la ACF mide la correlación entre los puntos en un instante t con los puntos en un instante t-2, mientras que PACF mide la misma correlación, pero después de controlar la correlación entre los puntos en un instante t con los que están en t-1.

Gracias a ACF y PACF podemos entender la dinámica temporal de una serie temporal individual.

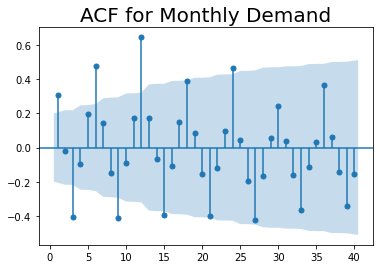


Fig. 5: Autocorrelación para la distribución mensual de la demanda energética (2014-2021)

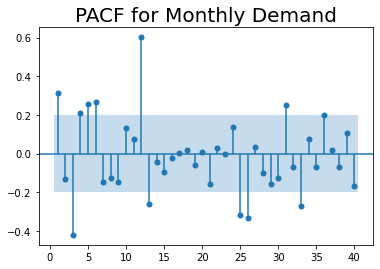


Fig. 6: Autocorrelación parcial para la distribución mensual de la demanda energética (2014-2021)

Los valores dentro del área azul en ambas gráficas indica que son estadísticamente cercanos a cero, mientras que sí que es estadísticamente significativo todo lo que hay fuera.

En ACF podemos ver una fuerte correlación positiva con el primer y tercer valor. Más adelante podemos ver fuertes correlaciones negativas y positivas cada 3 valores (3, 6, 9, etc) destacando 12 y 24.

En PACF, por otro lado, vemos una fuerte correlación positiva con el valor 1, negativa en el valor 3, en los valores 4, 5 y 6 y de nuevo destaca el valor 12.

Podemos explicar la fuerte correlación con 12, sus múltiplos y divisores, por el marcado período de 12 meses que tiene la serie.

**2.3.- Evolución de la demanda eléctrica**

**2.3.1.- Evolución de la demanda eléctrica diaria y semanal**

Vamos a centrarnos en la distribución de la demanda energética a lo largo de 2 semanas para ver cómo se distribuye la energía. Seleccionamos dos semanas aleatorias (14-01-2019 - 28-01-2019) para hacernos una idea de si hay diferencias en la demanda energética según sea el día laboral.

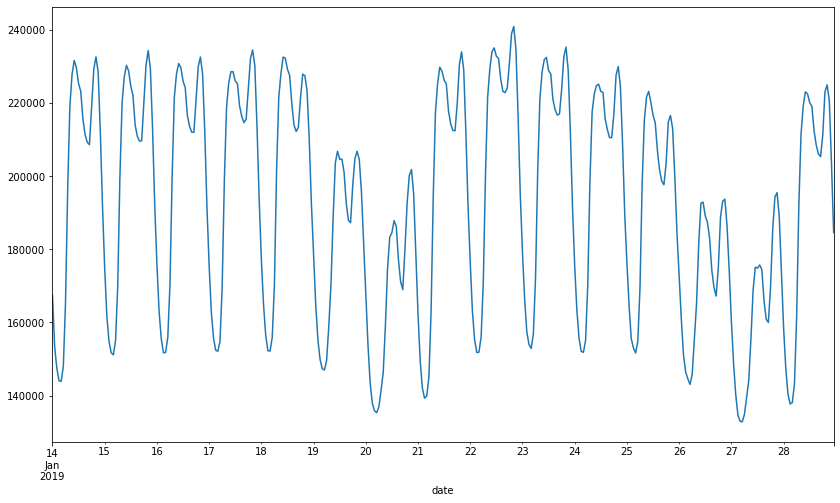


Fig. 7: Distribución diaria de la demanda energética (semanas 14-01-2019 - 28-01-2014)

Podemos ver claramente un comportamiento biestacional en el que los patrones de demanda de energía se repiten día a día y a lo largo de la semana.

Podemos ver cómo de lunes a viernes (días laborables) el patrón de demanda y sus valores son casi los mismos, para luego decaer en sábado y domingo (días no laborables), siendo el domingo el día de mayor demanda energética.

Vamos a ver si de verdad afectan los festivos a la demanda energética semanal. La Semana Santa es bastante ilustrativa, ya que siempre hay dos días laborables (jueves y viernes) festivos.

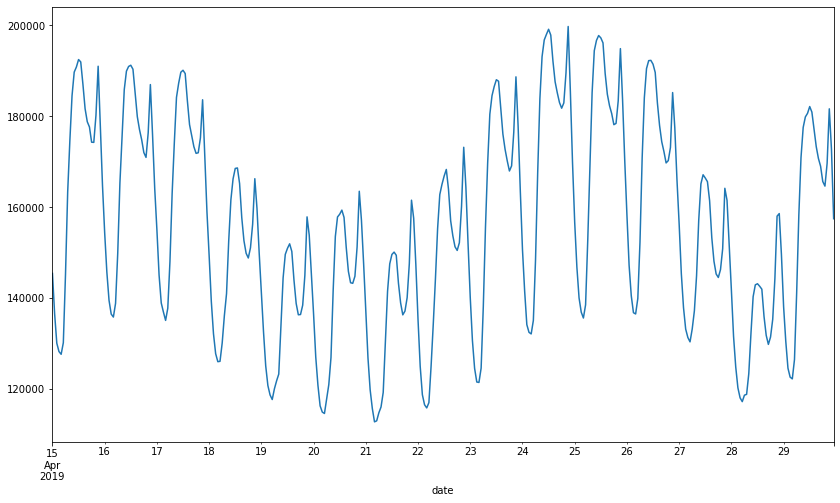


Fig. 8: Distribución diaria de la demanda energética (semanas 15-04-2019 - 29-04-2014)

Podemos ver cómo de lunes a miércoles el consumo es casi el mismo, para luego decaer durante 4 días (jueves, viernes, sábado y domingo) por ser festivos, para luego volver a aumentar durante la semana la siguiente semana a patrones habituales: 5 días de demanda elevada, 2 días de demanda reducida.

Para confirmar a nivel global que en los días festivos la demanda energética es menor, vamos a calcular la media de todos los valores de demanda energética según el tipo de día:

| **Tipo de día** | **Demanda media (MWh)** | **Porcentaje (%)** |
| --- | --- | --- |
| Festivo | 145177.96 | 46 |
| Laborable | 170217.51 | 54 |

Tabla 1: Demanda energética media para días laborables y festivos

Por lo tanto, podemos ver cómo la demanda promedio un día laborable es un 8% mayor que en un día festivo.

Vamos a ver a nivel global cómo es la distribución de la demanda energética según el día de la semana.

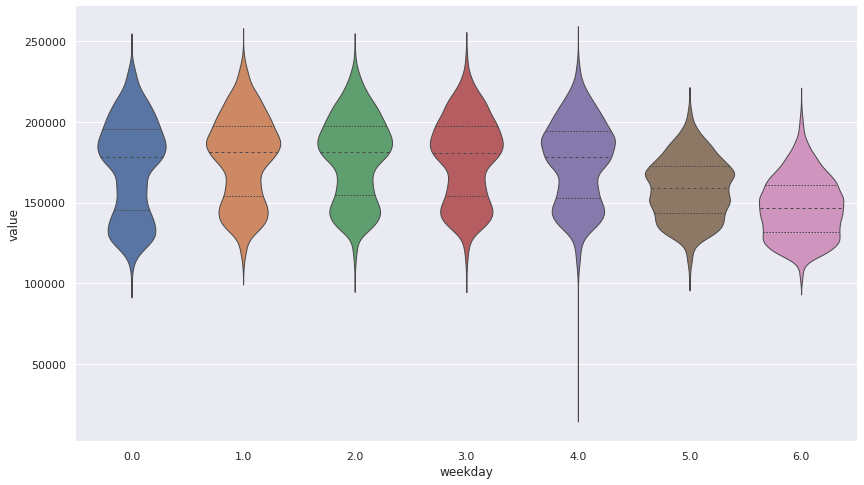


Fig. 9: Distribución promedio de la demanda energética por día de la semana (2014-2021)

Podemos ver cómo los días usualmente laborables (de lunes a viernes) concentran la mayor demanda energética, mientras que ésta decae los fines de semana (sábado y domingo), siendo el domingo el día de menor demanda. Martes, miércoles y jueves son los que tienen una mayor media y cabe destacar el viernes como el día de mayor variabilidad teniendo los picos más extremos de mayor y menor demanda.

**2.3.2.- Evolución de la demanda eléctrica mensual**

Vamos a centrarnos ahora en la distribución por hora-día a lo largo de los meses.

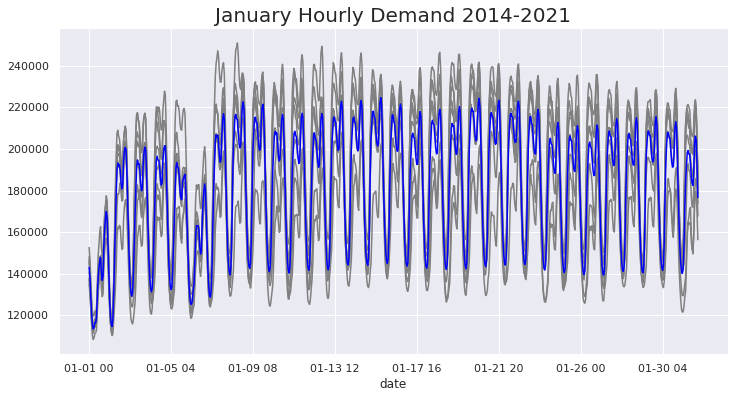


Fig. 10: Distribución promedio de la demanda energética durante el mes de enero (2014-2021)

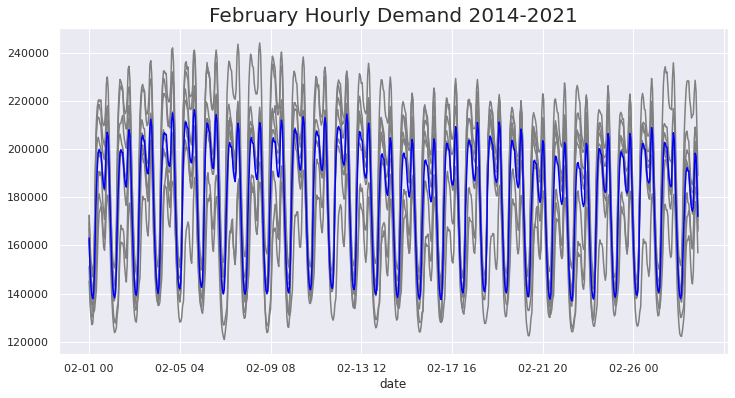


Fig. 11: Distribución promedio de la demanda energética durante el mes de febrero (2014-2021)

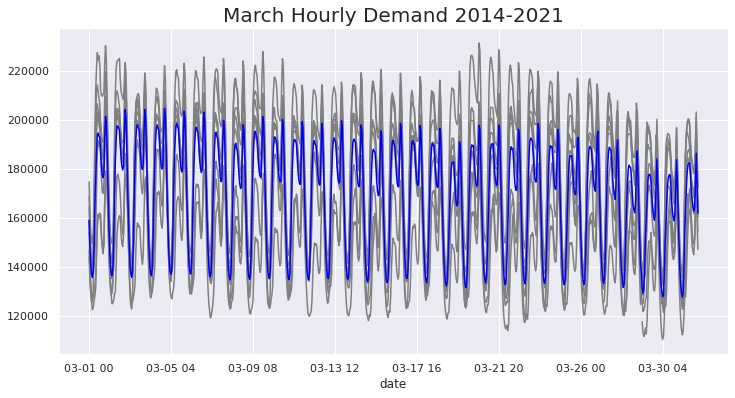


Fig. 12: Distribución promedio de la demanda energética durante el mes de marzo (2014-2021)

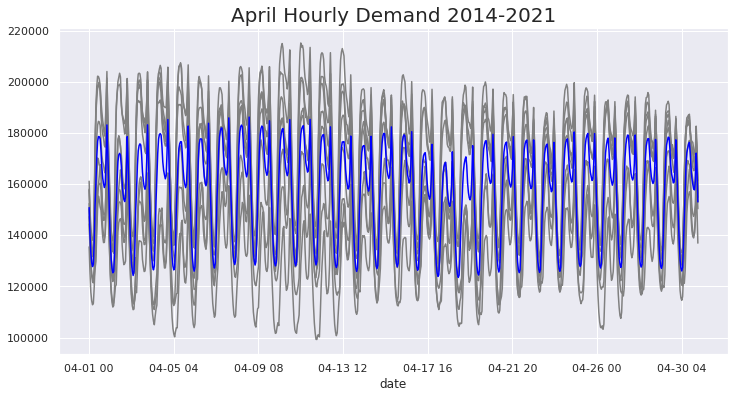


Fig. 13: Distribución promedio de la demanda energética durante el mes de abril (2014-2021)

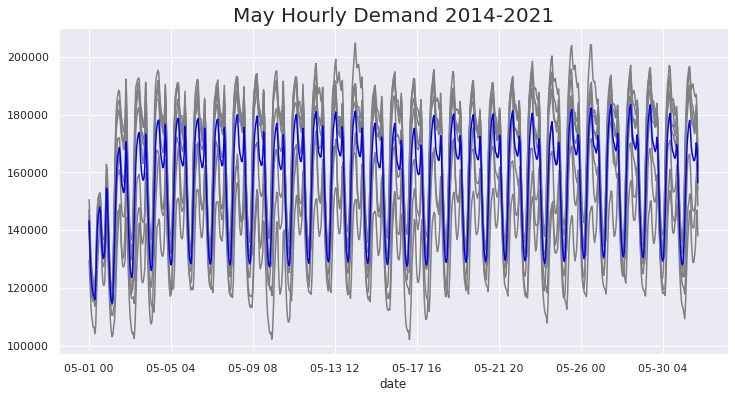


Fig. 14: Distribución promedio de la demanda energética durante el mes de mayo (2014-2021)

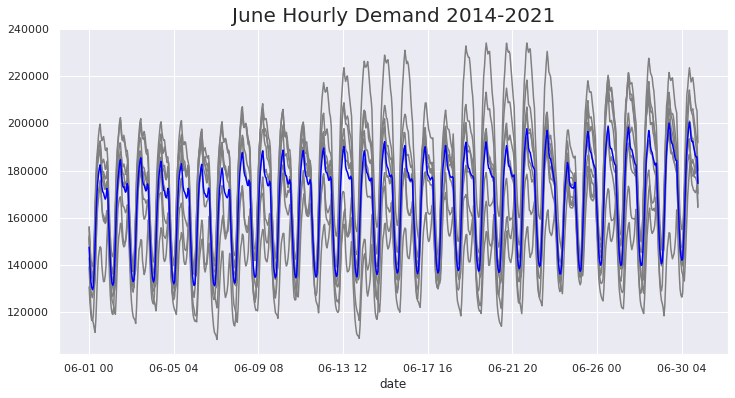


Fig. 15: Distribución promedio de la demanda energética durante el mes de junio (2014-2021)

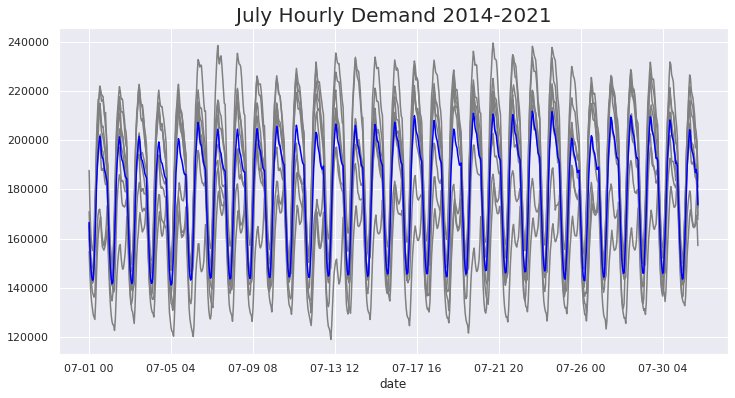


Fig. 16: Distribución promedio de la demanda energética durante el mes de julio (2014-2021)

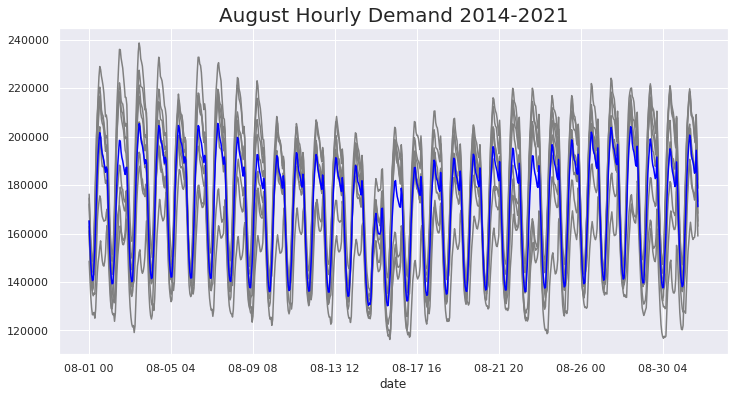


Fig. 17: Distribución promedio de la demanda energética durante el mes de agosto (2014-2021)

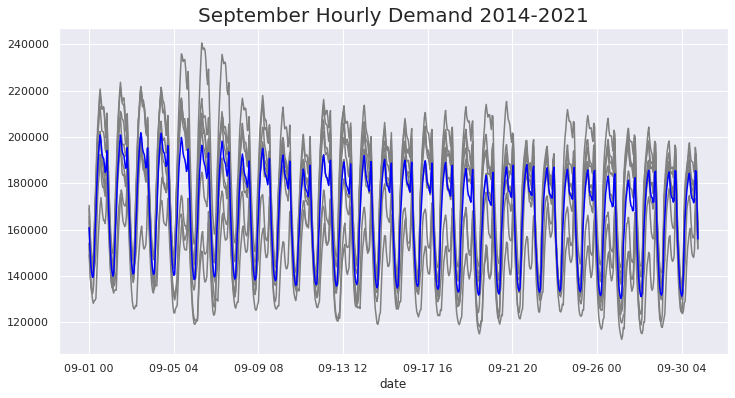


Fig. 18: Distribución promedio de la demanda energética durante el mes de septiembre (2014-2021)

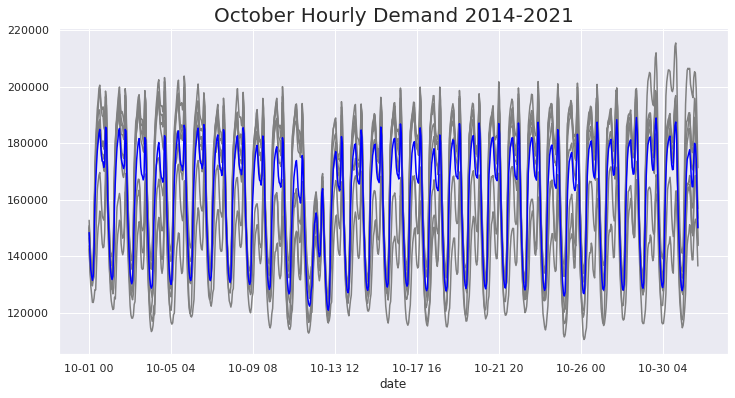


Fig. 19: Distribución promedio de la demanda energética durante el mes de octubre (2014-2021)

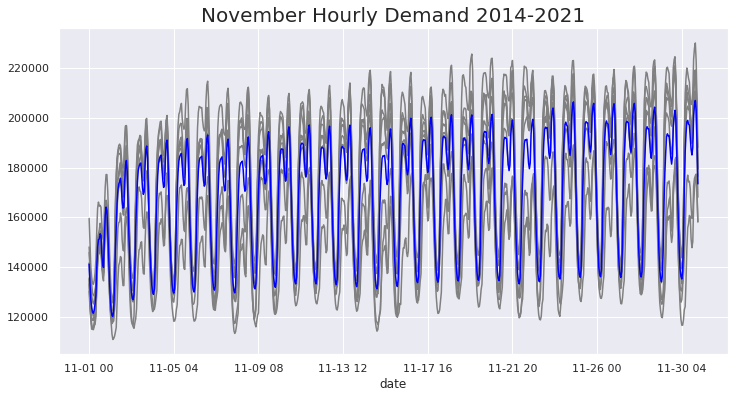


Fig. 20: Distribución promedio de la demanda energética durante el mes de noviembre (2014-2021)

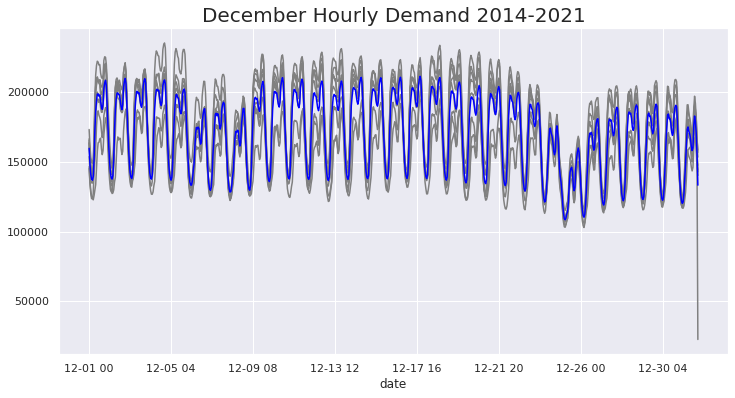


Fig. 21: Distribución promedio de la demanda energética durante el mes de diciembre (2014-2021)

Si atendemos a los festivos nacionales en el Estado español, sobre todo a los que se producen en días fijos, podemos ver y confirmar que los festivos producen una menor demanda energética. Los festivos de enero y diciembre, los dos primeros días de mayo, el 15 de agosto, 12 de octubre y los dos primeros días de noviembre son un claro ejemplo.

Por otro lado, a nivel de mes podemos ver cómo los valores se mueven siempre en unos márgenes bastante definidos, pero que varían de amplitud entre los diferentes meses. En los meses de transición de primavera a verano y otoño a invierno podemos ver tendencias de decrecimiento/crecimiento de la demanda debido principalmente al factor climático, así como los menores consumos.

En cuanto a los rangos de las medias de demanda podemos ver cómo enero es el mes donde mayor amplitud de demanda se alcanza, siendo el valor mínimo 140.000 MWh y el máximo 200.000 aproximadamente.

Agosto es el mes donde se puede apreciar bien una tendencia de caída de la primera semana a la segunda-tercera, no solo por el festivo del 15 de agosto sino también por ser el mes en el que mayor cantidad de gente toma sus vacaciones, para luego retomar rápidamente a niveles previos. En noviembre también podemos observar una tendencia clara de aumento de la demanda a medida que avanza el mes hacia diciembre.

Por último vamos a ver si la serie es estacionaria en el período de un mes.

**ADF Test Statistic:** -2.4717952359371846

**p-value:** 0.12248851951568468

**Lags Used:** 20

**Number of Observations Used:** 699

**Conclusion:** weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary

Podemos ver que la serie tampoco es estacionaria si se atiende a la distribución de la demanda por horas. Vamos a ver ahora cómo se correlacionan los valores entre sí.

(imágenes ACF y PACF)

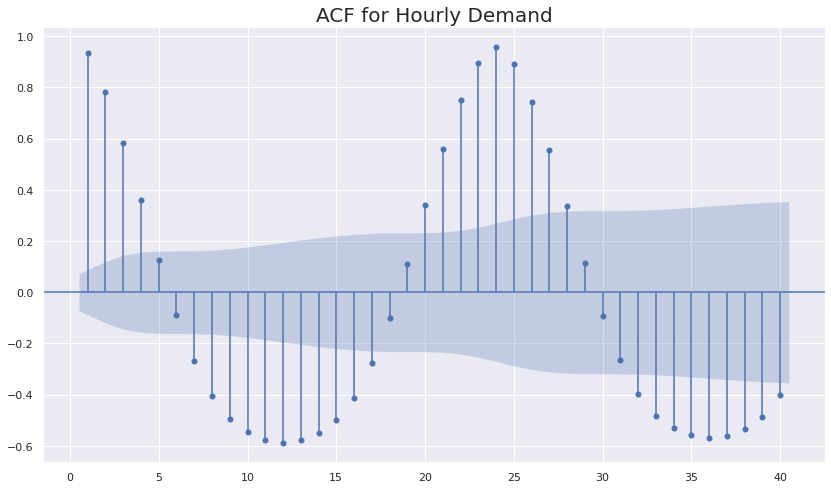


Fig. 22: Autocorrelación de los valores diarios de demanda energética (2014-2021)

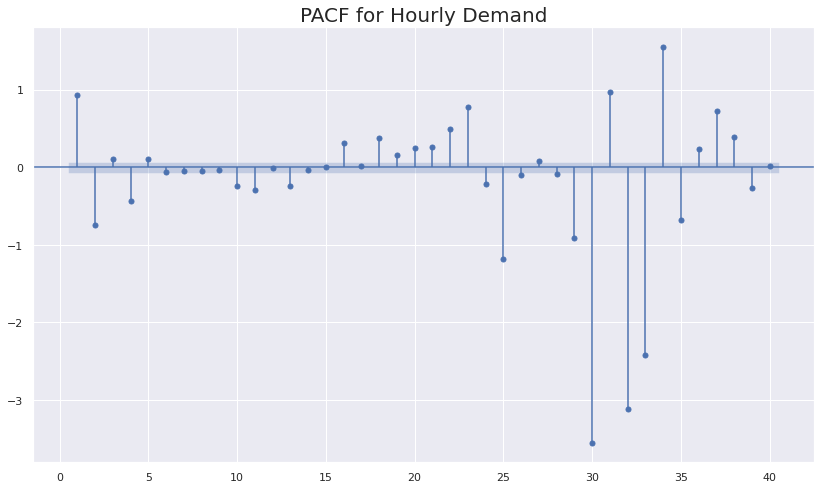


Fig. 23: Autocorrelación parcial de los valores diarios de demanda energética (2014-2021)

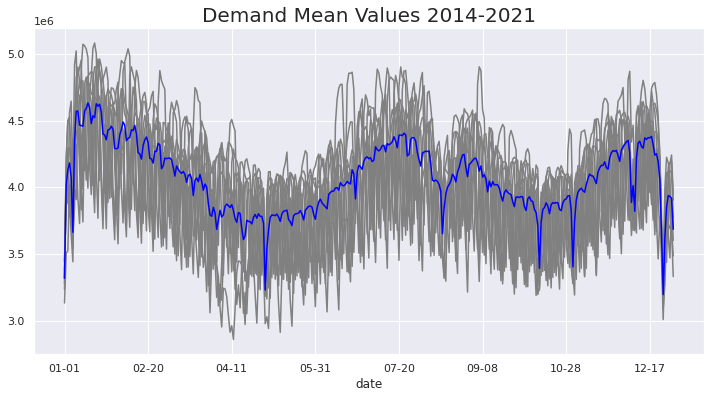
Podemos ver en la gráfica ACF cómo hay una fuerte dependencia de los 6 primeros valores, para luego tener una serie de valores estadísticamente no significativos y volver a tener valores significativos, aunque de menor importancia que los 3 primeros valores. Destaca cómo los valores vuelven a ser significativos en el valor 24, como corresponde a una serie de con un período de 24h.

Por otro lado, en PACF vemos que parece ser que todos los valores son estadísticamente significativos. La correlación es más importante en los 2 primeros valores, decae y vuelve a crecer mucho en valores bastante lejanos (a partir del valor 25) lo que nos puede indicar la presencia de ruido.

**2.3.3.- Influencia de la temperatura en la demanda eléctrica**

Para entender la influencia de la temperatura en la demanda eléctrica, vamos a comparar los gráficos demanda y temperatura media durante el intervalo de un año.

(gráficos temp-demanda)



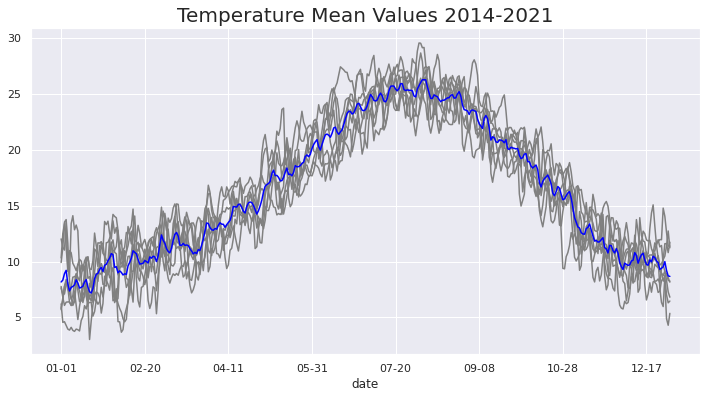


Fig. 24: Comparativa entre demanda energética media y temperaturas medias (2014-2021)

Podemos confirmar lo que explicamos al inicio sobre la influencia que ejerce la temperatura en la demanda energética: para bajas temperaturas (meses de invierno) y para meses de altas temperaturas (meses de verano) la demanda crece considerablemente, mientras que en los meses de otoño y primavera la demanda se reduce.

En los períodos extremos de invierno y verano podemos ver que se producen las mayores demandas, pero a la par grandes caídas. Esto se debe a que ambos períodos coinciden con festivos o con la costumbre generalizada de tomar vacaciones en esos espacios de tiempo.

Por lo tanto, para reflejar mejor en nuestro dataset esta influencia de la temperatura para valores extremos, vamos a añadir 2 columnas al dataset que notifiquen si el día y mes de cada registro está en verano o en invierno.

**3.- Extracción y preparación de los datos**

**3.1.-Introducción**

Los datos que se analizan en el presente trabajo se han extraído de 2 fuentes:

* **Demanda eléctrica:** los datos se han extraído de la API que ofrece el operador del sistema eléctrico en España, Red Eléctrica, y que se conoce como *e.sios*.
* **Temperatura:** de la web de *Copernicus Climate Data Store.*

**3.2.-Conexión con la API de E-SIOS**

Para poder extraer los datos de la API de *e.sios,* lo primero que se tiene que hacer es pedir un *token* de acceso que tiene que acompañar todas las consultas que se hagan. El *token* se solicita a través de correo electrónico a la dirección [consultasios@ree.es](mailto:consultasios@ree.es) . Una vez conseguido el token, las consultas se han realizado usando la librería *requests* de Python que permite realizar la gestión de la conexión *http* y la descarga de sus datos.

La API de *e.sios* ofrece información de una amplia gama de indicadores. Los indicadores se extraen con la siguiente consulta:

[*https://api.esios.ree.es/indicators*](https://api.esios.ree.es/indicators)

Con la respuesta se identifica el indicador de demanda eléctrica que tiene como índice *1293*. Por lo tanto, para extraer los datos de demanda eléctrica realizamos una consulta a la API de e.sios usando la siguiente url.

[https://api.esios.ree.es/indicators/1293/?start\_date='+start\_date+'&end\_date='+end\_date+'&time\_trunc=[VARIABLE\_TEMPORAL](https://api.esios.ree.es/indicators/1293/?start_date='+start_date+'&end_date='+end_date+'&time_trunc=%5BVARIABLE_TEMPORAL)]

Donde *start\_date* y *end\_date* son campos de fechas con los que fijar el intervalo de tiempo para el cual se solicita información de la demanda eléctrica. Por otro lado, el campo [VARIABLE\_TEMPORAL] es un campo que puede tomar los valores *hour, day, month, year*, etc para fijar en qué régimen de frecuencia queremos los datos (horario, diario, mensual, anual, etc).

**3.3.-Extracción de datos de temperatura**

Los datos de temperatura se extrajeron de la web de *Copernicus Climate Data Store.* Copernicus es una iniciativa de la Comisión Europea y de la Agencia Espacial Europea para construir un sistema autónomo de observación de la Tierra que permita la observación del medio ambiente y cómo le afectan los cambios ambientales, el origen de estos cambios y la influencia en la vida de las personas. Copernicus tiene como objetivo entonces ser parte fundamental en la protección del medio ambiente, la salud y seguridad de los ciudadanos de la Unión Europea.

La solicitud de datos se realiza en el siguiente enlace:

<https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=form>

Para poder solicitar datos es necesario hacerse una cuenta de usuario.

En el presente trabajo se han extraído datos de temperaturas para el período 2014-2021 de las 5 ciudades más importantes de España: Madrid, Barcelona, Bilbao, Valencia y Sevilla. Además de ser las ciudades más importantes de España, están bastante bien repartidas por su geografía, lo que permite captar con relativamente pocos datos la diversidad de temperaturas en todo el territorio en el mismo instante.

Para solicitar los datos de cada ciudad han sido necesarios indicar su ubicación geográfica, el período de tiempo (2014-2021) y la frecuencia de toma de datos, que ha sido horaria.

Una vez obtenidos los 5 datasets de temperaturas se han juntado junto con los de demanda para formar un único dataset.

**3.4.-Incorporación de variables exógenas**

Después de lo que se ha comentado en el análisis exploratorio sobre los factores que afectan a la demanda eléctrica, para los modelos se han escogido las siguientes variables exógenas que deben ayudar a realizar mejores predicciones:

* Día de la semana (0-6 lunes-domingo)
* Días laborables (1) / Festivos (0)
* Temperatura
* Estacionalidad (invierno / verano)

De la extracción de la temperatura se ha hablado en el apartado anterior. Los días de la semana se han obtenido a partir de las fechas de los registros de la base de datos, mientras que los días festivos en el Estado Español se han recopilado para el período 2014-2021 y después se han contrastado con las fechas que disponemos. Por último, con la estacionalidad se ha procedido de forma similar a los anteriores: si las fechas caen o no en invierno se marcan como 1 o 0 en la columna *is\_winter* y lo mismo para la columna *is\_summer* en caso del verano.



Fig. 25: Base de datos resumen

Modelos de series temporales

* Introducción
* Métricas escogidas
* Holt- Winters como modelo de partida
* SARIMAX
* XGBoost
* Prophet
* Comparación entre modelos

**4.- Modelos de series temporales**

**4.1.- Introducción**

Una vez obtenidos y analizados los datos de demanda eléctrica y los factores que pueden influir sobre ella, vamos a modelizar a través de diferentes técnicas para ver cómo afectan cada uno de estos factores sobre la precisión de los modelos.

Las técnicas de modelización empleadas han sido las siguientes:

* Holt-Winters
* SARIMAX
* XGBoost
* Prophet

El motivo de escoger estas técnicas ha sido la de probar si afectan por igual o no la misma los factores exógenos a la hora de hacer predicciones.

Holt-Winters se ha empleado como modelo de partida sobre el que se ha aplicado el dataset de la forma más sencilla posible: usando sólo datos de demanda a lo largo del tiempo y de ahí comparar con el resto de modelos y cómo cambian sus métricas con la adición de factores exógenos.

**4.2.- Holt- Winters como modelo de partida**

El método de Holt-Winters pertenece al conjunto de técnicas de suavizado exponencial y es una ampliación perfeccionada del método de suavización exponencial de Holt que considera solo dos exponentes suavizantes. Holt-Winters considera tres exponentes suavizantes (nivel, tendencia y estacional) de una serie temporal dada. A diferencia de otras técnicas de suavizado exponencial, Holt-Winters puede adaptarse fácilmente a cambios y tendencias, así como a patrones de estacionalidad.

Según el tipo de estacionalidad tenemos dos modelos:

* **Modelo Multiplicativo estacional:** este modelo supone que a medida que aumentan los datos, también aumenta el patrón estacional en la serie. La gran mayoría de series temporales muestran este patrón. En este modelo la tendencia y los componentes de estación se multiplican y luego se suman al componente de error.
* **Modelo Aditivo Estacional:** en este modelo los efectos de los factores individuales se diferencian y se agrupan para modelar los datos. Este método es el mejor para datos con tendencia y estacionalidad que no aumentan a lo largo del tiempo. Su resultado es una previsión de curva que muestra los cambios estacionales en los datos.

**4.3.- SARIMAX**

SARIMAX es el acrónimo de *Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors* que se traduciría como: Media móvil integrada autorregresiva estacional con factores exógenos.

SARIMAX supone una versión mejorada de los modelos ARIMA y SARIMA que incorpora tanto estacionalidad como factores exógenos a la elaboración del modelo. Otra ventaja es que a diferencia de ARIMA, SARIMAX se puede aplicar sobre series temporales no estacionarias, como ocurre en nuestro caso.

Los parámetros de entrada del modelo serían los siguientes:

model\_sarimax = SARIMAX(

train / test data,

exo train/ test,

order=(3, 2, 3),

seasonal\_order=(3, 0, 3, 4)

)

El primer parámetro serían los datos de entrenamiento/test, el segundo parámetro los datos exógenos; el parámetro order = (p, d, q) está formado por tres variables: p es el orden del componente AR no estacional, d el orden diferenciado no estacional y q el orden del componente MA no estacional; el parámetro seasonal\_order = (P, D, Q, s) está formado por 4 variables: P es el orden del componente RA estacional, D es el orden de integración de estacionalidad de las series temporales, Q es el orden del componente estacional MA y s es la duración de la estacionalidad.

**4.4.- XGBoost**

XGBoost es el acrónimo de eXtreme Gradient Boosting (Aumento del Gradiente Extremo) es un algoritmo de machine learning de tipo supervisado más popular a día de hoy.

Su popularidad se debe al poco esfuerzo relativo que necesita para obtener buenos resultados de predicción compitiendo muy bien frente a modelos complejos computacionalmente, sobre todo en lo relativo a tratar con datos heterogéneos.

El principio de boosting que emplea se basa en la generación de múltiples modelos de predicción que se pueden calificar como débiles secuencialmente, pero con la particularidad de que cada uno de esos modelos tome los resultados del anterior generando un modelo cada vez más fuerte en cuanto a mejor poder predictivo y resultados más estables. Los modelos débiles de partida son árboles de decisión de diferentes tipos que se pueden emplear en tareas de clasificación y regresión.

El algoritmo de optimización entre modelos es el del descenso del gradiente. El proceso para cuando se alcanza un número máximo de iteraciones o cuando la diferencia entre modelos es insignificante.

**4.5.- Prophet**

Prophet es una herramienta de código abierto creada por Facebook para el tratamiento y modelado de series temporales que se basa en un modelo aditivo descomponible donde las tendencias no lineales se ajustan a la estacionalidad, además de a eventos como las vacaciones y sus efectos en una serie.

Resulta muy útil para series que abarquen períodos de tiempo amplios con frecuencias de hora, día, semana, mes, etc; también para series que puedan llegar a tener varias estacionalidades muy marcadas, eventos irregulares, que les falten datos, así como series con tendencias de crecimiento no lineal que se aproximen a un límite.

La expresión en la que se basa el modelo es la siguiente:

**y(t) = g(t) + s(t) + h(t) + ϵₜ**

donde g(t) es la función que modeliza la tendencia del modelo, es decir, el comportamiento a largo plazo; s(t) modeliza la estacionalidad del modelo mediante series de Fourier; h(t) modela el efecto que tienen eventos como las vacaciones, ofertas, etc; y ϵₜ representa el término del error irreducible.

**4.6.- Métricas escogidas**

A la hora de comparar modelos entre sí aplicados a una misma serie temporal es importante saber que la selección puede afectar no solo a la optimización de los modelos, sino también al juicio e interpretación de los mismos.

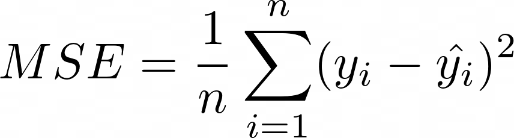
Cada vez son más las métricas de errores que se están proponiendo para el estudio de series temporales. Hay que tener en cuenta que no hay métrica mejor que otra, ya que cada una condensa información sobre una gran cantidad de datos en un solo valor, de forma que solo da una proyección de los errores de los modelos haciendo énfasis en un aspecto determinado de los errores de sus características.

Hay 3 grandes grupos de métricas de errores para series temporales:

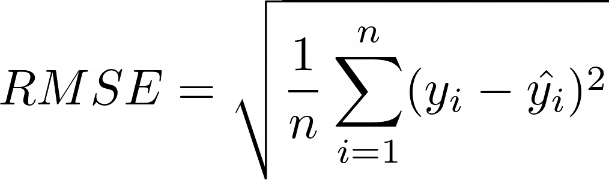
* **Dependientes de la escala:** métricas que vienen dadas en las mismas unidades que los datos que se usan. Su principal ventaja es que son fáciles de calcular y de interpretar, pero no se pueden usar para comparar diferentes series por la dependencia que tienen de las unidades. En este grupo entrarían métricas como el Error Absoluto Medio (MAE por sus siglas en inglés), el Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE).
* **Errores porcentuales:** dado que los errores dependientes de la escala no permiten la comparación entre diferentes series temporales, los errores porcentuales sí que permiten resolver este problema. Sin embargo, tienen el punto débil de que no admiten valores iguales a cero, ya que convierten a este tipo de errores en infinito o indefinido, lo que no permite interpretarlos. En este grupo de errores entrarían el Error Porcentual Absoluto (MAPE) y el Error Porcentual Absoluto Simétrico (sMAPE).
* **Errores libres de escala:** este tipo de errores viene a subsanar el problema de los errores dependientes de escala. En este grupo se encuentra el Error Absoluto Medio Escalado (MASE) que viene a mejorar el error MAE.

En el presente trabajo se han seleccionado las siguientes métricas:

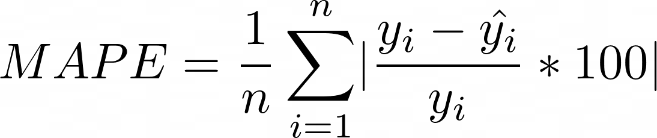
* **MAE (Mean Absolute Error):** se calcula calculando la media de las diferencias absolutas entre los valores reales de la serie y los predichos por el modelo. Este cálculo tan simple es su principal ventaja, pero, además de no permitir comparar series temporales diferentes, no se debe usar si se quieren penalizar los valores atípicos.
* **MSE (Mean Squared Error):** esta métrica nos sirve si se quiere poner atención en los valores atípicos. Se calcula como la media del cuadrado del MAE. Debido a que se eleva al cuadrado, da más peso a los errores grandes que a los pequeños, lo que puede ser una desventaja en algunas situaciones. Por lo tanto, el MSE es apropiado para situaciones en las que quieres poner el foco en errores grandes. Indicar que al estar elevado al cuadrado hace que pierda sus unidades.



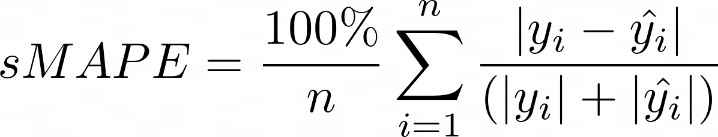
* **RMSE (Root Mean Squared Error):** para evitar la pérdida de unidades del MSE, podemos tomar su raíz cuadrada. Tiene las mismas ventajas que el MAE y el MSE, sin embargo, al igual que el MSE, es sensible a los valores atípicos.



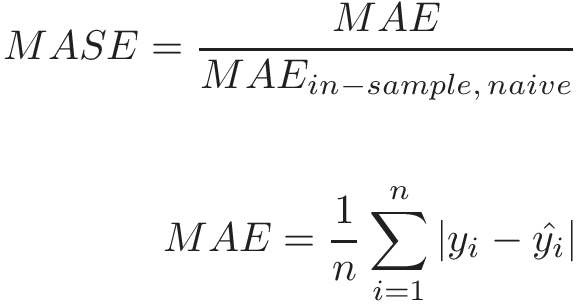
* **MAPE (Mean Absolute Percentage Error):** es una de las métricas más populares en series temporales. Se calcula como la media de la diferencia absoluta entre los valores reales y predichos dividido por los valores reales. La principal ventaja de MAPE es su independencia de las escalas de las unidades, así como su fácil interpretabilidad, lo que permite usarlo para comparar modelos de diferentes series. Sin embargo, MAPE tiene algunas desventajas como generar infinitos o valores indefinidos para valores iguales a cero o cercanos a cero. También penaliza fuertemente sobre errores negativos más que sobre errores positivos, lo que genera una asimetría. Por último, MAPE no se puede usar cuando el uso de porcentajes no tiene sentido, como en el caso de series de valores de temperaturas, en las que puede haber valores negativos y nulos.



* sMAPE (Symmetric Mean Absolute Percentage Error): esta métrica surge para evitar la asimetría de MAPE. Hay un debate tanto en qué expresión matemática debería definir este error, como si de verdad es simétrico, tal y como indica su nombre. El sMAPE es la media a lo largo de todas las predicciones hechas hasta un determinado límite. Tiene como ventajas que no tiene los mismos problemas que MAPE cuando los valores de la serie son cercanos a cero.



* **MASE (Mean Absolute Scaled Error):** se calcula tomando el MAE y dividiéndolo por el MAE de una submuestra. Los valores de MASE por encima de 1 indican que las predicciones son peores, en promedio, que en la submuestra escogida. Dado que MASE no depende de escalas, se puede usar para comparar modelos de series temporales diferentes. Sin embargo, hay controversia con esta métrica porque algunos estudios apuntan a que introduce un sesgo por ser sensible a valores atípicos.



**4.7.- Comparación entre modelos**

Comenzamos con las métricas de Holt-Winters que es el modelo que vamos a utilizar como modelo base (baseline) sobre el que comparar con los demás casos.

| **Métrica** | **Holt-Winters** |
| --- | --- |
| **MAE** | 4.36e+6 |
| **MSE** | 2.614362121e+11 |
| **RMSE** | 5.11e+6 |
| **MAPE** | 0.1139483934 |
| **sMAPE** | 10.55824629 |
| **MASE** | 1.793702407 |

Tabla XX: Resumen de métricas del modelo Holt-Winters

Los escenarios de variables exógenas que se han manejado para calcular los modelos en cada caso son los siguientes:

* **A:** sin variables exógenas
* **B:** día de la semana y día laboral/festivo.
* **C:** día de la semana, laboral/festivo y temperatura.
* **D:** día de la semana, laboral/festivo, temperatura y estacionalidad (invierno, verano).

**4.7.1.-SARIMAX**

| **SARIMAX** | | | | |
| --- | --- | --- | --- | --- |
| **Métrica** | **A** | **B** | **C** | **D** |
| **MAE** | 3.2e+6 | 2.37e+6 | 2.49e+6 | 2.5e+6 |
| **MSE** | 1.35e+13 | 1.03e+11 | 1.19e+11 | 1.20e+11 |
| **RMSE** | 3.67e+6 | 3.21e+6 | 3.44e+6 | 3.46e+6 |
| **MAPE** | 0.82 | 0.058 | 0.061 | 0.062 |
| **sMAPE** | 126.33 | 5.94 | 6.24 | 6.28 |
| **MASE** | 13.37 | 0.97 | 1.02 | 1.03 |

Tabla XX: Resumen de métricas del modelo SARIMAX

Podemos ver cómo el escenario A para SARIMAX es pésimo comparado con las métricas de los otros escenarios. Destacan especialmente las métricas MSE, sMAPE y MASE que tienen valores enormes. Esto es comprensible ya que SARIMAX está pensado para tener en cuenta variables exógenas. Todas las métricas, excepto MSE, son mejores que en Holt-Winters.

En cuanto a los escenarios B, C y D podemos ver cómo mejoran las métricas bastante con respecto al escenario A. Podemos ver cómo el escenario B (día de la semana y día laboral/festivo) es el que mejores métricas tiene de los cuatro escenarios, seguido de los escenarios C y D. Por lo tanto, podemos decir que la adición de variables exógenas más allá del día de la semana y festivos, penaliza negativamente las métricas del modelo.

**4.7.2.-XGBoost**

| **XGBoost** | | | | |
| --- | --- | --- | --- | --- |
| **Métrica** | **A** | **B** | **C** | **D** |
| **MAE** | 2.8e+6 | 2.8e+6 | 2.8e+6 | 2.8e+6 |
| **MSE** | 1.12e+11 | 1.12e+11 | 1.13e+11 | 1.13e+11 |
| **RMSE** | 3.35e+6 | 3.35e+6 | 3.37e+6 | 3.37e+6 |
| **MAPE** | 0.071 | 0.071 | 0.072 | 0.072 |
| **sMAPE** | 6.99 | 6.99 | 7.06 | 7.06 |
| **MASE** | 0.61 | 0.61 | 0.61 | 0.61 |

Tabla XX: Resumen de métricas del modelo XGBoost

En todos los escenarios las métricas son mejores que las del modelo base Holt-Winters.

Podemos ver en este caso unas métricas muy parecidas entre los diferentes escenarios, siendo los mejores los escenarios A y B, lo que indica que XGBoost funciona bien con pocas o ninguna variable exógena. Esto es comprensible por lo comentado de cómo funciona XGBoost para encontrar el mejor modelo posible para unos datos dados.

Por último decir, que XGBoost, como en SARIMAX, responde bien a la adición de días de la semana y festivos y laborables en el conjunto de datos. Por otro lado, temperatura y estacionalidad penalizan negativamente, pero a diferencia de SARIMAX, esta penalización es mínima por cómo XGBoost busca el mejor modelo posible.

**4.7.3.-Prophet**

| **PROPHET** | | | | |
| --- | --- | --- | --- | --- |
| **Métrica** | **A** | **B** | **C** | **D** |
| **MAE** | 2.56e+6 | 2.45e+6 | 2.33e+6 | 2.42e+6 |
| **MSE** | 8.93e+10 | 8.28e+10 | 7.62e+10 | 8.24e+10 |
| **RMSE** | 2.99e+6 | 2.88e+6 | 2.76e+6 | 2.87e+6 |
| **MAPE** | 0.064 | 0.061 | 0.056 | 0.06 |
| **sMAPE** | 6.61 | 6.33 | 6.00 | 6.24 |
| **MASE** | 1.05 | 1.01 | 0.96 | 1.00 |

Tabla XX: Resumen de métricas del modelo PROPHET

En este caso, como los anteriores modelos, tenemos mejores métricas que en el caso de Holt-Winters.

Podemos ver cómo la adición de variables exógenas ayuda a mejorar las métricas en todos los escenarios, menos en el D, donde la adición de la estacionalidad lejos de ayudar, empeora ligeramente el modelo. Por lo tanto, para Prophet, a diferencia del resto de modelos, la adición de la temperatura junto con los días de la semana, festivos y laborables, ayuda a la mejora de la precisión del modelo (escenario C). Sin embargo, penaliza la adición de la estacionalidad en todas las métricas (escenario D).

**5.- Conclusiones**

Después de haber comparado cada uno de los modelos en los escenarios planteados, podemos decir que, lejos de lo que se podría pensar a priori de que cuanto mayor número de variables exógenas, mayor precisión de los modelos, parece ser que con identificar las principales variables que tienen influencia sobre nuestra variable de interés es suficiente.

Para el caso de este estudio nuestra variable de interés ha sido la demanda eléctrica en el Estado Español y, aunque se ha demostrado en la visualización de los datos que la temperatura tiene influencia sobre la demanda, en concreto para valores extremos (por debajo de 15oC y por encima de 25oC), podemos decir que no tiene un peso decisivo a la hora de mejorar los modelos a diferencia de los días laborables y festivos.

Esto se debe a la dificultad de determinar cómo influye la temperatura para un mismo día en diferentes partes de España teniendo, por su diversidad geológica, una gran variedad de microclimas en un territorio relativamente pequeño.

Por otro lado, que las variables exógenas que tengan más peso sobre la demanda sean los días festivos y laborables, nos indica que la demanda eléctrica viene dada principalmente por patrones de consumo regulares que marca la industria.

**6.-Pasos a mejorar**

Una vez hemos conseguido un cierto conocimiento de cómo funciona la demanda eléctrica en el caso del Estado Español, cabe preguntarse si el presente estudio podría mejorarse de cara al futuro pudiendo ser en los siguientes ejes:

* **Temperatura:** investigar si el considerar la temperatura media de cada día en cada una de las cinco ciudades consideradas es acertado, teniendo en cuenta que hay épocas del año con variaciones de temperatura bastante amplias, siendo estas variaciones más acusadas según la ciudad.

Ver si conseguir otro tratamiento de lo que llamamos “estacionalidad” etiquetando los meses como invierno o verano podría ayudar que estas etiquetas pudieran ayudar a mejorar las métricas en vez de empeorarlas.

Otra línea en cuanto a la temperatura podría ser el tener en cuenta la temperatura por hora considerando, por tanto, esa misma frecuencia horaria en nuestros datos de demanda. Quizás en este sentido la temperatura cobre más peso como variable exógena que ayude a aumentar la precisión de los modelos en lugar de penalizar, como hemos visto para el caso diario.

Por último, el considerar otra ciudad en el noroeste atlántico, como Coruña, podría ayudar a tener una información más completa sobre la distribución de temperaturas y su influencia en el conjunto de la demanda eléctrica nacional.

* **Población**: podría ser interesante que junto con la temperatura se pudiera tener en cuenta la población de las ciudades en consideración. Que determinada densidad de población se vea afectada por las mismas condiciones climáticas y, por lo tanto, comparta patrones de consumo similares, podría ayudar a una mejora de las métricas. Haría falta investigar qué se ha hecho hasta ahora en este sentido.
* **Días laborables y festivos:** realizar una investigación más granular sobre la demanda energética de los viernes podría ser de ayuda ya que presenta mayor variabilidad que el resto de días laborales. Lo mismo ocurre con los sábados, que tienen un promedio de demanda mayor que los domingos. Por lo tanto, el considerar indicadores (tipo booleano o de otro tipo) que nos indiquen si el día es viernes, sábado o domingo, podría ayudar a mejorar aún más las métricas del modelo.

Por otro lado, sería interesante analizar qué comportamientos de demanda se producen según el festivo nacional caiga en un día de la semana u otro, permitiendo que haya puentes o no, etc.

Por último, un análisis del comportamiento de la demanda comparada entre los días laborables post-festivos y los no post-festivos podría ser interesante para tenerlo en cuenta.

* **Análisis pormenorizado por día:** encontrar los días en los que los contrastes entre los valores reales y las predicciones sean más grandes y analizar el por qué puede estar ocurriendo esto y si puede deberse a factores que no estamos teniendo en cuenta.
* **Empleo de otras técnicas de modelado:** otras técnicas como Redes Neuronales Convolucionales, Redes Neuronales Recurrentes, etc si además de tener interés en determinar variables exógenas y su peso sobre la demanda, queremos encontrar el mejor modelo posible.

**Bibliografía**

<https://docs.aws.amazon.com/es_es/forecast/latest/dg/aws-forecast-recipe-prophet.html>

<https://medium.com/@angel.r.chicote/forecasting-con-prophet-7fb36b25eb4b>

<https://es.acervolima.com/analisis-de-series-de-tiempo-usando-facebook-prophet/>

<https://support.numxl.com/hc/es/articles/214272666>

<https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/>

<https://support.numxl.com/hc/es/articles/215715323-SARIMAX-Definici%C3%B3n-de-un-modelo-SARIMAX>

<https://runebook.dev/es/docs/statsmodels/examples/notebooks/generated/statespace_sarimax_stata>

<https://www.energychisquared.com/post/c%C3%B3mo-conseguir-datos-de-esios-con-su-api/>

<https://api.esios.ree.es/>

<https://www.esios.ree.es/es/pvpc>

<https://github.com/bilbaomakers/PVPC-ESIOS>

<https://ipadbooks.contrataweb.com/2020/11/03/conocer-las-tarifas-de-luz-cada-hora-script-en-python/>

<https://demanda.ree.es/visiona/baleares/baleares/acumulada/2022-4-8>

<https://www.todobi.com/cuadros-de-mando-de-la-demanda-de/>

<https://rpubs.com/AlbertoFuente/408243>

<https://eprints.ucm.es/id/eprint/49444/1/2018-MIGUEL%20CABEZON%20Memoria.pdf>

<https://github.com/azogue/aiopvpc/issues/16>

<https://www.esios.ree.es/es/pagina/codigos-de-red>

<https://www.esios.ree.es/es/pvpc>

<https://github.com/jorgeatgu/apaga-luz>

<https://api.esios.ree.es/>

<https://rpubs.com/AlbertoFuente/408243>

<https://www.ree.es/es/apidatos>

<https://www.energias-renovables.com/panorama/redata-datos-datos-y-mas-datos-del-20190826>

<https://www.kaggle.com/code/apoorvabhide/energy-consumption-time-series-forecasting-in-r/notebook>

<https://www.kaggle.com/code/dinasamir/time-series-forecasting-for-energy-consumption/notebook>

<https://www.kaggle.com/code/nageshsingh/predict-electricity-consumption/notebook>

<https://www.kaggle.com/code/limyunjie1999/time-series-forecasting-on-energy-consumption-data/notebook>

<https://www.kdnuggets.com/2020/01/predict-electricity-consumption-time-series-analysis.html>

<https://towardsdatascience.com/end-to-end-time-series-analysis-and-forecasting-a-trio-of-sarimax-lstm-and-prophet-part-1-306367e57db8>

<https://medium.com/@minhsonle199/end-to-end-time-series-analysis-and-forecasting-a-trio-of-sarimax-lstm-and-prophet-part-2-4ca0046073ab>

<https://quan-possible.github.io/energy-demand-prediction/daily#2>.

<https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0391009>

<https://arxiv.org/pdf/2107.06174.pdf>

<https://towardsdatascience.com/how-to-forecast-sales-with-python-using-sarima-model-ba600992fa7d>

<https://analyticsindiamag.com/complete-guide-to-sarimax-in-python-for-time-series-modeling/>

<https://livebook.manning.com/book/time-series-forecasting-in-python-book/chapter-9/v-10/83>

<https://medium.com/@timonfloriangodt/forecasting-hourly-electricity-consumption-with-arimax-sarimax-and-lstm-part-i-cc652cdd905a>

<https://www.statsmodels.org/devel/examples/notebooks/generated/statespace_sarimax_faq.html>

<https://github.com/rishabh89007/China_Electricity_Consumption/blob/main/XGB_New.ipynb>

<https://12ft.io/proxy?q=https%3A%2F%2Fmedium.com%2Fmlearning-ai%2Fforecast-with-xgboost-a1abf870c1cc>

<https://www.kaggle.com/code/robikscube/tutorial-time-series-forecasting-with-xgboost/notebook>

<https://www.cienciadedatos.net/documentos/py29-forecasting-demanda-energia-electrica-python.html>

<https://oa.upm.es/43165/1/TFG_MARIA_FERNANDEZ_DE_MESA_BUSTELO.pdf>

<https://repository.libertadores.edu.co/bitstream/handle/11371/3097/P%C3%A9rez_Jeider_2020.pdf?sequence=1&isAllowed=y>

<http://openaccess.uoc.edu/webapps/o2/bitstream/10609/52984/6/franespigaTFM0716mem%C3%B2ria.pdf>

<https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html>

<https://12ft.io/proxy?q=https%3A%2F%2Ftowardsdatascience.com%2Ftime-series-forecasting-a-getting-started-guide-c435f9fa2216>

<https://medium.com/@shawanugya12/implementation-of-time-series-forecasting-methods-sarima-sarimax-and-prophet-ff8407b25aaa>

<https://www.cienciadedatos.net/documentos/py29-forecasting-electricity-power-demand-python.html>

<https://www.cienciadedatos.net/documentos/py29-forecasting-electricity-power-demand-python.html>

<https://es.wikipedia.org/wiki/Suavizamiento_exponencial#:~:text=El%20modelo%20Holt%2DWinters%20incorpora,as%C3%AD%20como%20a%20patrones%20estacionales>.

<https://docs.oracle.com/cloud/help/es/pbcs_common/CSPPU/holt-winters_additive.htm>

<https://rpubs.com/nanrosvil/283121>

<https://cienciauanl.uanl.mx/?p=7948#:~:text=MODELO%20HOLT%2DWINTERS&text=El%20m%C3%A9todo%20Holt%2DWinters%20es,de%20pron%C3%B3sticos%20a%20corto%20plazo>.

<https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73>

<http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html>

<https://github.com/CamDavidsonPilon/Python-Numerics/blob/master/TimeSeries/MASE.py>

<https://epftoolbox.readthedocs.io/en/latest/modules/metrics/mase.html>

<https://medium.com/@ashishdce/mean-absolute-scaled-error-mase-in-forecasting-8f3aecc21968>

<https://scikit-learn.org/stable/search.html?q=relative+error>

<https://campus.datacamp.com/courses/sampling-in-python/the-ns-justify-the-means-3?ex=2>

<https://gist.github.com/bshishov/5dc237f59f019b26145648e2124ca1c9>

<https://support.numxl.com/hc/en-us/articles/115001277723-MdRAE-Median-Relative-Absolute-Error>

<https://www.geeksforgeeks.org/how-to-calculate-smape-in-python/>

<https://www.statology.org/smape-python/>

<https://www.indeed.com/career-advice/career-development/smape-formula>