Práctica 2 Análisis de Datos y Visualización

Introducción

En esta práctica, se nos introduce al mundo de las series temporales mediante un *dataset* público proporcionado por la Red Eléctrica de España. Este *dataset* contiene datos relativos a la demanda y generación de electricidad en España. De todas las columnas nos centraremos en "Date" y "Demanda real".

Estos datos, al igual que en la práctica pasada, contienen imperfecciones de las que nos tendremos que hacer cargo con las técnicas aprendidas en clase. De la misma manera, tendremos que extraer conclusiones al respecto de los datos proporcionados.

Datos proporcionados

Como hemos explicado, nos centraremos en la evolución de la demanda real a través del tiempo. El intervalo de tiempo en cuestión va desde el 31 de diciembre de 2021 hasta el 30 de agosto de 2023.

A la hora de representar gráficamente dicha evolución, vemos que hay un hueco en torno a febrero de 2020, del cual hablaremos más tarde, y un salto notable a mediados de 2022. Esto se debe a que en esas fechas se puso un tope al precio de la electricidad, lo cual ocasionó que países vecinos nos comprasen electricidad, y por lo tanto un aumento en la demanda. Este salto es importante, ya que limita la cantidad de datos que podemos usar de cara a una posible predicción. Al tener dos flujos tan claramente diferenciados, no podremos inferir datos de uno de los flujos de cara al otro, como por ejemplo, en el hueco que hemos visto en torno a febrero de 2021.

Transformaciones de los datos

Los pasos que se han seguido en el desarrollo de la práctica han sido los siguientes:

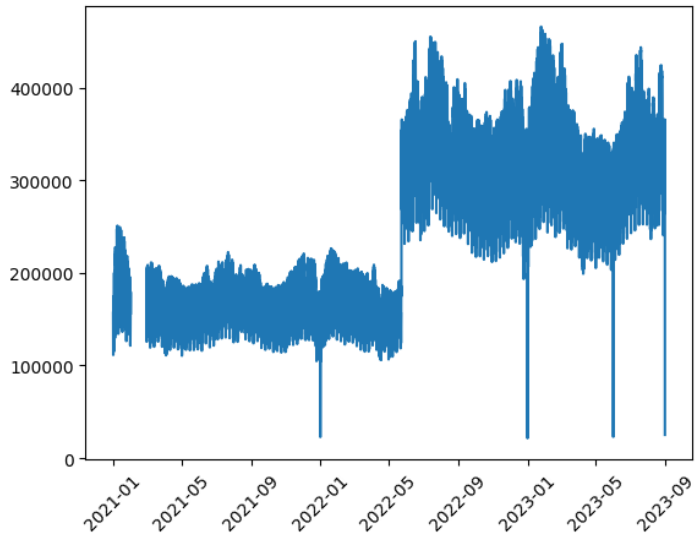
Indexación

En primer lugar, nada más leer los datos, estos han sido indexados por la columna "Date", siendo esta transformada de valor "str" a valor "datetime", de manera que sea más fácil tratarlos como serie temporal. Esto es importante, ya que en ese formato, se puede especificar la variable UTC, de manera que sea posible tratar con los cambios de hora, tanto entre estaciones (horario de verano/invierno) como distintas regiones de España (Islas Canarias).

Análisis cualitativo

A continuación, graficamos los datos de la columna "Demanda real" para ver qué aspecto tienen:

Autor: Martín de las Heras

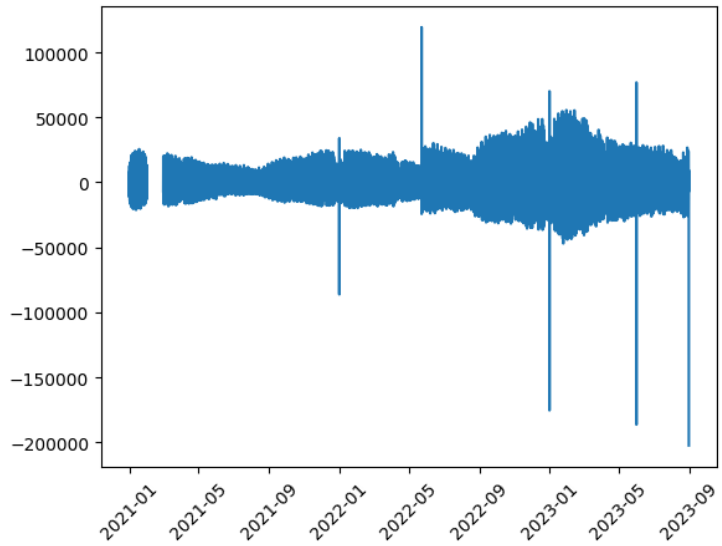


Podemos ver que hay un salto en torno a mayo de 2022, así como distintos valores anormalmente bajos y un hueco alrededor de febrero de 2021.

Podemos observar además la estacionalidad de los datos, la cual es tanto horaria como en función de las estaciones meteorológicas. Por ello, intentaremos trabajar con la granularidad más fina posible, a ser posible con los datos horarios.

Los valores bajos los extraemos y comprobamos que se tratan de fechas señaladas en las que se detiene al producción en las fábricas (como Nochevieja o Semana Santa), por lo que no necesitamos tratarlos.

Sin embargo, sí que nos interesa saber la fecha exacta del salto entre los dos flujos separados. De manera que, extraemos la diferencia entre los valores y los valores de la media de los 3 valores que la rodean, obteniendo la siguiente gráfica:



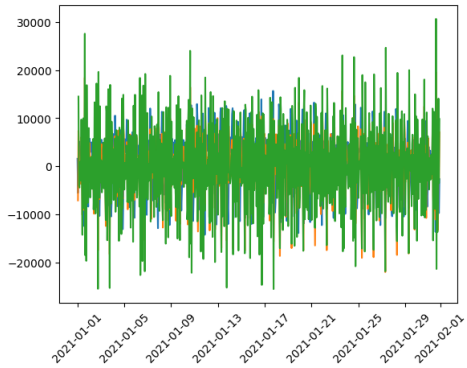
De aquí, vemos que hay valores que destacan por debajo (ya explicados) y un valor que destaca por arriba, es decir, el salto entre los dos flujos. Lo extraemos y vemos que es el valor con la fecha 23-05-2023 19:00:00+00:00.

De este modo, separamos por ese valor las dos series, y a partir de ese momento, trataremos solamente las patologías de la primera.

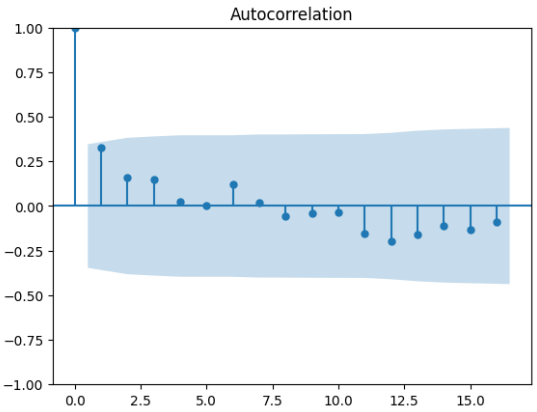
ARIMA

Como primer paso, vemos que los valores nulos corresponden al mes de febrero de 2021, por lo que sacamos las fechas del intervalo vacío, con el objetivo de realizar una predicción en base a una gráfica ARIMA entrenada con los valores previos a la fecha de corte. Para ello, es necesario obtener los tres valores **d**, **p** y **q**:

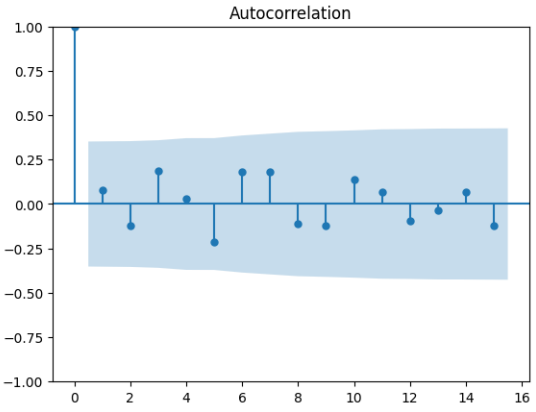
- **d**: para sacar el valor, derivamos los valores de "Demanda real" hasta conseguir una aproximación en torno al cero, como en esta gráfica para la derivada **d=3**, de manera que ya tenemos el valor



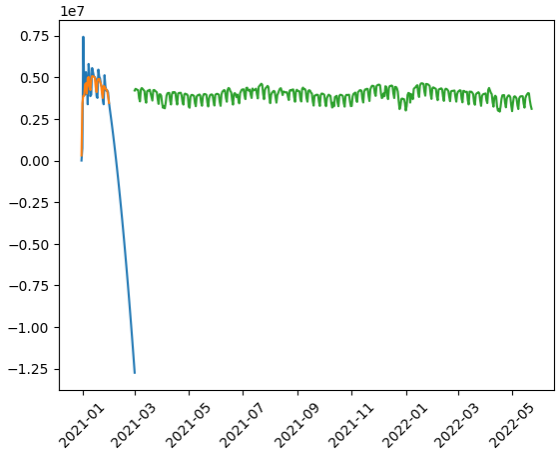
- **p**: para sacar el valor, realizamos el gráfico de autocorrelación y sacamos **p=1**



- **q**: para sacar el valor, realizamos el gráfico de autocorrelación sobre la derivada, obteniendo **q=1**



Tras esto, graficamos la ARIMA y obtenemos la siguiente predicción:



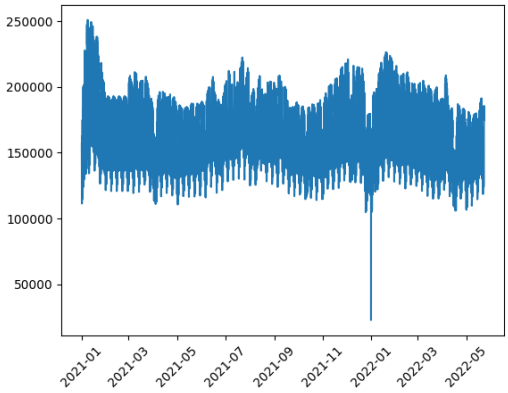
Conclusiones

El motivo principal de esta gráfica es la falta de datos de entrenamiento para la misma. Al no haber muchos datos previos a la fecha de corte (solamente enero de 2021), la predicción que realiza no cuadra con lo que esperamos.

Sin embargo, hemos podido comprobar cómo se hace una ARIMA, así como cómo tratar con series temporales dentro del mundo del análisis de datos.

Extra

Para rellenar los datos que faltan tanto en febrero de 2021, como en el segundo ciclo, como no hemos conseguido unos valores aceptables con la ARIMA, podemos estimarlos en función de los demás días de la semana a lo largo del año. Como no es lo mismo la demanda por la mañana que por la tarde. De esta manera, tomando los datos por hora, sacamos la media en función de esta y del día de la semana y rellenamos los huecos vacíos con las medias. Con este procedimiento, obtenemos la siguiente gráfica para el primer ciclo:



Para el segundo ciclo no se aprecian diferencias visibles, pero el procedimiento es el mismo. Si bien este proceso no nos proporciona datos reales, el hecho de rellenar esos huecos, nos puede servir de cara a una posible predicción a futuro, o un análisis estadístico de la misma, perdiendo el mínimo de información posible.

Autor: Martín de las Heras