

memoria

February 5, 2022

1 Libraries

2 Motivación

El Metro de Madrid fué inaugurado en 1919 por el rey Alfonso XIII, aquella primera “red” de Metro constaba únicamente con ocho paradas, desde la Puerta del Sol hasta Cuatro Caminos. Tuvo tal éxito el nuevo medio de transporte en la ciudad que fué usado por más de 14 millones de usuarios.

Actualmente el Metro de Madrid, es la segunda red de metro mas extensa de la Unión Europea y la cuarta del mundo, consta de 13 líneas con 278 paradas distribuidas por toda la ciudad, creando una gran red de transporte de casi 290km, estableciendo la red de transporte más eficiente de la capital.

| | Terminales | Longitud | Estaciones |
|-------|---|----------|------------|
| Línea | | | |
| 0 | Pinar de Chamartín - Valdecarros | 20,8 km | 31 |
| 1 | Las Rosas - Cuatro Caminos | 14 km | 20 |
| 2 | Villaverde Alto - Moncloa | 16,4 km | 18 |
| 3 | Argüelles - Pinar de Chamartín | 16 km | 23 |
| 4 | Alameda de Osuna - Casa de Campo | 23,2 km | 32 |
| 5 | Circular | 23,5 km | 28 |
| 6 | Hospital de Henares - Pitis | 31,2 km | 29 |
| 7 | Nuevos Ministerios - Aeropuerto T4 | 16,5 km | 8 |
| 8 | Paco de Lucía - Arganda del Rey | 38,0 km | 26 |
| 9 | Hospital Infanta Sofía - Puerta del Sur | 39,9 km | 31 |
| 10 | Plaza Elíptica - La Fortuna | 5,3 km | 7 |
| 11 | MetroSur (Circular) | 40,7 km | 28 |
| 12 | Ópera - Príncipe Pío | 1,1 km | 2 |
| 13 | - | 285,1 km | 278 |

Tabla extraída de wikipedia

En la Comunidad de Madrid (CAM) el transporte público preferido por los Madrileños es el Metro y en los tiempos que nos encontramos (de pandemia), resultaría interesante estimar el volumen de pasajeros que recibirá el Metro en distintos instantes del tiempo.

Apoyados en la temperatura, viento, presión atmosférica y cantidad de rayoUV, vamos a intentar predecir la cantidad de viajeros en el metro de manera mensual (debido a que no he encontrado datos diarios, o incluso por horas).

Tenemos 25 columnas y 396 filas

| fecha | volumenMetro | tmed | prec | racha |
|------------|--------------|-----------|-----------|------------|
| 2000-02-01 | 92.705 | 10.748276 | 5.448276 | 951.093103 |
| 2000-03-01 | 102.479 | 12.080645 | 6.777419 | 944.209677 |
| 2000-04-01 | 83.902 | 10.683333 | 6.273333 | 935.190000 |
| 2000-05-01 | 94.966 | 17.993750 | 12.743750 | 939.990323 |
| 2000-06-01 | 93.300 | 24.047917 | 16.800000 | 943.486667 |

Hemos recogido los datos desde Enero de 2000 hasta Diciembre de 2019 de volumen de pasajeros, datos climatológicos como temperatura, viento y presión medias y sus desviaciones típicas mensuales. Los datos han sido extraídos del banco de datos del ayuntamiento de Madrid y de la AEMET.

Hemos tenido que realizar una imputación de algunos de los datos, pues había datos nulos de presión atmosférica. El método utilizado, al tratarse de una serie temporal, ha sido la interpolación que ofrece la librería pandas.

3 EDA

Lo primero que debemos hacer es un pequeño análisis exploratorio de las variables input y de la variable objetivo. Decimos que una variable objetivo, cuando es la elegida para estimar y las variables input, son las que se basará nuestro modelo para realizar las predicciones. En nuestro conjunto de datos, tenemos diferenciadas tres tipos de variables.

Variable de tiempo(fecha), hasta el año 2020 ya que debido a la pandemia, los datos se han visto alterados de manera muy fuerte y tener un frecuencia mensual, no tendríamos datos suficientes para hacer una buena estimación, es por esto, que se ha optado recurrir a datos hasta el 2020, también a tener en cuenta esta frecuencia mensual, esto se debe a que los datos que disponemos sobre el volumen de pasajeros en el metro de Madrid es mensual. En una puesta en producción los datos podrían ser en streaming, con una actualización de minutos, probablemente gestionado en un entorno spark o con un procesado en batch. Hemos propuesto una pequeña base de datos SQL con tres tablas, que contienen el maestro de fechas, donde determina si es laborable, festivo, fin de semana... Otra tabla con los datos meteorológicos, agrupados por la variable tiempo, aunque en la ciudad de Madrid, por ejemplo existen 3 estaciones disponibles de donde sacar datos, estos datos los facilita la AEMET con su API y contiene un apartado para desarrollo de aplicaciones en streaming. Y por su puesto el volumen de pasajeros, que mediante los tornos ya instalados en las paradas de Metro se podría llevar el conteo casi en tiempo real. Por lo que, aunque la aplicación que propongo está gestionada en mensual, se podría extrapolar a un entorno en streaming o incluso para análisis en batch.

El modelo propuesto es un modelo ARIMA, ya que se ha comprobado en diversos estudios que para variables como la que vamos a estudiar tiene un gran rendimiento, además de su fácil interpretación. Además, al encontrarnos ante una serie temporal con carácter estacionario va el método clásico ARIMA tendrá un gran rendimiento.

3.1 Volumen de pasajeros

Estos datos han sido extraídos de [Banco de datos del ayuntamiento de Madrid](#) el dato viene informado miles de viajeros que están registrados en la agencia de viajeros de la CAM.

```

count      mean      std      min      25%      50%      75%  \
volumenMetro  239.0  88.589019  12.173598  48.479  84.0535  91.535  96.3655

max
volumenMetro  109.412

```

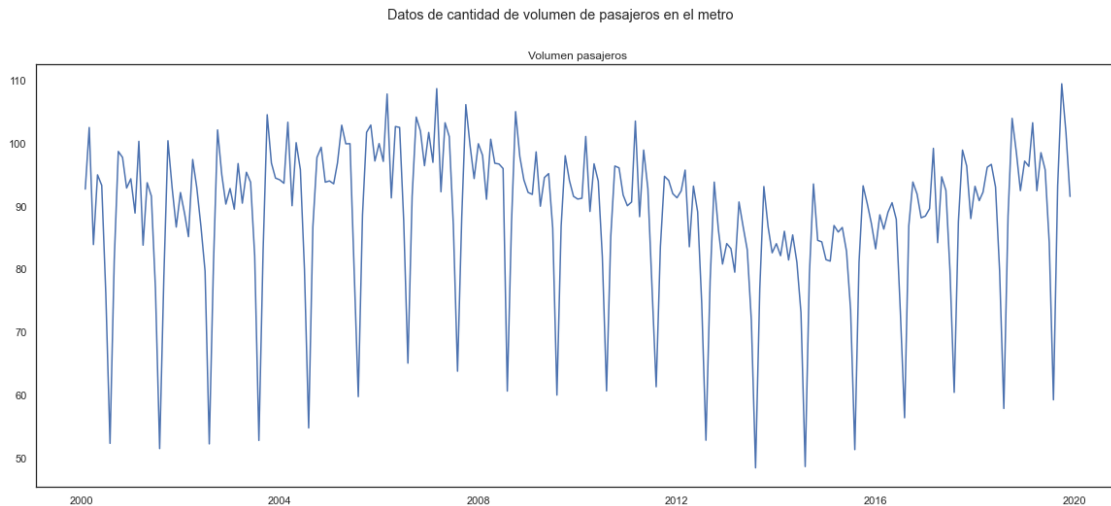
En el gráfico vemos como nos encontramos ante una serie casi en su totalidad estacionaria, hemos tenido que lidiar con una mala extracción de los datos, pues en el año 2010 aparecía una gran caída de pasajeros. La decisión ha sido utilizar la media del año anterior y posterior, es decir, la media entre los datos de 2019 y 2021.

Definición: Decimos que un proceso estocástico es estacionario en el sentido estricto cuando las distribuciones marginales de cualquier conjunto de k variables son idénticas, en distribución y en parámetros.

Para nuestro estudio, nos basta que el proceso sea estacionario en sentido débil, es decir

$$\begin{cases} \mu_t = \mu \quad \forall t \\ \sigma_t^2 = \sigma^2 \quad \forall t \\ Cov(X_t, X_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)] = \gamma_k \quad \forall k \end{cases}$$

Esto quiere decir, que tanto media como varianza permanecen constantes con el tiempo y la covariancia, entre dos variables de la serie depende sólo de su separación en el tiempo.



Vemos como la serie presenta una cierta periodicidad de los datos, pues la forma que toma la serie es similar en todo el tiempo. Presentando cierta tendencia sinusoidal, si queremos aplicar el modelo ARIMA, debemos “eliminar” esta componente estacional, mediante por ejemplo normalizando podríamos obtener un ruido blanco.

Decimos que un proceso estocástico es un ruido blanco si

$$E[X(t)] = 0 \quad V[X(t)] = \sigma^2 \quad \gamma_k = 0$$

El objetivo es ajustar el modelo ARIMA es que el error producido sea un ruido blanco, esto significará que nuestro modelo está bien ajustado.

Actualmente, vemos con un gráfico de cajas y bigotes y un histograma, que la distribución de nuestra serie no es simétrica y esto se debe a que aunque tenemos una periodicidad de los datos muy clara, existe tendencia, provocando que la distribución de la serie no sea simétrica, con una media distante a la mediana y una desviación típica elevada. Por lo que nuestra serie no es estacionaria.

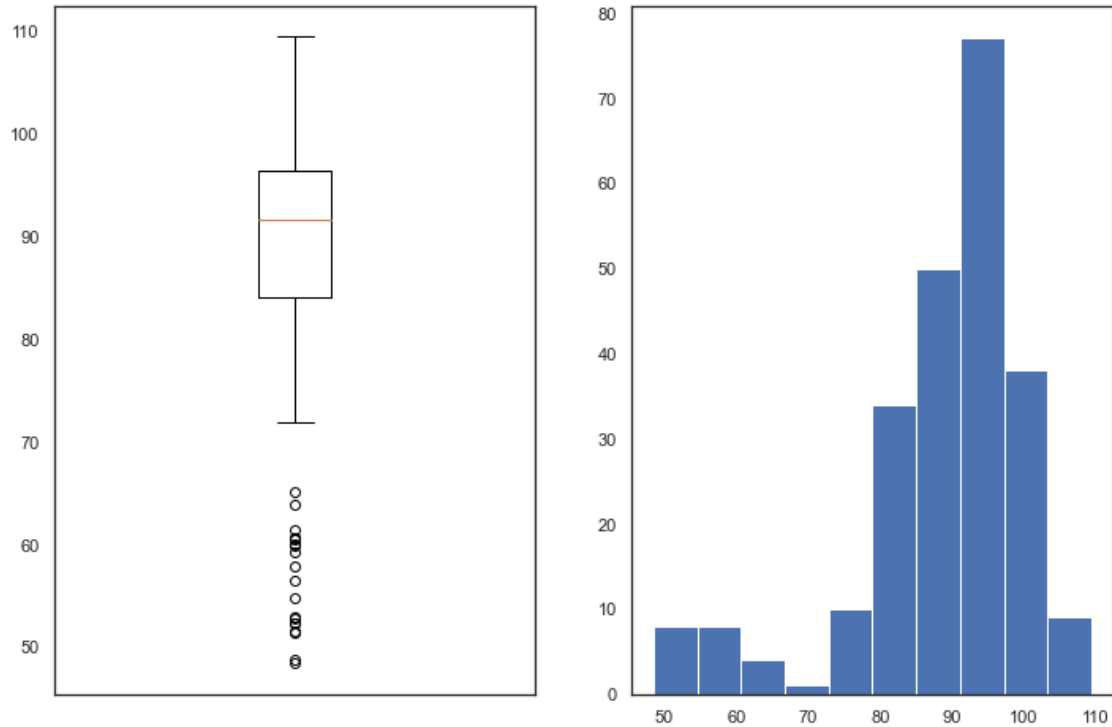
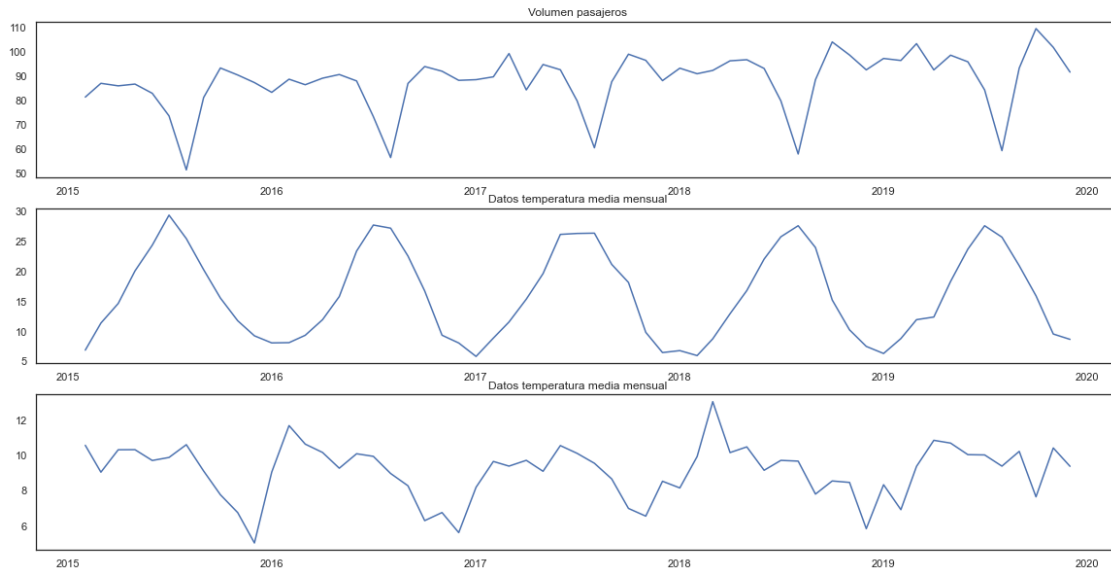


Fig 1.: Gráfico boxplot e histograma del volumen de Pasajeros del metro

3.2 Variables climatológicas

Vamos a exponer la comparación de un par de series de la variable objetivo con alguna exógena de apoyo, aunque el estudio se ha realizado en todas las variables. Vamos a aplicar un filtro de fecha de 5 años para poder comprar mejor.

Series temporales de volumen de pasajeros y temperatura media mensual ampliado a 3 años



Vemos como claramente en las épocas de mas calor, hay menos usuarios de Metro que en los meses más fríos. Parece que la temperatura media puede ayudarnos a estimar. Si nos fijamos en el viento sin embargo, cuando se producen disminuciones en la velocidad del viento, vemos como en general, se producen los aumentos de volumen de pasajeros. Esto es debido a que parece que va con un ligero retraso en su periodicidad, ya que en julio/agosto se produce un aumento de la temperatura, con una disminución del volumen de pasajeros, pero el viento se mantiene constante para estos meses. Es a partir de septiembre cuando se produce esta disminución de la velocidad media del viento y una disminución de la temperatura y el aumento del volumen de pasajeros.

4 Modelo ARIMA

Destacamos los modelos clásicos: - Decimos que un modelo es $AR(p)$ (Autoregresivo de orden p), cuando las autocorrelaciones simples decrecen de manera exponencial y existen p autocorrelaciones distintas de 0. - Decimos que un modelo es $MA(q)$ (Medias móviles), cuando las autocorrelaciones simples decaen y se cortan de forma rápida, sin embargo las autocorrelaciones parciales decrecen exponencialmente. - Decimos que un modelo es $ARMA(p,q)$, cuando comparten las características de ambos modelos.

Definimos modelo ARIMA (Autoregresivo integrado de medias móviles), como el modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Como ya hemos explicado, las series estacionarias son las que tienen media 0, por lo tanto, un proceso no estacionario lo llamaremos proceso integrado si al hacer una diferenciación se obtienen procesos estacionarios.

Decimos que aplicamos una diferenciación de orden k a una serie, cuando teniendo X_t le restamos la observación de k instantes anteriores, es decir, X_{t-k} .

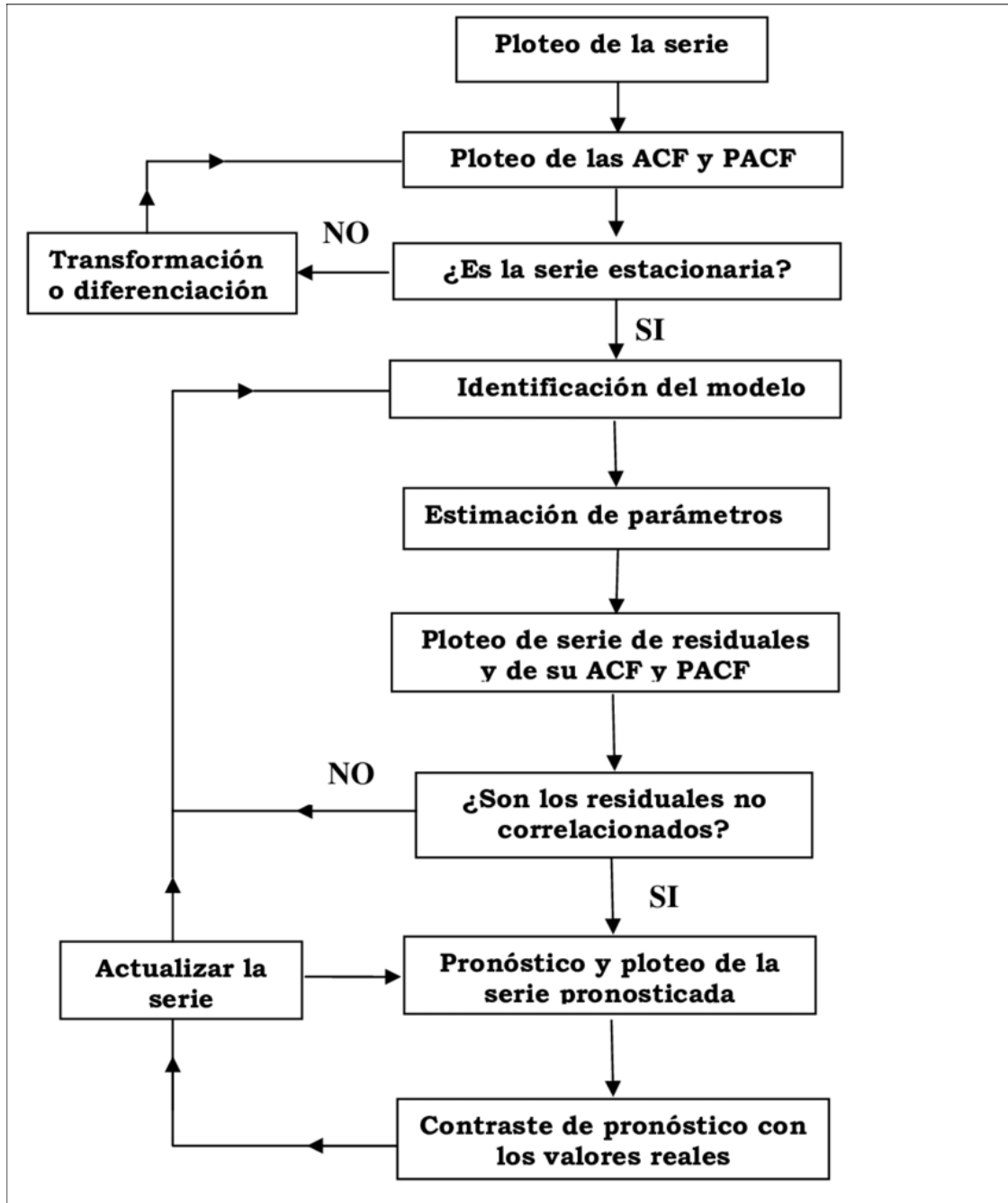
Vamos a hacer una diferenciación de la serie para ver si conseguimos eliminar la media igual a cero

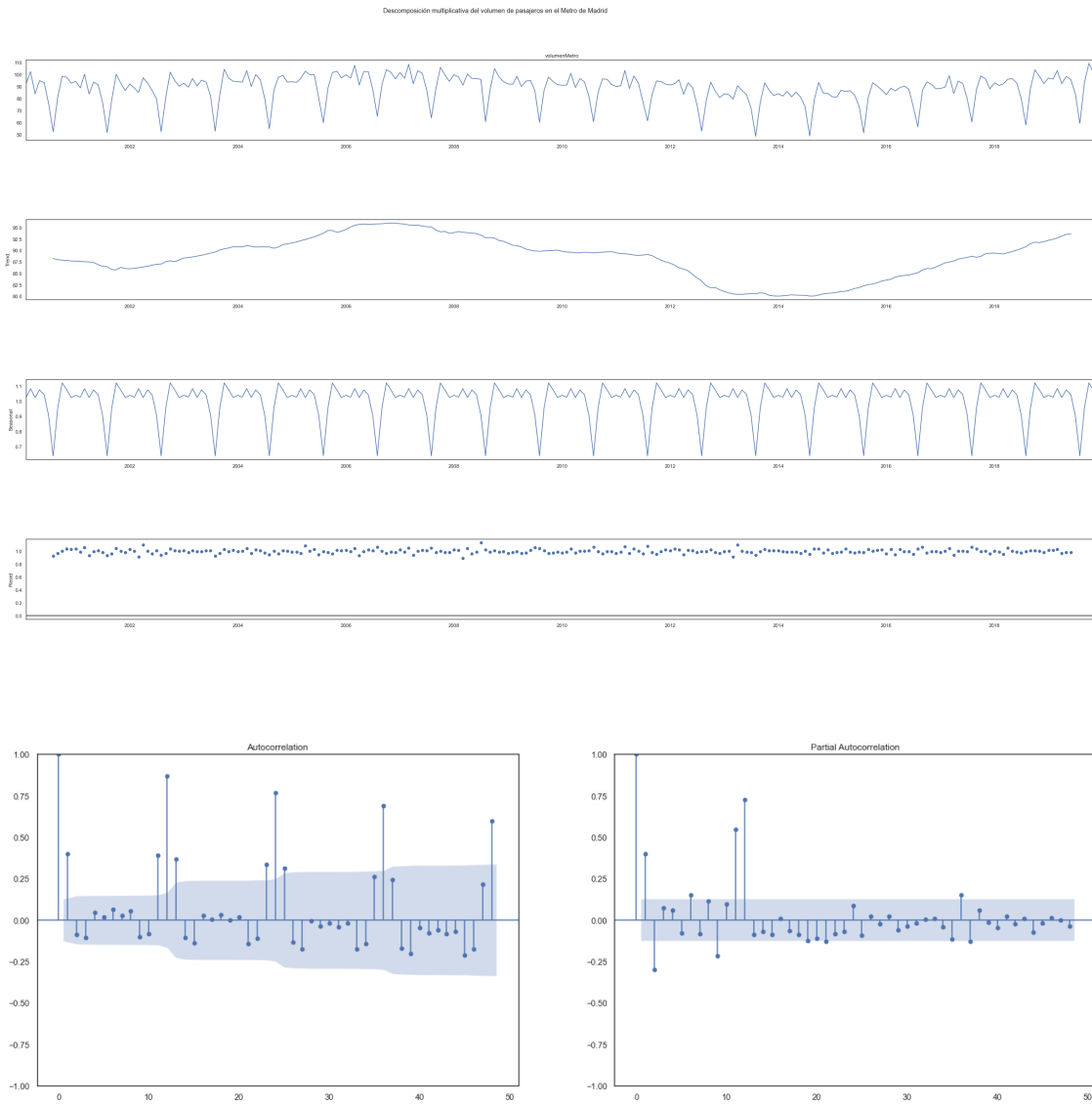
y así poder aplicar un modelo $ARIMA(p, d, q)(P, D, Q)_s$.

Vamos a utilizar la metodología Box-Jenkins para ajustar el modelo lo más fielmente posible, para realizar unas buenas predicciones.

Esquema de la metodología Box-Jenkins

[137]:





5 Bibliografía

<https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IASST-Libro.pdf>

```
[NbConvertApp] Converting notebook memoria.ipynb to pdf
[NbConvertApp] Support files will be in memoria_files\
[NbConvertApp] Making directory .\memoria_files
[NbConvertApp] Making directory .\memoria_files
[NbConvertApp] Making directory .\memoria_files
[NbConvertApp] Making directory .\memoria_files
[NbConvertApp] Making directory .\memoria_files
[NbConvertApp] Writing 31530 bytes to notebook.tex
[NbConvertApp] Building PDF
```

```
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']  
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']  
[NbConvertApp] WARNING | b had problems, most likely because there were no  
citations  
[NbConvertApp] PDF successfully created  
[NbConvertApp] Writing 453458 bytes to memoria.pdf
```