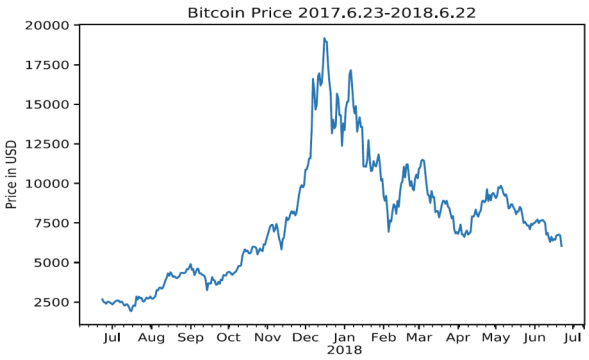
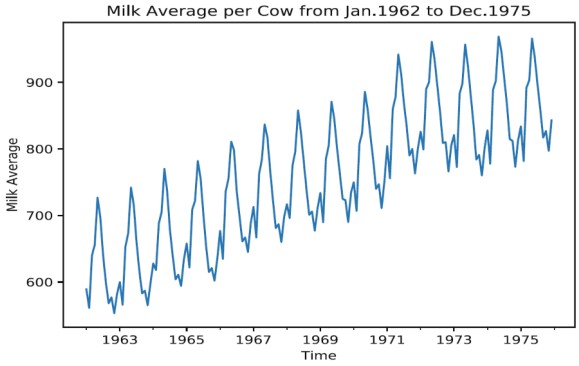
ANÁLISIS DE SERIES TEMPORALES

¿Qué son las series temporales?

Una serie temporal es una secuencia de datos de un fenómeno natural o social, observado a lo largo del tiempo. Se trata de una colección de registros separados por un intervalo de tiempo dado. Por lo tanto está ordenada temporalmente y no debemos intercambiar posiciones entre ningún par de valores de la serie temporal. Algunos ejemplos de esto son: el peso de una persona medido cada domingo; el precio del stock de una compañía al cierre de cada día; etc.





Debemos notar que no podemos obtener el valor de una serie temporal en el tiempo t antes que dicho momento llegue. A su vez, los valores de una serie temporal están afectados por varios factores, tanto deterministas como aleatorios, con lo cual dichos valores son aleatorios. Por lo tanto, las series temporales son referidas como procesos estocásticos y la variable a medir es una variable aleatoria en el sentido probabilístico.

Matemáticamente, se puede definir a una serie temporal como una serie de eventos $\{x_t\}$ de la variable aleatoria X y que están ordenados temporalmente según $t=1,2,\ldots,n$.

Existen varios modelos para describir las series temporales. Algunos de ellos son: ARMA (para series estacionarias), ARIMA (para series no estacionarias), SARIMA (para series con componente estacional), VAR (para series multivariadas), ARCH (útil para series volátiles como en el ámbito financiero), REGARMA, TBATS, aplicación de redes neuronales, etc.

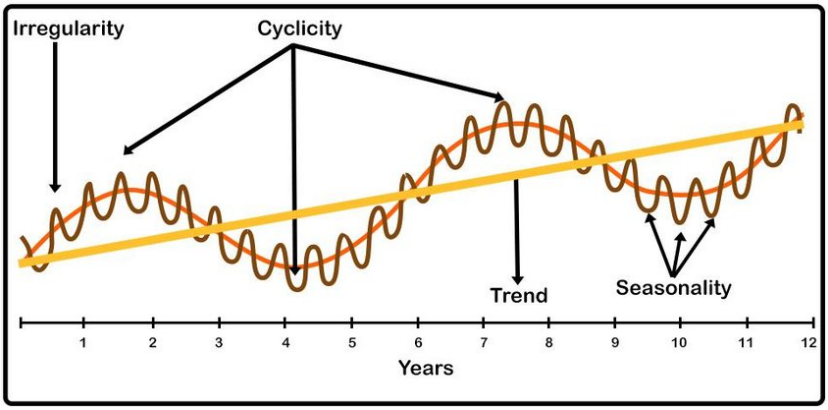
El objetivo de este tipo de análisis puede ser: (1) analizar la serie temporal en sí misma (analizar si tiene tendencia determinista, estacionalidad determinista, cambios dramáticos en su patrón de comportamiento, outliers); y/o (2) predecir comportamientos futuros, lo cual es muy útil en campos de economía, finanzas, sismología, meteorología, geofísica, etc.

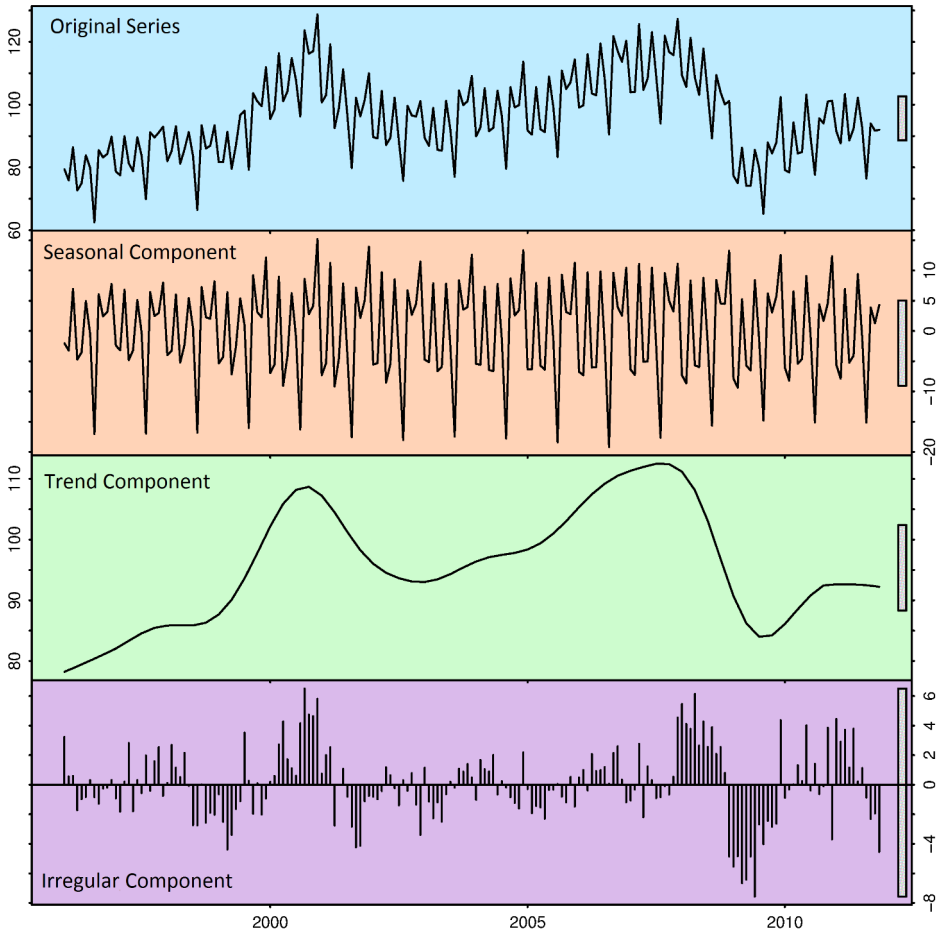
Componentes de una serie temporal

Las series temporales se pueden descomponer como una suma o multiplicación de varios términos o factores. Esto se realiza para comprender mejor la naturaleza de la serie. En la práctica, muchas series temporales reales poseen una componente de estacionalidad determinista (S_t), o una componente de tendencia determinista (T_t). Algunas tienen ambas. Luego de extraer las componentes de tendencia y estacionalidad, el remanente que queda es la componente de variación aleatoria o estocástica (R_t), también conocida como residuo. En algunas bibliografías se extrae otra componente asociada a la ciclicidad (C_t), la cual tiene que ver con variaciones estacionales de periodo irregular, usualmente mayores a un año. Aquí se la incorporará junto con la tendencia (ciclo-tendencia).

Existen dos modelos para descomponer las series temporales: (1) aditivo: $X_t=T_t+S_t+R_t$; (2) multiplicativo: $X_t=T_t S_t R_t$. También pueden darse modelos mixtos. El modelo aditivo se ajusta mejor si la magnitud de la viariación estacional no varía con el tiempo; en caso contrario, el modelo más indicado es el multiplicativo, con el cuidado de que los valores de la serie temporal sean no nulos.

Para extraer las diferentes componentes de una serie temporal se utilizan métodos de alisado como lo son el alisado exponencial, el alisado de Holt-Winters, o el alisado por medio de medias móviles.





Funciones momento

Debido a que la variable de una serie temporal es una variable aleatoria, podemos definir sus funciones momento:

- Media: $\mu_t = E(X_t)$

- Varianza: $\sigma_t^2 = Var(X_t) = E[(X_t - \mu_t)^2]$

- Autocovarianza: $\gamma(s,t) = Cov(X_s, X_t) = E[(X_s - \mu_s)(X_t - \mu_t)]$

- Autocorrelación simple: $\rho(s,t)=Corr(X_s,X_t)=\frac{\gamma(s,t)}{\sigma_s\sigma_t}$

- Autocorrelación parcial: Siendo $\hat{X}_t$ una predicción de X_t dada por $\hat{X}_t=\alpha_1 X_{t-1}+...+\alpha_{k-1}X_{t-k+1}$, si se define el error en la predicción como $\hat{Z_{t-k}}=X_{t-k}-\hat{X_{t-k}}$ entonces la función de autocorrelación parcial se define como $\phi_{kk}=Corr(\hat{Z_{t-k}},\hat{Z}_t)$.

Tanto la autocovarianza como la autocorrelación simple miden la correlación (lineal) entre dos puntos X_s y X_t en la misma serie temporal, pero la última es adimensional y más fácil de usar e interpretar. La autocorrelación parcial mide la correlación que existe entre dos valores separados por k retrasos y que no se explican por medio de los valores intermedios.

A menos que se conozca la naturaleza de la variable aleatoria, es imposible calcular estas funciones momento. Por lo tanto debemos hacerlo desde los propios datos de la serie temporal $\{x_t; 1\le t\le n\}$, de la siguiente manera:

- Media: $\bar{x}=\frac{1}{n}\sum_{t=1}^{n}x_t$

- Autocovarianza: $c_k=\frac{1}{n}\sum_{t=1}^{n-k}(x_{t+k}-\bar{x})(x_t-\bar{x})$

- Varianza: $\sigma^2=c_0$

- Autocorrelación simple: $r_k=\frac{c_k}{c_0}$

- Autocorrelación parcial: Es igual al último coeficiente de un modelo autorregresivo AR(k) (se verá más adelante) y se pude estimar por medio del algoritmo de recursión de Durbin-Levinson:

$$\hat{\phi_{kk}}=\frac{r_k-\sum_{j=1}^{k-1}\hat{\phi_{k-1,j}}r_{k-j}}{1-\sum_{j=1}^{k-1}\hat{\phi_{k-1,j}}r_j},\hat{\phi_{11}}=r_1$$

Otro concepto a tener en cuenta es el de **medias móviles**, las cuales se utilizan para evaluar gráficamente la tendencia de una serie temporal o extraer dicha componente en el proceso de descomposición. Las medias móviles pueden ser centradas o asimétricas y se calculan de la siguiente manera:

$$MM_C=\frac{X_{t-d}+X_{t-d+1}+...+X_{t+d-1}+X_{t+d}}{2d+1}\quad \text{y}\quad MM_A=\frac{X_{t-d+1}+X_{t-d+2}+...+X_t}{d}$$

Se debe tener en cuenta que con este método se pierden d datos y que si la serie tiene una estacionalidad, entonces d debe ser igual al periodo de la serie.

Estacionariedad

Si una serie temporal posee esta propiedad, entonces se encuentra en algún tipo de equilibrio estadístico. Hay dos tipos de estacionariedad: fuerte y débil. En este caso usaremos la débil ya que es más fácil de aplicar. Su definición es la siguiente:

Una serie temporal es débilmente estacionaria si (1) su media es constante y (2) para todo tiempo t , $E(X_t^2)<\infty$ y $\gamma(t,t+k)=\gamma(k)$ es independiente de t para cualquier entero k .

Que una serie temporal sea estacionaria quiere decir que sus valores son independientes del tiempo. Además, una serie temporal estacionaria presenta media y varianza constante, y su función de autocorrelación (en función de k) debe caer abruptamente a cero para $k>0$. Una serie temporal estacionaria también es independiente de efectos de estacionalidad.

Se puede comprobar la estacionariedad de una serie temporal mediante diferentes estrategias: (1) el gráfico de la serie temporal misma; (2) gráfico de medias móviles y desviación estándar; (3) pruebas estadísticas. Dentro de este último se encuentran los tests ADF (Augmented Dickey-Fuller) y KPSS (Kwiatkowski-Phillips-Schmidt-Shin).

Tanto el test ADF como el KPSS son tests de hipótesis. El primero asume como hipótesis nula que la serie es no estacionaria y su hipótesis alternativa es que la serie es estacionaria. Por el contrario, el test KPSS asume como hipótesis nula que la serie es estacionaria y su hipótesis alternativa es que la serie es no estacionaria. Como estos tests son complementarios, es recomendable efectuar ambos para estar seguros del resultado obtenido.

Una serie temporal no estacionaria puede convertirse en estacionaria al aplicarle alguna transformación. Un ejemplo de ella es el cómputo de las diferencias entre sus valores consecutivos. Este método se denomina diferenciación y puede ayudar a estabilizar la media y varianza de la serie temporal y, por lo tanto, a remover o reducir la tendencia y estacionalidad.

Notar que: (1) si la diferencia se computa entre valores consecutivos, se habla de una diferenciación de retraso 1, es decir $Y_t = X_t - X_{t-1}$; (2) si se toman valores separados por k puntos, se habla de diferenciación de retraso k , es decir $Y_t = X_t - X_{t-k}$; (3) si k coincide con el periodo de estacionalidad de la serie temporal, entonces se dice que se computa una diferenciación estacional. Notar además que se pueden realizar múltiples diferenciaciones hasta obtener una serie estacionaria.

Existen otros métodos o transformaciones que permiten hacer estacionaria una serie temporal. Algunos de ellos son: tomar el logaritmo de la serie, tomar la n-ésima raíz de la serie o combinaciones de métodos. No profundizaremos en estas transformaciones ya que la más utilizada es la diferenciación.

Análisis exploratorio de las series temporales

En primera medida se debe realizar una exploración del conjunto de datos para determinar si la serie temporal está completa. El conjunto de instantes temporales $\{t_1, t_2, \ldots, t_n\}$ no debe tener huecos, como así tampoco deben haber datos nulos. En caso de que existan datos nulos, se recurre a la interpolación de los datos como método para completar la serie.

Una vez que la serie está completa, el primer paso en cualquier análisis de series temporales es siempre realizar el gráfico de la misma y examinar cuidadosamente este gráfico, ya que muestra el patrón de evolución de la serie temporal. Este patrón incluye la tendencia y/o estacionalidad de la serie temporal. Se puede incluir en el mismo gráfico a las medias móviles y su desviación estándar, lo cual es un indicador de si la serie es estacionaria o no. Además de realizar los tests ADF y/o KPSS.

Además se deben realizar los correlogramas (gráfico en función de k) de las funciones de autocorrelación simple y parcial. Esto nos ayudará a modelar luego la serie temporal.

Otro paso para entender mejor la naturaleza de la serie temporal y mejorar su modelado, es su descomposición aditiva o multiplicativa en sus componentes de tendencia, estacionalidad y variación aleatoria.

Modelo de medias móviles MA(q)

Dada una serie temporal estacionaria, es posible representarla mediante un modelo de medias móviles de orden q, MA(q), de la siguiente manera:

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

donde ε_t es una función de ruido blanco (distribución normal con media nula), y μ, θ_i son los coeficientes del modelo. Es decir, se quiere hallar al conjunto de coeficientes que haga posible tal descomposición y, por lo tanto, predicción de los valores de la serie temporal.

Modelo autorregresivo AR(p)

Dada una serie temporal estacionaria, es posible representarla mediante un modelo autorregresivo de orden p, AR(p), de la siguiente manera:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + ... + \varphi_p X_{t-p} + \varepsilon_t$$

donde ε_t es una función de ruido blanco con $E(X_s \varepsilon_t) = 0$ si $s < t$, y φ_i son los coeficientes del modelo. Notar que este método se trata esencialmente de una regresión lineal la cual utiliza p retrasos de la serie misma como predictores.

Modelo autorregresivo de medias móviles ARMA(p,q)

Dada una serie temporal estacionaria, es posible representarla mediante un modelo autorregresivo de medias móviles de órdenes p,q, ARMA(p,q), de la siguiente manera:

$$X_t = \varphi_0 + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

donde ε_t es una función de ruido blanco con $E(X_s \varepsilon_t) = 0$ si $s < t$, y φ_i, θ_j son los coeficientes del modelo. El objetivo de este método es hallar estos coeficientes que hagan posible predecir los valores de la serie temporal mediante la ecuación antes mencionada. Dichos coeficientes se encuentran computacionalmente mediante diversos métodos, los cuales no discutiremos: algoritmo de innovaciones, método de los momentos, método de mínimos cuadrados condicionales, método de máxima probabilidad. Notar que los modelos AR(p) y MA(q) son casos especiales de este modelo: AR(p) = ARMA(p,0) y MA(q) = ARMA(0,q).

Para construir un modelo ARMA(p,q), primero debemos comprobar que la serie temporal es estacionaria. En caso de que no lo sea, se construirá un modelo ARIMA(p,d,q), el cual se discutirá a continuación. Luego, debemos graficar las funciones de autocorrelación simple (fas) y pacial (fap) para determinar los órdenes p y q de nuestro modelo. Estas funciones dan una idea de los posibles órdenes p y q de nuestro modelo. Se puede considerar que p=k donde la fap se hace (casi) nula; y q=k donde la fas se hace (casi) nula.

En caso de no poder determinar los ordenes p y q de las funciones momento, se debe buscar el modelo o los ordenes que minimicen alguno de los siguientes criterios estadísticos (los cuales no profundizaremos): AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) o HQIC (Hannan-Quinn Information Criterion). Estos valores dan cuenta del error cometido al usar el modelo elegido ARMA(p,q) para estimar los valores de nuestra serie temporal. Elegido uno de los tres criterios, se buscarán los órdenes p y q que minimicen dicho criterio.

Modelo autorregresivo de medias móviles integrado ARIMA(p,d,q)

En caso de que la serie sea no estacionaria pero sin componente estacional y, mediante d diferenciaciones (de retraso 1) se consiga hacerla estacionaria, entonces el modelo a construir es un modelo ARIMA(p,d,q). Éste es análogo al modelo ARMA(p,q), pero aplicado a la serie temporal diferenciada d veces para hacerla estacionaria. Por lo tanto, una vez determinada la cantidad de diferenciaciones necesarias para lograr la estacionariedad, el proceso para construir este modelo es totalmente análogo al del modelo ARMA(p,q). Lo único a tener en cuenta es que no es recomendable comparar los criterios estadísticos (AIC, BIC, HQIC) de modelos con diferente orden d.

Modelo autorregresivo de medias móviles integrado estacional SARIMA(p,d,q)(P,D,Q)s

En caso de que la serie temporal contenga una componente estacional de periodo s y, mediante d diferenciaciones (de retraso 1) y D diferenciaciones estacionales (de retraso s) se consiga hacerla estacionaria, entonces el modelo a construir es un modelo SARIMA(p,d,q)(P,D,Q)s. Este modelo es equivalente a realizar una descomposición multiplicativa de las componentes estacional y no estacional de

la serie y, a tales componentes, aplicarles un modelo ARIMA a cada una. Es decir SARIMA(p,d,q)(P,D,Q)s = ARIMA(p,d,q)ARIMA(P,D,Q), donde este último factor da cuenta de la componente estacional.

El proceso para construir este tipo de modelos es análogo a los modelos ARIMA, sólo que se debe determinar los parámetros s, d y D como primer paso. Luego, con la ayuda de las funciones fas y fap y/o los criterios estadísticos, se deben determinar los órdenes p,q,P,Q del modelo.

TBATS

TBATS es un modelo un poco diferente de lo que venimos discutiendo, cuyo objetivo es modelar series temporales con patrones complejos de estacionalidad (por ejemplo, que la serie tenga dos componentes estacionales con diferente periodo) utilizando alisado exponencial. El nombre de este modelo refiere a las siguientes estrategias matemáticas:

- **T**rigonometric seasonality: Cada componente estacional de la serie temporal es modelada por medio de series de Fourier.
- **B**ox-Cox transformation: Es una transformación que se utiliza para hacer que la serie temporal sea estacionaria, convirtiendo su distribución de valores en una distribución normal.
- **A**RMA errors: Los residuos de la descomposición se modelan por medio de un modelo ARMA(p,q).
- **T**rend components: Si la serie temporal presenta una componente de tendencia, esta puede ser con o sin amortiguamiento.
- **S**easonal components: Puede haber una, varias o ninguna componente estacional.

La estrategia es buscar el modelo con la combinación de todas o algunas de las componentes anteriores y que mejor se ajuste a la serie temporal a analizar. El algoritmo TBATS prueba y evalúa diferentes modelos: con y sin transformación Box-Cox; con y sin tendencia; con y sin amortiguamiento; con errores modelados o no mediante un modelo ARMA; sin componente estacional o con diferentes armónicos que la modelen. Finalmente elige el mejor modelo utilizando el criterio estadístico AIC.

Model:

$$y_t^{(\lambda)} = l_{t-1} + \phi\, b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = \phi\, b_{t-1} + \beta\, d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t$$

Seasonal part:

$$s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos(\omega_i) + s_{j,t-1}^{*(i)} \sin(\omega_i) + \gamma_1^{(i)} d_t$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin(\omega_i) + s_{j,t-1}^{*(i)} \cos(\omega_i) + \gamma_2^{(i)} d_t$$

$$\omega_i = 2\pi j / m_i$$

Where:

$y_t^{(\lambda)}$ - time series at moment t (Box-Cox transformed)

$s_t^{(i)}$ - ith seasonal component

l_t - local level

b_t - trend with damping

d_t - ARMA(p,q) process for residuals

e_t - Gaussian white noise

Model parameters:

T - Amount of seasonalities

m_i - Length of ith seasonal period

k_i - Amount of harmonics for ith seasonal period

λ - Box-Cox transformation

α, β - Smoothing

ϕ - Trend damping

φ_i, θ_i - ARMA(p, q) coefficients

$\gamma_1^{(i)}, \gamma_2^{(i)}$ - Seasonal smoothing (two for each period)

Predicción y evaluación

Una vez seleccionado nuestro modelo, podemos pasar a la etapa de predicción de nuestra serie temporal. Para ello, dividiremos a nuestro conjunto de datos en entrenamiento y testeo, pero no de forma aleatoria como se hace para otros algoritmos de machine learning, sino que lo haremos particionando a la serie temporal en dos "sub-series", a partir de cierto instante de tiempo. Los primeros valores serán el conjunto de entrenamiento, y los valores restantes serán el conjunto de prueba.

A partir del conjunto de entrenamiento se entrenará al modelo elegido y se podrá realizar la predicción de los valores de la serie temporal. Las predicciones hechas para los valores pertenecientes al conjunto de entrenamiento se denominan predicción sobre muestra (on-sample prediction) y sirven para evidenciar qué tan bien "aprende" nuestro modelo del conjunto de entrenamiento dado; en otras palabras, muestra si el modelo se ajusta bien a nuestros datos.

Por otro lado se pueden predecir los valores correspondientes al conjunto de prueba. Esto mostrará la capacidad de nuestro modelo para predecir a futuro, sin contar con información.

El gráfico de la serie temporal, la predicción y los errores en la predicción sirven para dar cuenta de qué tan bueno es nuestro modelo al predecir dichos valores. Los errores no deberían presentar ninguna tendencia y contar con baja desviación estándar. Es decir, los errores en la predicción deberían tener una distribución normal o de ruido blanco.

BIBLIOGRAFÍA

Huang C., Petukhina, A. *Applied Time Series Analysis and Forecasting with Python.* Springer, 2022.

https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python/notebook

https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

https://medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model-ce8c429442a9

Alysha M. De Livera, Rob J. Hyndman and Ralph D. Snyder (2011): *Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing*, Journal of the American Statistical Association, 106:496, 1513-1527. http://dx.doi.org/10.1198/jasa.2011.tm09771