

Análisis de Series de Tiempo

Carrera de Especialización en Inteligencia Artificial

Clase 1

Ing. Magdalena Bouza, Esp. Ing. Carlos German Carreño Romano

Acerca del curso

- Modelos clásicos
- Práctica matemática
- Análisis usando Python
- Aplicaciones con redes neuronales
- Aplicaciones en temas de interés de los alumnos
- Repo: www.github.com/charlieromano/timeseries

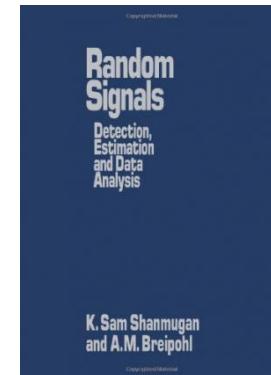
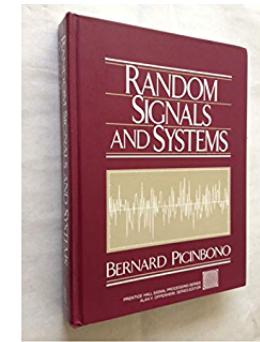
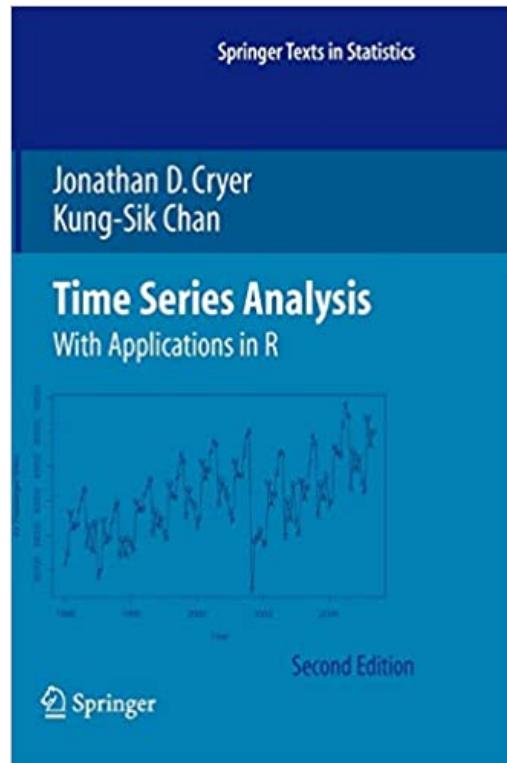
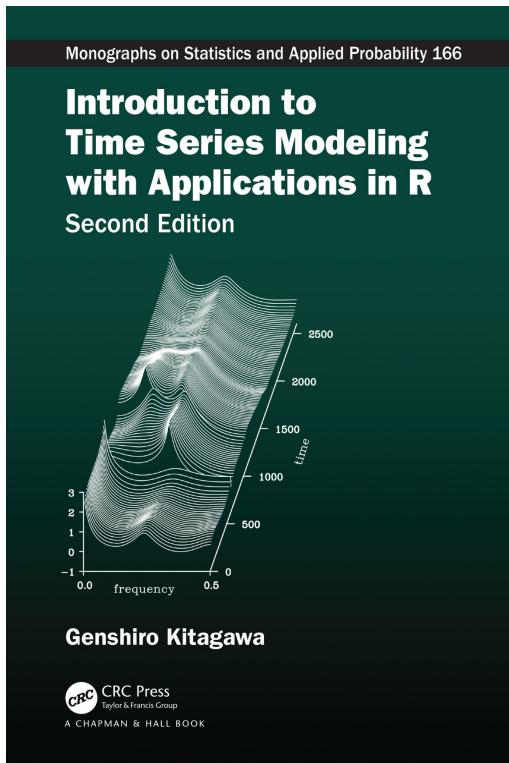
Cronograma

Clase 1	Introducción a series de tiempo, nociones básicas de modelado, técnicas de preprocesamiento.
Clase 2	Tendencia determinística, presentación de TP, modelos estacionarios AR, MA, ARMA.
Clase 3	Estacionariedad, Estacionalidad, Modelos ARIMA y SARIMA, Criterios de bondad de modelos. Primer entrega TP y consultas.
Clase 4	Caso de estudio, tendencia estocástica, estacionalidad, modelos SARIMA.
Clase 5	Predicciones, tratamiento de intervenciones y outliers. Segunda entrega. Consultas.
Clase 6	Aplicaciones. Análisis espectral, Heteroscedasticidad. Frameworks.
Clase 7	Presentaciones trabajo final.

Criterio de aprobación

- Primera entrega parcial TP Final (Clase 3).
- Segunda entrega parcial TP Final (Clase 5).
- Presentación final (Clase 7).
- Las entregas son sincrónicas.

Bibliografía



Repo

<https://github.com/charlieromano/TimeSeries>



- Datasets
- Docs
- Pics
- Scripts
- README

charlieromano / TimeSeries (Public)

Code Issues Pull requests Actions Projects Wiki Security Insights

main · 1 branch · 0 tags Go to file Code

charlieromano practica 01 6ba3185 39 seconds ago 14 commits

File	Commit Message	Time Ago
Datasets	practica 01	39 seconds ago
Docs	update README	3 months ago
Pics	practica 01	39 seconds ago
Scripts	practica 01	39 seconds ago
LICENSE	Initial commit	4 months ago
README.md	update README	3 months ago
STEM-RNN.Rmd	agrego datasets para actividad 1	2 hours ago

About

Este es el repo de la materia de Análisis de Series de Tiempo

Readme

BSD-2-Clause License

Releases

No releases published

Packages

No packages published

Languages

Python 100.0%

README.md

TimeSeries

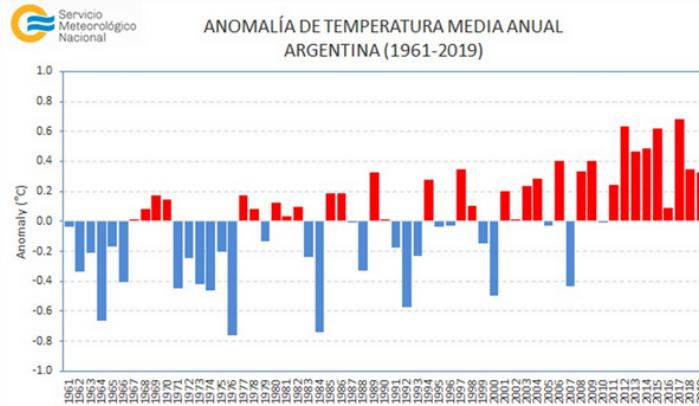
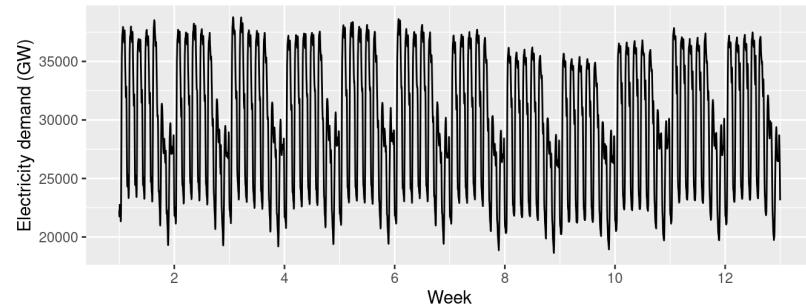
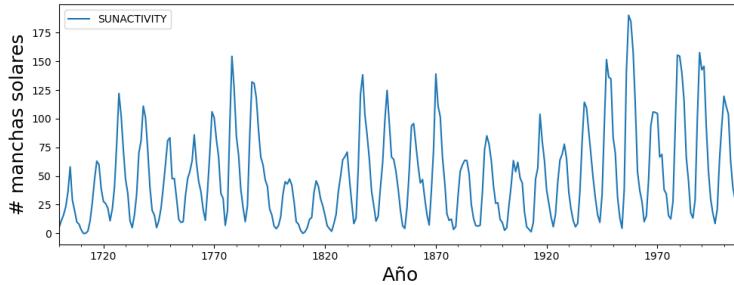
Este es el repo de la materia de Análisis de Series de Tiempo



Introducción

¿Qué es una serie de tiempo?

Un registro de un fenómeno que varía en el tiempo de forma irregular es una serie de tiempo.



¿Para qué estudiar Series de Tiempo?

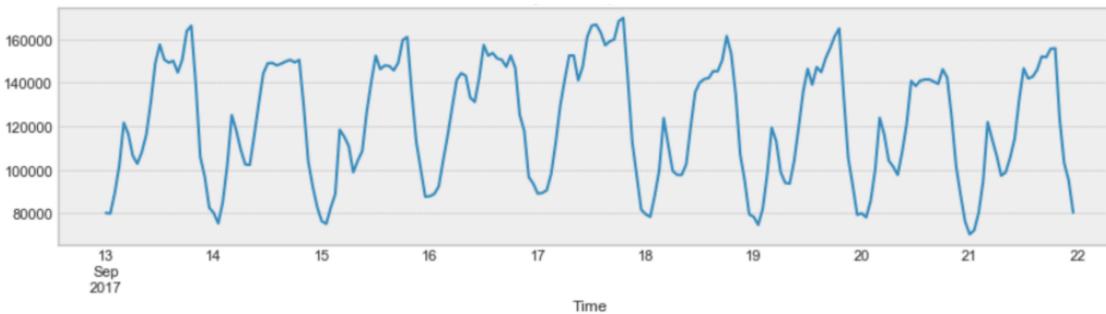
En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA. Para hacer esto solemos seguir un pipeline:



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

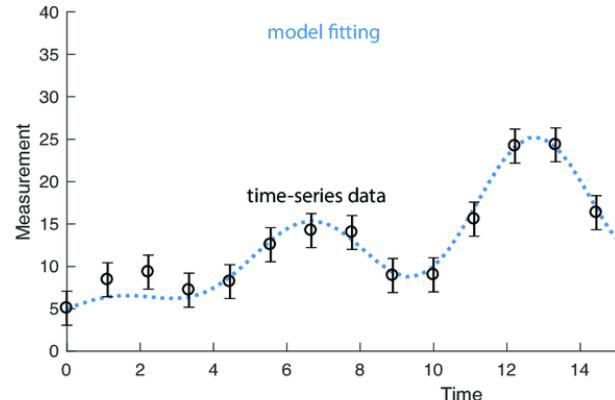
- **1: Descripción:** se desea resumir las características de una serie de tiempo.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

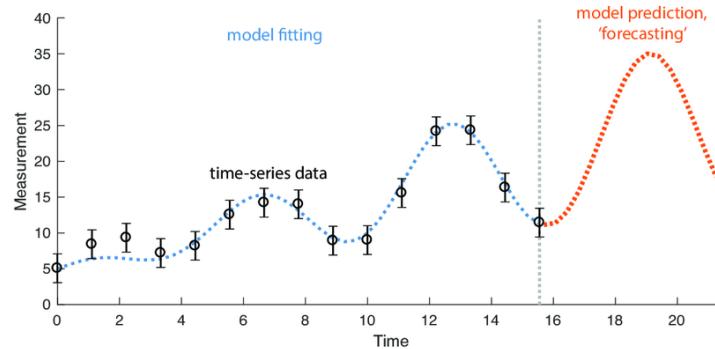
- **2: Modelado:** se busca identificar un modelo aproximado del problema. Es necesario identificar un modelo correcto, así como los parámetros asociados al mismo.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

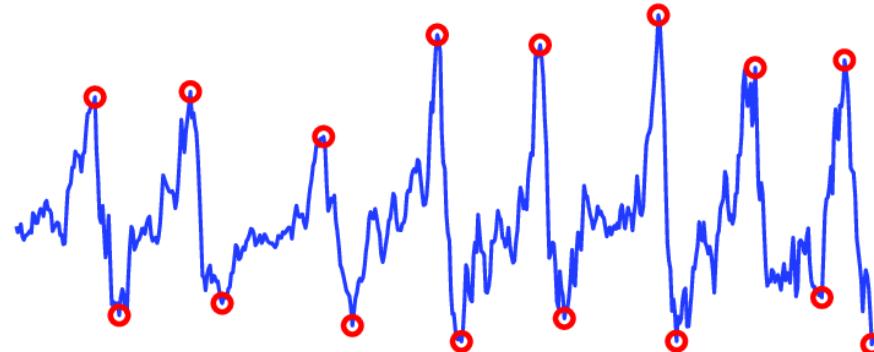
- **3: Predicción:** el objetivo es conocer, o predecir, el comportamiento futuro del sistema.



¿Para qué estudiar Series de Tiempo?

En general no podemos asumir que la **observación** surge independientemente de una población común. Estudiar modelos que incorporen **dependencia** es la clave en TSA.

- 4: **Extracción de señales**: de todo el modelo, se busca extraer ciertas señales que resultan de interés para el problema en cuestión

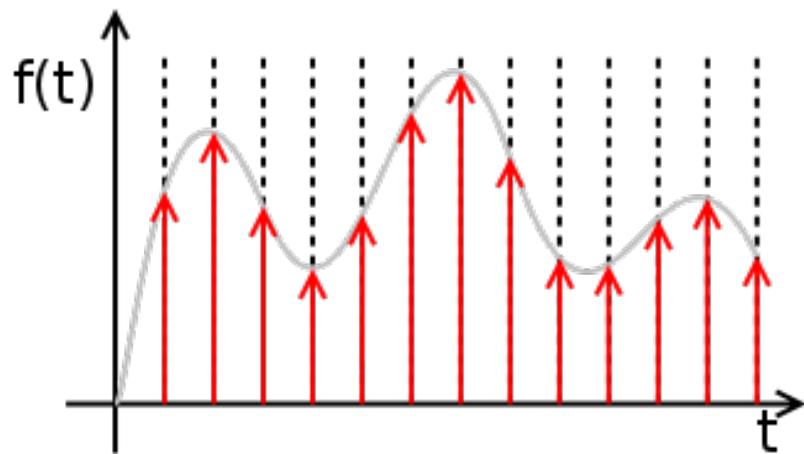


Cómo clasificar Series de Tiempo

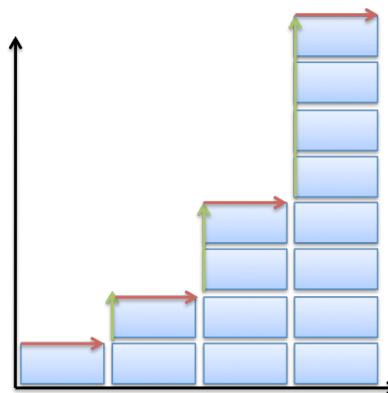
- Continuas o discretas
- Univariadas o multivariadas
- Estacionarias o no estacionarias (Stationary)
- Estacionales o no estacionales (Seasonal)
- Observaciones faltantes y outliers

Cómo clasificar Series de Tiempo

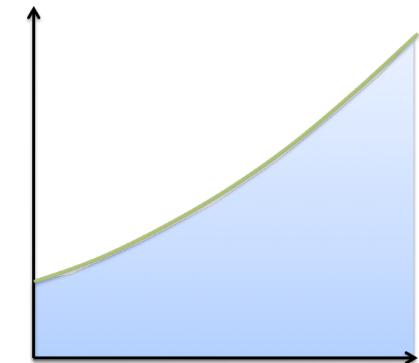
- **Continuas o discretas:** En general se consideran series de tiempo discretas, ya que los datos suelen medirse en intervalos de tiempo



Discrete Growth (2^n)

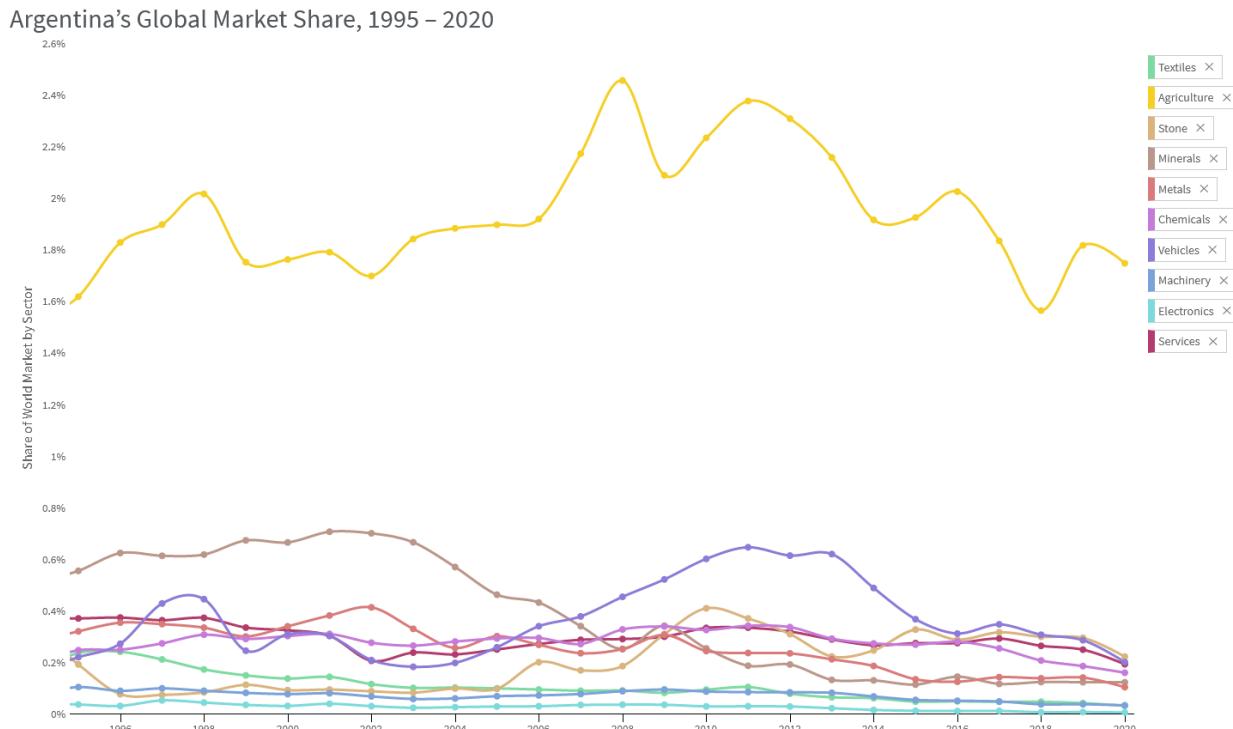


Continuous Growth (e^x)



Cómo clasificar Series de Tiempo

- Univariadas vs. multivariadas



Ref: <https://atlas.cid.harvard.edu/explore>

Dashboard: <https://atlas.cid.harvard.edu/explore/>

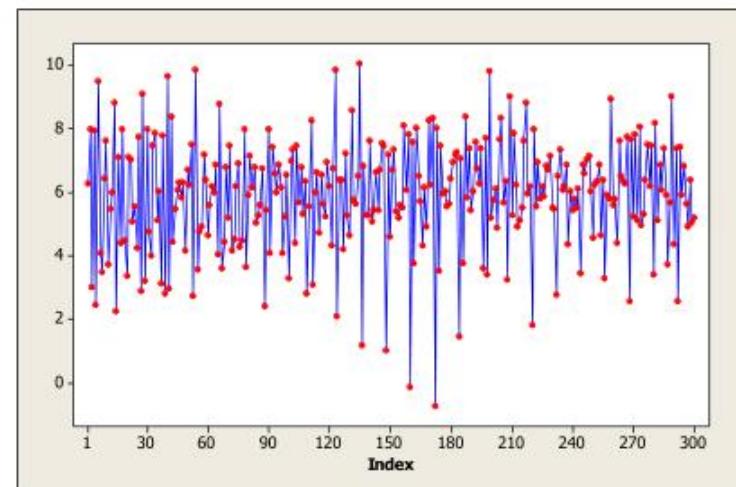
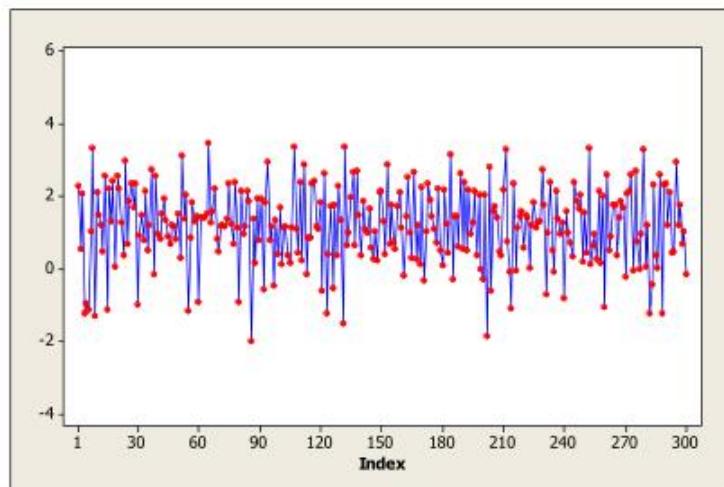
market?

country=8&product=undefined&year=2020&queryLevel=location&productClass=HS&target=Product&partner=undefined&startYear=1995

Cómo clasificar Series de Tiempo

- **Estacionarias (no) estacionarias**

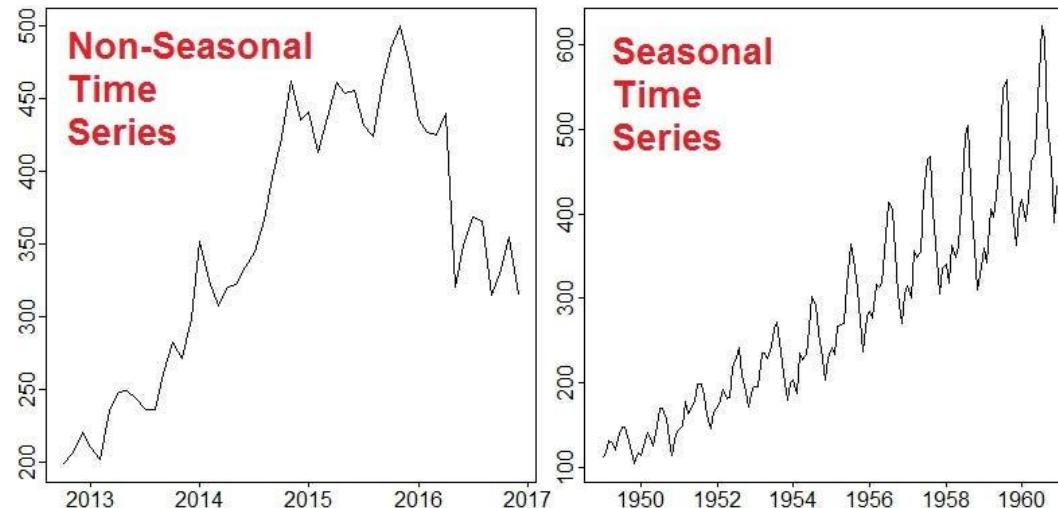
La serie de tiempo estudiada proviene de realizaciones de un proceso estocástico con una estructura invariante (media y desvio) en el tiempo se lo llama estacionario.



Cómo clasificar Series de Tiempo

- **Estacionales vs no estacionales**

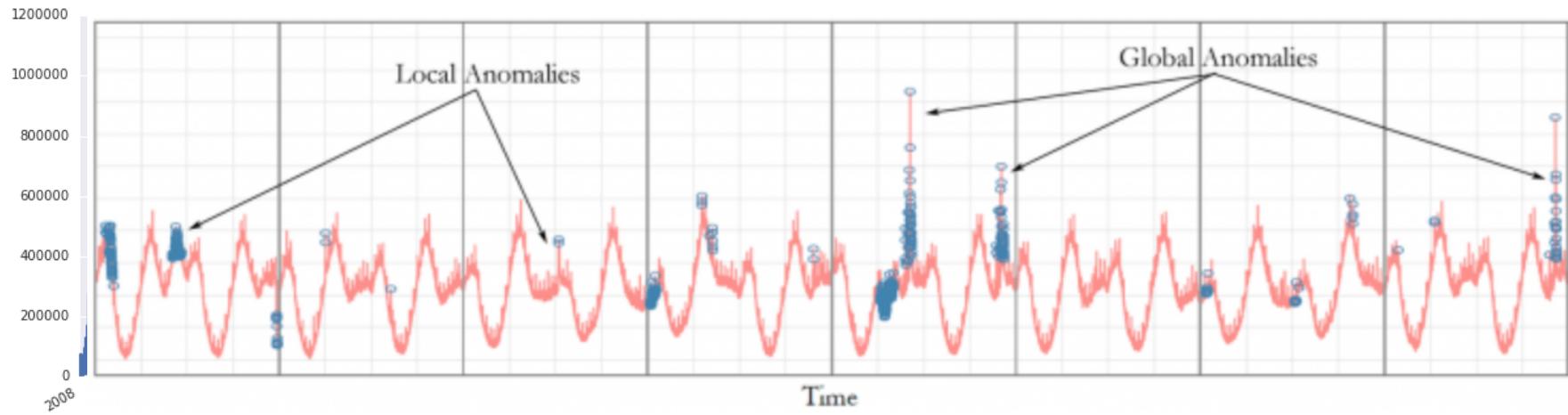
La serie presenta ciclos que tienen una frecuencia determinada, por ejemplo años o estaciones anuales (invierno, primavera, verano, otoño), mientras que la ausencia de ciclos a lo largo de toda la serie puede ser indicio de no estacionalidad.



Cómo clasificar Series de Tiempo

- **Observaciones faltantes y outliers**

Es muy común en series de tiempo la aparición de valores fuera del rango esperado, o de anomalías locales o globales. Una aplicación general con técnicas de análisis de datos es la detección de anomalías.



Cómo clasificar Series de Tiempo

Hasta acá hemos visto cómo estudiar y clasificar series de tiempo. La clasificación es parte del primer paso del pipeline de análisis y modelado que corresponde a la descripción.

Describir una serie de tiempo tiene un grado muy fuerte de dependencia con la naturaleza de la serie. Siempre que sea posible debemos buscar y comprender las causas naturales de los movimientos de las series de tiempo para poder describirlas, y no quedarnos simplemente en una descripción técnica. En algunas disciplinas esto se diferencia entre análisis técnico y análisis fundamental.

Actividad

1. Buscar un dataset de interes para el TP final.
2. Graficar una serie de tiempo de los ejemplos del repositorio:
3. Redactar un informe describiendo tres series de tiempo distintas (máx. 3 páginas)

Series de tiempo y Procesos estocásticos

Procesos estocásticos

Las series de tiempo forman parte de lo que se conoce como procesos estocásticos.

Así como las variables aleatorias mapean los posibles resultados de un experimento aleatorio a un número (real), los procesos estocásticos mapean los resultados de un experimento aleatorio a un conjunto de funciones en el tiempo.

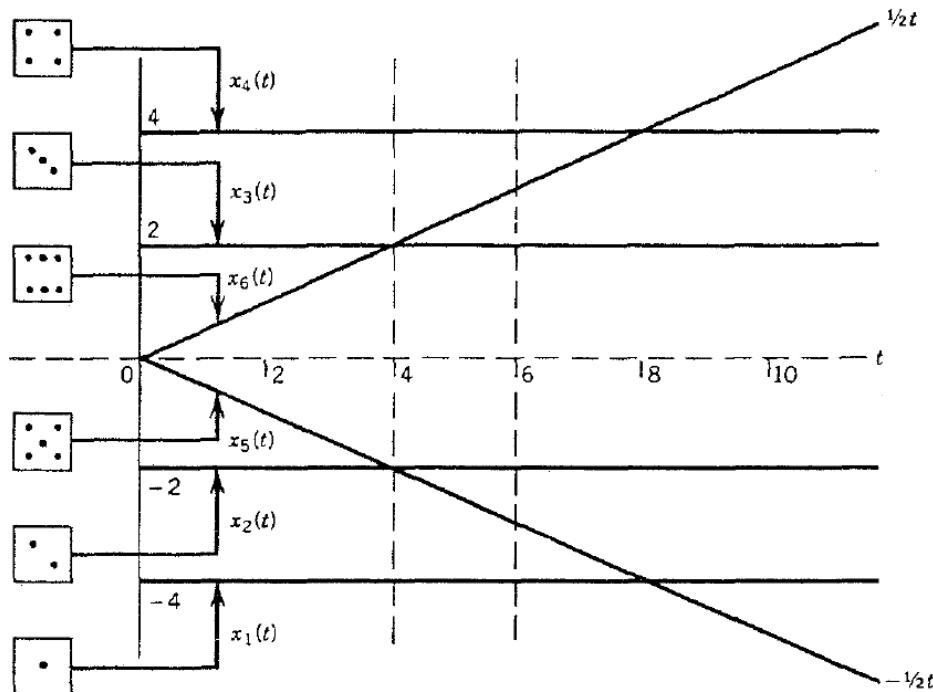


Figure 3.2 Example of a random process.

Procesos estocásticos

- Ensamble de funciones de tiempo

$$X(t, \Omega) = \{X(t, \omega_i) | \omega_i \in \Omega\} = \{x_1(t), x_2(t), \dots\}$$

- Función de tiempo específica

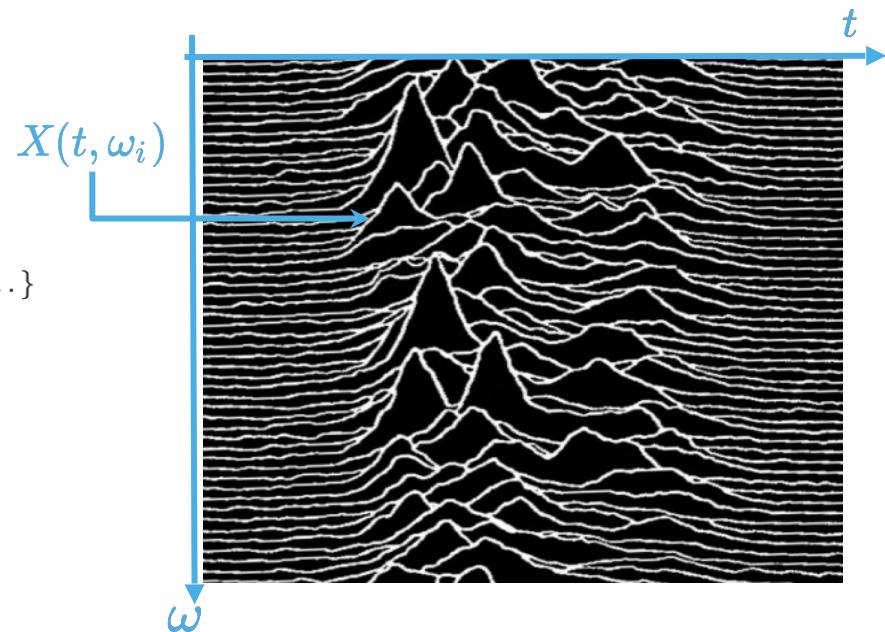
$$X(t, \omega_i) = x_i(t)$$

- Variable aleatoria

$$X(t_0, \Omega) = \{X(t, \omega_i) | \omega_i \in \Omega\} = \{x_1(t_0), x_2(t_0), \dots\}$$

- Valor numérico de una función en un tiempo dado

$$X(t_0, \omega_i) = x_i(t_0)$$



Una serie de tiempo es una sucesión de variables aleatorias correlacionadas entre sí a lo largo del tiempo.

Estrategia de modelado

Hallar modelos para las series de tiempo no es algo trivial. En general el proceso consiste de 3 pasos:

- **Especificación** (o identificación) del modelo: se seleccionan algunos modelos que podrían ser apropiados para modelar la serie que estamos estudiando
- **Ajuste del modelo**: dado un modelo propuesto, se busca estimar los mejores parámetros de ese modelo
- **Diagnóstico** del modelo: analizar mediante diversos tests la bondad o calidad del modelo

En este curso nos vamos a encargar de presentar distintos modelos y estudiar cómo ejecutar cada uno de estos pasos.

Análisis de Series de Tiempo

Box and Jenkins plantean tres etapas de **modelado**:

1. Especificaciones

- a. Gráficos
- b. Estadísticas
- c. Contexto

2. Ajuste

- a. Parámetros
- b. Valores

3. Diagnóstico

- a. Testing

“everything should be made as simple as possible but no simpler”

Algunos modelos

Algunos Modelos

1. Ruido blanco (white noise)
2. Caminante aleatorio (Random walk)
3. Coseno aleatorio (Random Cosine wave)

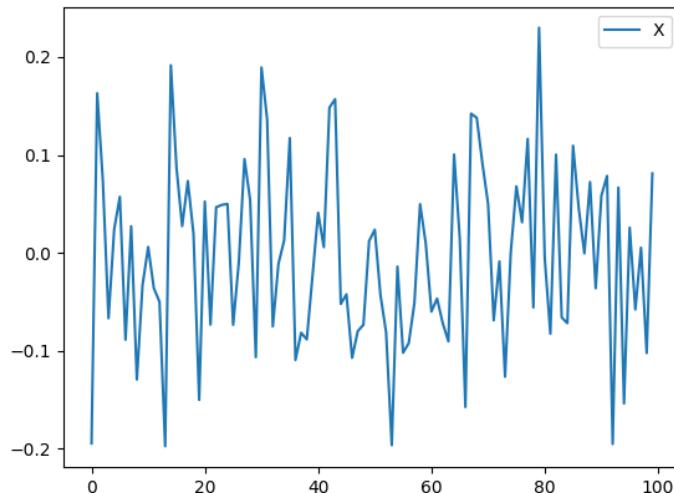
Ruido blanco

El proceso de **ruido blanco** se define como una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $\{e_t\}$

```
12 # random values data series
13 X = np.random.normal(mu, sigma, N)
```

$$X \sim \mathcal{N}(\mu, \sigma) \rightarrow f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

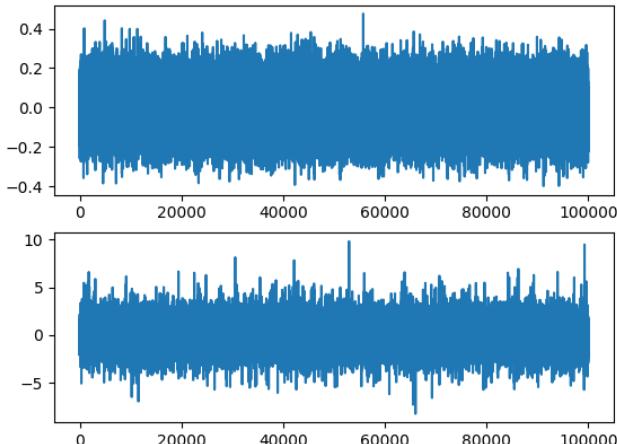
```
## stats
df.X.describe
df.Y.describe()
dt = pd.DataFrame(df.X.describe())
dt=pd.concat([df.X.describe(), df.Y.describe()], axis=1)
```



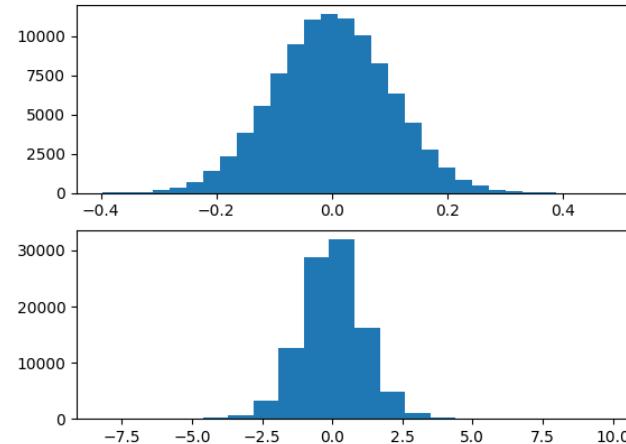
Stat	X	Y
count	100000	100000
mean	6.733E-05	6.608E-03
std	1.001E-01	1.129E+00
min	-3.980E-01	-8.216E+00
25%	-6.742E-02	-6.949E-01
50%	-3.210E-05	4.958E-03
75%	6.800E-02	7.056E-01
max	4.761E-01	9.797E+00

Gráficos

```
## data series
fig, (ax1, ax2) = plt.subplots(2)
fig.subtitle('X vs. Y timeseries')
ax1.plot(X)
ax2.plot(Y)
plt.show()
```



```
## histograms
fig, (ax1, ax2) = plt.subplots(2)
fig.subtitle('X vs. Y histograms')
ax1.hist(X, bins=30)
ax2.hist(Y, bins=20)
plt.show()
```

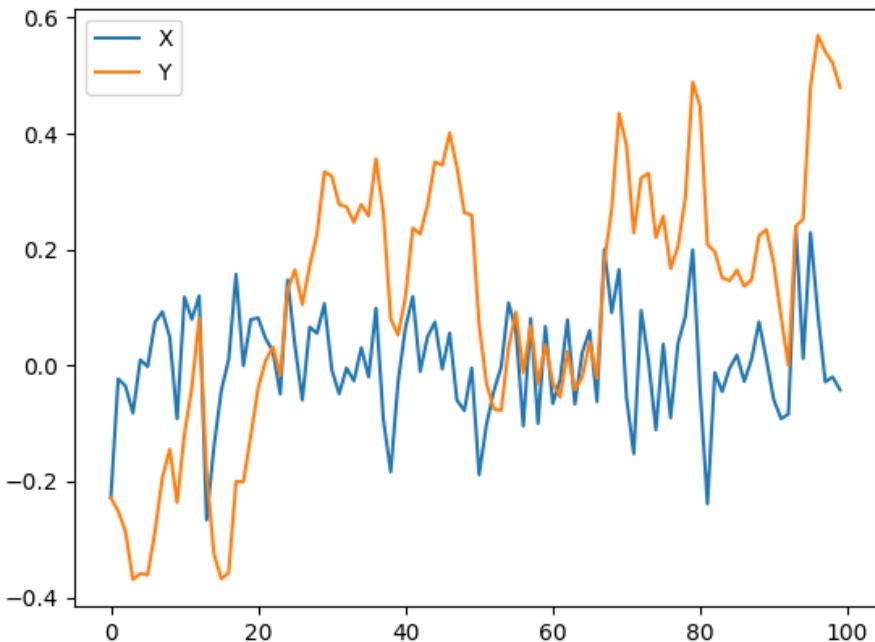


Caminata aleatoria (Random walk)

Sean e_1, e_2, \dots una secuencia de v.a. i.i.d., de media nula y varianza σ_e^2 . Se construye la serie de tiempo como:

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= e_1 + e_2 \\ &\vdots \\ Y_t &= e_1 + e_2 + \dots + e_t \end{aligned} \quad \Rightarrow Y_t = Y_{t-1} + e_t$$

```
12 # random values data series
13 X = np.random.normal(mu, sigma, N)
14
15 # Random walk
16 Y = np.cumsum(X)
```



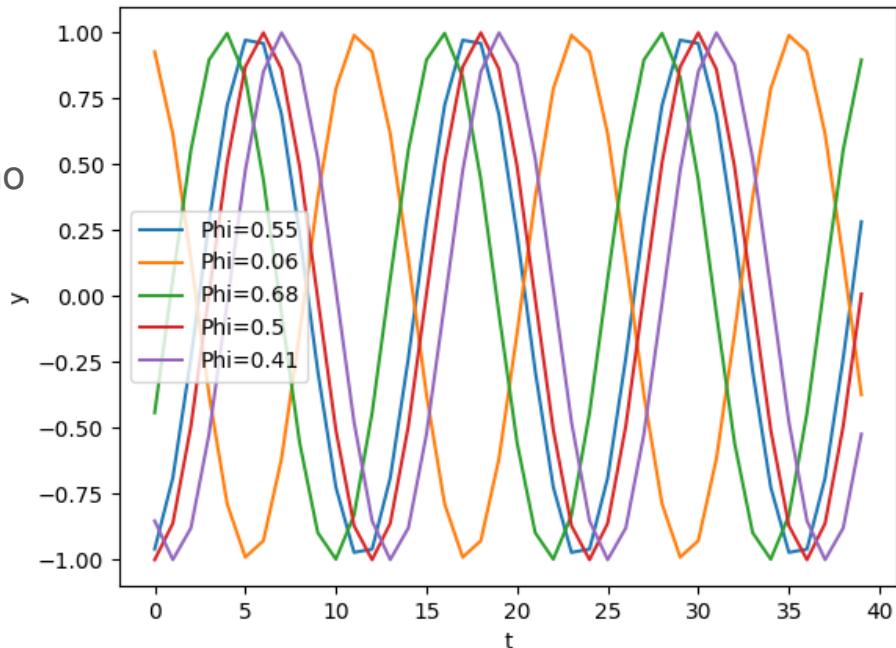
Coseno aleatorio

Podemos definir un proceso aleatorio como

$$Y_t = \cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right], \quad t = 0, \pm 1, \pm 2, \dots$$

donde Φ se elige **una única vez** de una distribución uniforme en $(0,1)$. Observar que una vez definido el valor de Φ , Y_t es un coseno período 12.

```
3  N=40
4  T=12
5  t=np.arange(N)
6  Phi=np.random.rand(5)
7  for phi in Phi:
8      plt.plot(np.cos(2*np.pi*(t/T + phi)))
9
```



¿preguntas?

Nociones de modelado: momentos

Algunos momentos

Una serie de tiempo $Y(t)$, puede entonces modelarse a través de la secuencia de variables $\{Y(t), t = t_k, t_{k+1}, \dots, t_{k+n}\}$ que definen un proceso estocástico. Gran parte de la información de la distribución de estos procesos está contenida en los momentos de dicha distribución: medias, varianzas y covarianzas. A diferencia de las variables aleatorias, estos momentos deberán definirse como función del tiempo.

- **Media (o esperanza):** $\mu_t = \mathbb{E}[Y_t]$
- **Función de autocovarianza:** $C_{t,s} = Cov(Y_t, Y_s) = \mathbb{E} [(Y_t - \mu_t)(Y_s - \mu_s)]$
- **Función de autocorrelación:** $R_{t,s} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{var(Y_t)var(Y_s)}}$

Algunos momentos (caso multivariado)

Análogamente, si se tiene una serie de tiempo multivariada, donde $Y_t = [Y_t^{(1)}, \dots, Y_t^{(l)}]$, definimos

- **Media (o esperanza):** $\mu_t = [\mu_t^{(1)}, \dots, \mu_t^{(l)}] = [\mathbb{E}[Y_t^{(1)}], \dots, \mathbb{E}[Y_t^{(l)}]]$
- **Matriz de cross-covarianza:** $C_{t,s} = \begin{bmatrix} C_{t,s}^{(1,1)} & \dots & C_{t,s}^{(1,l)} \\ \vdots & \ddots & \vdots \\ C_{t,s}^{(l,1)} & \dots & C_{t,s}^{(l,l)} \end{bmatrix}$, donde
 $C_{t,s}^{(i,j)} = Cov(Y_t^{(i)}, Y_s^{(j)}) = \mathbb{E}[(Y_t^{(i)} - \mu_t^{(i)})(Y_s^{(j)} - \mu_s^{(j)})]$
- **Matriz de cross-correlación:** $R_{t,s} = \begin{bmatrix} R_{t,s}^{(1,1)} & \dots & R_{t,s}^{(1,l)} \\ \vdots & \ddots & \vdots \\ R_{t,s}^{(l,1)} & \dots & R_{t,s}^{(l,l)} \end{bmatrix}$

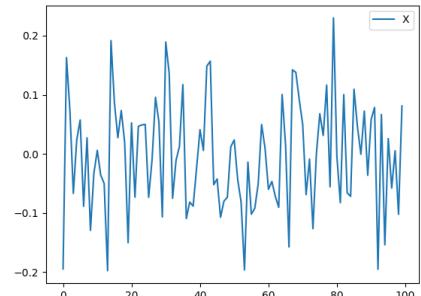


Ruido blanco

El proceso de **ruido blanco** se define como una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) $\{e_t\}$.

Algunas propiedades:

- $\mathbb{P}(e_{t_1} \leq x_1, \dots, e_{t_n} \leq x_n) = \prod_{i=1}^n \mathbb{P}(e_{t_i} \leq x_i) = \mathbb{P}(e_{t_1-k} \leq x_1, \dots, e_{t_n-k} \leq x_n)$
- $\mu_t = \mu$ y $\text{var}(e_t) = \sigma_e^2 \quad \forall t$
- $C_{t,s} = \sigma_e^2 \mathbf{1}\{t = s\}$





Caminante aleatorio (Random Walk)

Sean e_1, e_2, \dots una secuencia de v.a. i.i.d., de media nula y varianza σ_e^2 . Se construye la serie de tiempo como:

$$Y_1 = e_1$$

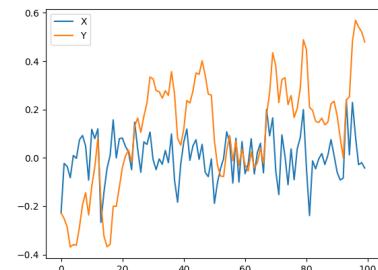
$$Y_2 = e_1 + e_2$$

$$\vdots$$

$$Y_t = e_1 + e_2 + \dots + e_t$$

~~Y~~

$$\Rightarrow Y_t = Y_{t-1} + e_t$$



Podemos calcular distintos momentos para Y_t :

$$\mu_t = \mathbb{E}[Y_t] = \mathbb{E}[e_1 + \dots + e_t] = 0 \quad \forall t$$

$$\sigma_t^2 = \text{var}(Y_t) = \text{var}(e_1 + \dots + e_t) = t\sigma_e^2 \quad \forall t$$

$$C_{t,s} = \text{Cov}(Y_t, Y_s) = t\sigma_e^2 \quad (1 \leq t \leq s)$$

$$R_{t,s} = \sqrt{\frac{t}{s}}$$

¿Qué podemos concluir?

$$E[Y_t] = E[\underbrace{e_1}_0 + \underbrace{te_2}_0 + \dots + \underbrace{te_t}_0] = 0$$

$$e_1 + e_2 + \dots + e_t$$

$$\text{Var}(Y_t) = t \text{Var}(e_i) = t \sigma_e^2$$

indep

$$\begin{aligned} \text{Cov}(Y_3, Y_5) &= \text{Cov}(\underbrace{e_1 + e_2 + e_3}_3, \underbrace{e_1 + e_2 + e_3 + e_4 + e_5}_5) \\ &= \underbrace{\text{cov}(e_1, e_1)}_{\sigma_e^2} + \text{cov}(e_1, e_3) + \text{cov}(e_1, e_5) + \dots + \\ &= 3 \sigma_e^2 \end{aligned}$$

Coseno aleatorio (Random Cosine Wave)

Podemos definir un proceso aleatorio como

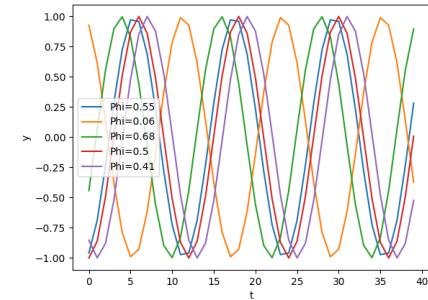
$$Y_t = \cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right], \quad t = 0, \pm 1, \pm 2, \dots$$

donde Φ se elige una única vez de una distribución uniforme en $(0, 1)$. Observar que una vez definido el valor de Φ , Y_t es un coseno período 12.

$$\mathbb{E}[Y_t] = \mathbb{E}\left\{\cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right]\right\} = \frac{1}{2\pi} \left[\sin\left(2\pi\frac{t}{12} + 2\pi\right) - \sin\left(2\pi\frac{t}{12}\right) \right] = 0$$

$$C_{t,s} = \text{Cov}(Y_t, Y_s) = \mathbb{E}\left\{\cos\left[2\pi\left(\frac{t}{12} + \Phi\right)\right] \cos\left[2\pi\left(\frac{s}{12} + \Phi\right)\right]\right\} = \frac{1}{2} \cos\left[2\pi\left(\frac{|t-s|}{12}\right)\right]$$

Vemos que se corresponde con un proceso estacionario

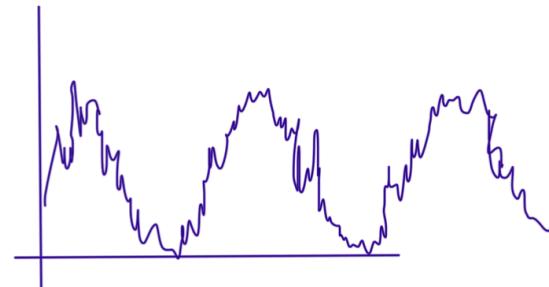




Estacionariedad

Estacionariedad vs. Estacionalidad

- **Estacionariedad** es una propiedad de algunas series de tiempo que tiene que ver con **invarianza** a lo largo del tiempo.
- **Estacionalidad** se refiere a ciclos o períodos que se pueden observar en una serie de tiempo, por ejemplo las **estaciones** del año



Estacionariedad

Una suposición que suele hacerse a la hora de modelar series de tiempo es la estacionariedad.

Se dice que un proceso es **estacionario (stationary)** si la distribución del proceso no cambia a lo largo del tiempo. Matemáticamente, quiere decir que

$$f_{Y_{t_1}, \dots, Y_{t_n}} = f_{Y_{t_1+\delta}, \dots, Y_{t_n+\delta}}, \quad \forall t_1, \dots, t_n, \forall \delta$$

Una conclusión que se desprende de esta definición es que si los procesos son estacionarios, su media es constante en el tiempo y las funciones de autocorrelación y autocovarianza dependen sólo de la diferencia de tiempos:

$$\mu_t = \mu_s = \mu \quad \forall t, s$$

$$\underline{C_{t,s}} = Cov(Y_t, Y_s) = Cov(Y_{t+\delta}, Y_{s+\delta}) = \underline{C_{s-t}}, \quad \forall s, t, \delta$$

$$\underline{R_{t,s}} = Corr(Y_t, Y_s) = Corr(Y_{t+\delta}, Y_{s+\delta}) = \underline{R_{s-t}}, \quad \forall s, t, \delta$$

Estacionariedad débil

Se dice que un proceso es **débilmente estacionario (DE)** si sólo se cumple

$$\mu_t = \mu_s = \mu \quad \forall t, s$$

Es decir que sólo se pide que sean estacionarios los momentos hasta de segundo orden.

$$C_{t,s} = Cov(Y_t, Y_s) = Cov(Y_{t+\delta}, Y_{s+\delta}) = C_{s-t}, \quad \forall s, t, \delta$$

$$R_{t,s} = Corr(Y_t, Y_s) = Corr(Y_{t+\delta}, Y_{s+\delta}) = R_{s-t}, \quad \forall s, t, \delta$$

En general cuando hablamos de series de tiempo estacionarias nos estaremos refiriendo a este tipo de estacionariedad.

Estimación de momentos

Dadas N muestras de una serie de tiempo DE $\{y_1, \dots, y_N\}$, entonces podemos estimar los momentos como:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N y_i$$

$$\hat{C}_k = \frac{1}{N} \sum_{n=k+1}^N (y_n - \hat{\mu})(y_{n-k} - \hat{\mu})$$

$$\hat{R}_k = \frac{\hat{C}_k}{\hat{C}_0}$$

Preprocesamiento

¿ Cuándo aplicar un preprocesamiento?

Cuando la distribución de la serie de tiempo cambia a lo largo del tiempo se dice que es **no estacionaria**.

En estos casos, pueden aplicarse transformaciones para que la serie de tiempo resultante sea aproximadamente estacionaria.

Algunos métodos:

1. Transformación de variables
2. Diferenciación
3. Promedio móvil

1. Transformación de variables

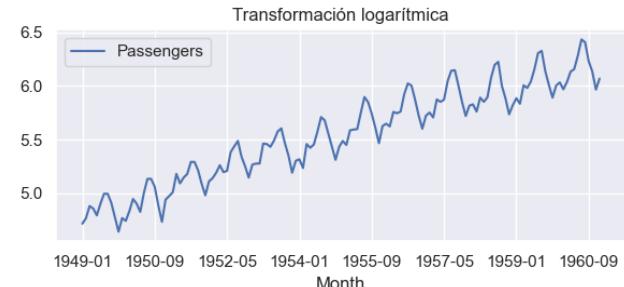
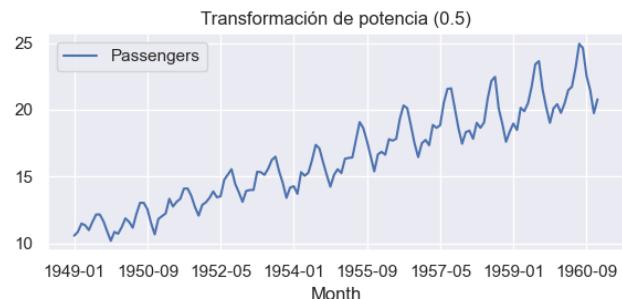
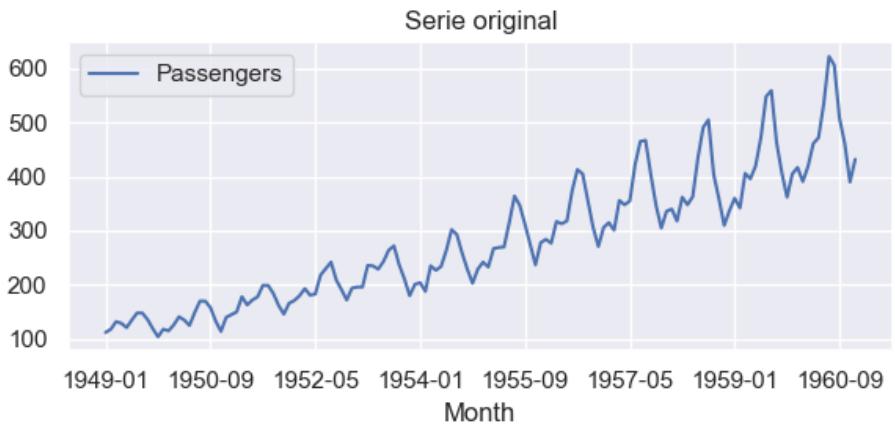
Muchas series de tiempo tienen la característica de que sus varianzas van aumentando a medida que avanza el tiempo.

Se pueden aplicar transformaciones a los puntos de la serie de tiempo.

- Transformación Box-Cox

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0, \\ \ln y_i & \text{si } \lambda = 0 \end{cases}$$

1. Ejemplos



2. Diferenciación

Si la serie presenta un tendencia se puede analizar en su lugar la serie diferenciada. Dada una serie de tiempo y_n , $n = 0, 1, \dots$ definimos

$$z_n = \Delta y_n = y_n - y_{n-1}$$

Motivación: si $y_n = an + b$, al diferenciar se obtiene una constante.

Observación: En general, si y_n se corresponde con un polinomio de grado n , diferenciando n veces se recupera una constante.

Ejemplos

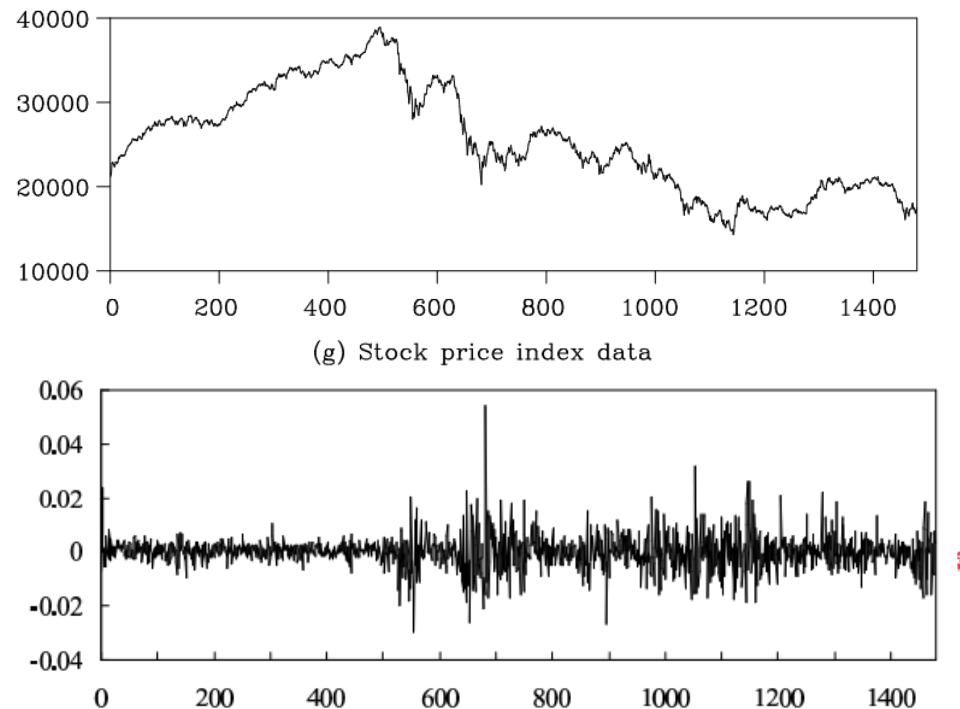


Figure 1.2: *Difference of the logarithm of the Nikkei 225 data.*

3. Promedio móvil

Para una serie de tiempo y_n , el promedio móvil de $(2k + 1)$ términos está dado por

$$T_n = \frac{1}{2k+1} \sum_{j=-k}^k y_{n+j}$$

Si modificamos la definición intercambiando la media por la mediana obtenemos la mediana móvil, definida como

$$T_n = \text{mediana}\{y_{-k}, \dots, y_k\}$$

En general, la mediana móvil puede capturar cambios en la tendencia más rápido que el promedio móvil.

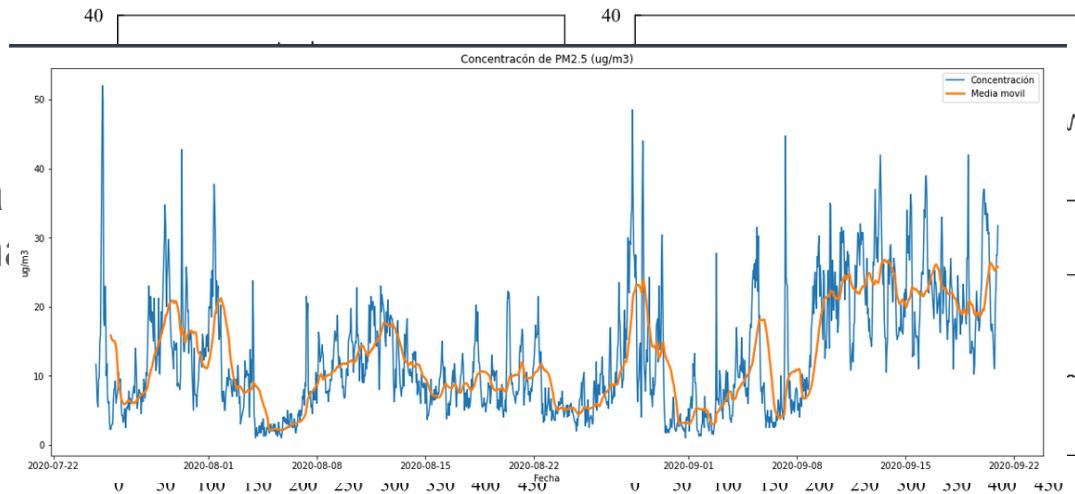


Figure 1.4 Maximum temperature data and its moving average. Top left: original data, top right: moving average with $k = 5$, bottom left: $k = 17$, bottom right: $k = 29$.

Transformación logarítmica

¿Cuándo usar la transformación logarítmica?

- Cuando observamos que la varianza del proceso parece aumentar con el tiempo.
En particular si $\mathbb{E}[Y_t] = \mu_t$ y $\sqrt{\text{var}(Y_t)} = \mu_t\sigma$, luego $\mathbb{E}[\log(Y_t)] \approx \log(\mu_t)$ y $\text{var}(\log(Y_t)) \approx \sigma^2$
- Si Y_t tiene cambios porcentuales relativamente estables entre un instante de tiempo y otro, y supongamos que $Y_t = (1 + X_t)Y_{t-1}$. Luego tomando el log

$$\log(Y_t) = \log((1 + X_t)Y_{t-1}) = \log(1 + X_t) + \log(Y_{t-1}) \rightarrow \log(Y_t) - \log(Y_{t-1}) = \log\left(\frac{Y_t}{Y_{t-1}}\right) = \log(1 + X_t)$$

Si además suponemos que X_t está acotado, $|X_t| < 0.2$, sucede que

$\log(1 + X_t) \approx X_t$ y $\nabla(\log Y_t) \approx X_t$ va a ser relativamente estable y posiblemente se encuentre bien modelada por un poc. estacionario.

Resumen

- Qué es una serie de tiempo, para qué sirve estudiarlas y cómo modelarlas.
- Momentos útiles para analizar series de tiempo: media, varianza, autocovarianza y autocorrelación.
- Algunos modelos de series de tiempo: ruido blanco, caminante aleatorio, coseno aleatorio.
- Estacionariedad.
- Técnicas de preprocessamiento: transformación Box-Cox, diferenciación, promedio móvil.

