Almacenes y Minería de Datos

Prácticas

Minería de Datos I

Series Temporales

Joaquín Ángel Triñanes Fernández
Instituto de Investigacións Tecnolóxicas
Joaquin.Trinanes@usc.es Ext: 16001 Externo: 881816001

R

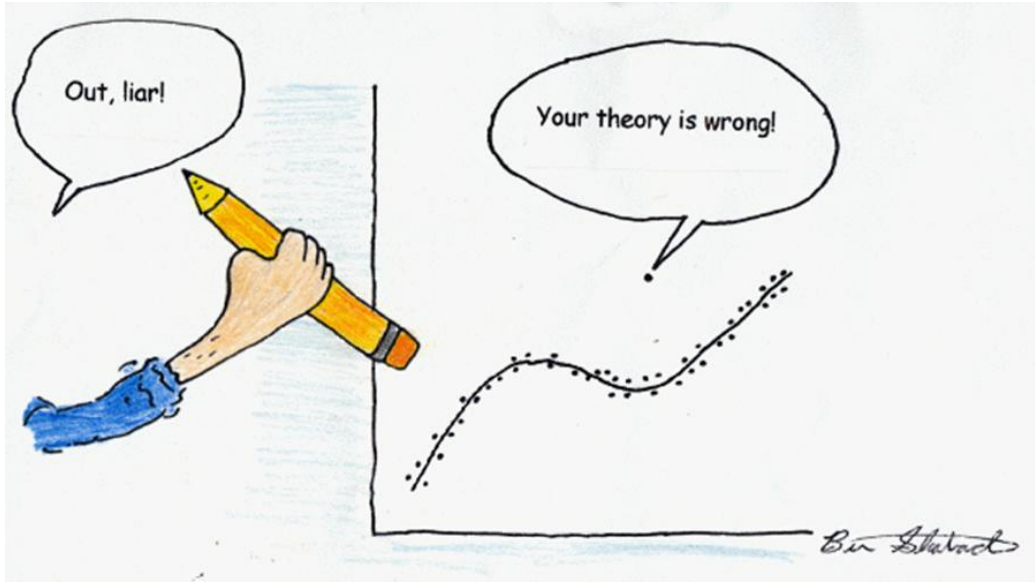
- Entorno de trabajo muy popular en los campos de la Estadística y los Análisis de Datos
- help.start() help(max) o ?max
- ls() Variables, rm(list=ls()) borra las variables, mcsv<-read.csv("fichero.csv"), write,csv(…), read.table…
- Existen muchos paquetes ('packages') que se pueden usar para expandir las carácterísticas de R. Son conjuntos de librerías.Para saber las librerías instaladas: library()
- Instalar: http://www.r-project.org/
- Sin coste y fácil de aprender. Se puede combinar con Tableau y Power BI.
- RStudio (www.rstudio.com) es una IDE para R.

R

- Estructuras de datos:
  - Vectores-Conjunto de elementos del mismo tipo: estudiantes<-c(“Pepe”,”Pepa”, “Jose”, “Josefa”), edad<-c(1,2,3,4)
  - Factores- Variables categóricas que se almacenan como enteros que tienen asociadas etiquetas. Las etiquetas se almacenan una sola vez y tiene asociados un vector de enteros. Categoria<-factor(c(“Amateur”,”Profesional”))
  - Matrices- Elementos del mismo tipo en filas y columnas. mimatriz<- matrix(c(1,2,3,4),nrow=2)
  - Listas- Parecidos a los vectores pero los tipos de datos puedes ser diferentes. Incluso pueden contener objetos de naturaleza diferente. Lista_ejemplo<-list(vector_estudiantes, matriz notas)
  - Data frames- Semejente a matriz pero con datos de diferentes tipos. Se usa mucho en gráficos. Podemos tener una columna numérica, otra con caracteres, etc. Ejem_df<-data.frame(vector1,vector2)

Valores atípicos

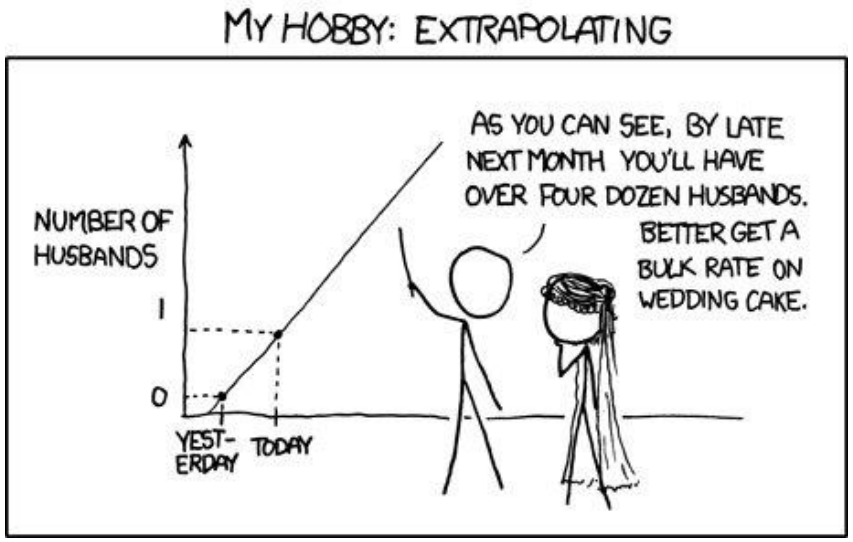
- Es un valor que se diferencia de forma significativa de los demás valores
- Ruido != Valores atípicos
- Nos interesan
- Origen: errores o varianza de los datos
- El contexto es importante
- Existen múltiples métodos para detectarlos

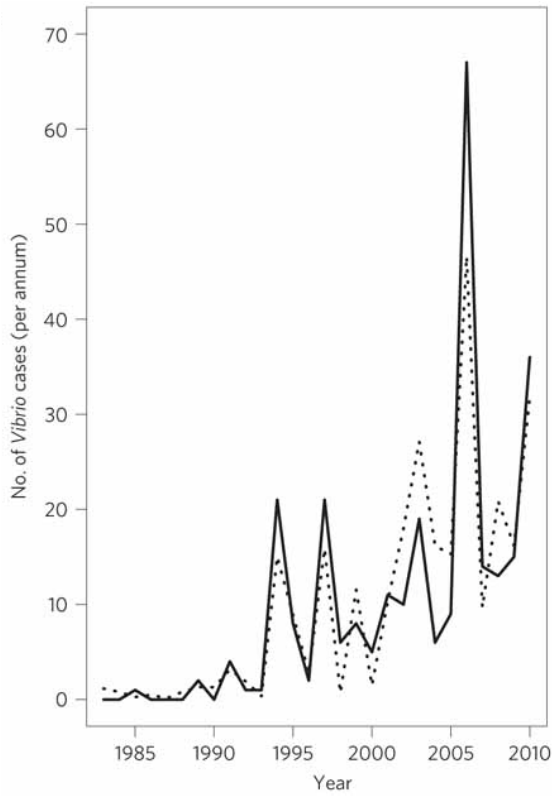


Series temporales

- Podemos construirlas de variables geofísicas pero también pueden ser económicas, demográficas, salud, etc.
- Deseamos desarrollar un modelo matemático que nos permita describir de forma consistente los datos.
- Se suelen graficar con la variable en el eje Y y el tiempo en la abscisa.

Ref: Time Series Analysis and Its Applications (Shumway&Stoffer, Ed. Springer, 2011)





Series temporales

- En muchas ocasiones, se tratan conceptualmente como series contínuas, pero la naturaleza de los procesos de medida las hacen discretas.
- Procesos tipicos de tratamiento pueden implicar un suavizado de la serie: serie_filtrada=filter(serie_sin_filtrar, ..
- Si suavizamos la serie a través de un filtro, de forma general, la media no cambia
- Muchos modelos asumen la serie como la suma de una señal y un ruido aleatorio (S/N)
- R: ts
  - Series<-ts(datos,start=c(2021,1),end=c(2021,12),frequency=12)
- A menudo, descomponemos la serie en componentes (ej. función decompose en R)

Series temporales

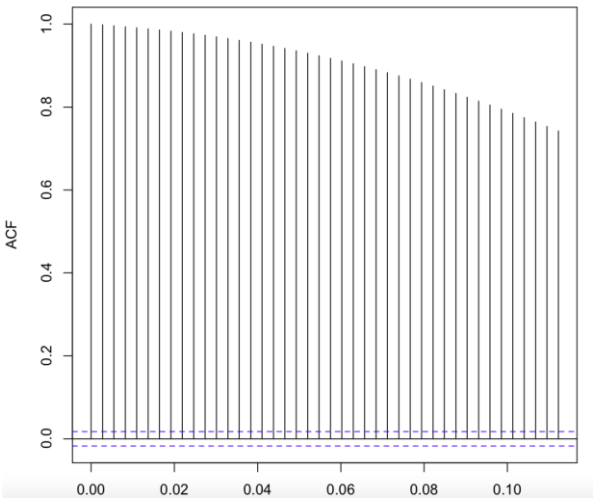
- Factores a tener en cuenta: funciones de autocovarianza, autocorrelaciones (ACF), covarianza cruzada, correlación cruzada (CCF).
- Procesos:
  - Estacionarios: sus propiedades estadísticas no cambian en el tiempo, no dependen el tiempo de observación.
- ¿Y si tenemos una serie con un componente estacional o una tendencia?
  - no es estacionaria

Series temporales

- ¿Cómo podemos hacer una serie estacionaria?
  - Calculando las diferencias entre valores consecutivos.
  - En la serie anterior, ayudaría a eliminar tendencia y estacionalidad
- Podemos identificar series no estacionarias con los diagramas ACF
    - Para este tipo de series el ACF decae lentamente
  - R: acf()

Python:

```
fromstatsmodels.graphics.tsaplots import plot_acf
```



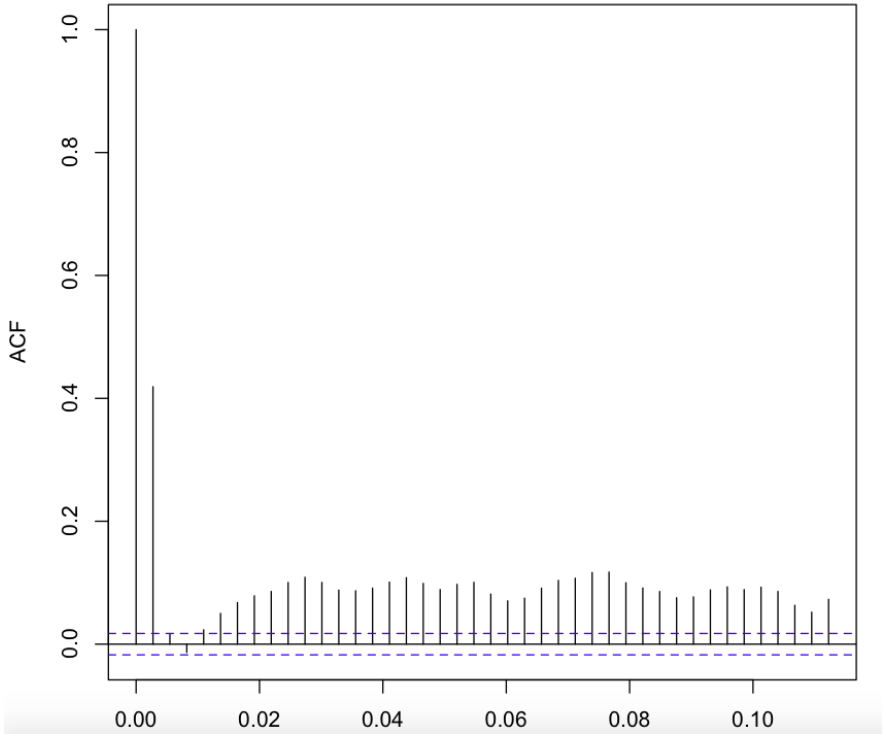
Series temporales

- ¿Por qué queremos una serie estacionaria?
  - Una serie estacionaria es fácil de predecir (sus propiedades estadísticas futuras son las mismas que las de ahora)
  - Una vez que modelamos la serie estacionaria, siempre podremos revertir los métodos aplicados para poder modelar la serie original
  - Al hacerla estacionaria, podemos comparar con otras variables(medias, correlaciones, varianzas)

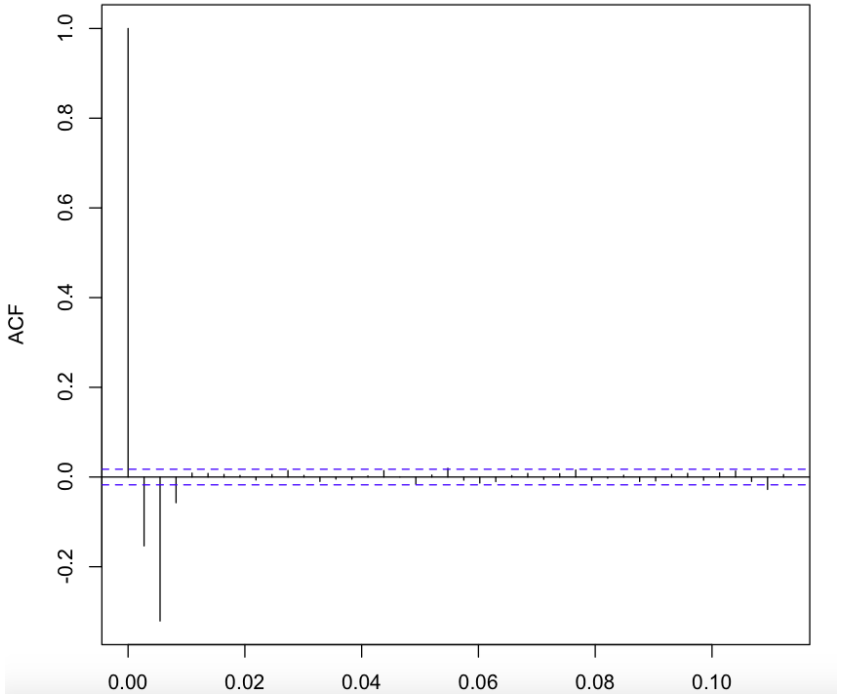
Series temporales

- En ocasiones, una diferencia no hace la serie estacionaria y tenemos que volver a calcular. R:diff Python: numpy.diff()

Primera diferencia

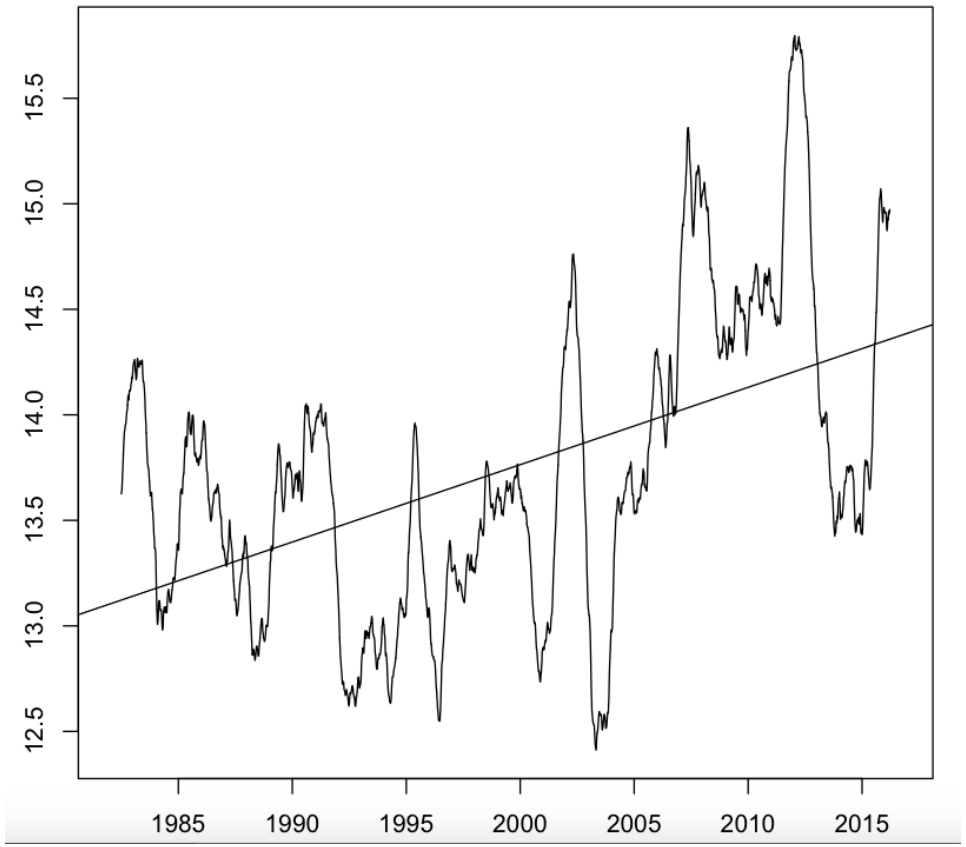


Segunda diferencia



Series temporales

- Tendencia lineal



Incremento: 0.03667 ºC/año

Queremos hacer predicciones

Queremos “comprenderla”

Series temporales

1-Cargar datos de series temporales desde los archivos csv presentes en el Campus Virtual. Haced una valoración de los valores atípicos que pudieran estar presentes.

2-Crear serie de tiempos para cada variable.

- Calcular primera y segunda diferencia y ACF respectivos
- Descomponer la serie en componentes estacionales y tendencia

3-Interpolar linealmente el trend

4-Calcular la correlación cruzada entre las series temporales (ej: TM e IRRA)

5-Filtrarlas en tendencia con una media móvil

6- Intentad realizar una predicción (ej. Paquete forecast)

Se valorará la calidad en la interpretación/descripción de los resultados obtenidos.

Series temporales

Esta práctica será realizada en equipos de 3 componentes y consistirán en un documento PDF con el código y la descripción de la práctica.