DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada



# Análisis de series temporales

© Fernando Berzal, berzal@acm.org

## Análisis de series temporales

- Características de las series temporales
- Visualización de series temporales
- Filtrado de series temporales
  - Medias móviles
  - Suavizado exponencial
- Técnicas de regresión
  - Regresión lineal
  - Coeficiente de correlación de Pearson
- Función de autocorrelación
- Caso práctico: Una sesión de análisis

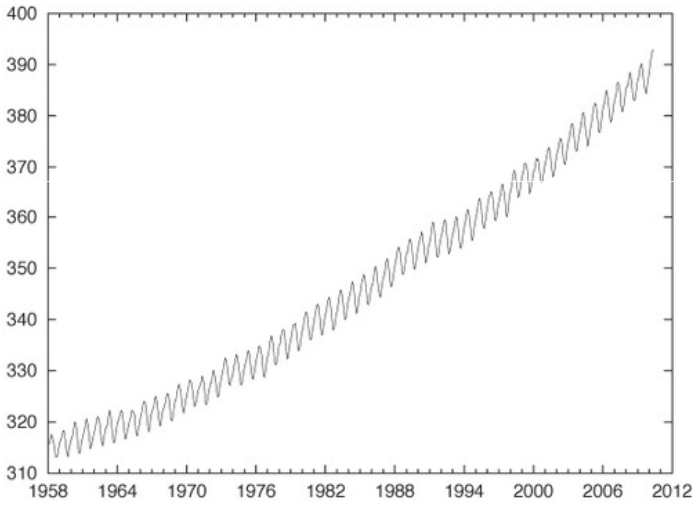
# Características

- Tendencias
- Estacionalidad (comportamientos periódicos)
- Ruido
- Otros, p.ej. cambios bruscos de comportamiento



# Ejemplos

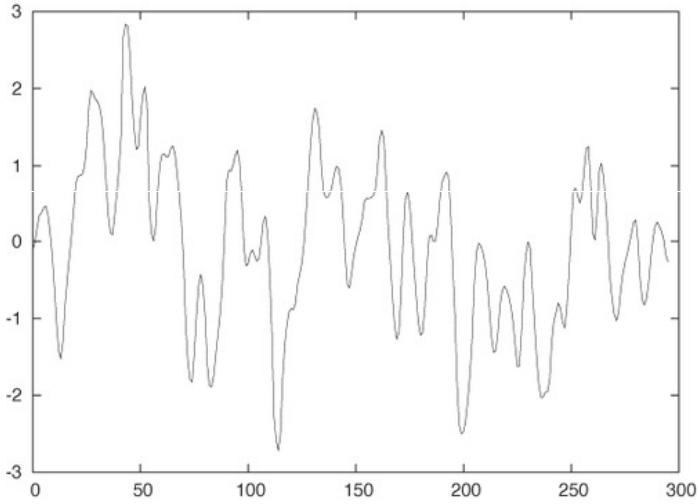
**Tendencia y estacionalidad**




Concentración de $CO_2$
medida en el observatorio de Mauna Loa, Hawaii.

# Ejemplos

**Variación "suave" pero sin tendencia a largo plazo**

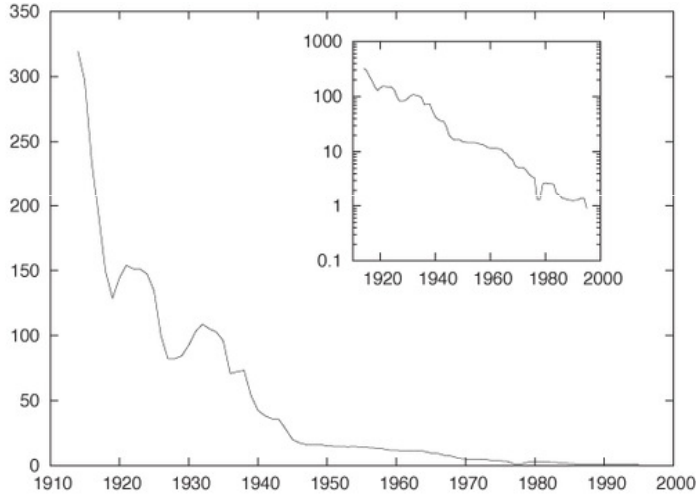



Concentración de gas a la salida de una caldera



# Ejemplos

**Tendencia no lineal**




Coste de llamadas telefónicas de larga distancia (USA)

# Ejemplos

**Cambios "bruscos" de comportamiento**

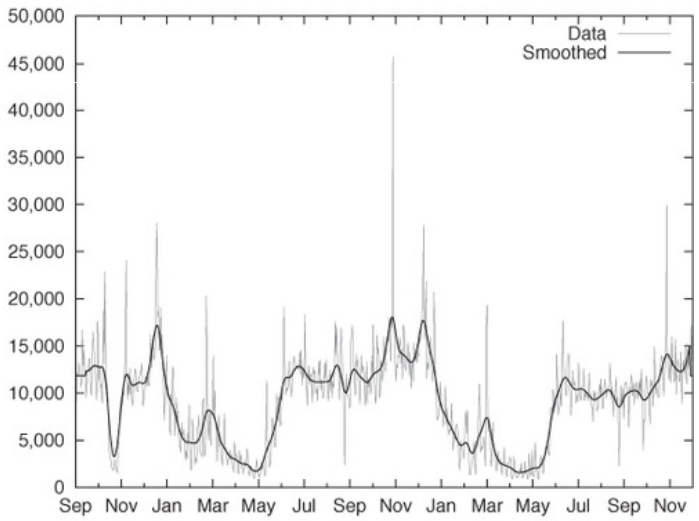

Índice Nikkei (Bolsa de Tokyo)



# Ejemplos

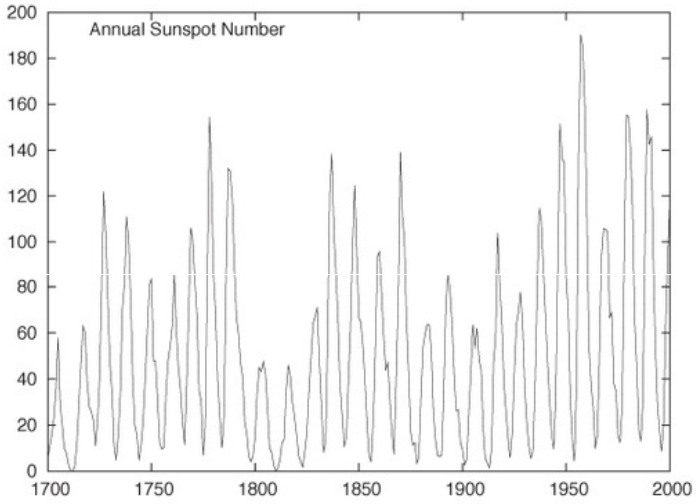
**Conjuntos de datos reales…**

Estacionalidad a corto y largo plazo,
posibles cambios de comportamiento
**y ruido**




Llamadas diarias
a un call-center

# Visualización




Número anual de manchas solares durante 300 años

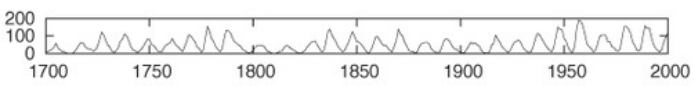
**Una relación de aspecto incorrecta**
**hace difícil reconocer los detalles de cada ciclo.**



# Visualización

**Banking** [Banking to 45 degrees]

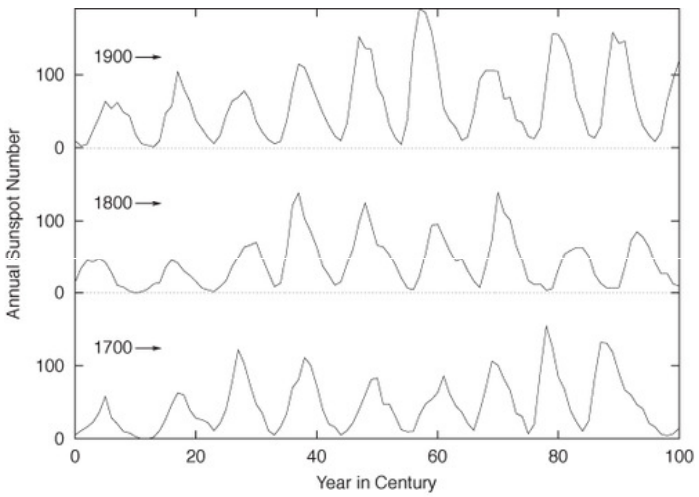
Los cambios casi verticales de la figura anterior nos cuesta trabajo apreciarlos. Sin embargo, reconocemos mejor los cambios en una serie cuando se dibujan con un ángulo de 45º:




Ahora podemos apreciar que las "subidas" son más rápidas que las bajadas, aunque la figura es tan pequeña que apenas se pueden analizar detalles…

# Visualización

**Stacking**




Dividiendo el eje temporal en 3 fragmentos, mantenemos el "banking" y generar un gráfico con unas dimensiones más razonables (p.ej. 4:3).



# Filtrado de series temporales

**Medias móviles** [moving averages]

IDEA: Reemplazar el punto central de una serie de un número impar de números consecutivos por su media aritmética (filtro "paso bajo").

$$s_i = \frac{1}{2k+1} \sum_{j=-k}^{k} x_{i+j}$$

# Filtrado de series temporales

**Medias móviles** [moving averages]

PROBLEMA: La presencia de un pico en la ventana [i-k, i+k] distorsiona la media móvil.

POSIBLE SOLUCIÓN: Utilización de pesos (menores en los extremos de la ventana).

$$s_i = \sum_{j=-k}^{k} w_j x_{i+j} \quad donde \quad \sum_{j=-k}^{k} w_j = 1$$

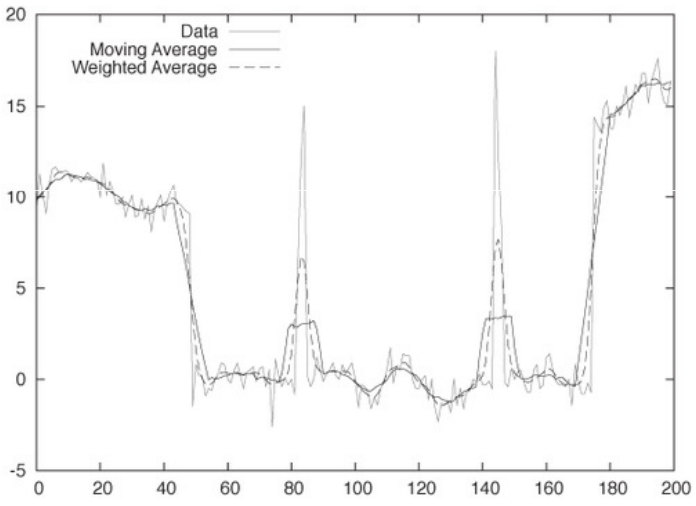
Ejemplos: Gaussiana, ventana de Hamming…

http://en.wikipedia.org/wiki/Window_function#Hann_window



# Filtrado de series temporales

**Medias móviles** [moving averages]




k=5

# Filtrado de series temporales

**Medias móviles** [moving averages]

Limitaciones de las medias móviles:

- "Costosas" de calcular: Cuando se utilizan pesos, el cálculo hay que hacerlo desde cero para cada valor.
- Problemáticas en los extremos de las series de datos (dada la anchura de la ventana, no se pueden extender hasta el final de la serie, que suele ser lo más interesante).
- No se pueden definir fuera de la serie temporal, por lo que no se pueden utilizar para realizar predicciones.



# Filtrado de series temporales

**Suavizado exponencial** [exponential smoothing]

Proporciona un filtrado fácil de calcular,
además evita los problemas de las medias móviles:

- **Suavizado exponencial simple**
(para series sin tendencia ni estacionalidad).
- **Suavizado exponencial doble**
(para series con tendencia pero no estacionalidad).
- **Suavizado exponencial triple**
(para series con tendencia y estacionalidad).

# Filtrado de series temporales

**Suavizado exponencial simple**

$$s_i = \alpha x_i + (1-\alpha) s_{i-1}$$

Los distintos métodos de suavizado exponencial actualizan el resultado del anterior valor con el último dato de la serie original (combinando la información ya disponible con la aportada por el nuevo dato mediante un parámetro, $0<\alpha<1$).



# Filtrado de series temporales

**Suavizado exponencial simple**

¿Por qué se llama suavizado exponencial?

Si expandimos la recurrencia, obtenemos:

$$s_i = \alpha \sum_{j=0}^{i} (1-\alpha)^j x_{i-j}$$

Todas las observaciones previas contribuyen al valor suavizado, pero su contribución se suprime por el exponente creciente del parámetro $\alpha$.

# Filtrado de series temporales

**Suavizado exponencial simple**

"Uso" en predicción: Si extendemos el suavizado más allá del final de los datos disponibles, la predicción es extremadamente simple :-(

$$x_{i+h} = s_i$$

Ante la presencia de tendencias, la señal suavizada tiene ir retrasada con respecto a los datos originales salvo que utilicemos un valor de $\alpha$ cercano a 1.



# Filtrado de series temporales

**Suavizado exponencial doble**

$$s_i = \alpha x_i + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1}$$

El suavizado exponencial doble retiene información acerca de la tendencia: la señal suavizada $s_i$ y la tendencia suavizada $t_i$.

El parámetro $\beta$ se utiliza para realizar un suavizado exponencial sobre la tendencia.

# Filtrado de series temporales

**Suavizado exponencial doble**

"Uso" en predicción:

Si extendemos el suavizado más allá del final de los datos disponibles, la predicción es la siguiente:

$$x_{i+h} = s_i + ht_i$$



# Filtrado de series temporales

**Suavizado exponencial triple**

(a.k.a. método de Holt-Winters)

Una tercera cantidad se utiliza para describir la estacionalidad, que puede ser aditiva o multiplicativa según nos interese.

NOTA:

$p_i$ modela el componente periódico de la señal, donde k es el período observado.

# Filtrado de series temporales

**Suavizado exponencial triple**

(a.k.a. método de Holt-Winters)

ESTACIONALIDAD ADITIVA

$$s_i = \alpha(x_i - p_{i-k}) + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1}$$
$$p_i = \gamma(x_i - s_i) + (1-\gamma)p_{i-k}$$
$$x_{i+h} = s_i + ht_i + p_{i-k+h}$$



# Filtrado de series temporales

**Suavizado exponencial triple**

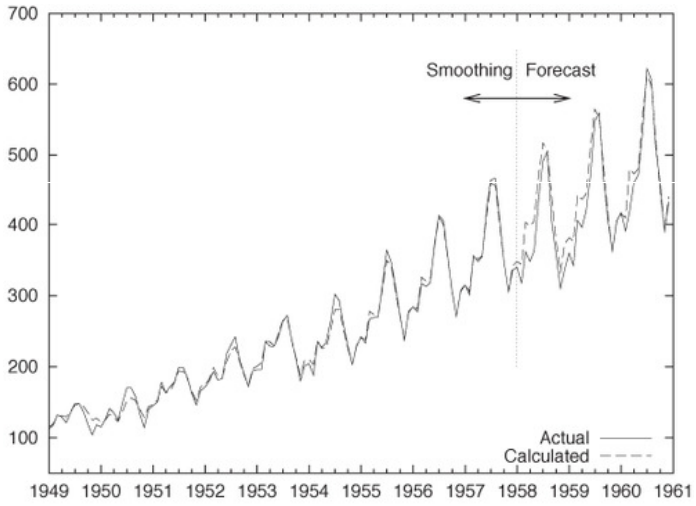
(a.k.a. método de Holt-Winters)

ESTACIONALIDAD MULTIPLICATIVA

$$s_i = \alpha\frac{x_i}{p_{i-k}} + (1-\alpha)(s_{i-1} + t_{i-1})$$
$$t_i = \beta(s_i - s_{i-1}) + (1-\beta)t_{i-1}$$
$$p_i = \gamma\frac{x_i}{s_i} + (1-\gamma)p_{i-k}$$
$$x_{i+h} = (s_i + ht_i)p_{i-k+h}$$

# Filtrado de series temporales

**Suavizado exponencial** [exponential smoothing]




Número mensual de pasajeros (en miles).



# Técnicas de regresión

La predicción (numérica) es…

- Similar a la clasificación:
  - Se construye un modelo a partir de un conjunto de entrenamiento.
  - Se utiliza el modelo para predecir el valor de una variable (continua u ordenada).
- Diferente a la clasificación:
  - El modelo define una función continua.

Método más empleado: **Regresión**

# Técnicas de regresión

Las técnicas de regresión modelan la relación entre una o más variables independiente (predictores) y una variable dependiente (variable de respuesta).

**Métodos de regresión**

- Regresión lineal
- Regresión no lineal
- Árboles de regresión (p.ej. CART)
- …



# Técnicas de regresión

**Regresión lineal simple**

Una única variable independiente:

$$y = w_0 + w_1 x$$

donde $w_0$ (desplazamiento) y $w_1$ (pendiente) son los coeficientes de regresión.

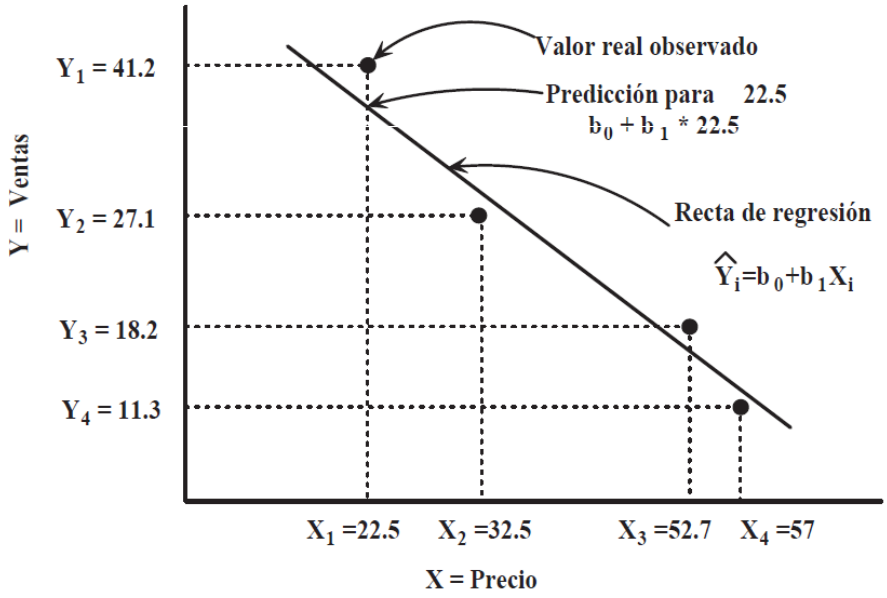
- **Método de los mínimos cuadrados**
  (estima la línea recta que mejor se ajusta a los datos):

$$w_0 = \bar{y} - w_1 \bar{x} \qquad w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2}$$

# Técnicas de regresión

## Regresión lineal simple

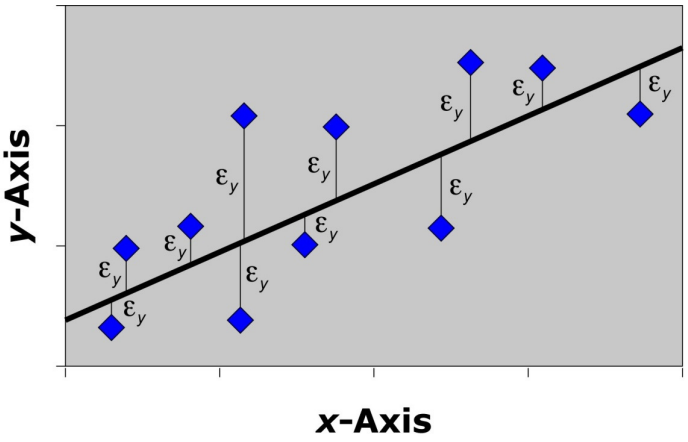
$$\hat{Y}_i = b_0 + b_1 X_i$$






# Técnicas de regresión

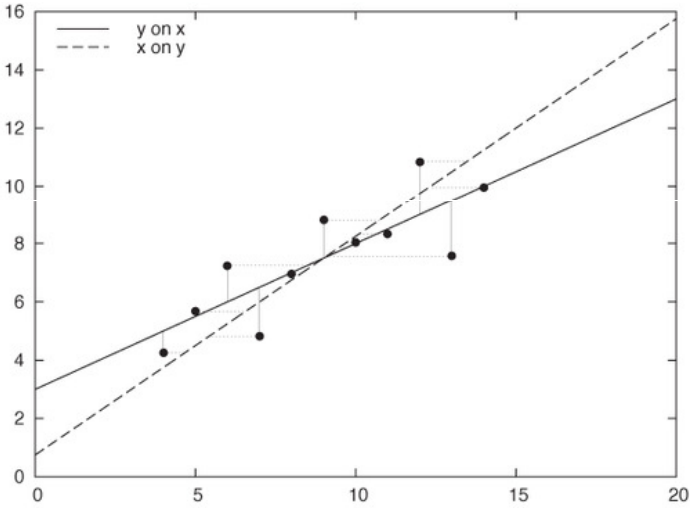
## Regresión lineal simple




El método de los mínimos cuadrados minimiza la suma de los cuadrados de los residuos $\varepsilon_i$ (las diferencias entre las predicciones y los valores observados).

# Técnicas de regresión

## Regresión lineal simple




¡OJO! Al utilizar regresión lineal, la recta y=f(x) que se obtiene es distinta a la que obtenemos si x=f(y).



# Técnicas de regresión

## Regresión lineal múltiple

Varias variables independientes:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots$$

- Resoluble por métodos numéricos de optimización.
- Muchas funciones no lineales pueden transformarse en una expresión lineal.

p.ej. Un modelo de regresión polinomial
$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$
puede transformarse en un modelo lineal definiendo las variables $x_2 = x^2$, $x_3 = x^3$:
$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$

# Técnicas de regresión

## Regresión lineal

Condiciones necesarias para aplicar regresión lineal:

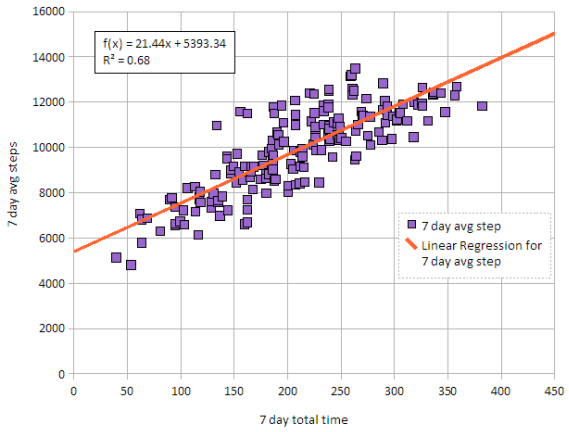
- Obviamente, la muestra ha de ser aleatoria.
- El tipo de dependencia descrita ha de ser lineal.
- Fijado un valor de la(s) variable(s) independiente(s), la variable dependiente se distribuye según una distribución normal.
- Los errores han de tener la misma varianza (nube de puntos homogénea).



# Técnicas de regresión

## Regresión lineal simple

1. Mediante un diagrama de dispersión, comprobamos visualmente si existe una relación lineal entre las variables X (predictor) e Y (respuesta):



# Técnicas de regresión

## Regresión lineal simple

2. Cuantificamos la relación construyendo la recta que resume la dependencia y damos una medida de cómo se ajusta la recta a los datos (correlación):

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1]$$



# Técnicas de regresión

## Coeficiente de correlación

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1]$$

**r=+1** Dependencia lineal total en sentido positivo (cuanto mayor es X, mayor es Y).

**r=-1** Dependencia lineal total en sentido negativo (cuanto mayor es X, menor es Y ).

# Técnicas de regresión

## Coeficiente de correlación

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1]$$

**r > 0** Existe una dependencia positiva.
Cuanto más se acerque a 1, mayor es ésta.

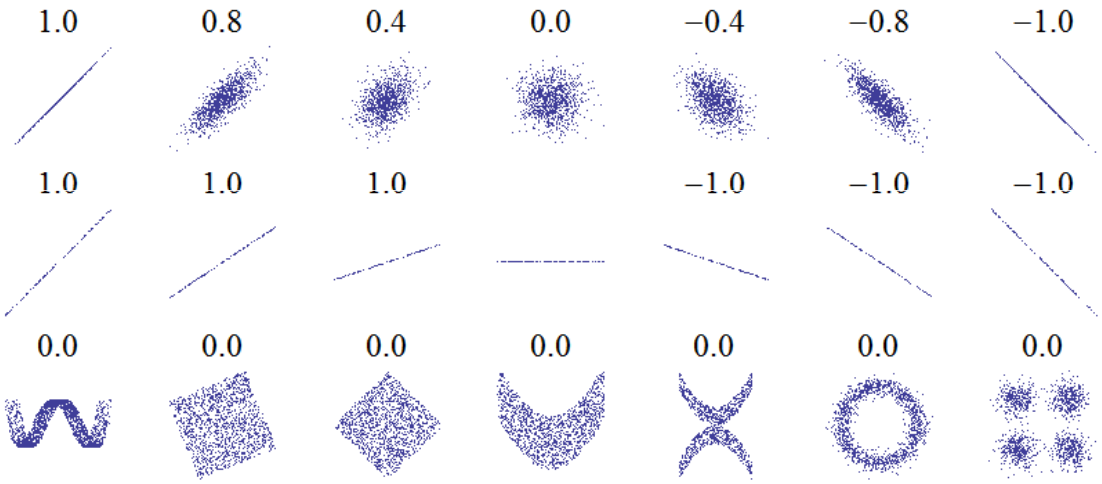
**r < 0** Existe una dependencia negativa.
Cuanto más se acerque a -1, mayor será.

**r = 0** No podemos afirmar nada.



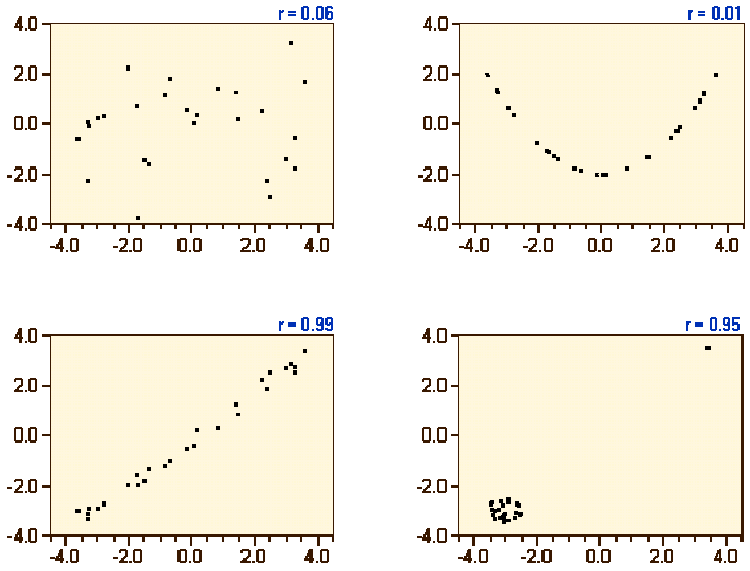
# Técnicas de regresión

## Coeficiente de correlación



# Técnicas de regresión

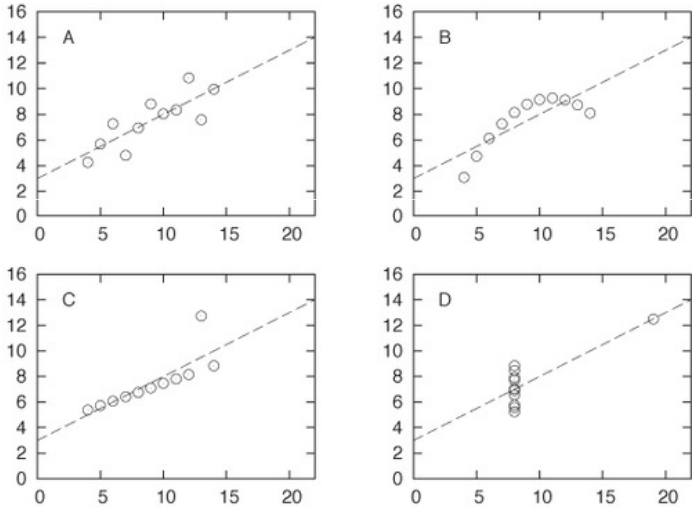
## Coeficiente de correlación






# Técnicas de regresión

## Coeficiente de correlación




El cuarteto de Anscombe

(4 conjuntos de datos con el mismo coeficiente de correlación)

# Técnicas de regresión

## Coeficiente de correlación

### Ventaja de r

- No depende de las unidades usadas en la medición.

### Limitaciones de r

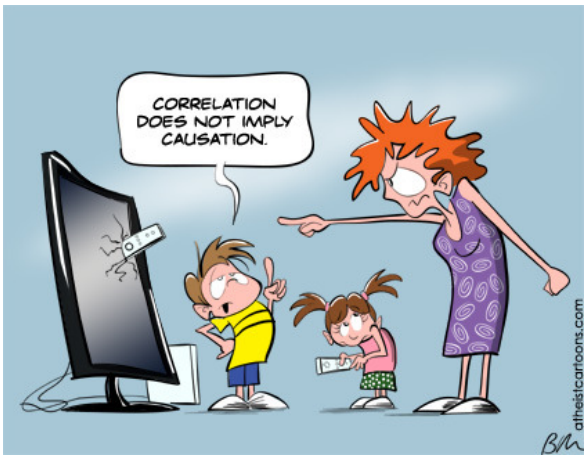
- Sólo mide dependencia lineal entre las variables.

¡OJO! La correlación no implica causalidad…



# Técnicas de regresión

## Coeficiente de correlación



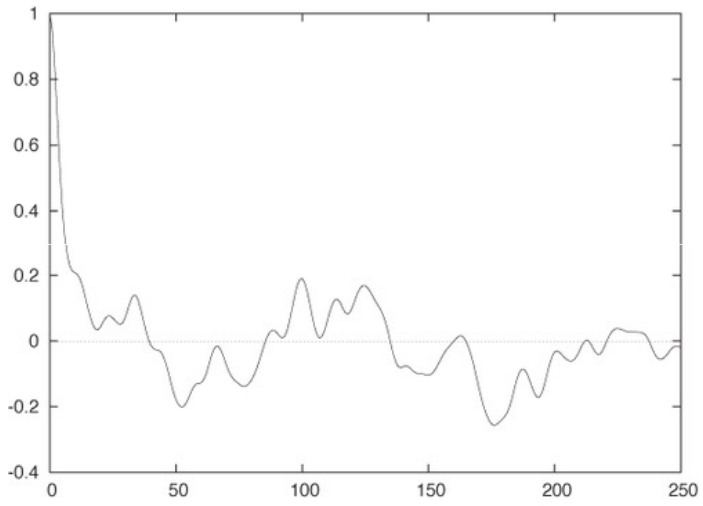

"Correlation is not causation but it sure is a hint."
-- Edward Tufte

# Función de autocorrelación

$$c(k) = \frac{\sum_{i=1}^{N-k}(x_i - \mu)(x_{i+k} - \mu)}{\sum_{i=1}^{N}(x_i - \mu)^2} \quad con \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

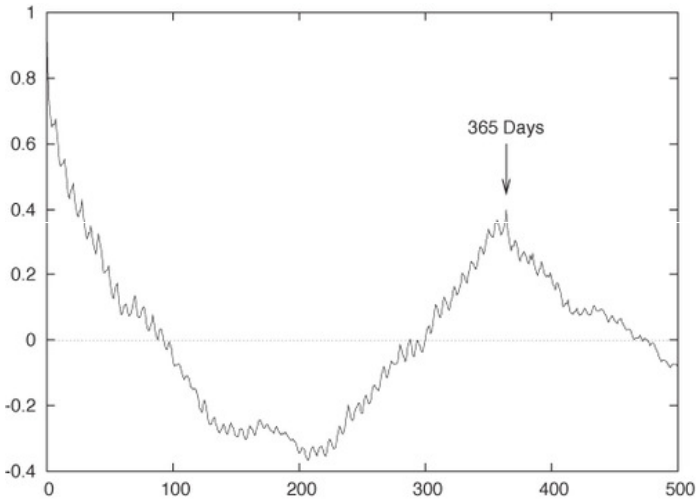


# Función de autocorrelación




Autocorrelación para la salida de gas de una caldera

# Función de autocorrelación



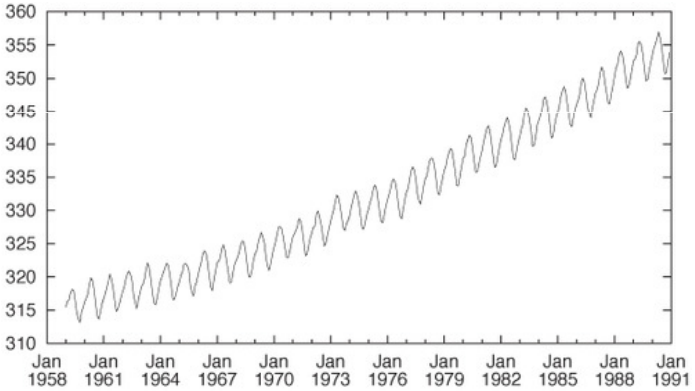

Autocorrelación en las llamadas a un call-center



# Caso práctico

**Conjunto de datos**

Mediciones de $CO_2$ en Mauna Loa (Hawaii)

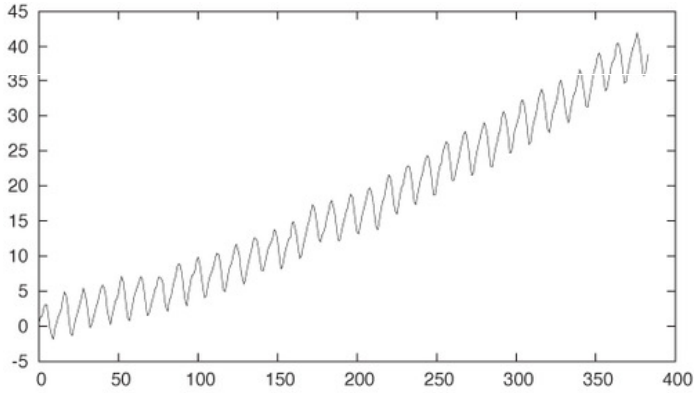



Adaptado de Philipp K. Jannert:

"Intermezzo: A Data Analysis Session" [capítulo 6]

# Caso práctico

A partir de las mediciones mensuales (1959-1991), eliminamos las fechas del eje X
y hacemos que la serie empiece de cero:



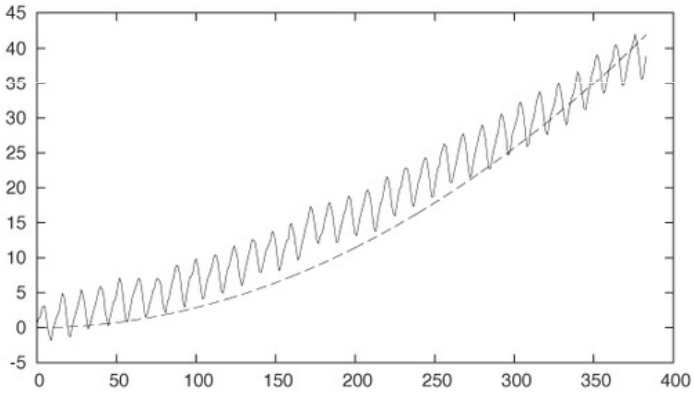



# Caso práctico

**Tendencia**: Apreciamos una tendencia no lineal:

Intentamos ajustarla con una función de la forma $x^k$

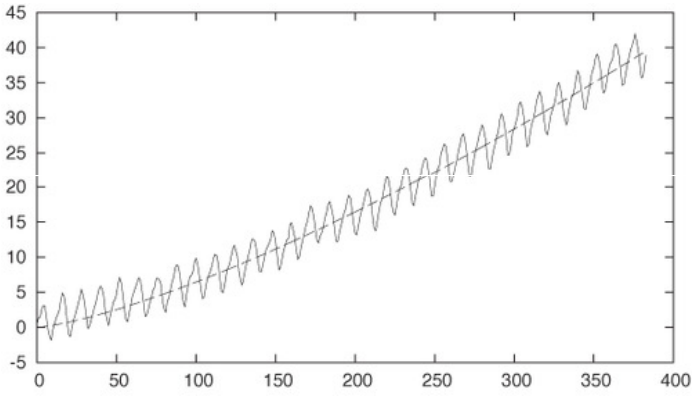
Nota: Todas las curvas de ese tipo pasan por (0,0) y (1,1)




Con k=2, tenemos 35*(x/350)^2, pero parece que nos hemos pasado…

# Caso práctico

Afinamos un poco más y usamos un valor menor:




OK!

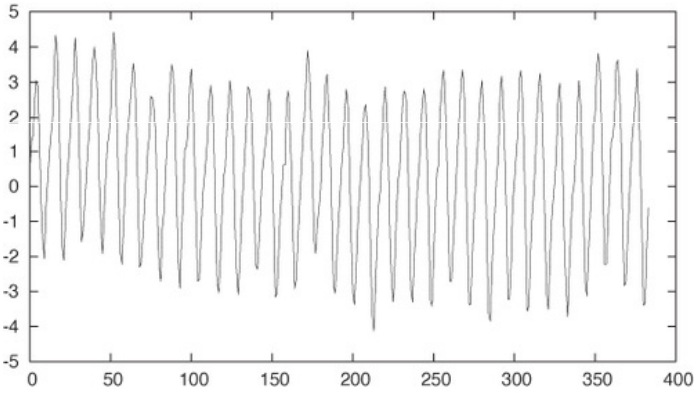
k=1.35

Ajuste de la función f(x) = 35*(x/350)^1.35



# Caso práctico

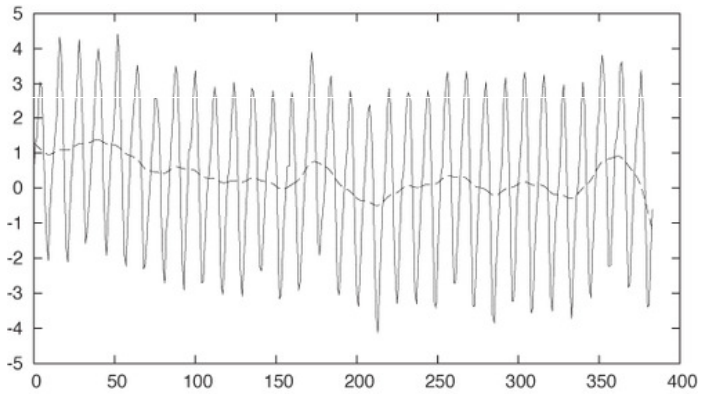
Para comprobar que no vamos mal,
calculamos los residuos (valor original – aproximación):




Residuos del ajuste f(x) = 35*(x/350)^1.35

# Caso práctico

Si nuestro ajuste de la tendencia es correcto, los residuos no deben exhibir tendencia alguna (deberían aparecer balanceados en torno a y=0):



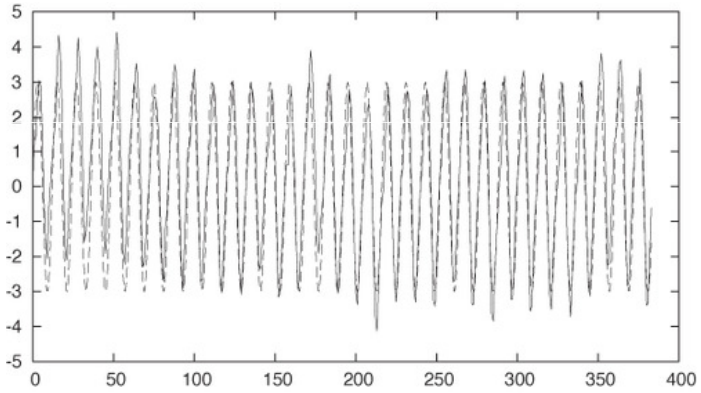

Suavizamos los residuos para comprobar si aún existe algún tipo de tendencia en los residuos…



# Caso práctico

**Estacionalidad**:

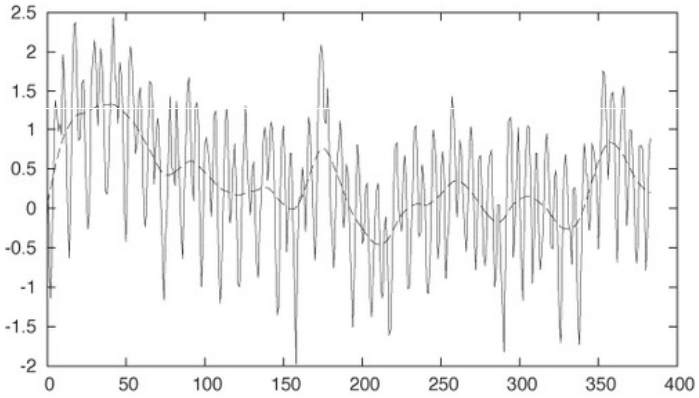
Apreciamos una periodicidad anual (cada 12 valores)




Ajustamos con una función senoidad 3*sin(2*pi*x/12)

# Caso práctico

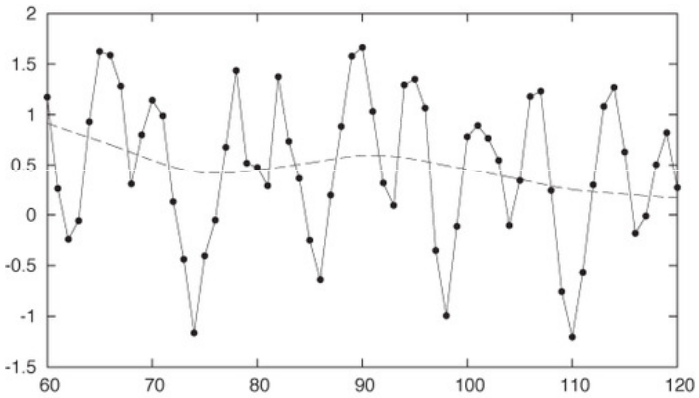
Calculamos los residuos tras nuestras aproximaciones (valor original – tendencia – estacionalidad)






# Caso práctico

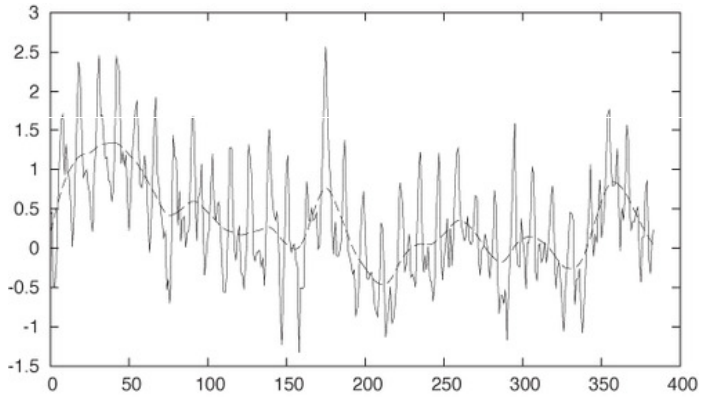
En la figura anterior no se ve mucho… hacemos zoom:




Se sigue apreciando cierta periodicidad, por lo que usamos un segundo armónico -0.75*sin(2*pi*x/6)

# Caso práctico

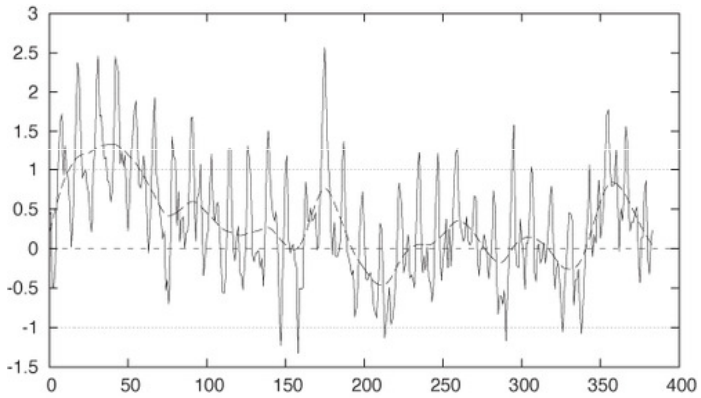
Residuos tras eliminar la tendencia y los dos primeros armónicos correspondientes a la estacionalidad:






# Caso práctico

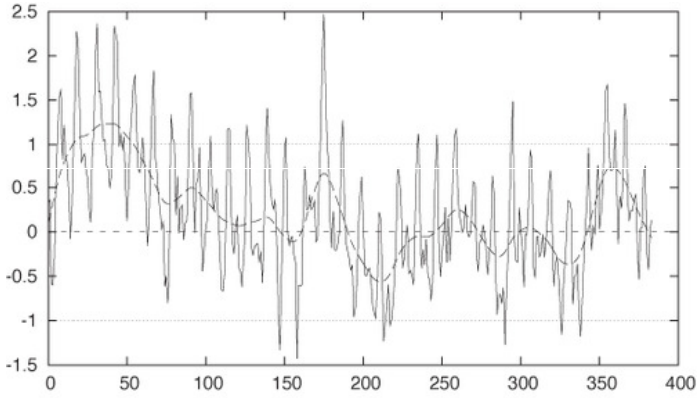
Añadimos líneas que nos ayuden a ver si los residuos están sesgados:




Parece sesgado hacia arriba,
por lo que añadimos un desplazamiento de +0.1

# Caso práctico

Los residuos de nuestra aproximación final:

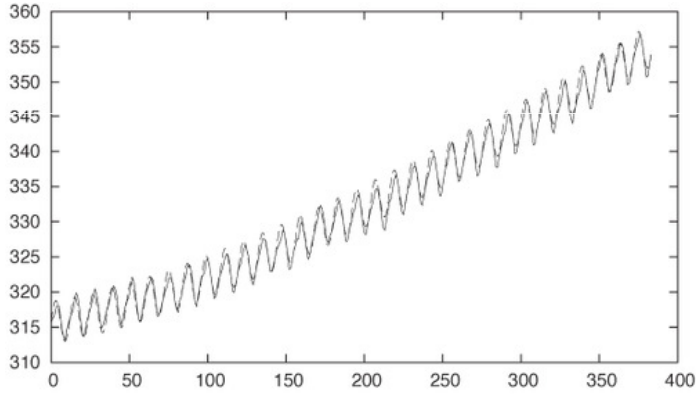



f(x) = 315 + 35*(x/350)**1.35
+ 3*sin(2*pi*x/12) – 0.75*sin(2*pi*x/6)
+ 0.1



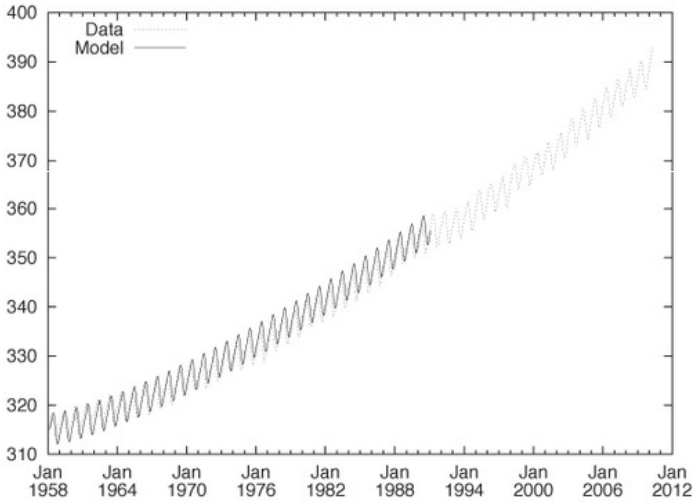
# Caso práctico

El ajuste que hemos realizado (1959-1990):



# Caso práctico

Nuestra predicción del futuro (1991-2010)





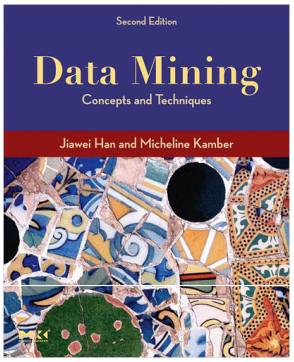

# Más técnicas de análisis


**Forecasting**

http://en.wikipedia.org/wiki/Forecasting

# Bibliografía


- Jiawei Han
  & Micheline Kamber:
  **Data Mining:**
  **Concepts and Techniques** [8.2]
  Morgan Kaufmann, 2006.
  ISBN 1558609016




- Philipp K. Janert:
  **Data Analysis**
  **with Open Source Tools** [Part I]
  O'Reilly, 2010.
  ISBN 0596802358