Análisis de Series Temporales (Breve Introducción)

Juan A. Botía
juanbot@um.es

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

TIIA, Primer Cuatrimestre, 2009/2010

1. Introduction
2. El análisis de series temporales
3. Modelos
4. Estimando elementos en una serie temporal
5. Relación entre valores de una serie
6. Relación entre dos series

Datos Secuenciales

Hasta ahora, hemos asumido en los datos

- Observaciones de un fenómeno concreto
- Independientes, sin un orden entre ellos asumido
- Idénticamente distribuidos (i.e. responden a la misma distribución de probabilidad subyacente)

En las series temporales no es el caso

- Dejan de tener validez las técnicas de estimación paramétrica

Algunas definiciones

Serie Temporal

Está compuesta de una secuencia de valores o eventos que cambian con el tiempo. Típicamente, sus valores se miden en intervalos iguales (e.g. la evolución diaria de un índice bursátil).

Secuencia

Responde al concepto genérico de hilera de datos en la que el orden es importante. No se considera aquí la noción del tiempo de manera explícita (e.g. secuencia de clickeos de ratón de usuario en un portal Web).

El análisis de series temporales

Objetivo: explicación de la serie bajo estudio, para ...

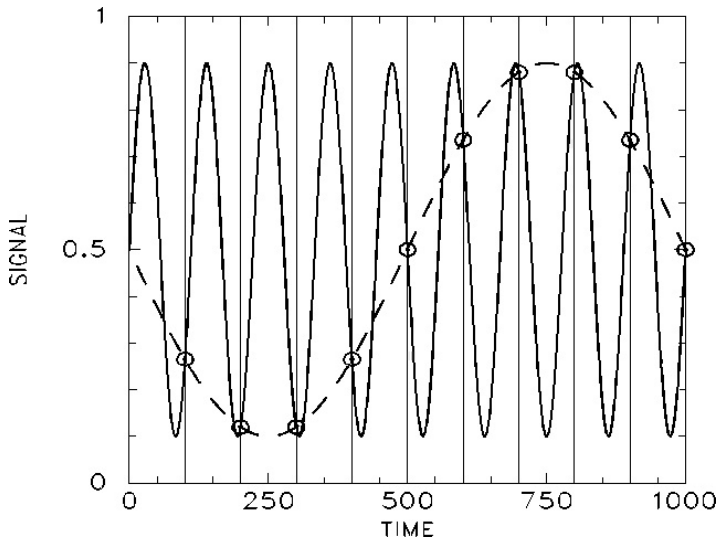
- Predicción
  - Generar valores futuros para una serie, basándonos en valores ya conocidos de la misma u otras series
- Estudiar el impacto de un suceso
  - Impacto de una ley antialcohol en el número de gente beoda al volante
  - Cantidad de coches que descongestionarían una parte de la ciudad al colocar un puente estratégicamente.
- Estudiar patrones causales (i.e.efecto de varias variables en una serie)
  - Con más de una serie
  - Si detectamos que el aumento puntual del número de desempleados ocurre antes que un aumento parecido en los índices de crimen, podríamos inferir que el paro es casua del aumento del crimen.

Representación, concepto

- Desde el punto de vista estadístico (i.e. si basamos nuestro modelo de serie en v.a.'s), una serie temporal es una colección de variables aleatorias indexadas de acuerdo al orden en el que han sido obtenidas en el tiempo, $x_1, x_2, x_3, \ldots$, en donde x_i se refiere a la v.a. obtenida en el instante iésimo.
- Un cjto. de variables aleatorias $\{x_t\}$ indexadas en el tiempo es un proceso estocástico.
- Típicamente, una serie temporal se representa graficamente mediante la representación de los valores de las v.a. en el eje vertical (ordenadas) y el tiempo en el eje horizontal (abscisa).
- Valores adyacentes se conectan para realizar así una reconstrucción visual hipotética de la serie.

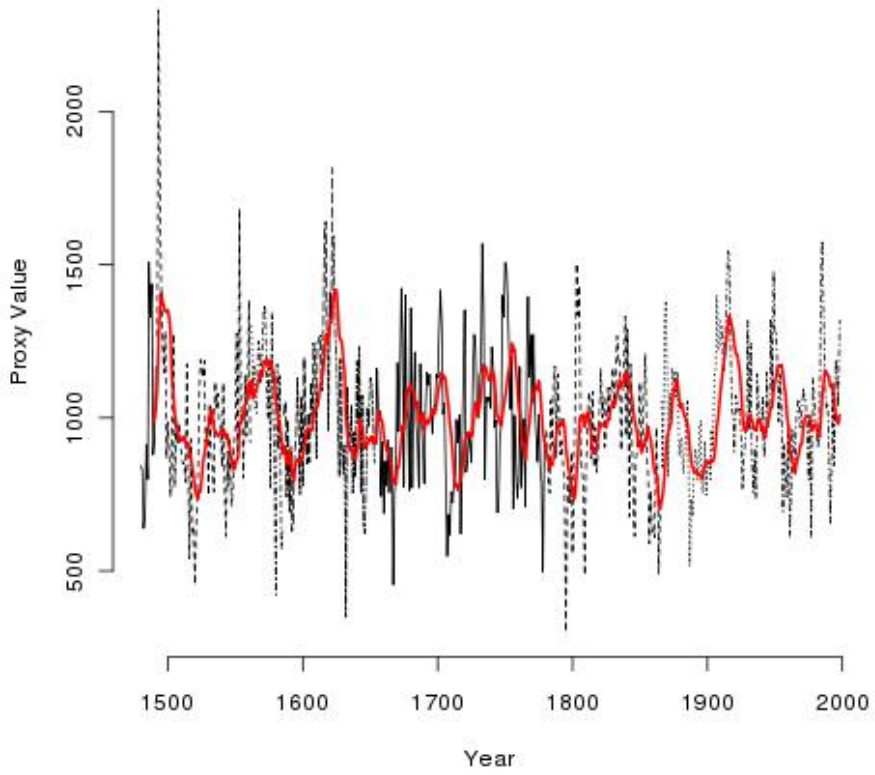
Representación, problemática

- Representar una serie implica transformar una señal continua en una muestra obtenida a intervalos regulares de tiempo
- Concecuencia: si la frecuencia no es lo suficientemente alta, la serie no reflejará la señal real (i.e. aliasing)
- En la figura, ambas curvas representan la serie al contener todos los puntos de la muestra (están aliased)



Idea básica que subyace en el análisis de series

- Cada serie tiene un grado de suavidad (i.e. frecuencia, magnitud y regularidad de los dientes de sierra)
- La suavidad viene derivada de que el valor de la serie en x_t depende de los valores $x_{t-1}, x_{t-2}, \ldots$ (idea básica)
- La curva en negro son datos reales, en rojo suavizados



Datos reales vs. suavizados

- Con el suavizado del ejemplo anterior, aumentamos la correlación de los datos (disminuyen frecuencia y amplitud de señal)
- Si tenemos datos reales, eso es lo que queremos estudiar
  - Solo suavizaremos cuando querramos realizar predicciones, por ejemplo
- Los patrones en las series los esconde el ruido
  - Si queremos hacer descubrimiento de patrones, lo primero es eliminarlo
  - Ojo!! Solo en aquellas series en las que realmente lo hay
  - Es posible que tengamos series sin ruido y aun sin patrón evidente

Caracterización del ruido

Ruido blanco

El el contexto de las series, definimos el ruido como una colección de v.a. no correladas w_t, con media 0 y varianza σ_w^2. Lo denotaremos como

$$w_t \sim wn(0, \sigma_w^2).$$

Este tipo de serie temporal se usa como modelo para el ruido en aplicaciones de ingeniería convencionales

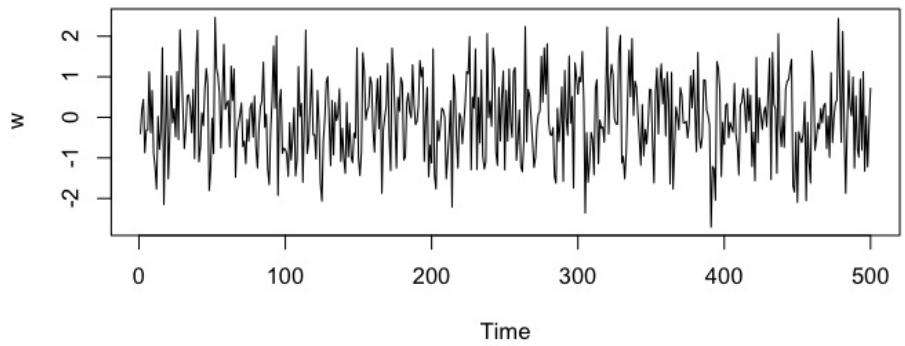
Eliminación del ruido

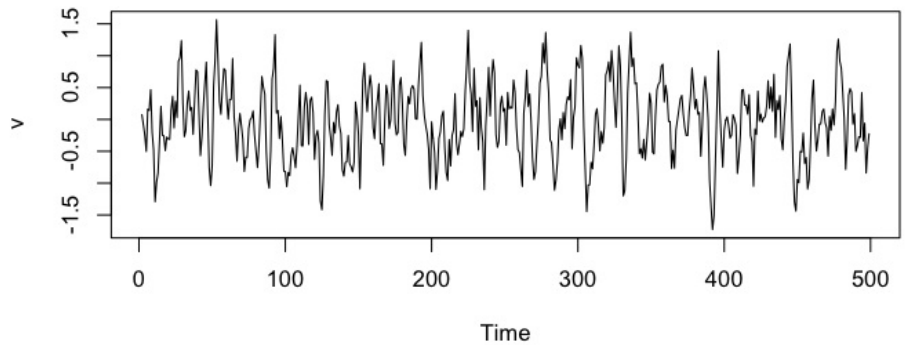
- Podemos remplazar el ruido blanco mediante una media móvil
- Para una serie w_t

$$v_t = \frac{1}{3}(w_{t-1} + w_t + t_{t+1}).$$

- Arriba tenemos la serie original (500 v.a. del tipo $iidN(0,1)$, representadas en el orden que fueron obtenidas)
- Abajo la misma serie suavizada. y obtenida con los comandos *R* siguientes

```
> w = rnorm(500,0,1)
> v = filter(w,sides=2,rep(1,3)/3)
> par(mfrow=c(2,1))
> plot.ts(w)
> plot.ts(v)
```





Modelos AR (*Auto-Regressive*)

- Un modelo AR consiste en una forma alternativa de suavizado de series mediante autoregresiones, e.g.

$$x_t = x_{t-1} - 0.9x_{t-2} + w_t,$$

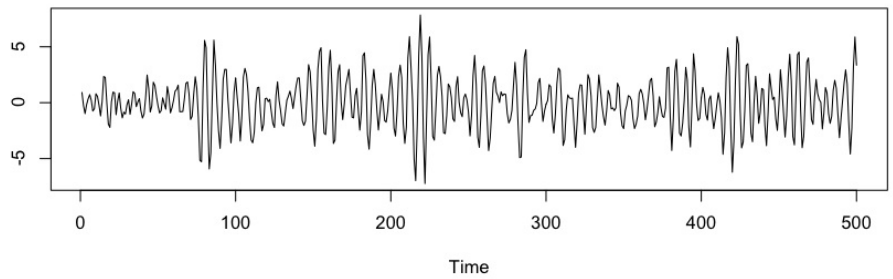
- Los comandos en R

```
> w = rnorm(550, 0, 1)
> z = filter(w,filter=c(1,-.9),
  method="recursive")
> plot.ts(z[51:550])
```

- En todo caso, la forma genérica de un modelo de este tipo, para p parámetros es

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t$$

en donde los φ_i, $i = 1, \ldots, p$ son los parámetros del modelo y ϵ representa el ruido blanco.



Suelen usarse para modelar fenómenos oscilatorios como la voz

El modelo Random Walk

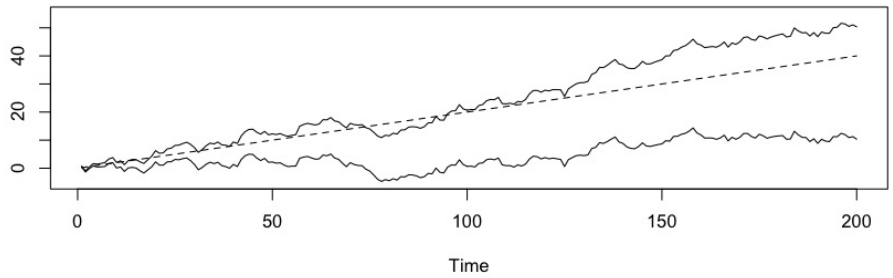
- Usado para el análisis de tendencias, según

$$x_t = \delta + x_{t-1} + w_t$$

en donde $x_0 = 0$ y w_t es ruido blanco

- δ se denomina la deriva y cuando $\delta = 0$, la expresión se denomina *random walk*
- Los comandos en R

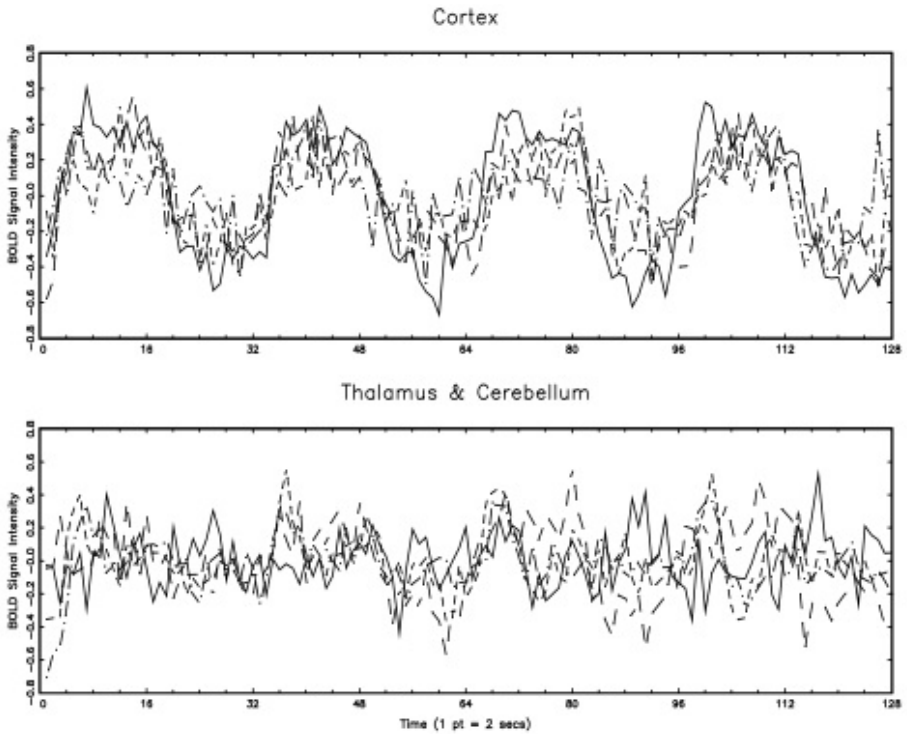
```
> set.seed(154)
> w = rnorm(200,0,1)
> x = cumsum(w)
> wd = w + 0.2
> xd = cumsum(wd)
> plot.ts(xd,ylim=c(-5,55))
> lines(x)
> lines(.2*(1:200), lty="dashed")
```



- Curva superior, $\delta = 0.2$, su tendencia queda reflejada con una recta $y = 0.2x$
- La curva inferior es un random walk (la deriva se usa para reproducir análisis bursátiles, por ejemplo)
- La idea subyacente es que los valores bursátiles no pueden predecirse, debido a las características especiales del mercado
- Inventados a partir de una moneda, un grupo de alumnos y un broker despistado

Series con un componente estacional

- Consideramos señales con variaciones periódicas y ruido blanco
- En la figura, arriba señales registradas en puntos del cortex, abajo en tálamo y cerebelo
- Cinco sujetos, sometidos a cepillado periódico de las manos (caa cepillado durante 32 segundos, y se paraba durante otros 32)



Reproduciendo la serie anterior

- Sea el modelo siguiente

$$x_t = 2\cos(2\pi t/50 + .6\pi) + w_t$$

para $t = 1, \ldots, 500$.

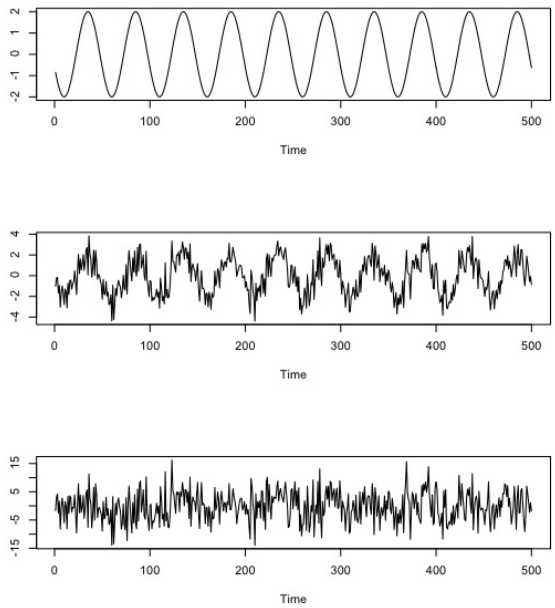
- Una onda puede escribirse como

$$A\cos(2\pi wt + \phi),$$

en donde A es la amplitud, w es la frecuencia de oscilación y ϕ es el desplazamiento de la fase.

- Los comandos en R

```
> t = 1:500
> c = 2*cos(2*pi*t/50 + .6*pi)
> w=rnorm(500,0,1)
> par(mfrow=c(3,1))
> plot.ts(c)
> plot.ts(c+w)
> plot.ts(c + 5*w)
```



La curva inferior oculta más la señal (i.e. arriba) que la curva de en medio.

El grado con el que se obscurece una curva se denomina relación señal-ruido (SNR) y es el cociente entre la amplitud de la señal y σ_w

Estimadores puntuales

Vamos a estudiar cómo obtener un valor representativo de una serie, según los modelos que acabamos de ver

Media

Vamos a definir la función media como

$$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx,$$

siempre que esta existe en donde E es la esperanza y $f_t(x)$ es la función de densidad de probabilidad para la variable x. El primer subíndice de μ_{xt} hace referencia a una serie concreta.

Media de una señal de ruido blanco

La media de una serie w_t con ruido blanco es $\mu_{wt} = E(w_t) = 0$ para todo t.

Esto es lógico ya que todos sus valores fluctúan por encima y debajo de cero.

Si suavizamos ese ruido blanco, la media no cambia ya que su comportamiento fluctuante al rededor de cero no cambia.

Media de un Random Walk

Sea un RW con deriva, según

$$x_t = \delta t + \sum_{j=1}^{t} w_j,\ t = 1, 2, \ldots$$

Como $E(w_t)$ sigue siendo cero para todo t, y δ es una cte., tenemos que

$$\mu_{xt} = E(x_t) = \delta t + \sum_{j=1}^{t} E(w_j) = \delta t,$$

es decir, una línea recta con pendiente δ (i.e. la derivada con respecto a t).

Media de una señal básica más ruido

Por último, si estudiamos la media en una serie temporal formada por una señal básica más el ruido, como la que ya hemos visto arriba,

$$x_t = 2\cos(2\pi t/50 + .6\pi) + w_t$$

para $t = 1, \ldots, 500$, vamos a tener que

$$\begin{aligned} \mu_{xt} = E(x_t) &= E[2\cos(2\pi t/50 + .6\pi) + w_t] \\ &= 2\cos(2\pi t/50 + .6\pi) + E(w_t) \\ &= 2\cos(2\pi t/50 + .6\pi) \end{aligned}$$

con lo que podemos comprobar que la media es justo la onda del coseno.

Estacionaridad en las series

- Estudiamos estimadores de parámetros de s.t.
  - Requisito: la serie temporal es estacionaria, (i.e. sus propiedades estadísticas, media, varianza, autocorrelación, etc.) no cambian con el tiempo
  - E.g. una random walk con deriva no es estacionaria
- Estacionaridad de las series es importante en el contexto de la predicción

Media, covarianza, correlación

- Si una serie es estacionaria, podemos estimar su media mediante la media de la muestra, con

$$\hat{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

- La covarianza de la serie mediante la estimamos con la de la muestra

$$\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \hat{x})(x_t - \hat{x}),$$

en donde $\hat{\gamma}(-h) = \hat{\gamma}(h)$, para $h = 0, 1, \ldots, n-1$.

- Definimos entonces la autocorrelación de una serie como sigue

$$\hat{\rho}(h) = \frac{\hat{\gamma(h)}}{\hat{\gamma(0)}}.$$

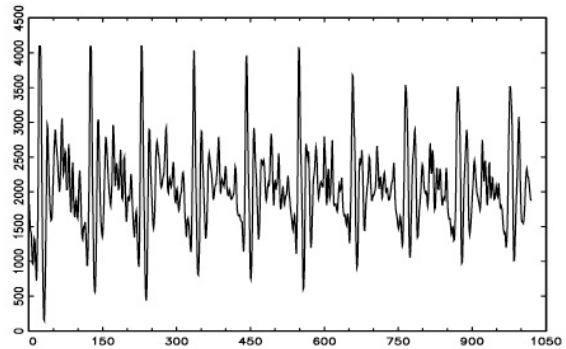
Interpretación de la correlación

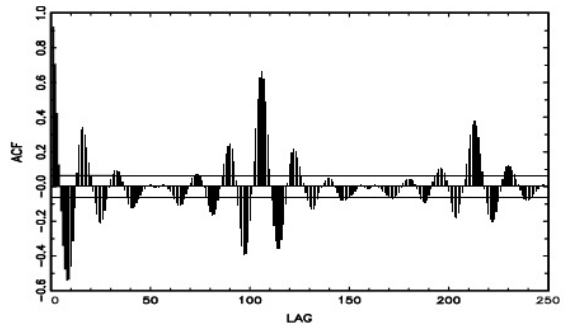
- Correlaciones, signos
  - Una correlación positiva indica persistencia, o tendencia de un sistema a permanecer en el mismo estado de una observación a otra (i.e. la probabilidad a que llueva mañana es más alta si hoy está lloviendo que si está seco).
  - En una correlación negativa, es usual ver varios valores enla serie consecutivos aumentando, seguidos posteriormente por una serie de valores negativos consecutivos.
- Correlaciones, valores según los lags
  - Los valores de ACF serán significativos, si están fuera del intervalo $\pm 2/\sqrt{n}$ en donde n es el tamaño de la muestra (es decir, el 95% de los valores de ACF[1] debe estar en ese intervalo).

[1] Autocorrelation function

Correlación a diferentes valores de h

- hay que considerar distintos valores para h ya que, si la señal no es conocida, no sabemos su periodo
- Vamos a trabajar con una muestra de voz que pronuncia una serie de a's y luego termina con unas cuantas h's. (serie de arriba)
- Si calculamos con R el ACF $(1 \leq h \leq 250)$, tenemos la serie de abajo
- La curva inferior oculta más la señal (i.e. arriba) que la curva de en medio.
- El grado con el que se obscurece una curva se denomina relación señal-ruido (SNR) y es el cociente entre la amplitud de la señal y σ_w





Correlación a diferentes valores de h, análisis

- Hay unos picos para el índice entre 106 y 109
- En el caso de las señales de voz, la distancia entre señales que se repiten se denomina *pitch* y es un parámetro fundamental en este campo
  - Dado que la señal se muestrea 10000 veces por segundo, el periodo de pitch está entre 106×10^{-5} y 109×10^{-5}
- En R, lo hacemos

```
> speech = scan("/mydata/speech.dat")
> acf(speech,250)
```

Correlación cruzada entre dos series

- Podemos también estimar la función de covarianza cruzada de una muestra para un par de series x e y
- Es la función de la covarianza cruzada de la muestra

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \overline{x})(y_t - \overline{y}),$$

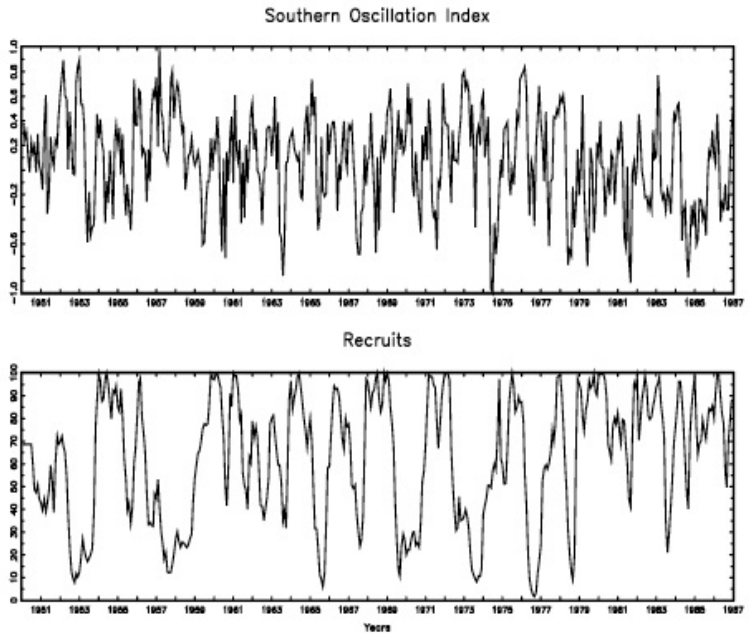
- La *sample cross-correlation fuction* es

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}.$$

- Dado que $-1 \leq \hat{\rho}_{xy}(h) \leq 1$, podemos comparar de manera adecuada la magnitud de los picos para diferentes valores de $\hat{\rho}_{xy}(h)$.

Ejemplo de uso

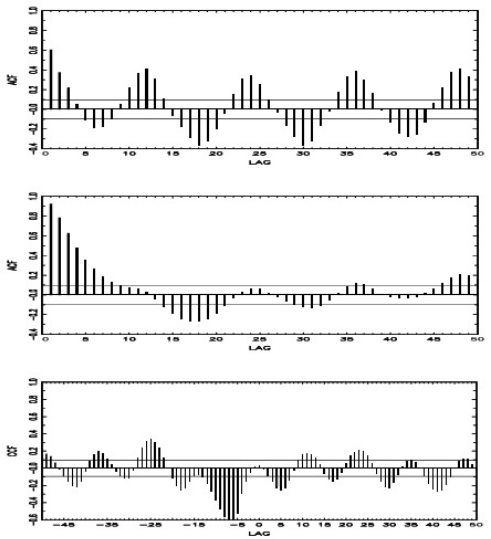
- Serie arriba 453 meses, durante los años de 1950-1987, muestra cambios en la presión del aire, con respecto a temperaturas de la superficie del mar en el Pacífico central (el mar se caliente con una periodicidad que oscila entre tres y siete años, efecto del niño)
- Serie abajo mide (mismo sitio y lugar) una determinada cantidad de capturas nuevas de peces.
- Una de las oscilaciones de esta serie parece repetirse cada doce meses
- La más lenta se repite cada 50 meses.
- ¿Podemos decir que la cantidad de capturas depende de los valores de SOI?



Ejemplo de uso, Análisis

- Para ambas series, se perciben correlaciones para valores de h de 12 unidades. Y que se repiten en sus múltiplos.
- CCF muestra un pico para un valor de $h = -6$, lo cual muestra que SOI, va por delante en seis meses de Recruitment
- Dado que el signo es negativo, ambas se mueven en direcciones diferentes
- Las líneas rectas trazadas nos sirven para medir si la serie es una serie basada en ruido blanco
- En R

```
> soi=scan("/mydata/soi.dat")
> rec=scan("/mydata/recruit.dat")
> par(mfrow=c(3,1))
> acf(soi, 50)
> acf(rec, 50)
> ccf(soi, rec, 50)
```



Conclusiones iniciales

- Las series temporales son radicalmente diferentes en su análisis a lo que hemos visto hasta ahora
- El carácter estacionario es importante
- Su interpretación se basa en la existencia de un modelo subyacente