Análisis Estadístico de Datos Climáticos

SERIES TEMPORALES 3

(Análisis espectral)

2015

Dominio temporal vs. dominio de frecuencias

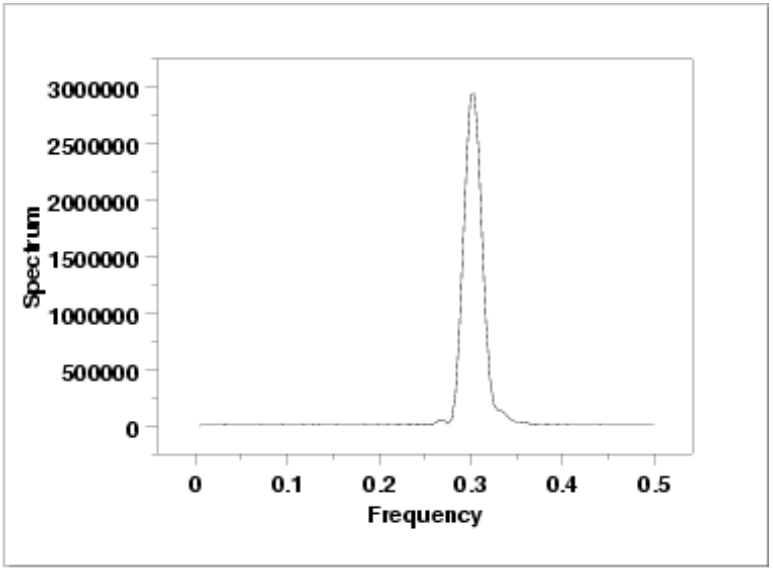
Son dos enfoques para encarar el análisis de las series temporales, aparentemente muy distintos, pero vinculados entre sí.

Los métodos en el dominio temporal buscan caracterizar las series de datos en los mismos términos en que son observados, en función del tiempo.

Por ejemplo, la media y la desviación estándar se calculan en el dominio temporal.

Una herramienta básica para caracterizar las relaciones entre los datos en el enfoque del dominio temporal es la función de autocorrelación, que ya hemos visto.

El análisis en el dominio de frecuencias representa las series en cuanto a la contribución a su variabilidad, que se tiene en diferentes *escalas temporales*, o frecuencias características.



P. ej., si tenemos una serie de tres meses de datos horarios consecutivos de temperatura del aire en una localidad, el análisis en el dominio de frecuencia debería mostrar una contribución relativamente importante en la escala diaria, o sea para la frecuencia de (1/24) h^{-1} = 0.042 h^{-1}

Veremos que el análisis en el espacio de frecuencias ocurre en el espacio definido por funciones trigonométricas (senos y cosenos).

En principio, trabajaremos con series temporales discretas, y supondremos que los datos están equi-espaciados en el tiempo, siendo Δt el intervalo entre observaciones.

(A veces supondremos Δt =1, en las unidades que corresponda.)

Llamaremos frecuencia f al número de ciclos por unidad de tiempo. P. ej., para el ciclo diario tenemos la frecuencia $f = (1/24)\ h^{-1} = 0.042\ h^{-1} = 0.042$ ciclos /hora.

Si T es el período, es f = 1/T.

Llamaremos frecuencia angular a $\omega = 2*\pi*f$ (que se mide en radianes por unidad de tiempo).

El espectro de potencia

Tanto los procesos determinísticos como estocásticos pueden, en principio, ser caracterizados por una función f de la frecuencia (en vez del tiempo). Esta función S(f) se llama espectro de potencia o densidad espectral (o simplemente espectro).

Así, una serie con variabilidad temporal muy irregular tiene un espectro suave y continuo, indicando que todas las frecuencias en un cierto rango o banda de frecuencias son excitadas por ese proceso.

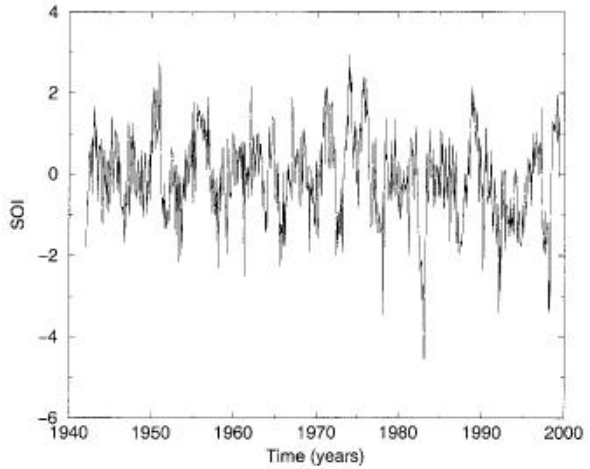
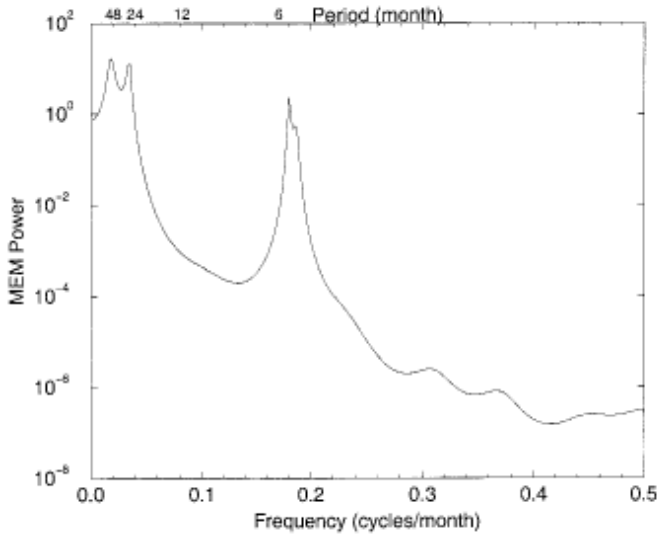
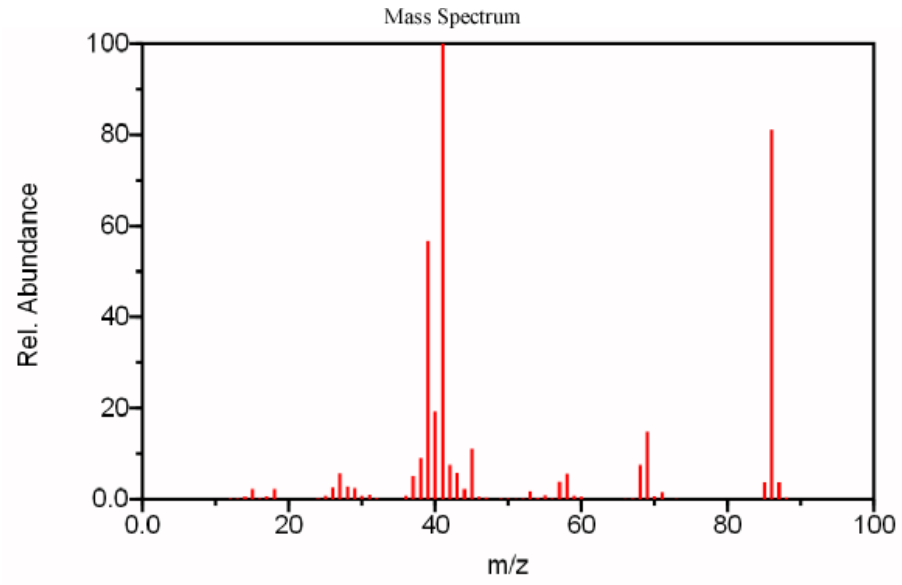


Figure 2. Variations of the Southern Oscillation Index (SOI) between January 1942 and June 1999. Time on the abscissa is in calendar years, and SOI on the ordinate is normalized by its standard deviation.



Por el contrario, un proceso puramente periódico o cuasi-periódico, o superposición de ellos, queda descripto por una sola "línea" o un número finito de "líneas" en el dominio de frecuencias.



Entre estos dos extremos, los procesos determinísticos no lineales caóticos pueden presentar picos superpuestos a un fondo continuo y con muchas ondulaciones.

En la práctica, la distinción entre procesos determinísticos caóticos y procesos aleatorios a través del análisis espectral puede ser delicada, debido a que las series climáticas son cortas y ruidosas.

Los métodos espectrales intentan estimar la parte continua del espectro o las “líneas”, o ambos.

A menudo, las líneas se estiman a partir de datos discretos y ruidosos, y aparecen como “picos” más o menos agudos.

La estimación e interpretación dinámica de estos últimos, cuando aparecen, son a menudo más robustas y de fácil comprensión que la naturaleza de los procesos que podrían generar el fondo de banda ancha, sea determinístico o estocástico.

Frecuencia más alta (Nyquist) y más baja (fundamental) asociada a una serie discreta

Previo:

Si k y t son enteros, $\cos[(\omega + k\pi)t] = \cos \omega t$, si k es par.

$= \cos(\pi - \omega)t$ si k es impar.

Entonces, a la variación a una frecuencia angular mayor que π, le corresponde una variación idéntica a una frecuencia angular en [0, π], por lo que alcanza con considerar frecuencias angulares en ese intervalo, o frecuencias f en [0, ½].

Supongamos que tenemos una serie discreta de observaciones, espaciadas un intervalo Δt entre sí.

La frecuencia angular $\omega = \pi/ \Delta t$ se llama frecuencia de Nyquist asociada a la serie. Lo mismo vale para la frecuencia $f = 1/ (2 \Delta t)$.

Frecuencia más alta (Nyquist) y más baja (fundamental) asociada a una serie discreta

La frecuencia de Nyquist de una serie discreta de datos es la mayor frecuencia sobre la que se puede obtener información.

Ej.: supongamos que en una localidad se toman medidas de temperatura todos los días a mediodía (Δt = 1 día). Es claro que estas observaciones no nos informarán nada sobre la variabilidad de la temperatura *dentro* de un día. En particular, no nos dirán nada sobre si las noches son más cálidas o frías que los días.

En este caso, $\omega_{Ny} = \pi$ radianes por día o f_N = ½ ciclo por día (o 1 ciclo cada 2 días), o T = 2 días.

Estas frecuencias son más bajas que las frecuencias correspondientes a la variabilidad *dentro de 1 día*. P. ej., la variabilidad correspondiente a un período T = 1 día tiene una frecuencia de $\omega = 2\pi$ radianes por día, o f = 1 ciclo por día (o sea, Δt = ½ día).

Para obtener información sobre la variabilidad dentro de un día, debemos aumentar la frecuencia de medidas, tomando 2 o más observaciones por día.

Frecuencia más alta (Nyquist) y más baja (fundamental)

En el otro extremo del espectro, existe una frecuencia por debajo de la cual no tiene sentido tratar de obtener información a partir de un conjunto de datos dado.

P.ej., si tenemos 6 meses de datos de temperatura, de invierno y primavera, no se podría decidir si los veranos son más cálidos que los inviernos. Sin embargo, con un año de datos, se podría discernir eso.

Con un año de datos, la frecuencia más baja que podemos ajustar es de 1 ciclo por año.

Si tenemos observaciones semanales, un año de datos son N = 52, con $\Delta t = 1$ semana y la frecuencia más baja es $1/ (N \Delta t)$ ciclos por semana.

(Aquí N es la longitud de la serie.)

Esta frecuencia ($1/ (N \Delta t)$) es llamada a veces frecuencia fundamental de Fourier.

Los múltiplos de esta frecuencia: $k / (N \Delta t)$ con $k=1,2,\ldots,N/2$,

se llaman armónicos y los re-encontraremos enseguida.

Estimaciones no paramétricas del espectro

Transformada discreta de Fourier (DFT)

Cualquier serie discreta Y_t con N puntos se puede representar exactamente como una función armónica, o sea como una combinación lineal de senos y cosenos de las frecuencias armónicas.
Suponemos, para simplificar, que N es par, y que $\Delta t = 1$.

$$Y_t = \bar{y} + \sum_{k=1}^{N/2} \left\{ A_k \cos\left(\frac{2\pi kt}{N}\right) + B_k \operatorname{sen}\left(\frac{2\pi kt}{N}\right) \right\} =$$

$$\bar{y} + \sum_{k=1}^{N/2} \left[C_k \cos\left(\frac{2\pi kt}{N} - \theta_k\right) \right]$$

t =1,2,…,N

Notar que al variar k se obtienen funciones que cubren k ciclos en todo el intervalo.

Se obtiene, para k = 1,2,….,N/2 :

$$A_k = \frac{2}{N} \sum_{t=1}^{N} Y_t \cos\left(\frac{2\pi kt}{N}\right)$$

$$B_k = \frac{2}{N} \sum_{t=1}^{N} Y_t \operatorname{sen}\left(\frac{2\pi kt}{N}\right)$$

$$C_k = \sqrt{A_k^2 + B_k^2}$$

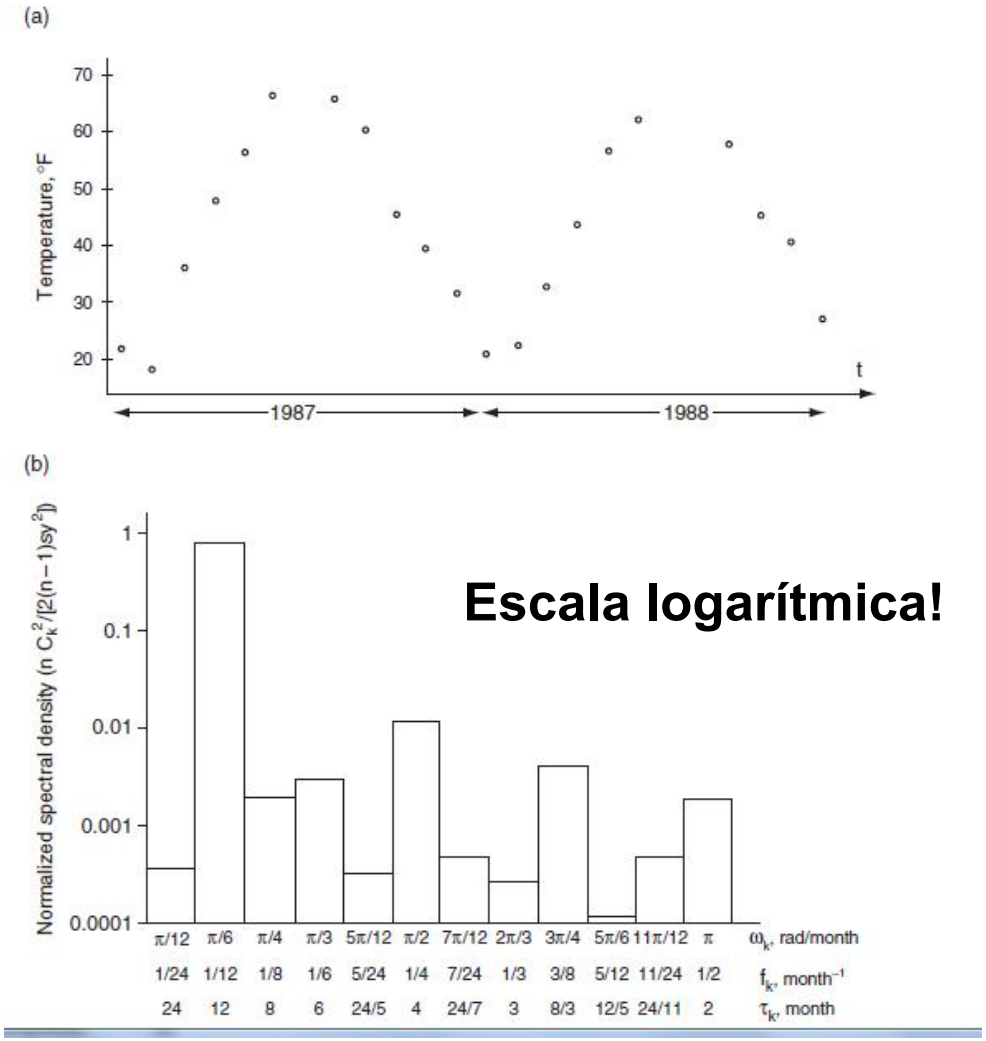
Se llama <u>transformada</u> discreta porque pasa de los Y_t a los C_k y Φ_k (o A_k y B_k).

Notar que las frecuencias armónicas son un conjunto discreto y finito, y dependen de N (la longitud de la serie), y no tienen por qué coincidir con frecuencias que tengan significado físico.
Por eso, si se conoce alguna frecuencia natural (ej., asociada al ciclo diario o anual) conviene tomar un N tal que esa sea una de los armónicos.

Ejemplo con N=24 (pocos datos!!)
Wilks: p. 384-385

TABLE 8.6 Average monthly temperatures, °F, at Ithaca, New York, for 1987–1988, and their discrete Fourier transform.

Month	1987	1988	k	τ_k, months	A_k	B_k	C_k
1	21.4	20.6	1	24.00	−0.14	0.44	0.46
2	17.9	22.5	2	12.00	−23.76	−2.20	23.86
3	35.9	32.9	3	8.00	−0.99	0.39	1.06
4	47.7	43.6	4	6.00	−0.46	−1.25	1.33
5	56.4	56.5	5	4.80	−0.02	−0.43	0.43
6	66.3	61.9	6	4.00	−1.49	−2.15	2.62
7	70.9	71.6	7	3.43	−0.53	−0.07	0.53
8	65.8	69.9	8	3.00	−0.34	−0.21	0.40
9	60.1	57.9	9	2.67	1.56	0.07	1.56
10	45.4	45.2	10	2.40	0.13	0.22	0.26
11	39.5	40.5	11	2.18	0.52	0.11	0.53
12	31.3	26.7	12	2.00	0.79	—	0.79



Escala logarítmica!

Se grafica C_k^2 normalizado para k = 1,…,12

El espectro que se obtiene se llama espectro de ”línea”.

Cada valor C_k^2 es proporcional a la parte de varianza de la serie Y_t a la que contribuye la frecuencia $f_k = k/(N\,\Delta t)$.

Más precisamente, se tiene el teorema de Parseval:

$$\frac{1}{N}\sum_{t=1}^{N}(y_t-\bar{y})^2 = \sum_{k=1}^{(N/2)-1} C_k^2/2 + A_{N/2}^2$$

****

Si la serie presenta una trend (creciente o decreciente), se recomienda removerla.
De lo contrario, puede aparecer como una baja frecuencia en el espectro, y podría ser dominante respecto de otras variaciones que se estén buscando.

Fast Fourier Transform (FFT)

Las ecuaciones dadas no son la forma más eficiente de calcular A_k y B_k, ya que presentan muchas redundancias.

Existe la FFT que permite ahorrar mucho tiempo de cálculo, especialmente cuando N es alto.
El uso de la FFT permite hacer los cálculos aprox. $N/\log_2 N$ más rápido (15 veces para N=100, 750 veces para N=10000).

Habitualmente, la DFT y FFT se expresan utilizando números complejos; p. ej:

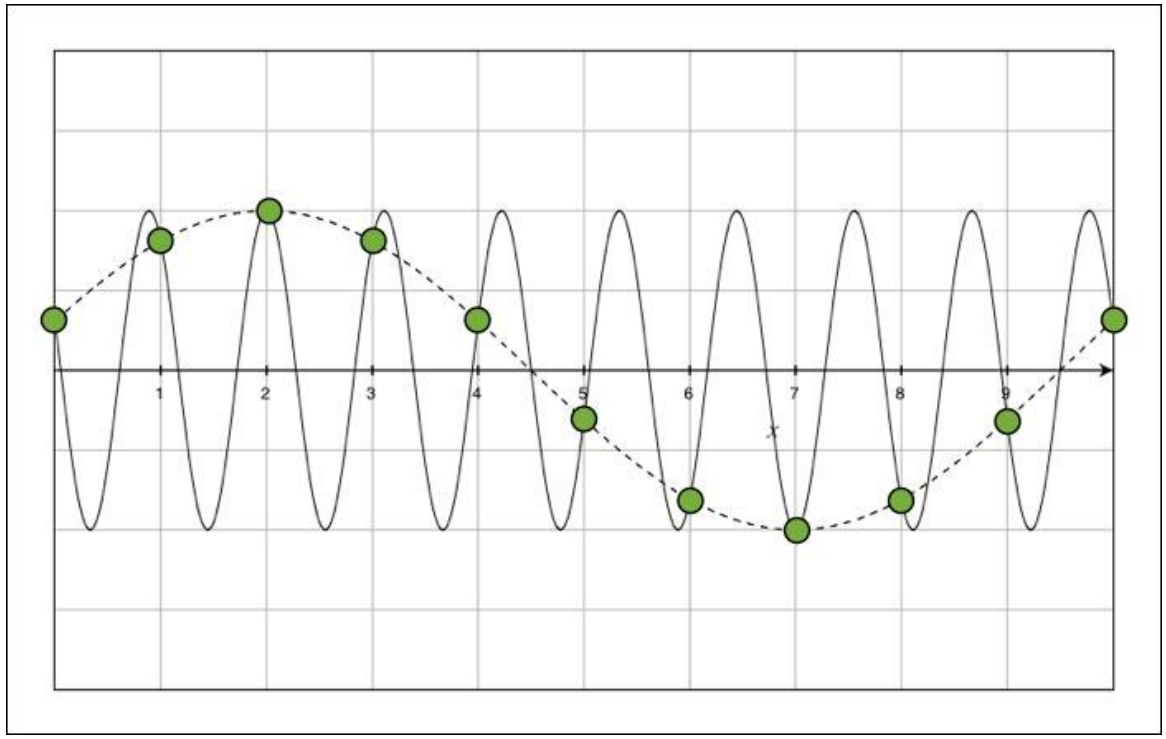
$$Y_t = \bar{y} + \sum_{k=1}^{N/2} H_k \, e^{i\,(2\pi k/n)\,t}$$

siendo $H_k = A_k + iB_k$

En Matlab, fft.m

Aliasing
Wilks 388-389

En la figura, los puntos son los datos observados. Si se hace un ajuste armónico, se obtiene la curva punteada (de “baja” frecuencia).
Pero, podría ocurrir que el proceso real sea el de curva llena (de “alta” frecuencia).
La alta frecuencia es mayor que la frecuencia de Nyquist (1/(2Δt)), o sea que las oscilaciones son muy rápidas para poder ser bien muestreadas con esa resolución temporal.



Como ya vimos, las oscilaciones que se pueden resolver deben cumplir $f \leq 1/(2\Delta t)$, o sea $T \geq 2\Delta t$.
Esto determina el Δt con el que se debe muestrear según los períodos que se quieran identificar.

Pero además, si no se hace eso, la variabilidad en frecuencias mayores que la de Nyquist no se pierde, sino que se agregan incorrectamente a frecuencias en el rango [0, 1/(2Δt)].

Tomando Δt=1,

si $f_A > 1/2$, esta frecuencia tendrá un "alias" en otra frecuencia f (con $0< f \leq 1/2$), tal que

$$f_A = j \pm f \qquad \text{(siendo j un entero cualquiera).}$$

Si k y t son enteros, $e^{i2\pi ft} = e^{i\,2\pi\,(f \pm k)t}$

El periodograma

A veces se da este nombre al resultado de la DFT que recién vimos, que asocia a los valores Y_t (t=1,2,…,N), los valores C_k cada uno a su vez asociado a la frecuencia k/(N Δt), (k=1,2,….,N/2).

Aquí vamos a usar la palabra periodograma para el caso en que se estima la función de densidad espectral S(f) para un conjunto continuo de frecuencias f ϵ [0, 1/2Δt].

Dicha estimación se basa en el teorema de Wiener-Khinchine (o de representación espectral), que establece que la función de autocovarianza y la función de densidad espectral son transformadas de Fourier una de la otra. En particular, se tiene:

$$S(f) = \Delta t \sum_{k=-\infty}^{\infty} \gamma(k)\, e^{-i\,2\pi\, f\, k\, \Delta t} = \Delta t \left[\gamma(0) + 2 \sum_{k=1}^{\infty} \gamma(k) \cos(2\pi\, f\, k\, \Delta t) \right]$$

A partir de allí y después de algunos cálculos (que implican truncar la suma entre - (N-1) y (N-1) y sustituir $\gamma(k)$ por su estimación) , se obtiene la estimación:

$$\hat{S}^{(p)}(f) = \frac{\Delta t}{N}\left|\sum_{t=1}^{N} Y_t\, e^{-i\,2\pi f\, t\,\Delta t}\right|^2$$

(Se llama periodograma aunque se estima una función de la frecuencia.

Observar que ahora la función $\hat{S}^{(p)}(f)$ está definida para una variable continua $f \in [0, ½]$

Además, para los valores de f que coinciden con las frecuencias armónicas (k/N), esta estimación coincide (a menos de un factor constante) con la ya obtenida para la DFT.

Propiedades del periodograma

1) El periodograma es asintóticamente insesgado, o sea:

$$E(\hat{S}^{p}(f)) \rightarrow S(f) \qquad \text{cuando } N \rightarrow \infty$$

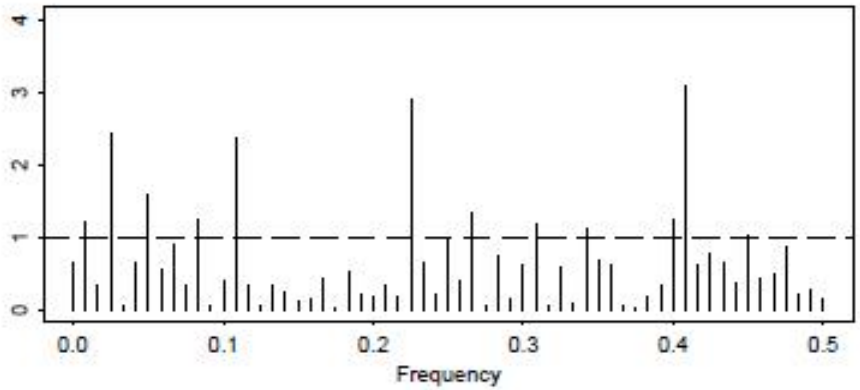
pero pueden necesitarse valores muy altos de N para lograr una aproximación razonable. Es decir que para valores de N habituales, el sesgo puede ser considerable.

2) La varianza del estimador $\hat{S}^{p}(f)$ **no tiende a 0 cuando** $N \rightarrow \infty$

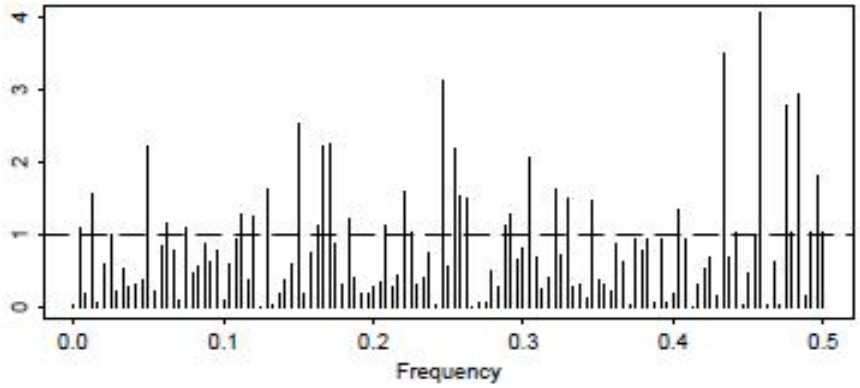
(Es más, la varianza no depende de N.)

O sea que el estimador no es consistente, lo cual es una propiedad muy poco deseable, que veremos cómo trata de solucionarse.
Una explicación intuitiva de la falta de consistencia del periodograma es que las autocovarianzas de mayor orden que se usan en su estimación, están muy mal estimadas, cualquiera sea N.

Estimación del periodograma



Ruido blanco: N = 120



Ruido blanco: N = 240

La varianza no disminuye al aumentar N

Obtención de estimadores alternativos consistentes del espectro

Mencionamos dos formas de estimar que tratan de disminuir el sesgo y evitar la falta de consistencia del periodograma:

1) Blackman-Tukey: trunca la estimación de la autocovarianza para utilizar únicamente aquellos valores que están mejor estimados.

2) Welch: calcular periodogramas de segmentos mas cortos de la serie y promediar en el dominio de frecuencias.

En ambos casos el efecto final es que reducimos la varianza del estimador pero a costa de aumentar su sesgo (menos resolución).

Blackman y Tukey: se utiliza la siguiente expresión, modificada de la ya vista para el espectro.

$$\hat{S}(f) = \Delta t \left[\lambda_0 \, c_o + 2 \sum_{k=1}^{M} \lambda_k \, c_k \cos(2 \pi \, f \, k \, \Delta t) \right]$$

donde los $\{c_k\}$ son estimadores de las covarianzas $\{\gamma(k)\}$, los $\{\lambda_k\}$ son coeficientes ("pesos") llamados "lag windows", y M < N, se llama punto de truncamiento.

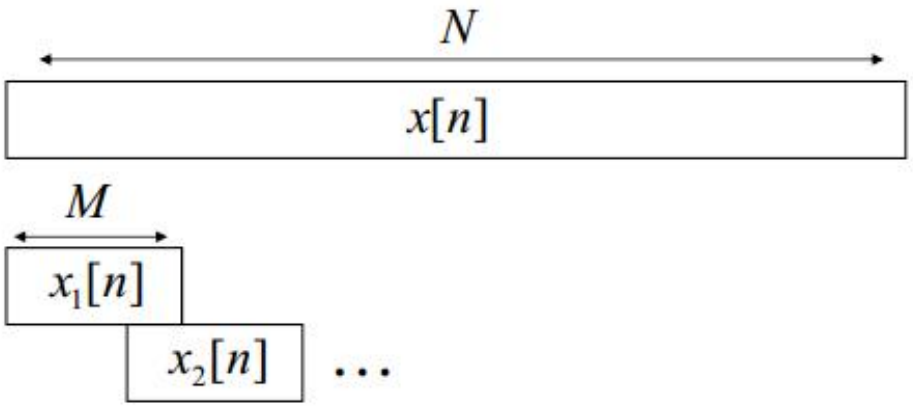
Como se ve, los valores de los c_k con M < k < N no se usan (justamente esas son las peores estimaciones, por ser calculados con cada vez menos términos, cuando k crece).

Hay varias "ventanas" que se utilizan (Tukey, Parzen, Bartlett, etc), con diferentes propiedades.

En cuanto a la elección de M, hay un compromiso entre sesgo (o resolución) y varianza: cuanto más pequeño sea M, menor será la varianza del estimador, pero mayor será el sesgo. Si M es muy pequeño, se suavizará demasiado detalles importantes de S(f), pero si M es muy grande, el comportamiento de S será más errático, parecido al del periodograma puro.

La idea fundamental del estimador de Welch o periodograma promediado consiste en dividir la serie original de N muestras en K registros de M < N muestras, calcular los periodogramas de cada uno de los segmentos y promediarlos. En la estimación de cada periodograma también se usan ventanas.

Los registros pueden tener un solapamiento



$$\hat{S}_x^W(\omega) = \frac{1}{K}\sum_{j=1}^{K} \hat{S}_{x_j}^p(\omega)$$

donde $\hat{S}_{x_j}^p(\omega)$ es el periodograma del segmento j-ésimo

Similares comentarios sobre el valor de M.
El solapamiento permite mejorar la reducción de la varianza.
Veremos luego una implementación en Matlab.

Estimación paramétrica del espectro

Veremos los casos del ruido blanco y ruido rojo (o sea AR(1)).

Para el caso de un ruido blanco, Z_t, como todas las autocovarianzas a partir del orden 1 son nulas se tiene que

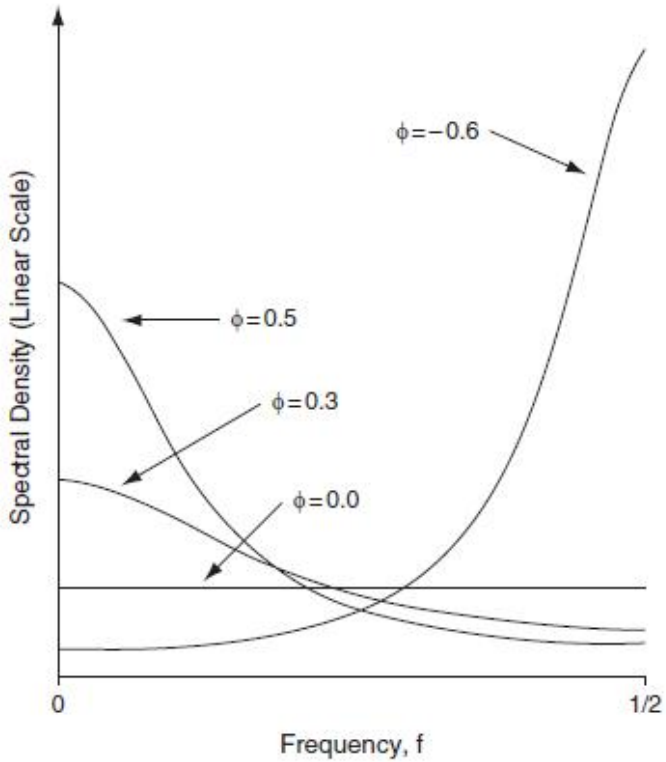
$S_Z(f)$ =constante $(0 \leq f \leq ½)$

(todas las frecuencias contribuyen igualmente a la varianza).

Para el caso de un ruido rojo:

$X_t = \Phi X_{t-1} + Z_t$, ($|\Phi| < 1$ y $Var(Z_t) = \sigma_\varepsilon^2$), se tiene:

$$S(f) = \frac{4\sigma_\varepsilon^2/n}{1 + \phi^2 - 2\phi\cos(2\pi f)}, \quad 0 \leq f \leq 1/2.$$



Como caso particular, para $\Phi = 0$, se tiene el caso del ruido blanco.

Intervalos de confianza para la estimación del espectro

Los intervalos de confianza para los C_k^2 (obtenidos por DFT) son muy grandes debido a que, adecuadamente escalados, su distribución es proporcional a χ_2^2. (El número de grados de libertad (2) es bajo.)

Cuando se usa la estimación del espectro suavizado (Blackman y Tukey), se aumenta el número de grados de libertad υ y se pueden obtener intervalos de confianza más pequeños.

$$\Pr\left[\frac{\nu C_k^2}{\chi_\nu^2\left(1-\frac{\alpha}{2}\right)} < S(f_k) \leq \frac{\nu C_k^2}{\chi_\nu^2\left(\frac{\alpha}{2}\right)}\right] = 1-\alpha,$$

P. ej., para $\alpha = 0.05$, tenemos un intervalo de confianza del 95%

Prueba de hipótesis para el espectro

Puede interesar comparar los valores de Ck2 que se obtengan con los que se obtendrían con un modelo paramétrico que se ajuste a la serie (ej. ruido blanco o rojo).

Hay 2 casos:

1) el valor de C_k^2 para una frecuencia f_k elegida de antemano, y

2) el máximo valor de todo los C_k^2

En ambos casos, para variables climáticas es razonable utilizar como modelo a ajustar un AR(1).

En el caso 1), si $S_0(f_k)$ es el espectro para el modelo AR(1), la hipótesis nula se rechaza al nivel α (prueba de un extremo) si

$$C_k^2 \geq \frac{S_0(f_k)}{\nu}\chi_\nu^2(1-\alpha),$$

siendo u = 2

En cambio, si se trata del máximo, que no es elegido de antemano, sino que depende de los datos con los que hace el test, en realidad estamos ante K tests Independientes.

Entonces, si α* es el nivel de la prueba (para el máximo) y α es el nivel (para un valor espectral elegido de antemano), hay una relación entre ellos:

$$\alpha^* = 1 - (1 - \alpha)^K$$

Se debe elegir el valor α (que es menor que α*) para el test.

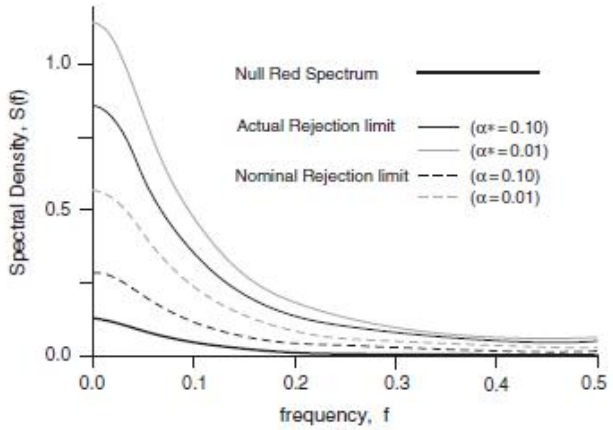


FIGURE 8.24 Red spectrum for $\phi_1 = 0.6$, $\sigma_z^2 = 1.0$, and $n = 200$ (heavy curve) with minimum values necessary to conclude that the largest of $K = 100$ periodogram estimates is significantly larger (lighter solid curves) at the 0.10 (black) and 0.01 (grey) levels. Dashed curves show erroneous minimum values resulting when test multiplicity is not accounted for.