Remuestreo en Problemas de Estadística Ambiental

Monitorización de variables ambientales y epidemiológicas

Andrés M. Alonso Fernández

Departamento de Estadística
Universidad Carlos III de Madrid

Cádiz - 19 de julio de 2005

Estructura

1. Introducción.

2. Métodos de remuestreo para datos i.i.d.

3. Métodos de remuestreo para series temporales.
   - Métodos basados en modelos.
   - Métodos no basados en modelos.

4. Aplicación a la monitorización de variables ambientales y epidemiológicas.

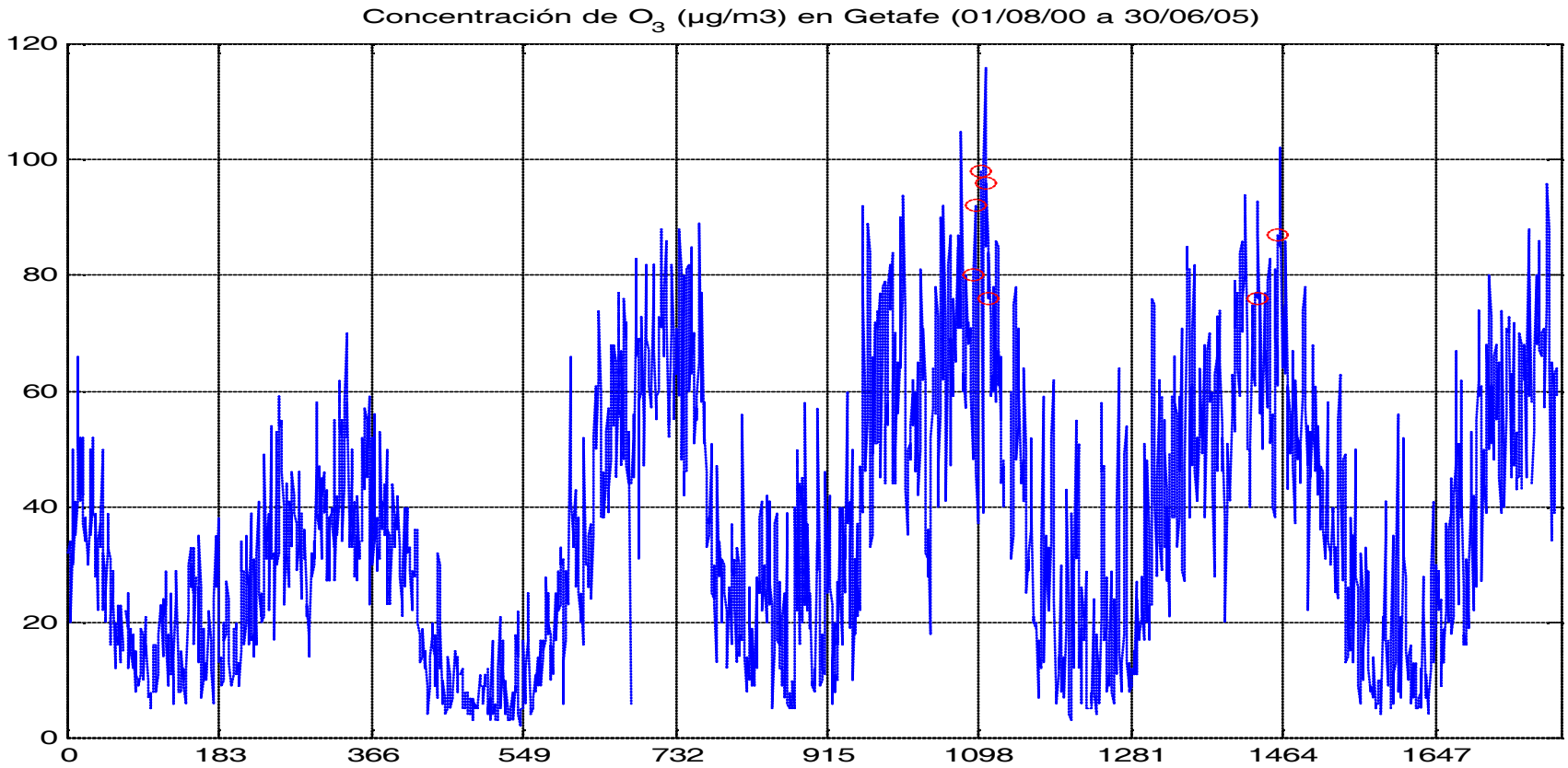
5. Conclusiones y Extensiones.

Introducción

Problema: Sean $\boldsymbol{X} = (X_1, \ldots, X_N)$ observaciones generadas por un modelo $\mathcal{P}$, y sea $T(\boldsymbol{X})$ el estadístico de interés que estima al parámetro poblacional θ. Se desea conocer:

- Sesgo: $b_T = \mathrm{E}\left[T(\boldsymbol{X})\right] - \theta$,
- Varianza: v_T,
- Distribución: $\mathcal{L}(T, \mathcal{P})$.

Niveles de ozono en el municipio de Getafe:



Fuente: Dirección General de Calidad y Evaluación Ambiental.

Niveles de ozono en el municipio de Getafe:

- ¿Cuál será el nivel medio de ozono la "siguiente" semana/mes/año?
- ¿Cuál es la probabilidad de que se alcance un nivel de información (180) en la "siguiente" semana/mes/año?
- ¿Cuál es la probabilidad de que se alcance un nivel de de alerta (240) en la "siguiente" semana/mes/año?
- ¿Cuál es el número medio de superaciones mensuales/anuales?

Niveles de ozono en el municipio de Getafe:

Superaciones en los años 2003 y 2004:

Fecha	Hora	Concentración de Ozono µg/m3
28/07/2003	17:00	182
29/07/2003	17:00	183
06/08/2003	14:00	181
06/08/2003	15:00	181
06/08/2003	16:00	193
06/08/2003	18:00	184
12/08/2003	15:00	191
12/08/2003	16:00	192
12/08/2003	17:00	190
13/08/2003	15:00	189
13/08/2003	16:00	189
14/08/2003	14:00	187
14/08/2003	15:00	192
14/08/2003	16:00	196

Fecha	Hora	Concentración de Ozono µg/m3
02/07/2004	15:00	186
02/07/2004	16:00	185
29/07/2004	17:00	182
29/07/2004	18:00	187

Estructura

1. Introducción.
2. Métodos de remuestreo para datos i.i.d.
3. Métodos de remuestreo para series temporales.
   - Métodos basados en modelos.
   - Métodos no basados en modelos.
4. Aplicación a la monitorización de variables ambientales y epidemiológicas.
5. Conclusiones y Extensiones.

Métodos de remuestreo para datos i.i.d.

Jackknife: Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ una muestra de tamaño N y sea $T_N = T_N(\boldsymbol{X})$ un estimador de θ.

En las *muestras jackknife* se excluye una observación de $\boldsymbol{X}$ cada vez: $\boldsymbol{X}_{(i)} = (X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$, para $i = 1, 2, \ldots, N$, y $T_{N-1,i} = T_{N-1}\left(\boldsymbol{X}_{(i)}\right)$ es la i-ésima *réplica jackknife*.

Estimador jackknife de sesgo:

$$b_{Jack} = (N-1)(\bar{T}_N - T_N)$$

Estimador jackknife de la varianza:

$$v_{Jack} = \frac{N-1}{N} \sum_{i=1}^{N} \left(T_{N-1,i} - \bar{T}_N\right)^2,$$

donde $\bar{T}_N = N^{-1} \sum_{i=1}^{N} T_{N-1,i}$

Jackknife

Esquema del procedimiento de remuestreo:

$$(X_1,\ldots,X_N) \Rrightarrow \left\{\begin{array}{ll} \boldsymbol{X}\backslash X_1 & \Rrightarrow T_{N-1,1} \\ \boldsymbol{X}\backslash X_2 & \Rrightarrow T_{N-1,2} \\ \vdots & \vdots \\ \boldsymbol{X}\backslash X_N & \Rrightarrow T_{N-1,N} \end{array}\right. \Rrightarrow \left\{\begin{array}{l} b_{Jack} = (N-1)(\bar{T}_N - T_N) \\ v_{Jack} = \frac{N-1}{N}\sum_{i=1}^{N}\left(T_{N-1,i} - \bar{T}_N\right)^2 \end{array}\right.$$

Usos posibles:

- Reducción del sesgo: $T_{Jack} = T_N - b_{Jack}$.
- Si $\sqrt{N}(T_N - \theta)$ es asintóticamente normal obtenemos la estimación de la varianza asintótica mediante v_{Jack}.

Ejemplo 1: Sean $X_1, X_2, \ldots, X_N$ observaciones i.i.d. $\mathcal{N}(\mu, \sigma^2)$ y $T_N = \bar{X}$ el estadístico de interés. En este caso, sabemos que:

$$b_T = 0, \quad v_T = \frac{\sigma^2}{N}, \quad \mathcal{L}(T) = \mathcal{N}(\mu, \frac{\sigma^2}{N}).$$

Resultados con los estimadores jackknife: $N = 100$, $\mu = 0$ y $\sigma^2 = 1$.

Que trabaje Matlab ...

```
> x = randn(100, 1);
> [s, b, v] = jackknife(x, 'mean')

s =
   -0.1270

b =
  2.7478e-015

v =
    0.0089
```

d-jackknife: Sea $S_{n,r}$ los subconjuntos de $\{1, 2, \ldots, N\}$ de tamaño r. Para cualquier $s = \{i_1, i_2, \ldots, i_r\} \in S_{N,r}$ obtenemos la *réplica d-jackknife* por $T_{r,s^c} = T_r(X_{i_1}, X_{i_2}, \ldots, X_{i_r})$.

Estimador d-jackknife de la varianza:

$$v_{Jack-d} = \frac{r}{dC} \sum\nolimits_{s \in S_{N,r}} \left(T_{r,s^c} - C^{-1} \sum\nolimits_{s \in S_{N,r}} T_{r,s^c} \right)^2,$$

donde $C = \binom{N}{d}$.

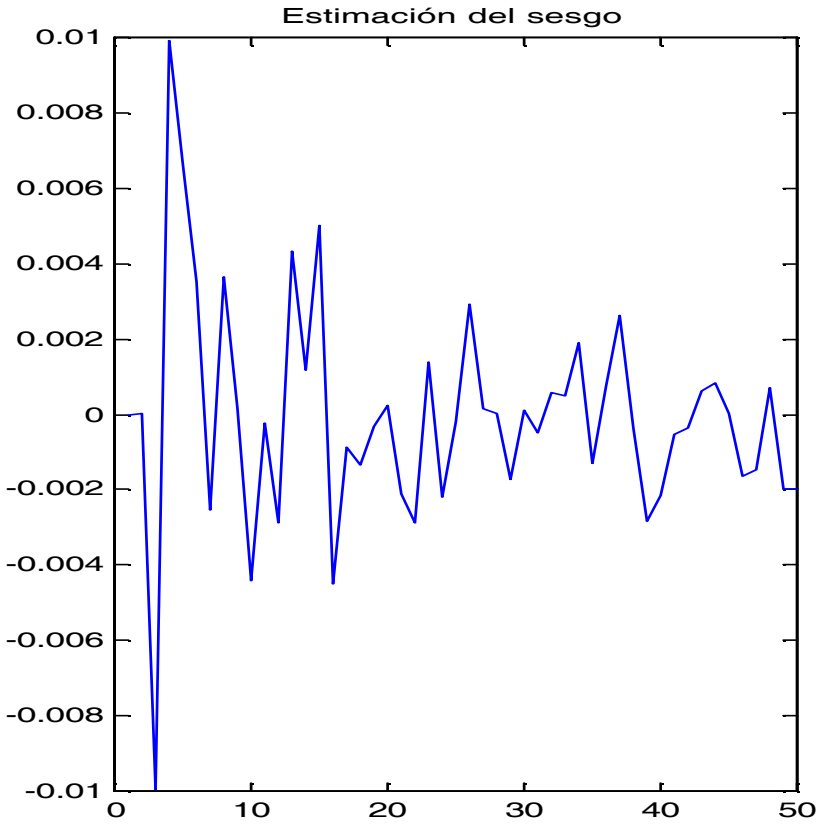
d-jackknife como estimador de la distribución de T_N: Sea $H_n(x) = \Pr\{\sqrt{n}(T_n - \theta) \leq x\}$ la distribución a estimar. Se define el *histograma jackknife* como:

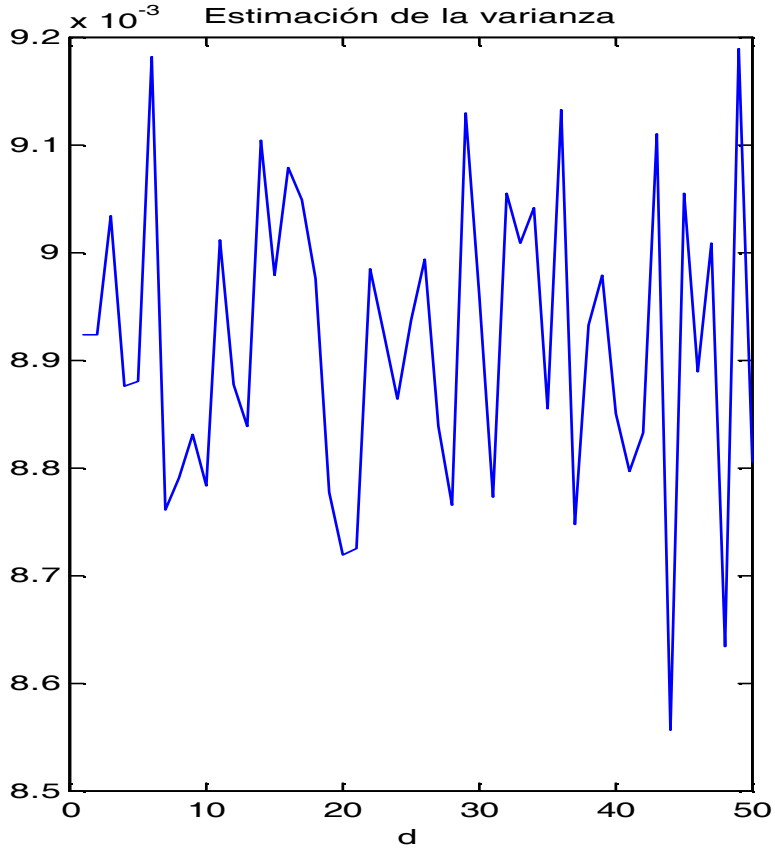
$$H_{Jack}(x) = \frac{1}{C} \sum\nolimits_{s \in S_{N,d}} I\left(\sqrt{Nr/d}(T_{r,s} - T_N) \leq x \right).$$

Referencias básicas: Shao y Wu 1989, Wu 1990 y Shao y Tu 1995

Ejemplo 1 (continuación): Estadístico = Media.

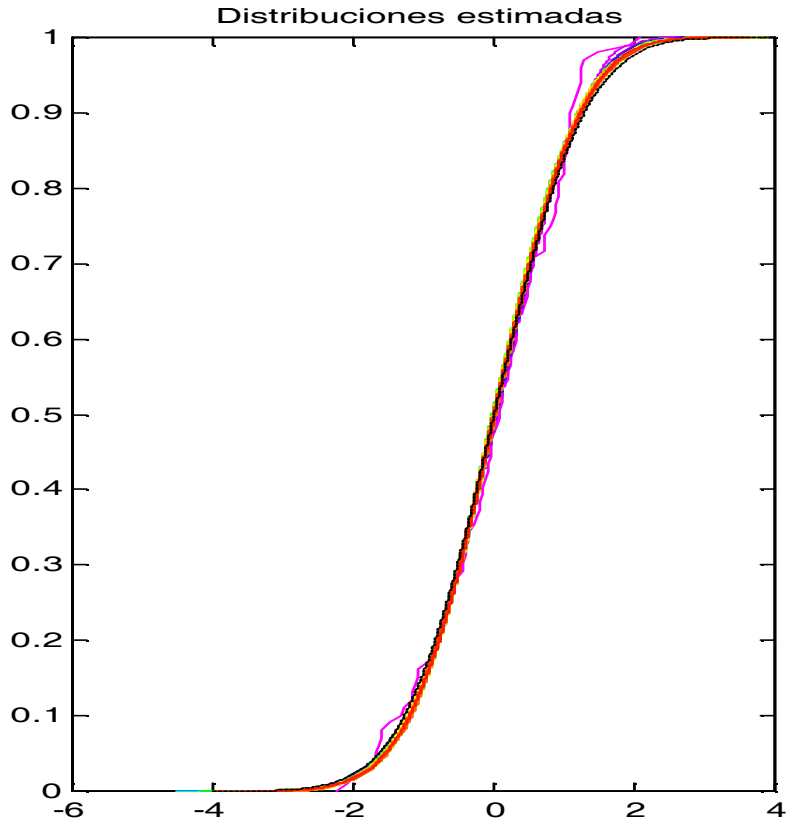
Que trabaje Matlab ...

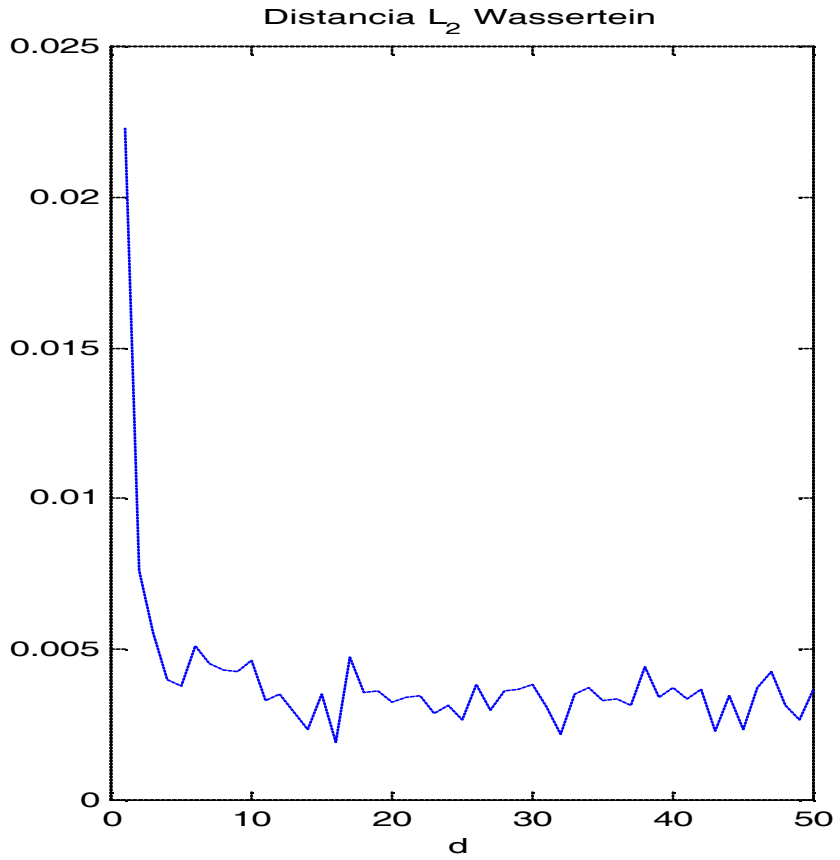




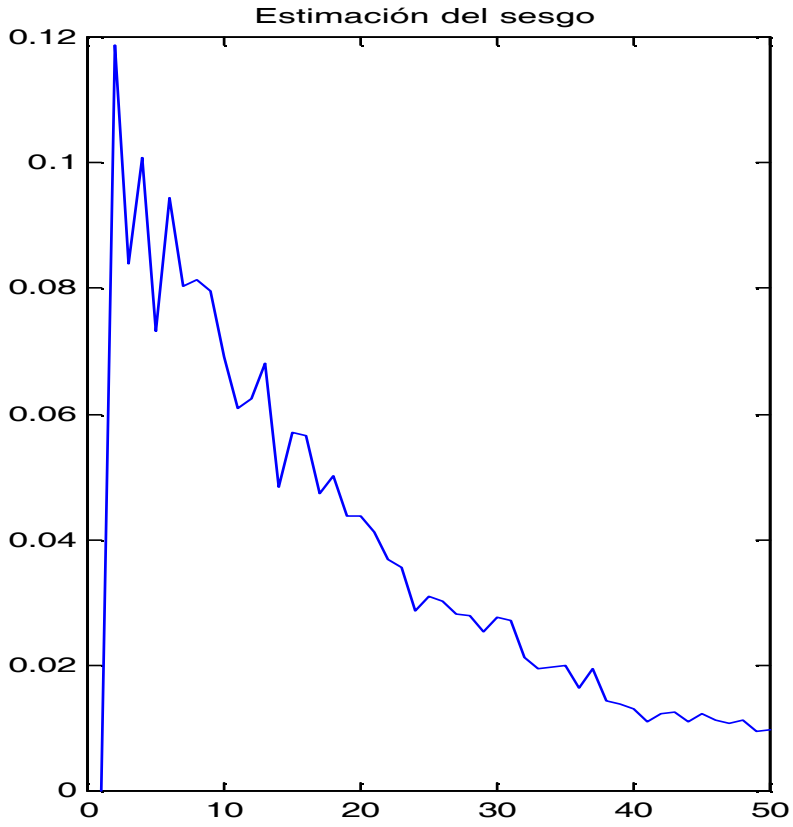
Ejemplo 1 (continuación): Estadístico = Media.

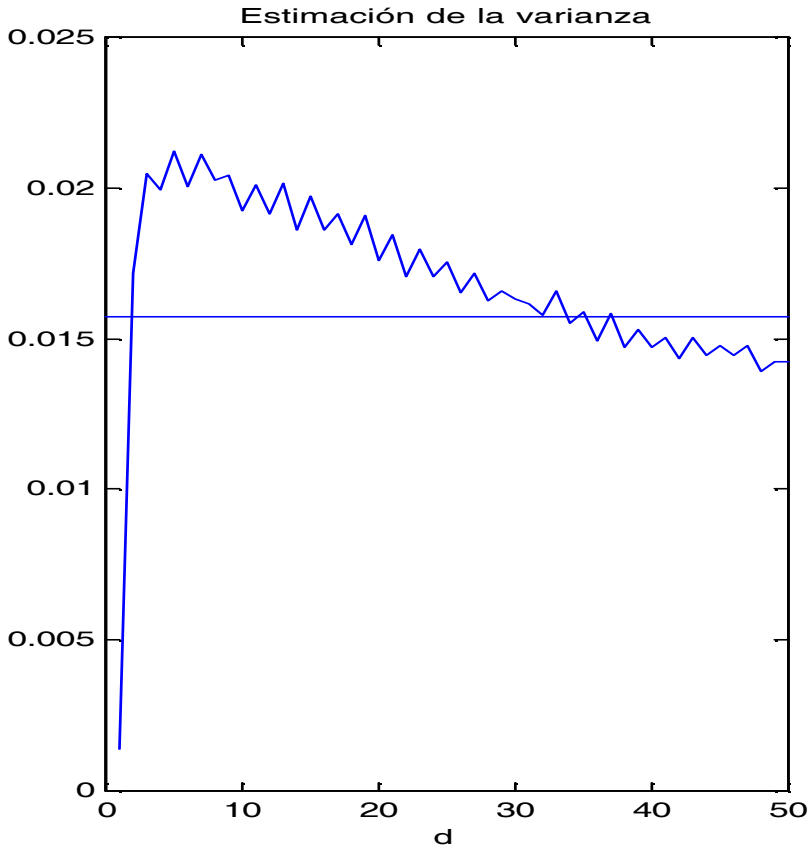
Que trabaje Matlab ...



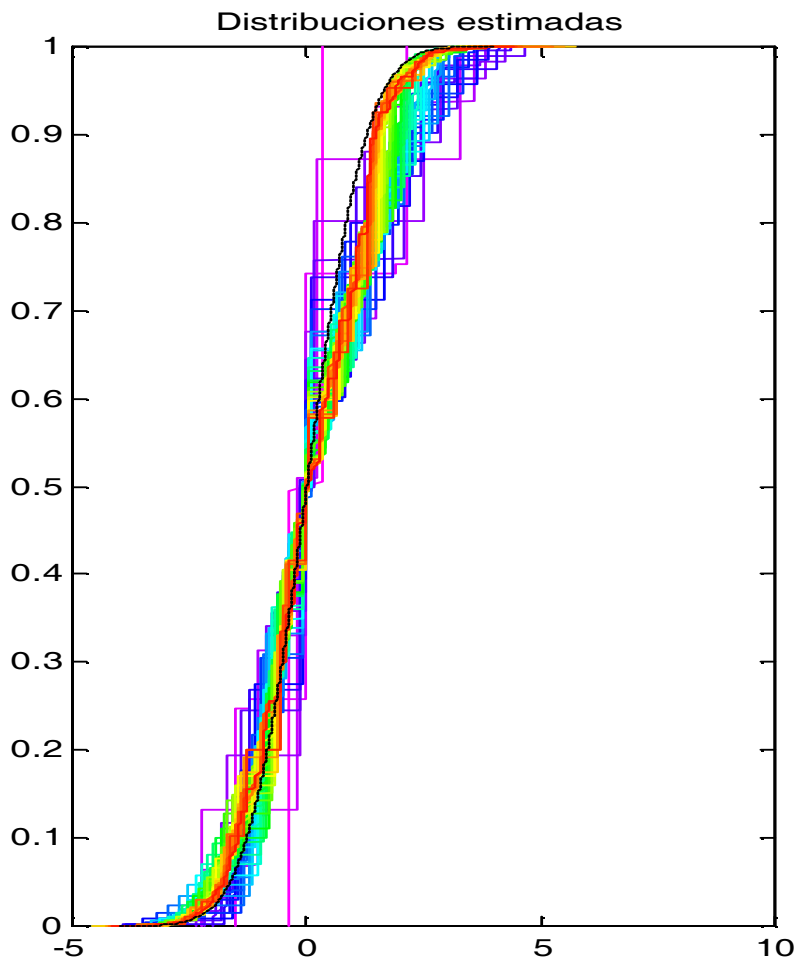


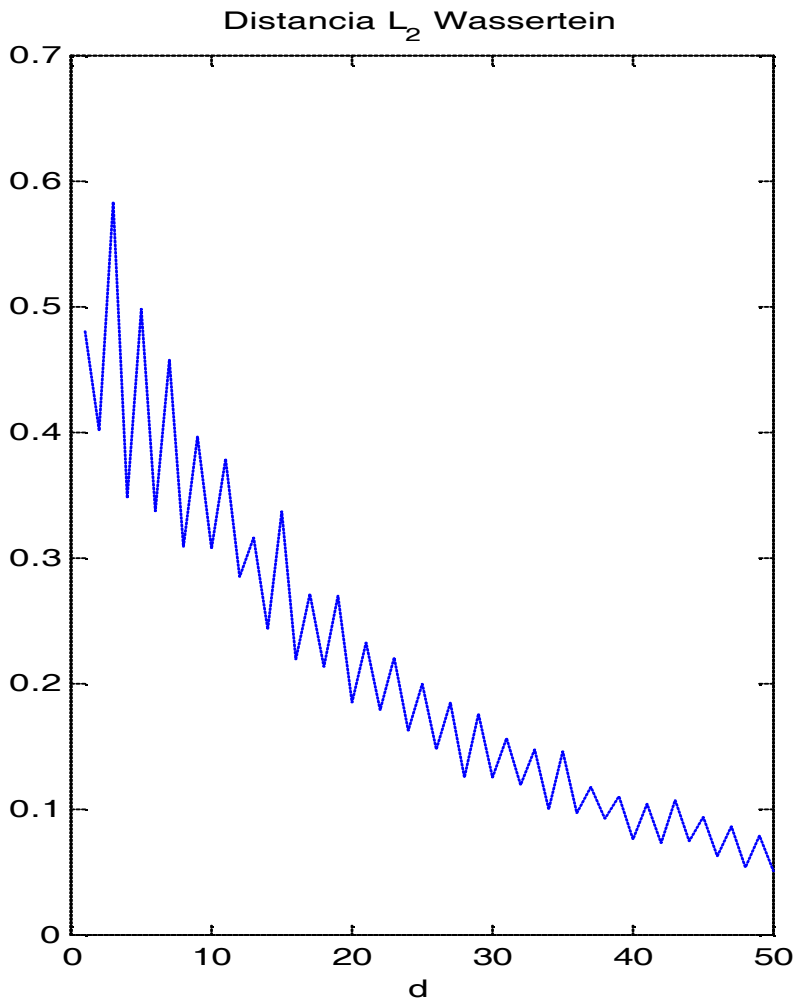
Ejemplo 2: Consideremos otro estadístico, la mediana muestral, $T_n = F_n^{-1}(\frac{1}{2}) = \widehat{Q}_2$. En este caso, sabemos que la distribución asintótica de $\sqrt{N}(\widehat{Q}_2 - Q_2)$ es $\mathcal{N}(0, 1/4\phi(0)^2)$. Con los datos del ejemplo, $v_T \approx 1/(4 \times 100 \times 0{,}3989423^2) = 0{,}01570796$.





Ejemplo 2 (continuación): Estadístico = Mediana.





Bootstrap: Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ un conjunto de datos generados por el modelo P, y sea $T(\boldsymbol{X})$ el estadístico cuya distribución $\mathcal{L}(T, P)$ deseamos estimar.

El método bootstrap propone como estimador de $\mathcal{L}(T, P)$ la distribución $\mathcal{L}^*(T^*; \widehat{P}_N)$ del estadístico $T^* = T(\boldsymbol{X}^*)$, donde $\boldsymbol{X}^*$ es un conjunto de datos generado por el modelo estimado $\widehat{\mathcal{P}}_N$.

- Posibles estimadores de P en el caso i.i.d.:
  - Bootstrap estándar: $F_N(x) = N^{-1} \sum_{i=1}^{N} I\,(X_i \leq x)$.
  - Bootstrap paramétrico: $F_{\widehat{\vartheta}}$ (supuesto subyacente $P = F_\vartheta$).
  - Bootstrap suavizado: $F_{n,h}$ un estimador kernel de la distribución F.

- Tamaño m del conjunto de datos $\boldsymbol{X}^*$, Bickel y Freedman 1981 y Bickel et al. 1997.

Referencias básicas: Efron 1979, Efron y Tibshirani 1993, Shao y Tu 1995 y Davison y Hinkley 1997.

Bootstrap estándar

Esquema del procedimiento de remuestreo:

$$(X_1, \ldots, X_N) \Rrightarrow F_N \Rrightarrow \begin{cases} (X_1^{*\,(1)}, \ldots, X_N^{*\,(1)}) & \Rrightarrow T_N^{*\,(1)} \\ (X_1^{*\,(2)}, \ldots, X_N^{*\,(2)}) & \Rrightarrow T_N^{*\,(2)} \\ \vdots & \vdots \\ (X_1^{*\,(B)}, \ldots, X_N^{*\,(B)}) & \Rrightarrow T_N^{*\,(B)} \end{cases}$$

$$\Rrightarrow \begin{cases} b_{Boot} = \mathrm{E}^* \left[T_N^*\right] - T_N \\ v_{Boot} = \mathrm{E}^* \left[(T_N^* - \mathrm{E}^*[T_N^*])^2\right] \\ \mathcal{L}_T^*(x) = \mathrm{Pr}^* \left\{(T_N^* - T_N) \leq x\right\} \end{cases}$$

donde E^* y Pr^* denotan la esperanza y la probabilidad bajo el esquema de remuestreo de bootstrap estándar.

Observación: X_i^* son observaciones i.i.d. F_N, $\therefore$ el esquema de remuestreo de bootstrap estándar puede interpretarse como una selección con reemplazamiento de la muestra original $(X_1, X_2, \ldots, X_N)$.

Ejemplo 1 (continuación): $T_N = \bar{X}$ y tomamos B $=$ 1000 remuestras de tamaño N.

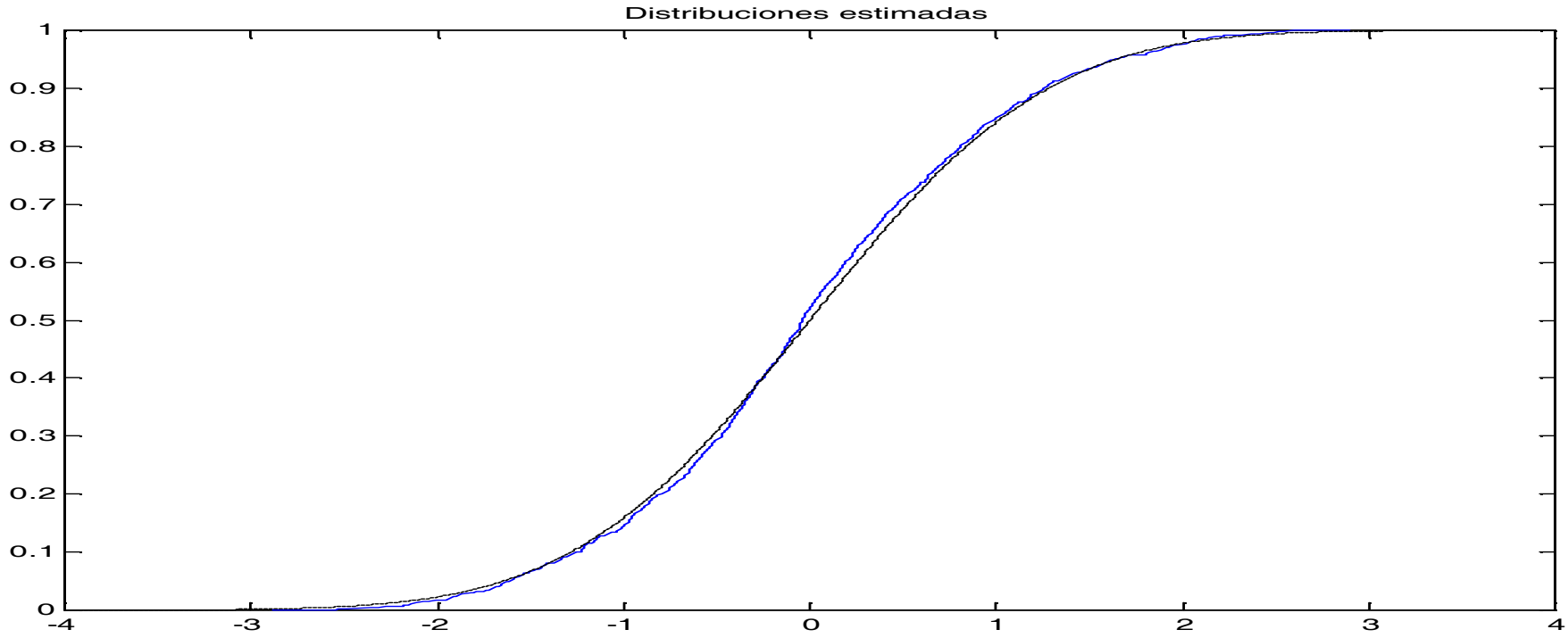
Que trabaje Matlab ...

```
> B = 1000;
> [d, b, v] = bootstrap(x, B, 'mean');

> [b v]

-0.0032    0.0082
```

$$b_{boot} = -0{,}0032 \approx 0 \text{ y } v_{boot} = 0{,}0082 \approx 0{,}01 = \frac{1}{100}.$$

Ejemplo 1 (continuación): $T_N = \bar{X}$ y tomamos B = 1000 remuestras de tamaño N.

Que siga trabajando Matlab ...



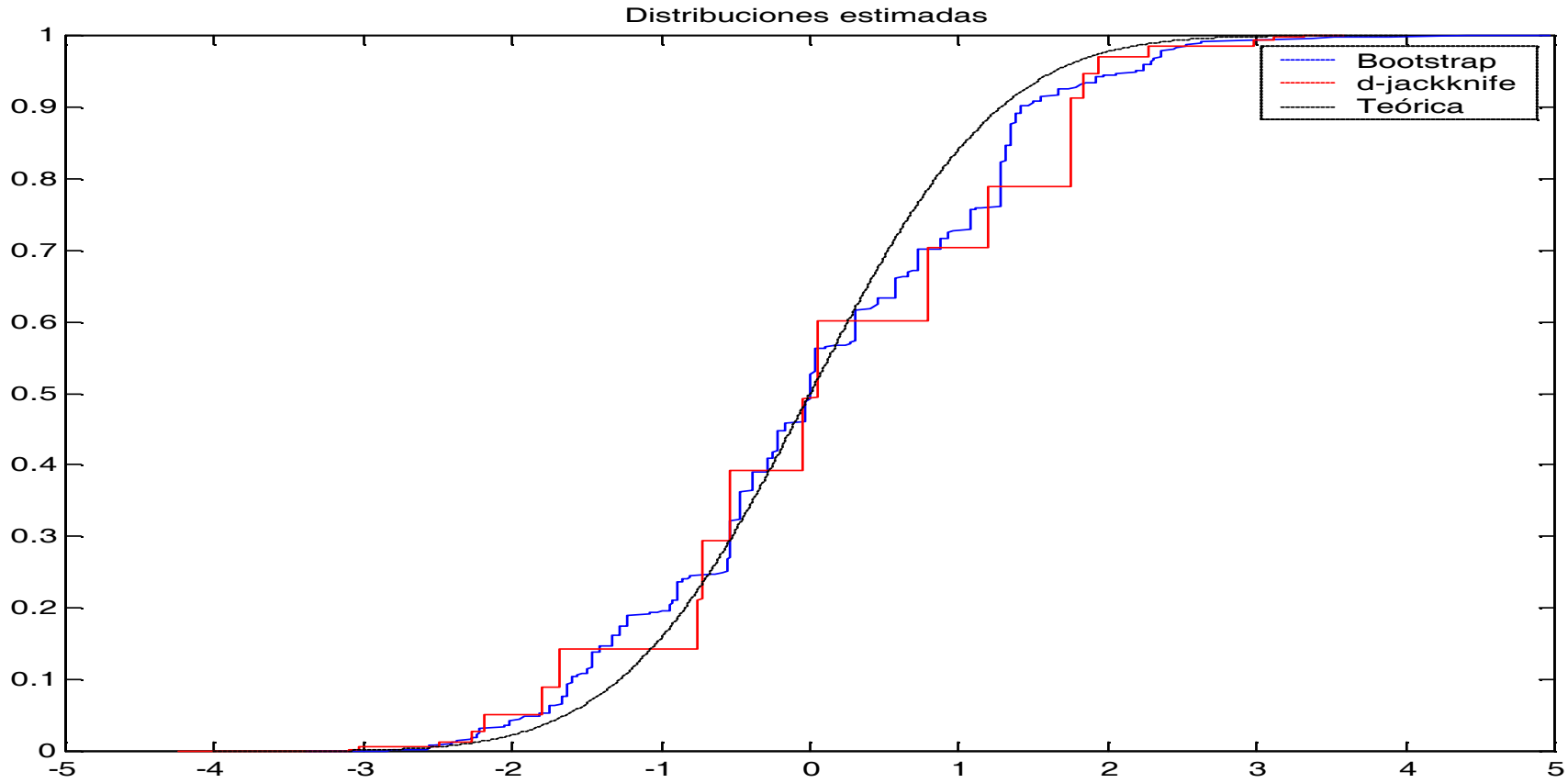
L2W = 0.0028

Ejemplo 2 (continuación): $T_N = Q_2$ y tomamos B $= 1000$ remuestras de tamaño N.

```
> B = 1000;
> [d, b, v] = bootstrap(x, B, 'mean');

> [b v]

0.0064    0.0144
```

$$b_{boot} = 0{,}0064 \text{ y } v_{boot} = 0{,}0144 \approx 0{,}0157.$$

Ejemplo 2 (continuación): $T_N = Q_2$ y tomamos B $= 1000$ remuestras de tamaño N.

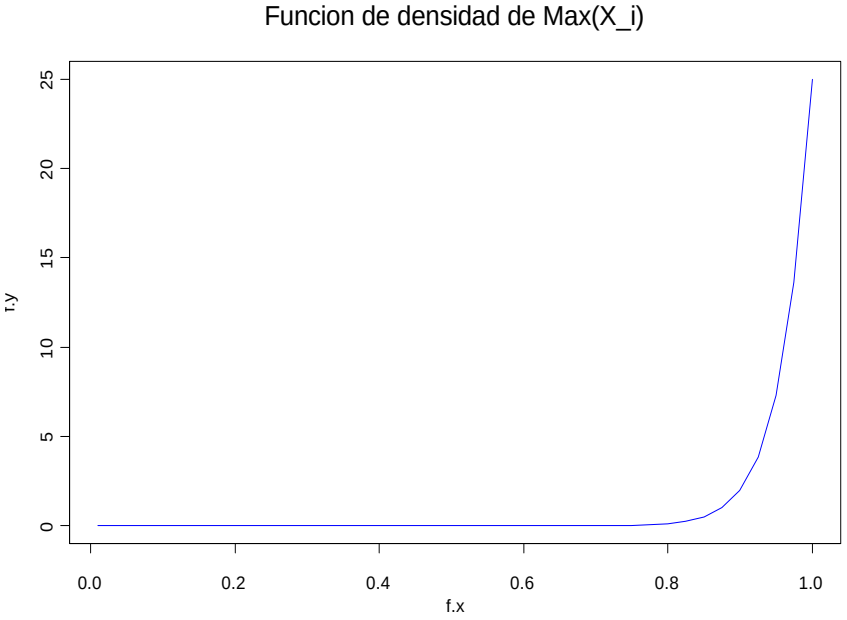


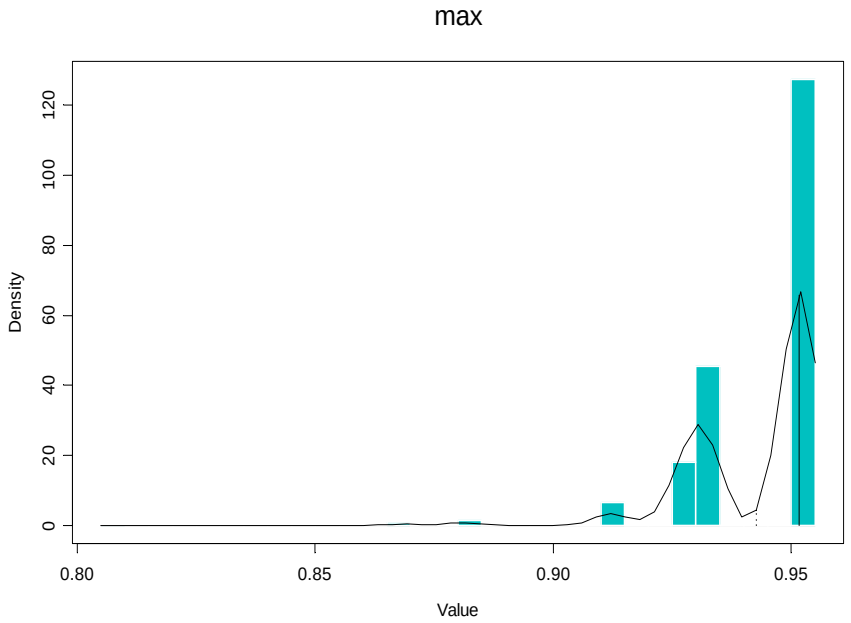
L2WB = 0.0551 y L2WJ = 0.1426

¿Bootstrap siempre?

Un ejemplo donde el bootstrap falla: Sean $X_1, X_2, \ldots, X_N$ una muestra i.i.d. $\mathcal{U}(0, \theta)$. Sabemos que el e.m.v de θ es $T_N = \widehat{\theta} = \text{máx}_{1 \leq i \leq N} X_i$ cuya función de densidad está dada por:

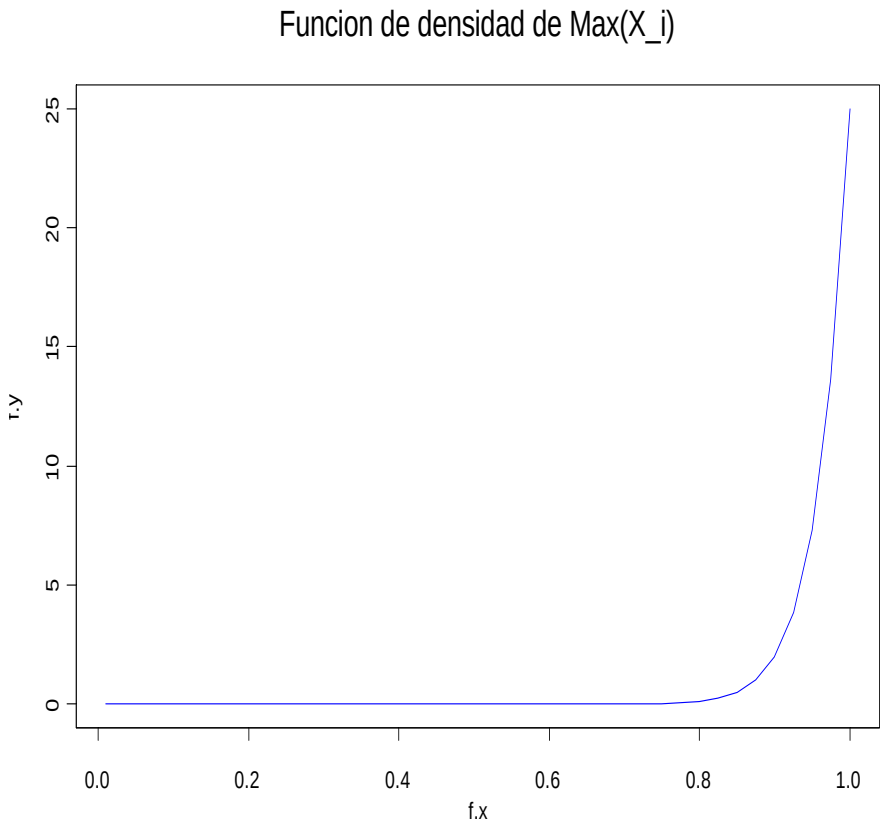
$$f_{\widehat{\theta}}(x) = \begin{cases} 0 & \text{si} \quad x < 0 \text{ ó } x > \theta \\ \frac{n x^{n-1}}{\theta^n} & \text{si} \quad 0 \leq x \leq \theta \end{cases}$$

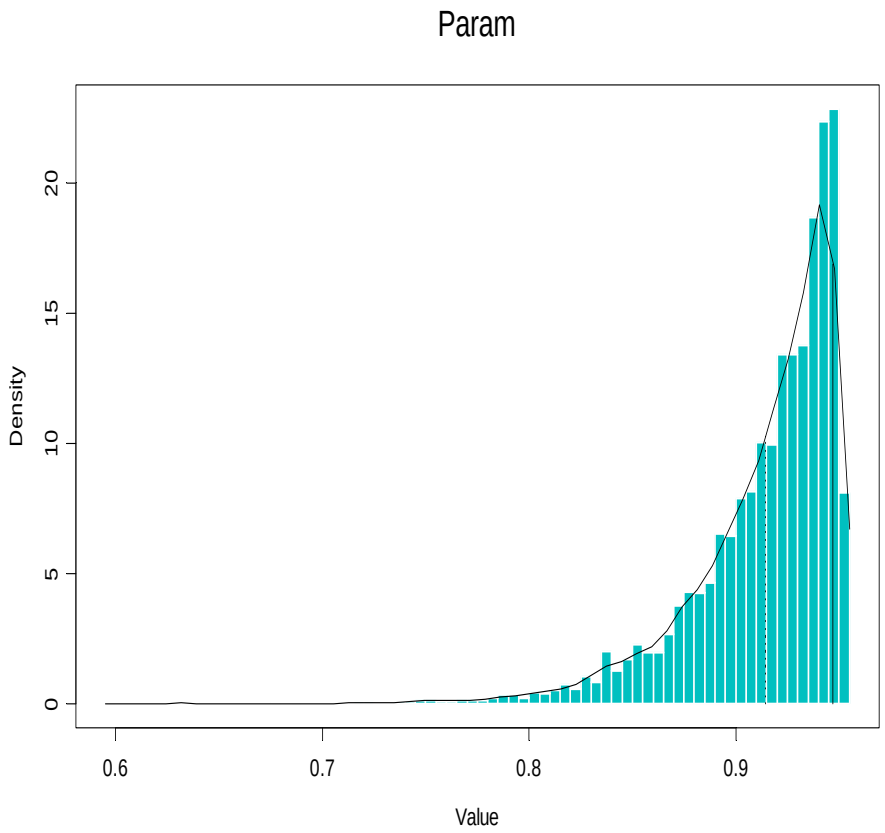




¿Bootstrap siempre?

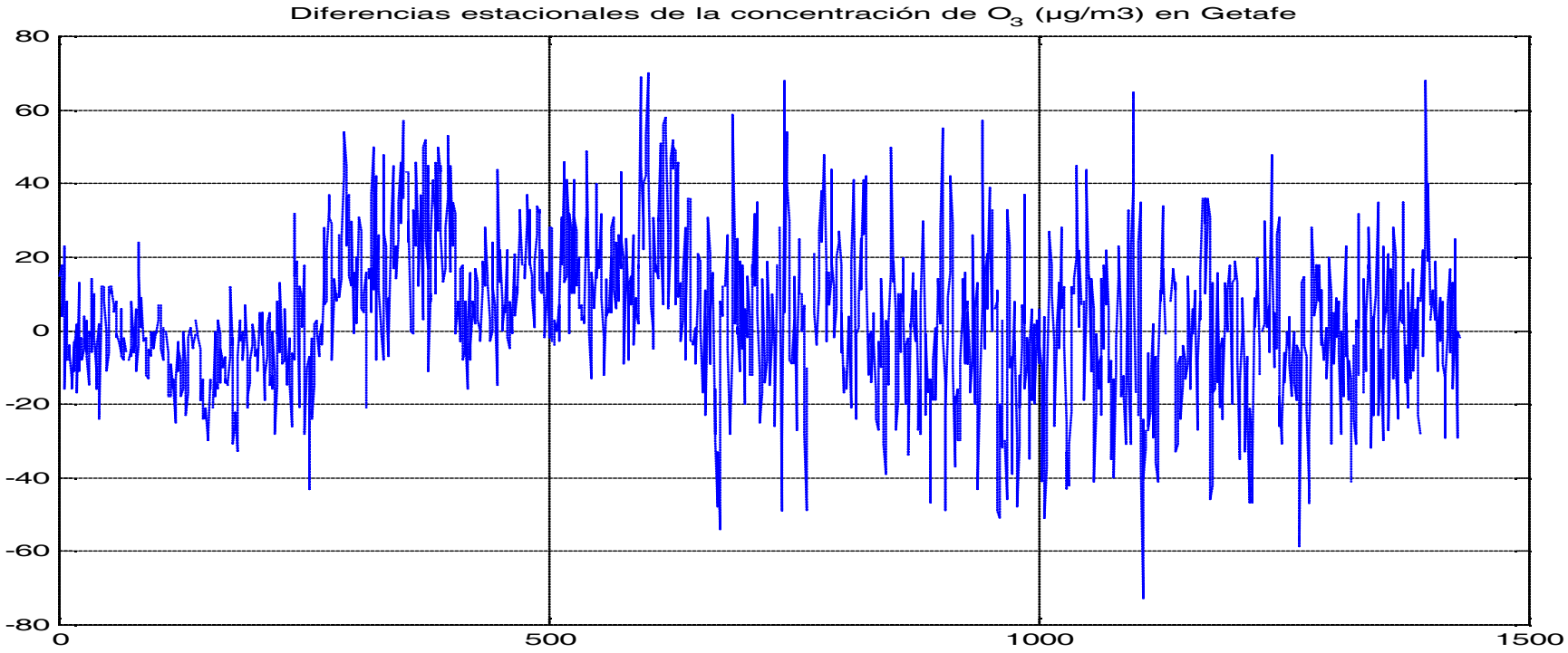
Un ejemplo donde el bootstrap falla (cont.): Una solución *bootstrap paramétrico*:





Aplicación a la monitorización - 1

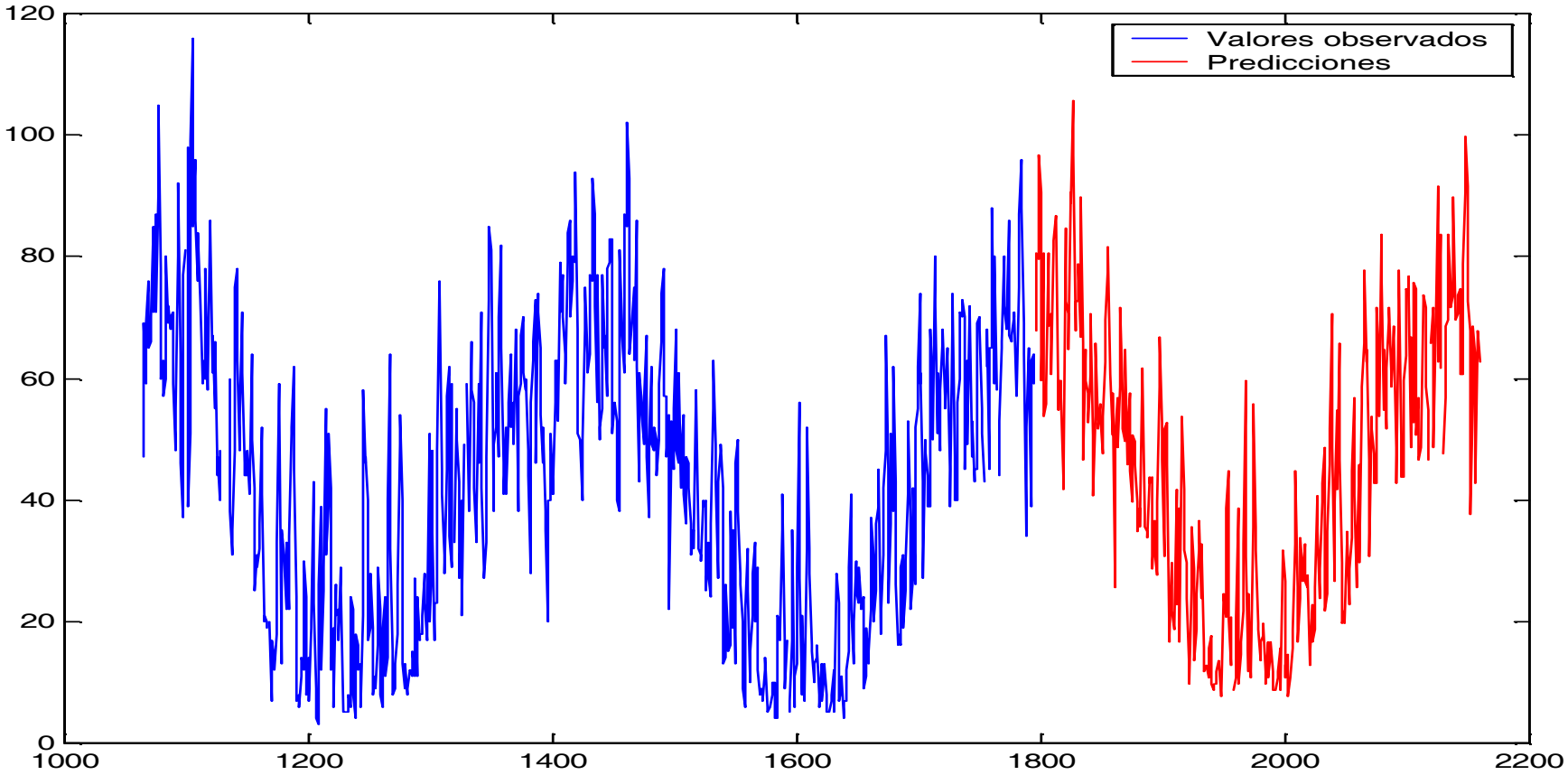
Supongamos que las series diferencias estacionales de la concentración de ozono, C_t es aproximadamente ruido blanco, es decir, $X_t = (1 - L^{365})C_t = C_t - C_{t-365}$ se supone i.i.d.:



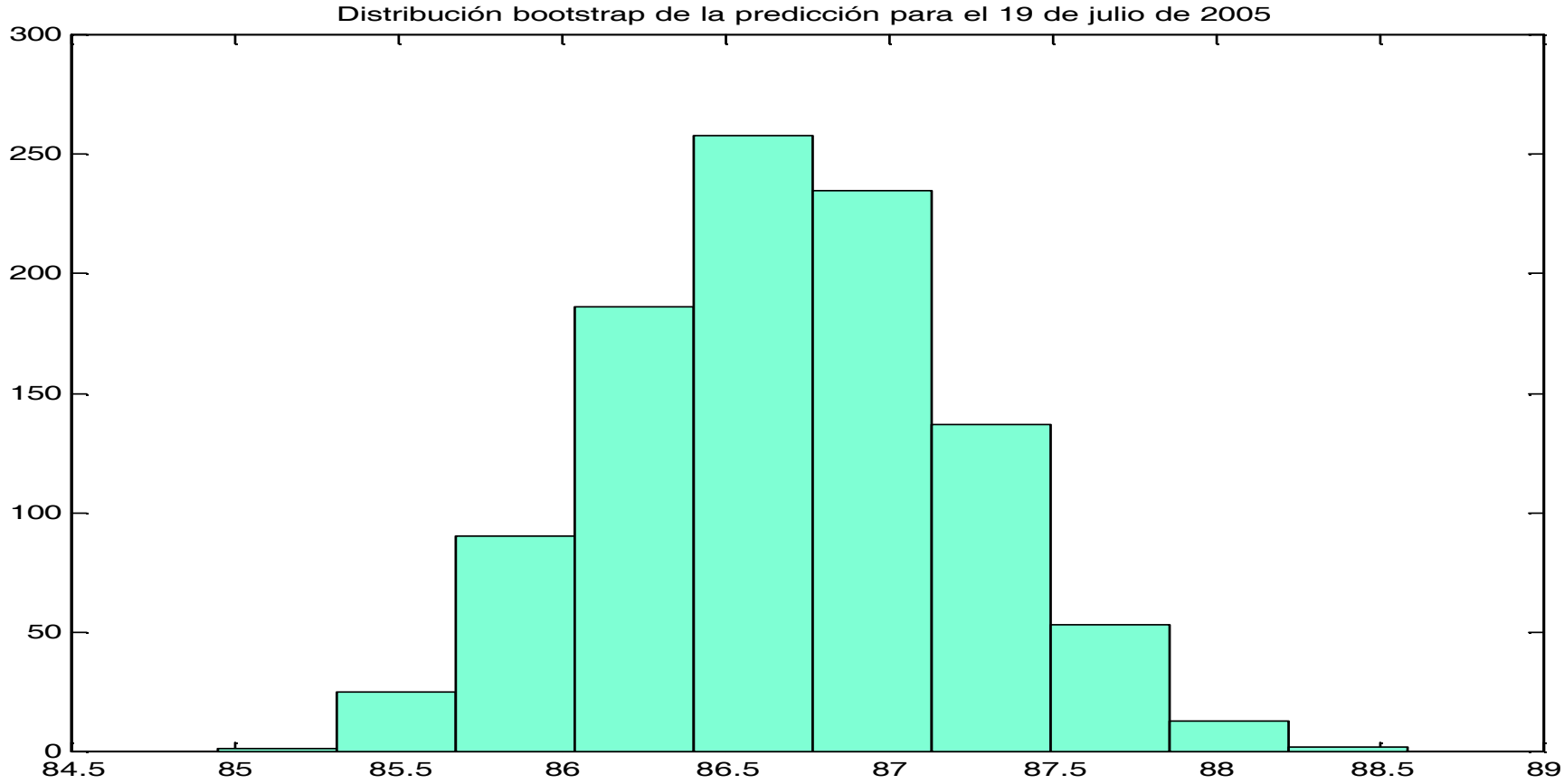
Este supuesto se relajará en breve ...

¿Cuál es el nivel medio esperable para los próximos 365 días?

$$C_t = C_{t-365} + \bar{X}_t$$



¿Cuál es el nivel medio esperable para HOY? ¿Y su distribución?

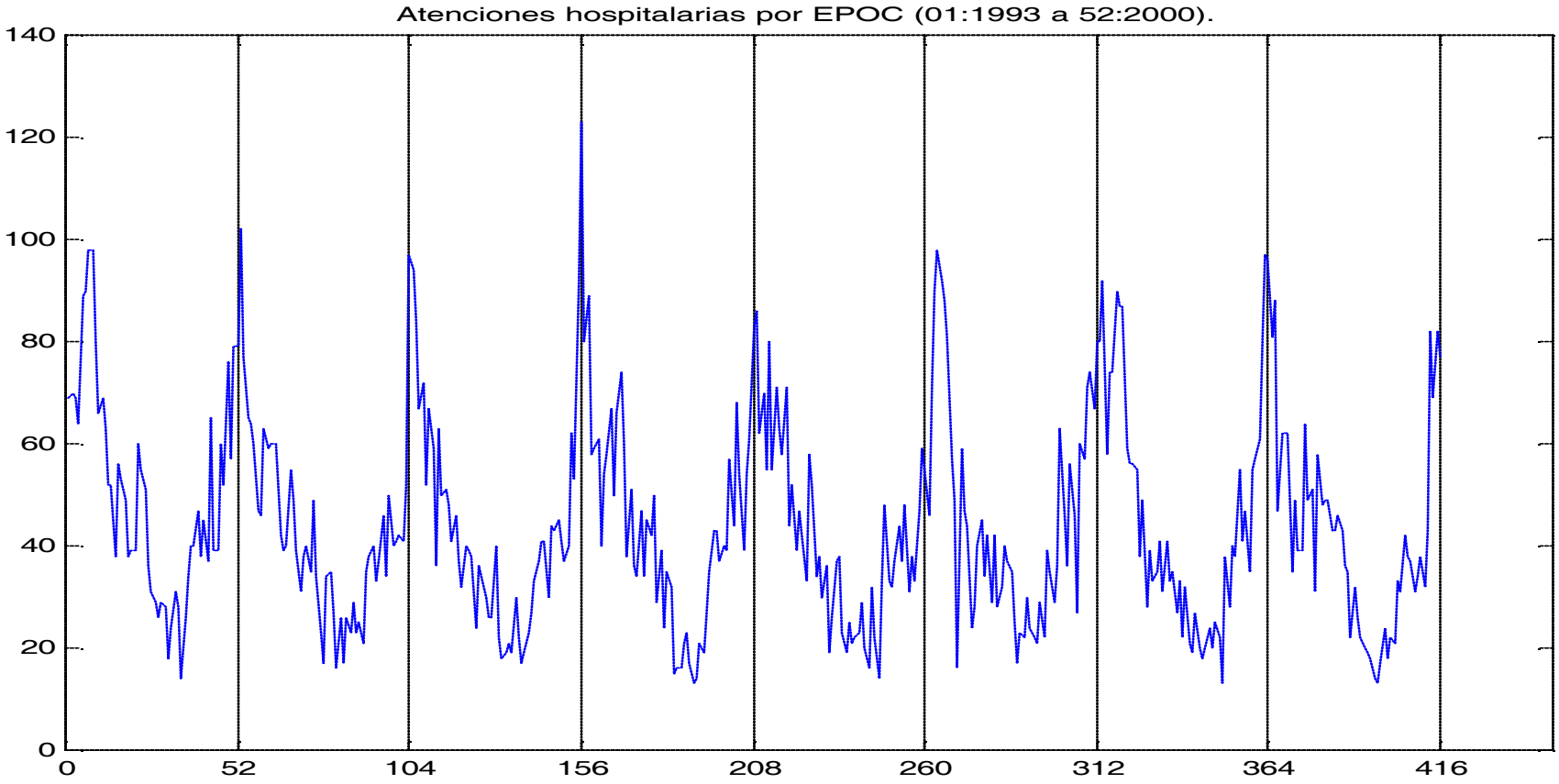


Podemos comprobar los resultados en:

`http://dgpea2.comadrid.es/areastematicas/atmosfera/p_atmosfera.html`

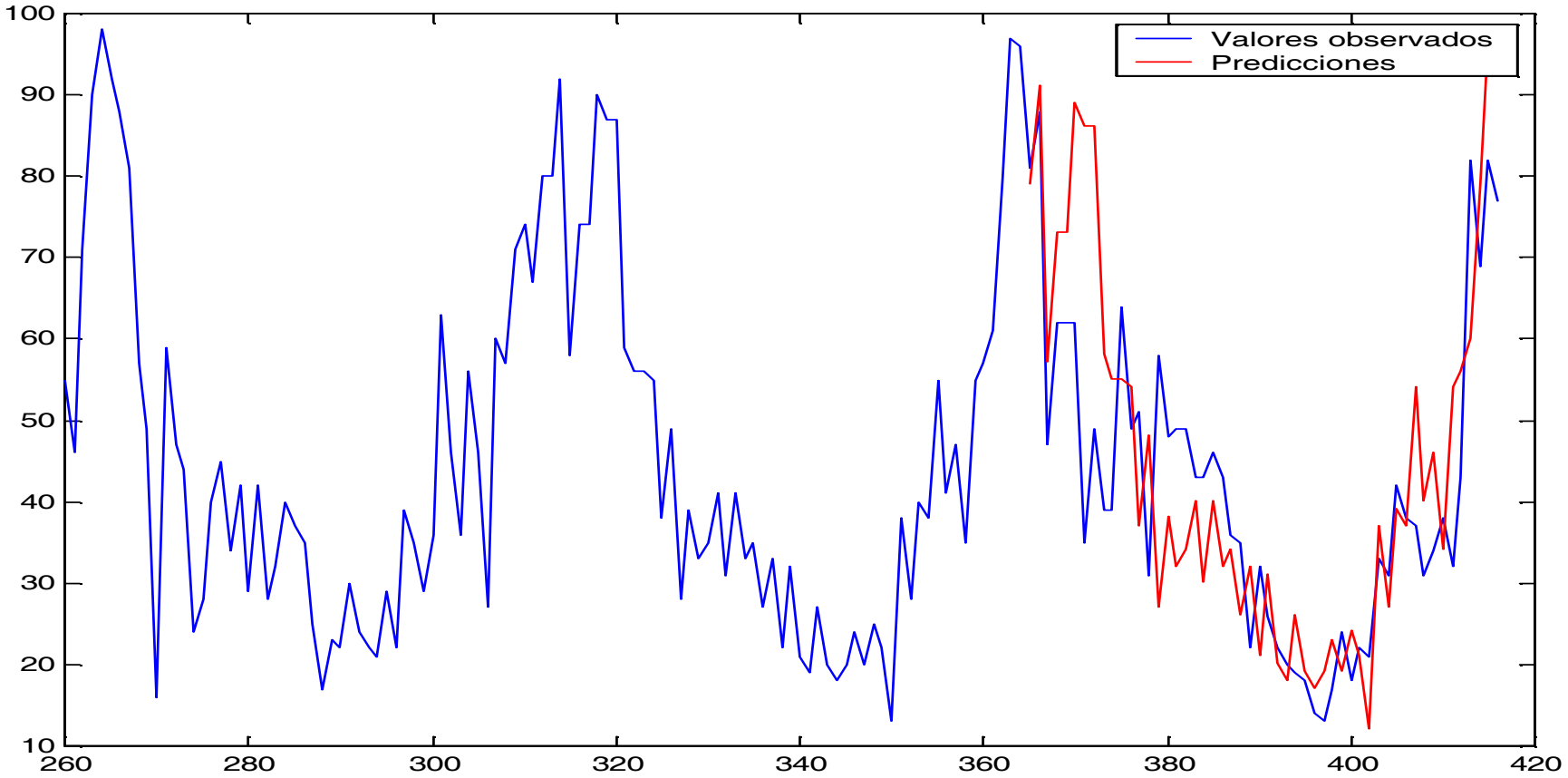
Aplicación a la monitorización - 2

Datos semanales del número de atenciones hospitalarias por EPOC en la Comunidad de Madrid, 1993 - 2000.

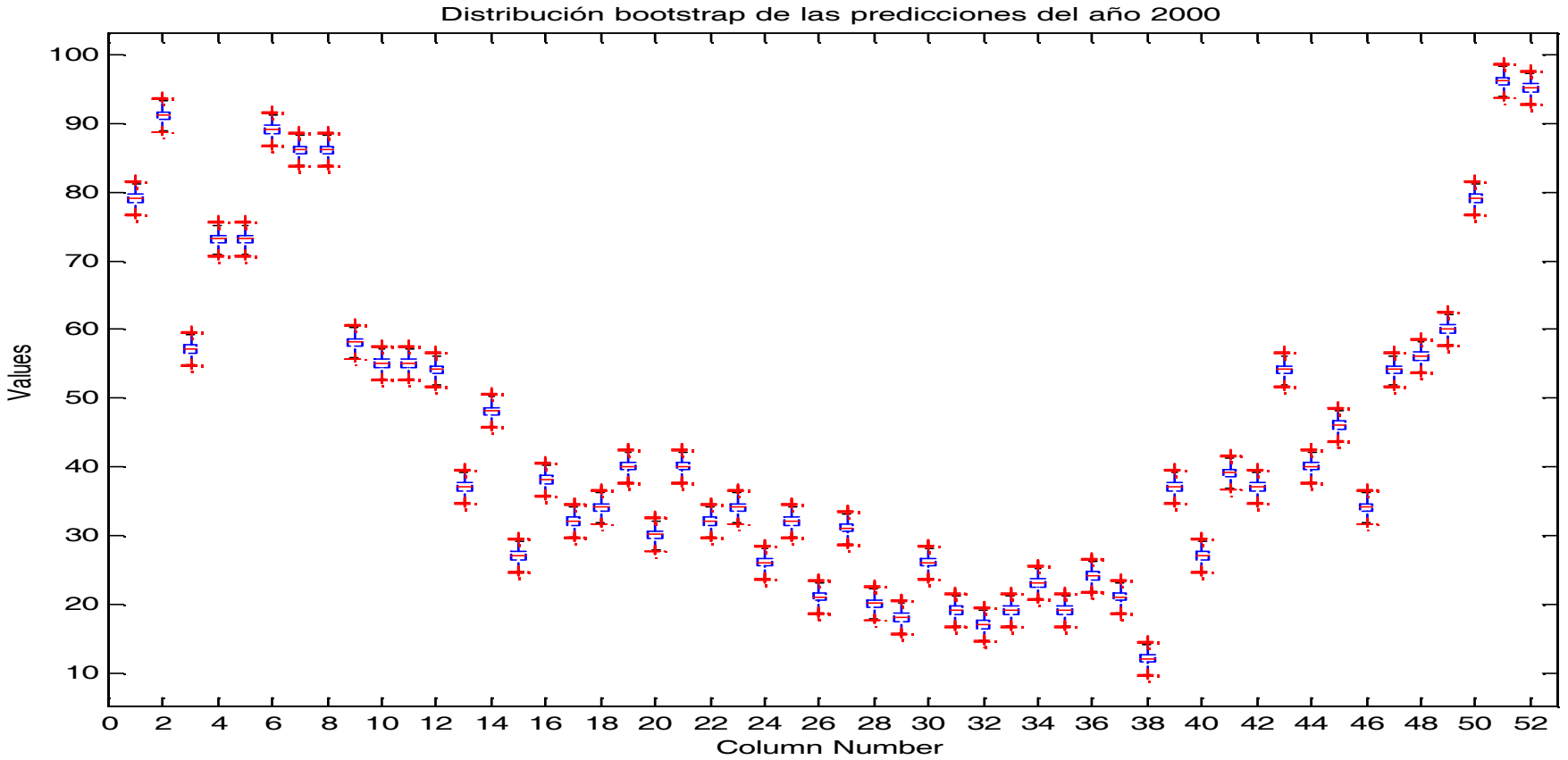


¿Cuál es el número de atenciones esperable para las próximas 52 semanas (utilizaremos datos hasta el año 1999)?

$$E_t = C_{t-52} + \bar{X}_t$$



Distribución de las predicciones:



Estructura

1. Introducción.

2. Métodos de remuestreo para datos i.i.d.

3. Métodos de remuestreo para series temporales.
   - Métodos basados en modelos.
   - Métodos no basados en modelos.

4. Aplicación a la monitorización de variables ambientales y epidemiológicas.

5. Conclusiones y Extensiones.

Métodos basados en modelos

$$X_t = g(\mathcal{F}_{-\infty}^{t-1}, \epsilon_t),$$

donde $\{\epsilon_t\}$ es una sucesión de v.a. i.i.d. $\sim F_\epsilon$, y g es una función de enlace.

Elemento común. La utilización de remuestras de los residuos estimados en el modelo $\mathcal{P} = (g, F_\epsilon)$ postulado.

Referencias básicas. Freedman 1984 y Efron y Tibshirani 1986 en modelos $AR(p)$, Bose 1990 en modelos $MA(q)$, Kreiss y Franke 1992 en modelos $ARMA(p, q)$, Franke et al. 1997 en modelos autorregresivos no paramétricos.

Procedimiento de remuestreo:

Consideremos un proceso autorregresivo de orden p:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

donde $\mathcal{P} = (g, F_\varepsilon)$ es $g(\mathcal{F}_{-\infty}^{t-1}, \epsilon_t) = g(\mathcal{F}_{t-p}^{t-1}, \epsilon_t) = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t$ y sobre F_ε suponemos que $E[|\varepsilon|^{2+\alpha}] < +\infty$ para algún $\alpha > 0$.

Esquema del procedimiento de remuestreo:

$$(X_1, \ldots, X_N) \Rrightarrow \widehat{AR(p)} \Rrightarrow \begin{cases} X_1^{*(1)}, \ldots, X_N^{*(1)} & \Rrightarrow & \widehat{AR(p)}^{*(1)} \\ \vdots & & \vdots \\ X_1^{*(B)}, \ldots, X_N^{*(B)} & \Rrightarrow & \widehat{AR(p)}^{*(B)} \end{cases}$$

Procedimiento de remuestreo:

1. Obtener estimaciones a partir de $\boldsymbol{X}$ de los parámetros autorregresivos: $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$.

2. Calcular los residuos: $\widehat{\varepsilon}_t = X_t - \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i}$, para $t = p+1, p+2, \ldots, N$.

3. Calcular la función de distribución empírica de los residuos centrados $\widetilde{\varepsilon}_t = \widehat{\varepsilon}_t - (N-p)^{-1} \sum_{t=p+1}^{N} \widehat{\varepsilon}_t$ por:

$$F_N^{\widetilde{\varepsilon}}(x) = \frac{1}{N-p} \sum_{t=p+1}^{N} I\left(\widetilde{\varepsilon}_t \leq x\right).$$

4. Obtener $N-p$ observaciones i.i.d. de $F_N^{\widetilde{\varepsilon}}$ que denotamos $(\varepsilon_{p+1}^{*(b)}, \ldots, \varepsilon_N^{*(b)})$.

Procedimiento de remuestreo:

5. Fijar los p primeros valores de la serie $(X_1^{*(b)}, \ldots, X_p^{*(b)})$ y calcular las restantes observaciones de la remuestra $(X_{p+1}^{*(b)}, \ldots, X_N^{*(b)})$ mediante:

$$X_t^{*(b)} = \sum_{i=1}^{p} \widehat{\phi}_i X_{t-i}^{*(b)} + \varepsilon_t^{*(b)}.$$

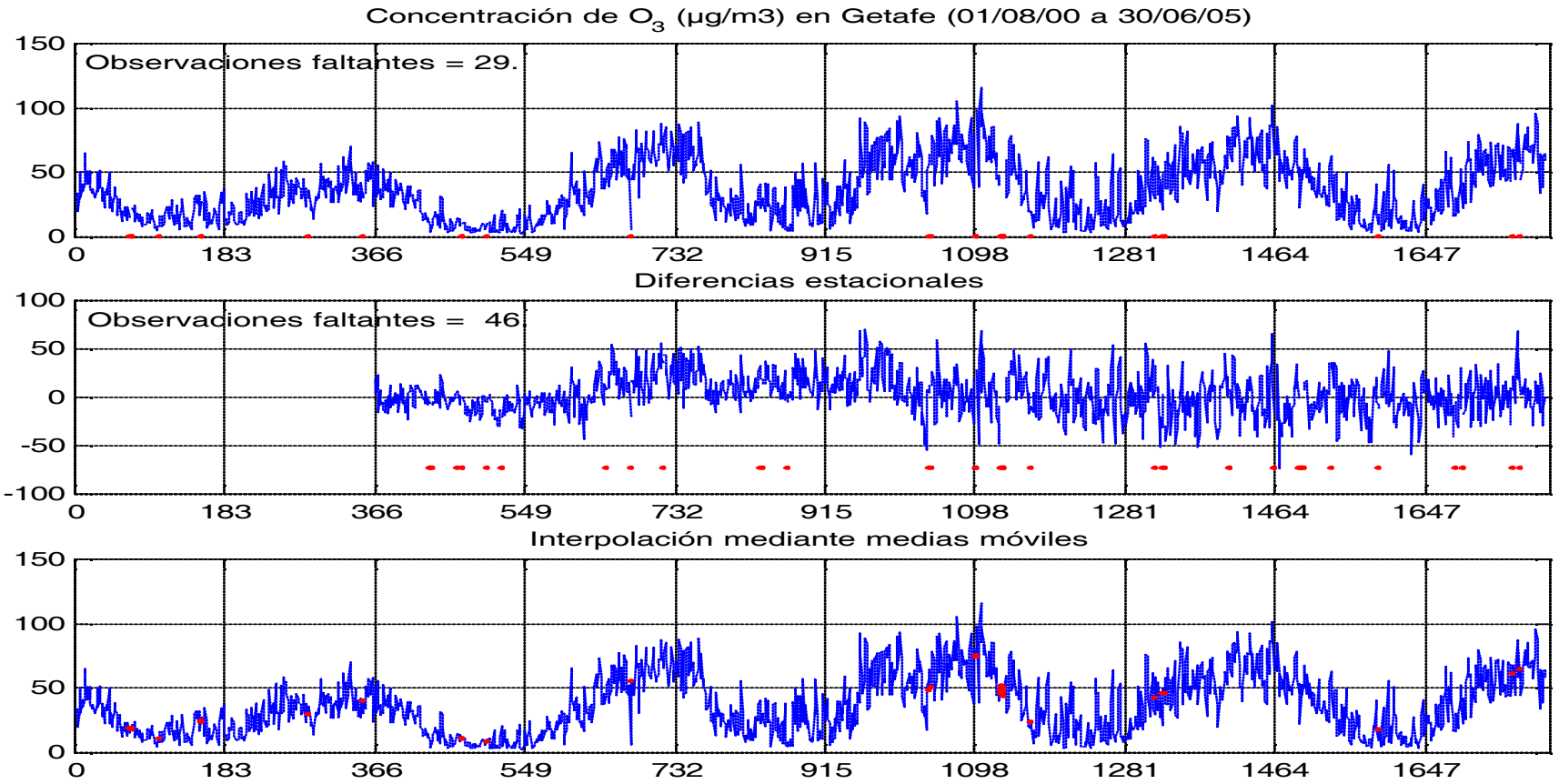
6. Utilizando $\boldsymbol{X}^{*(b)} = (X_1^{*(b)}, \ldots, X_n^{*(b)})$, calcular los análogos bootstrap $\widehat{\boldsymbol{\phi}}^{*(b)} = (\widehat{\phi}_1^{*(b)}, \ldots, \widehat{\phi}_p^{*(b)})$ de los estimadores $\widehat{\boldsymbol{\phi}}$.

Observación sobre el procedimiento de remuestreo:

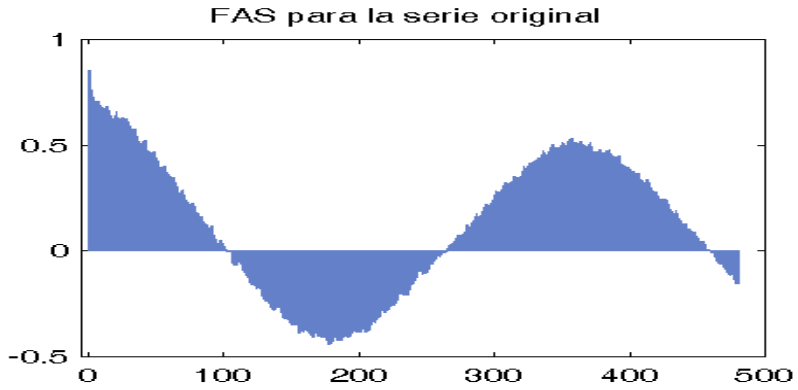
• Este método de remuestreo, como señalan Cao et al. 1997, debe modificarse para construir intervalos de predicción, pues no replica la distribución condicional de las observaciones futuras X_{T+h} dada la serie observada $\boldsymbol{X}$.

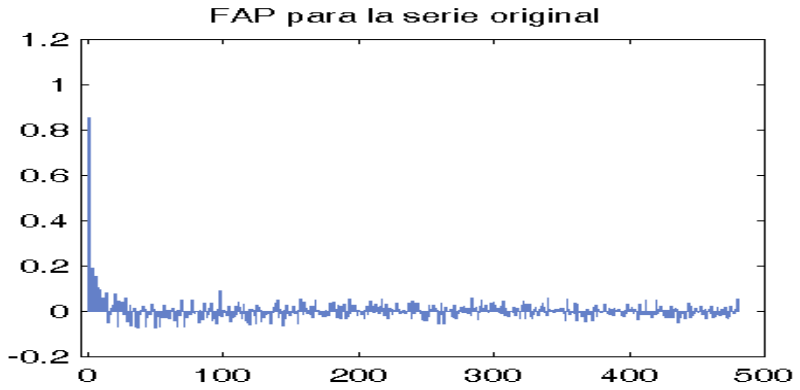
Aplicación a la monitorización - 1

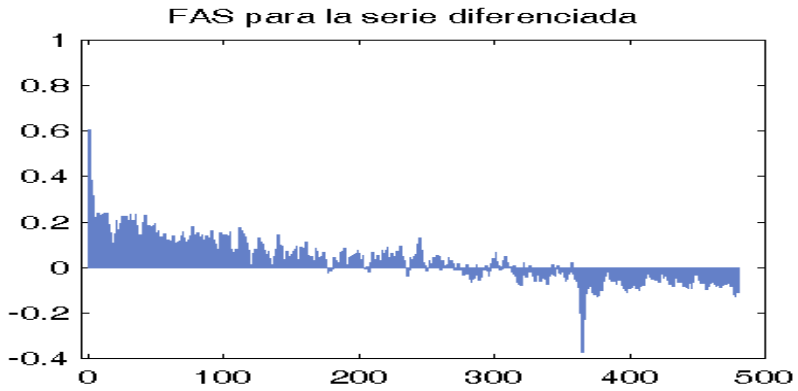
Interpolación de los valores faltantes:

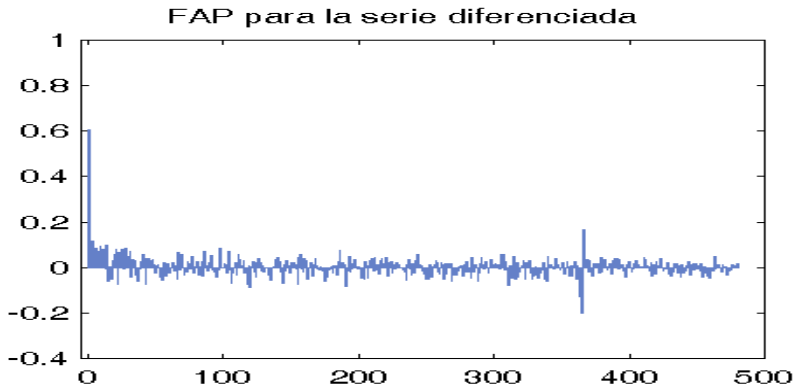


Funciones de autocorrelación simple (FAS) y parcial (FAP):



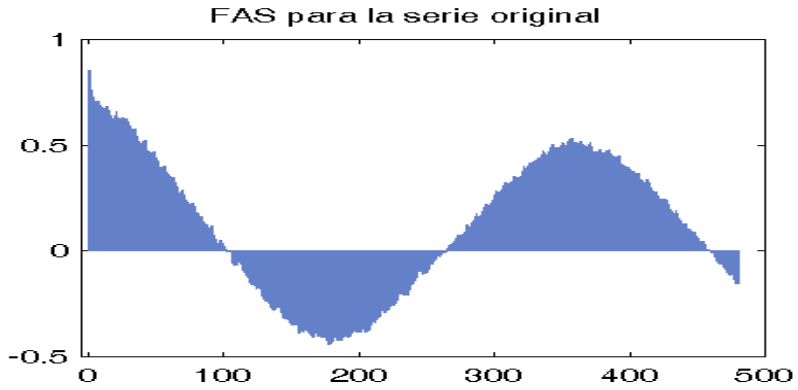


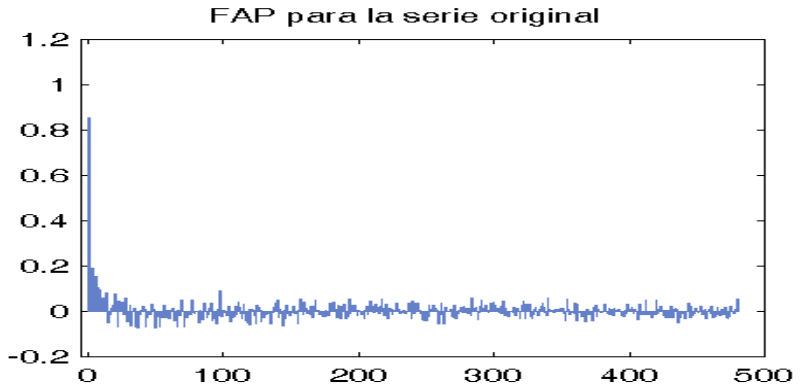


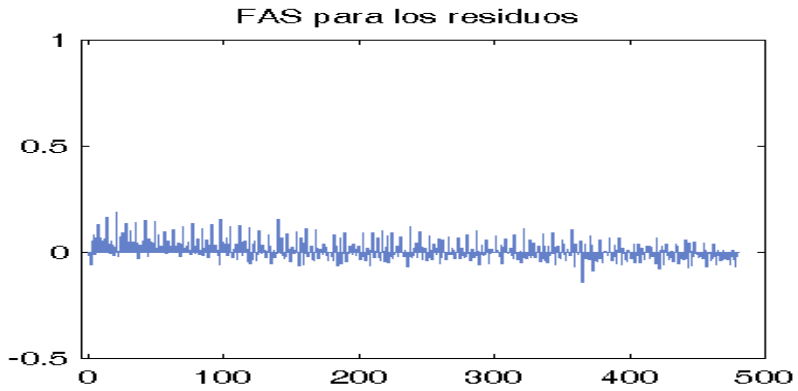


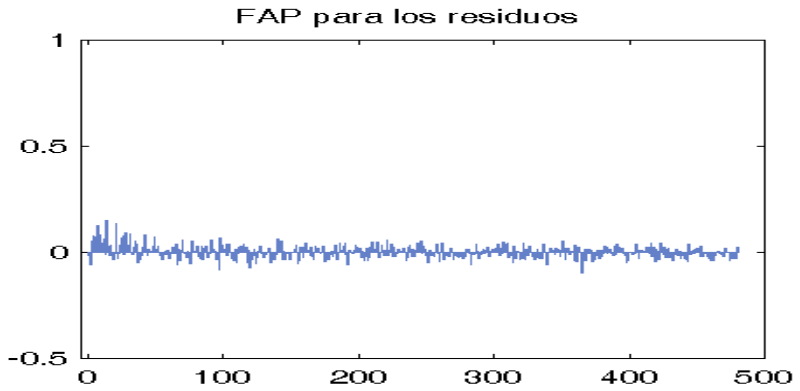
Sugiere (*al menos*) un modelo ARIMA(1,0,0)(1,1,0)$_{365}$.

Funciones de autocorrelación simple (FAS) y parcial (FAP):



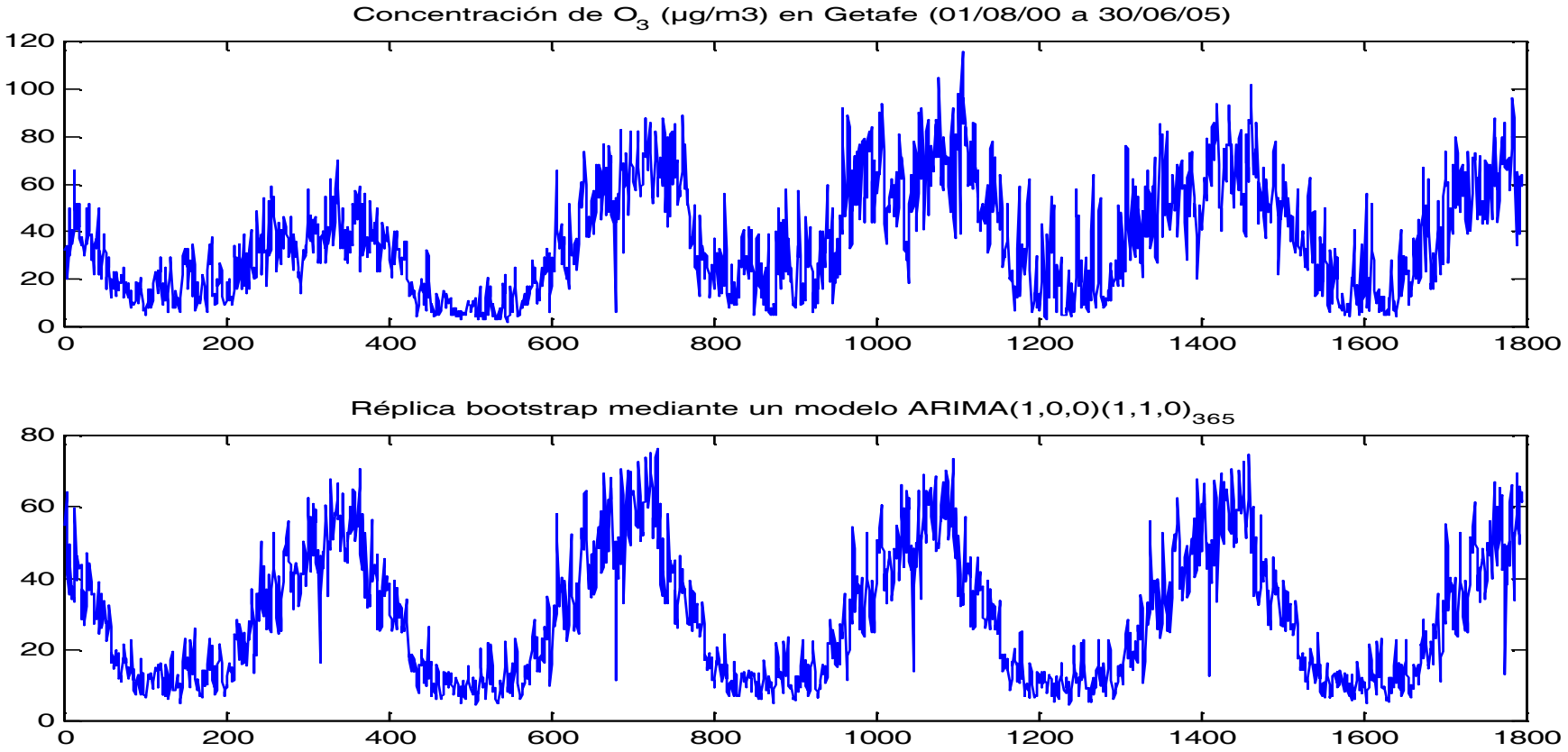






Se “capta razonablemente” la estructura de dependencia temporal.

Réplicas bootstrap suponiendo un modelo ARIMA$(1,0,0)(1,1,0)_{365}$:



Métodos de remuestreo basados en modelos

Aplicación a la predicción $\mathcal{L}(X_{T+h}|(X_1, \ldots, X_T)$

Sean $(X_1, X_2, \ldots, X_T)$ observaciones de un proceso estacionario lineal:

$$X_t - \mu_X = \sum_{j=0}^{+\infty} \psi_j \varepsilon_{t-j},$$

donde $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ son tales que $\mathrm{E}[\varepsilon_t] \equiv 0$, y $E[\varepsilon_t^2|] \equiv \sigma^2$.

Solución al problema de predicción:

1. Seleccionar un modelo apropiado.
2. Estimar el modelo seleccionado.
3. Obtener un estimador de la distribución de X_{T+h} dado $(X_1, X_2, \ldots, X_T)$.

Un enfoque clásico, asume que $\{X_t\}_{t\in\mathbb{Z}}$ sigue un modelo ARMA(p,q):

$$\sum_{i=0}^{p} \phi_i \left(X_{t-i} - \mu_X\right) = \sum_{j=0}^{q} \theta_j \, \varepsilon_{t-j},$$

donde $\phi_0 = \theta_0 = 1$, y los ε_t son i.i.d. $\mathcal{N}(0, \sigma^2)$.

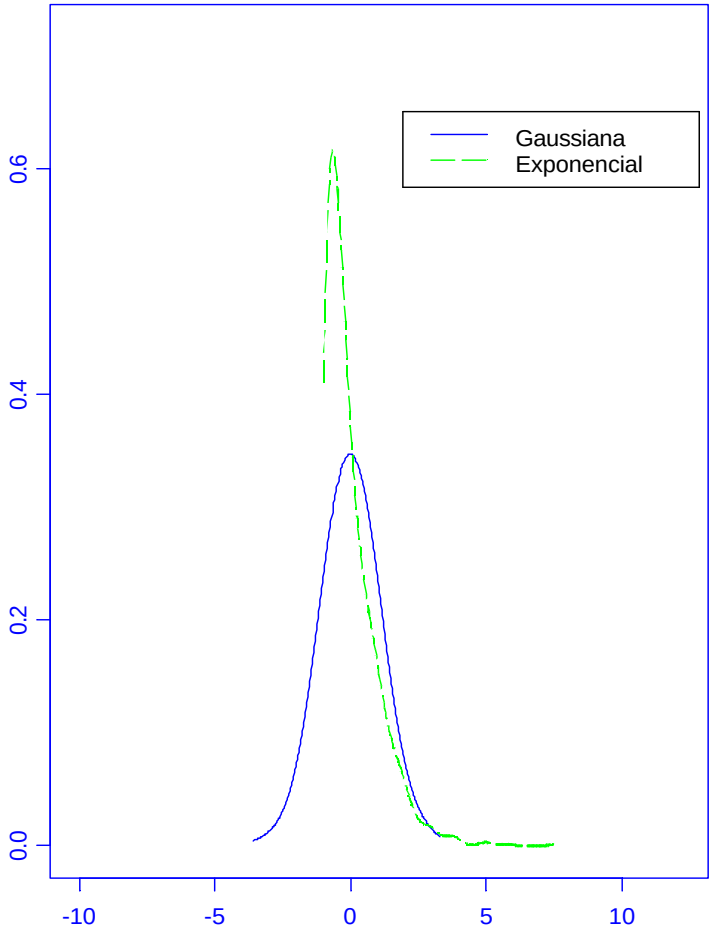
Dada esta hipótesis, tenemos un intervalo de predicción estimado:

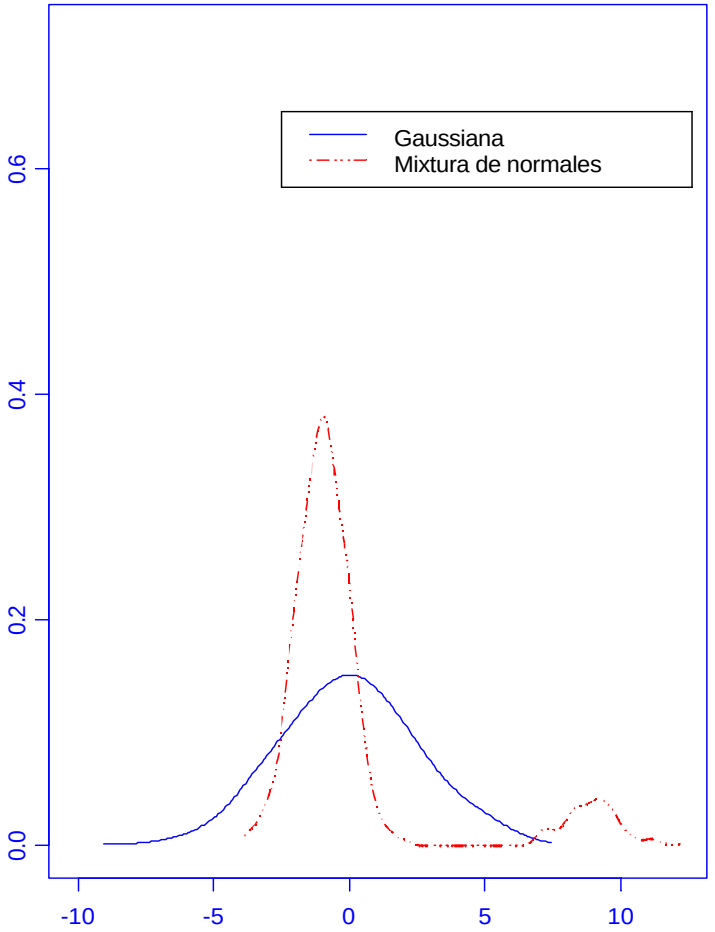
$$\widehat{E}[X_{T+h} | \boldsymbol{X}_{T-p}^{T}] \pm z_{\alpha/2} \left(\widehat{\sigma}^2 \sum_{j=0}^{h-1} \widehat{\psi}_j^2 \right)^{1/2}.$$

Observaciones sobre el intervalo de predicción anterior:

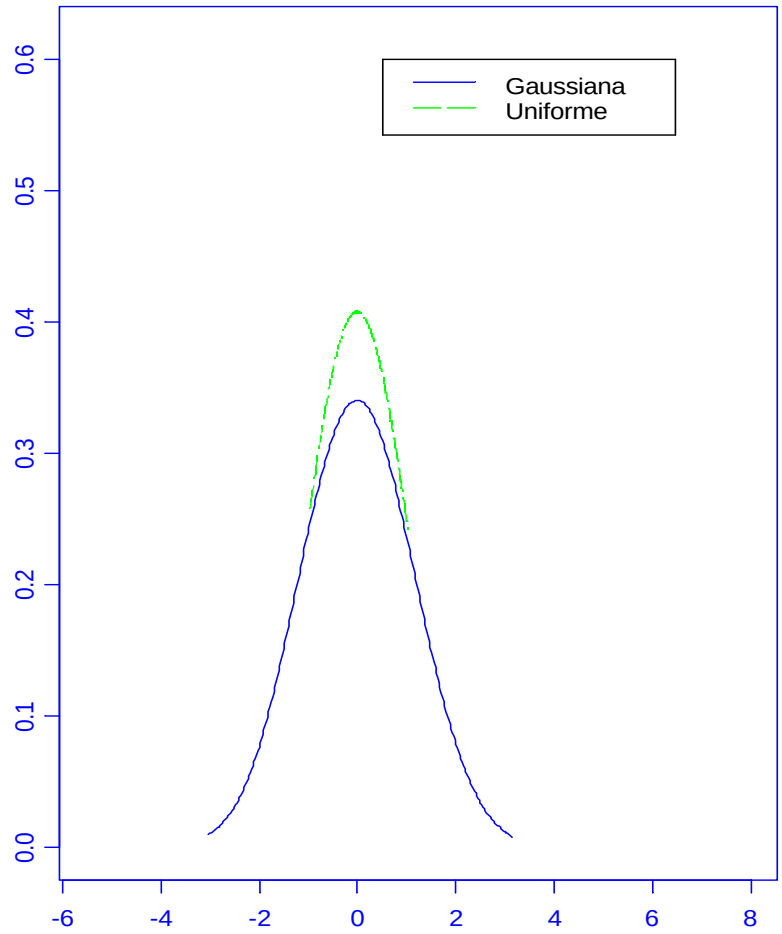
1. No tiene en cuenta la variabilidad de $\widehat{E}[\centerdot]$, $\widehat{\sigma}$, y $\widehat{\psi}$.
2. Puede verse afectado por desviaciones de la hipótesis de normalidad de ε.

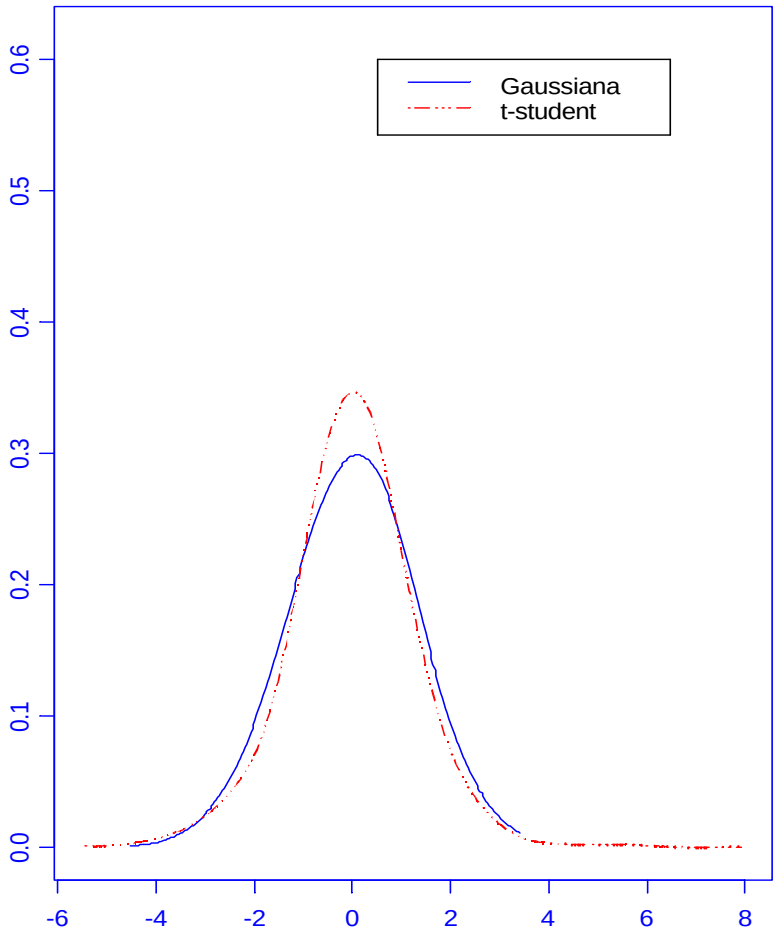
Densidades predictivas y su aproximación gaussiana





Densidades predictivas y su aproximación gaussiana





Soluciones bootstrap a (1) y (2)

Para modelos AR(p) con orden p conocido:

Esquema del procedimiento de remuestreo:

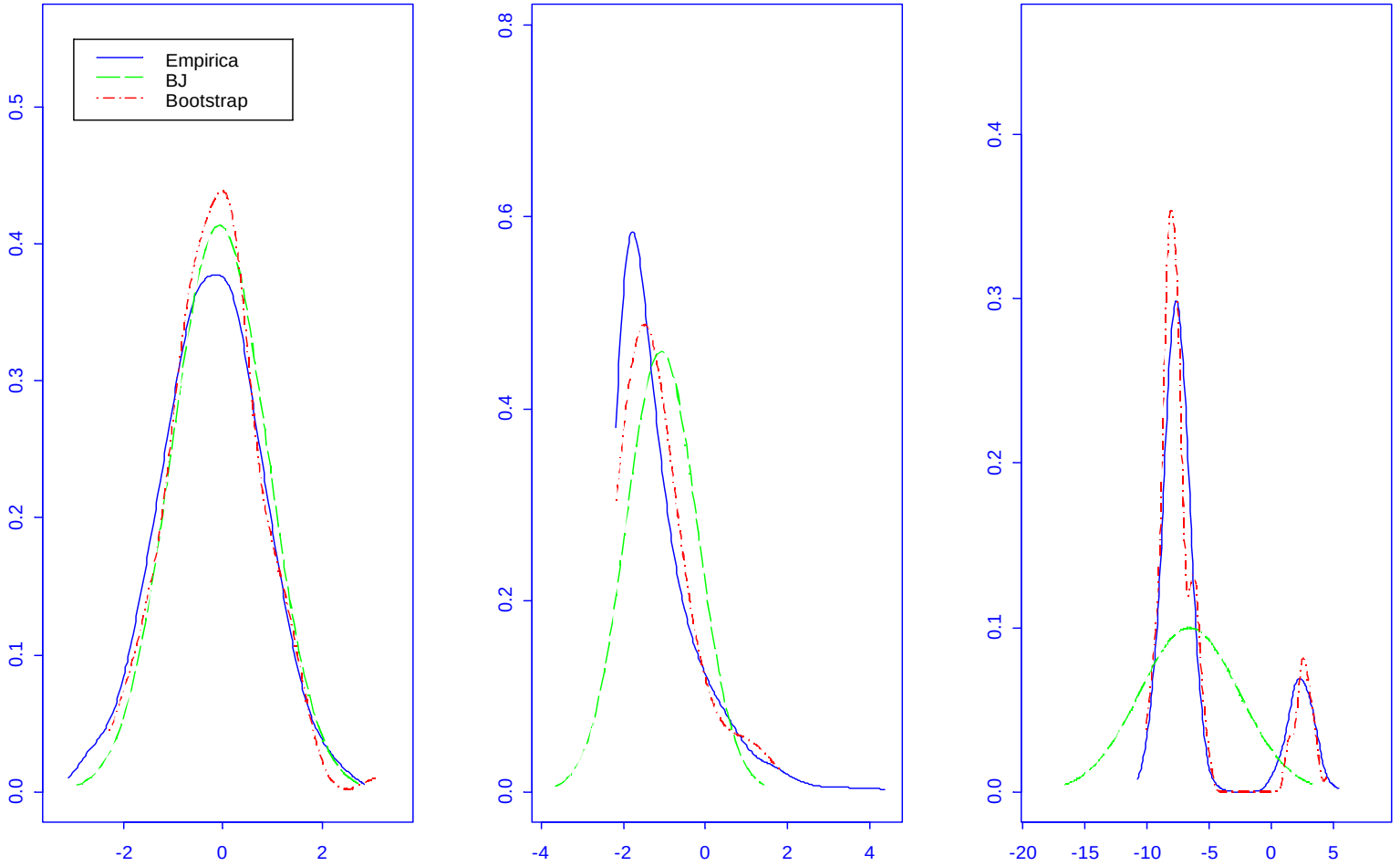
$$(X_1,\ldots,X_N) \Rrightarrow \widehat{AR(p)} \Rrightarrow \begin{cases} X_1^{*(1)},\ldots,X_N^{*(1)} & \Rrightarrow & \widehat{AR(p)}^{*(1)} \\ \vdots & & \vdots \\ X_1^{*(B)},\ldots,X_N^{*(B)} & \Rrightarrow & \widehat{AR(p)}^{*(B)} \end{cases}$$

$$\Rrightarrow \quad (X_{T-p},\ldots,X_T) \quad X_{T+1}^{*(1)},\ldots,X_{T+h}^{*(1)}$$
$$\vdots \qquad\qquad \vdots$$
$$\Rrightarrow \quad \underbrace{(X_{T-p},\ldots,X_T)}_{v.\ observados} \quad \underbrace{X_{T+1}^{*(B)},\ldots,X_{T+h}^{*(B)}}_{v.\ futuros}$$

Referencias básicas: Thombs y Schucany 1990, Cao et al. 1997, y Pascual et al. 1998.

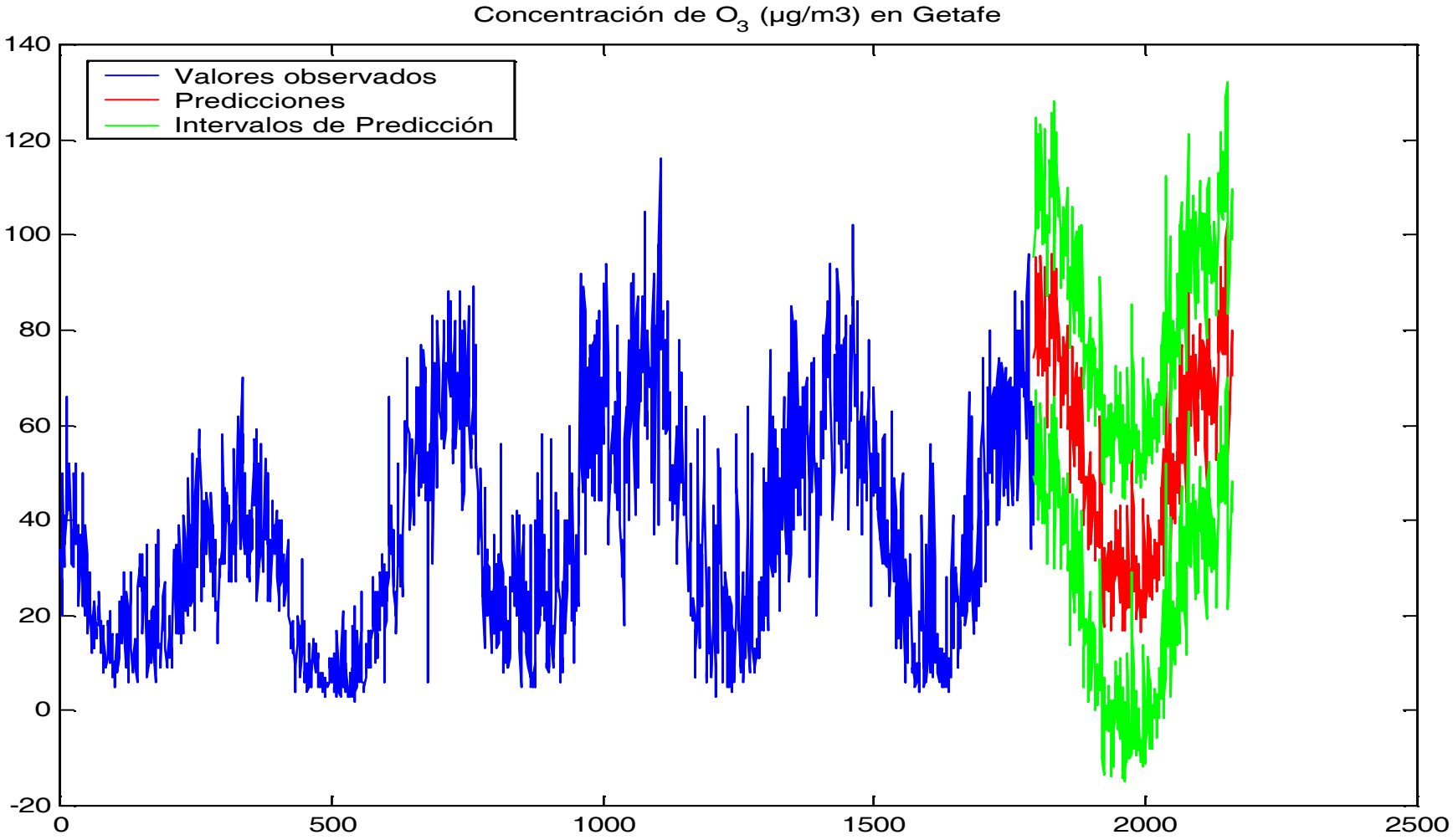
Extensiones: A modelos lineales generales en Alonso et al. 2002 y aplicaciones a epidemiología en Alonso y Romo 2005.

Estimadores kernel de las densidades predictivas



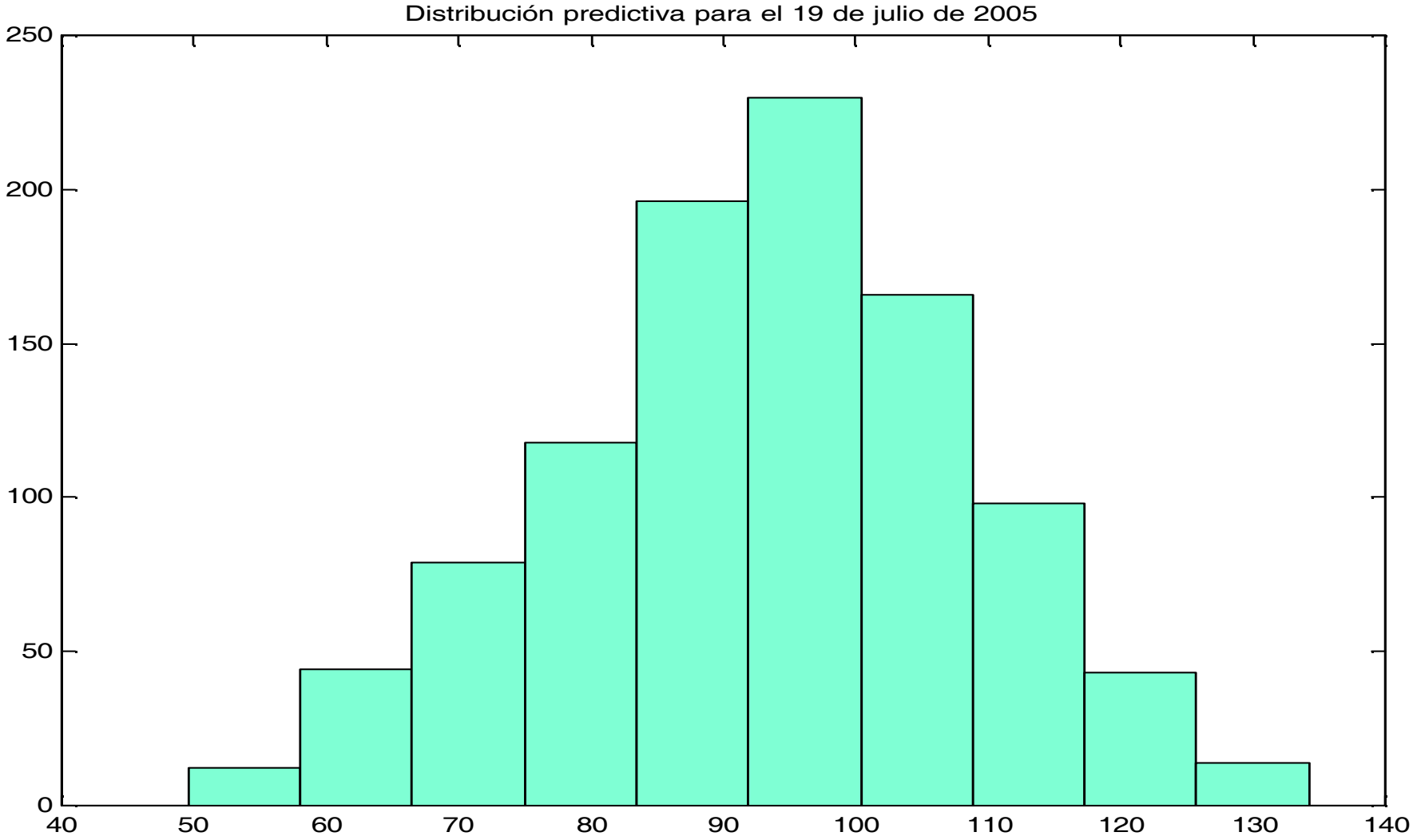
Aplicación a la monitorización - 1

Predicción de los niveles medios de ozono:



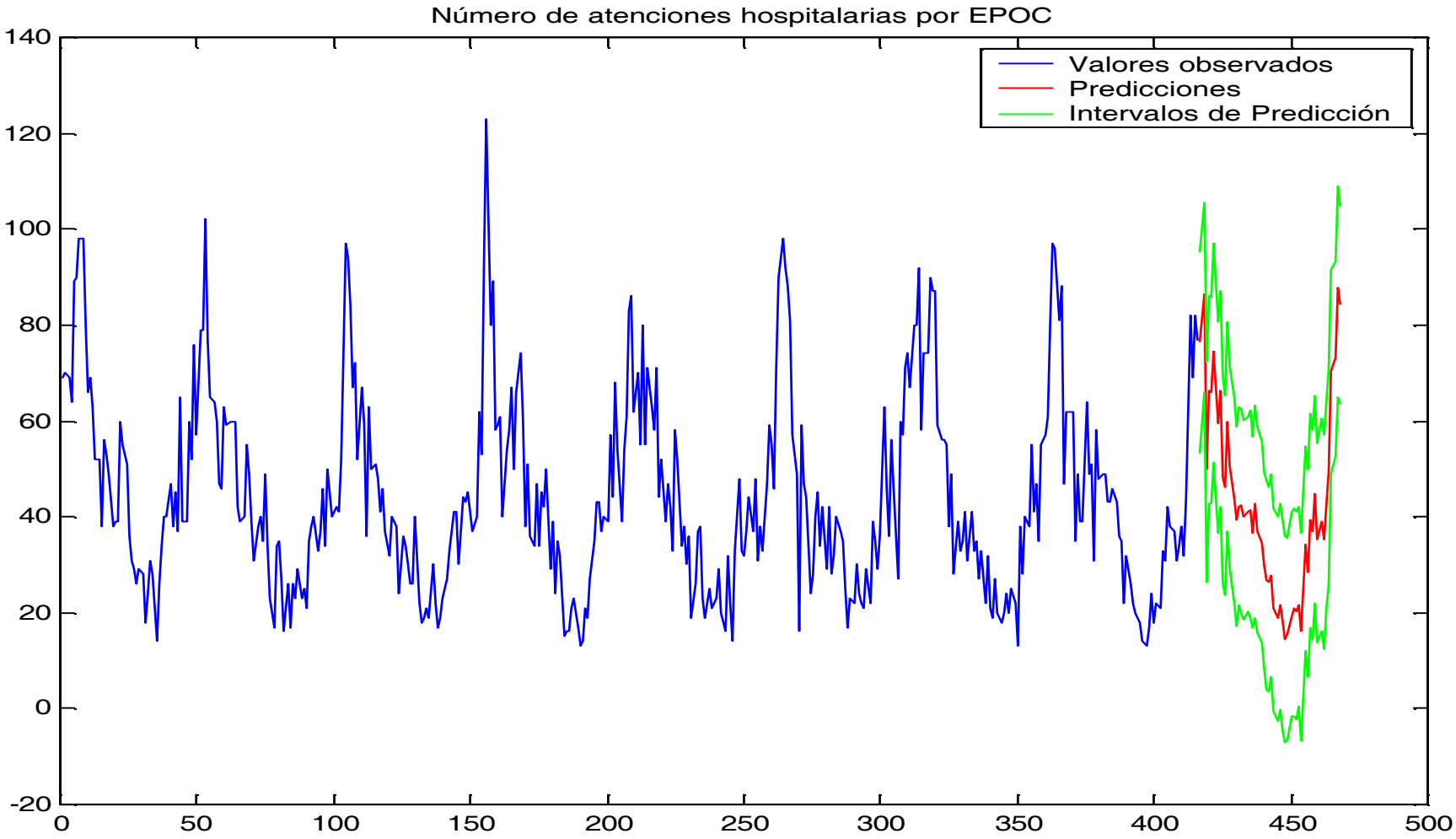
Aplicación a la monitorización - 1

Distribución de la predicción para HOY:

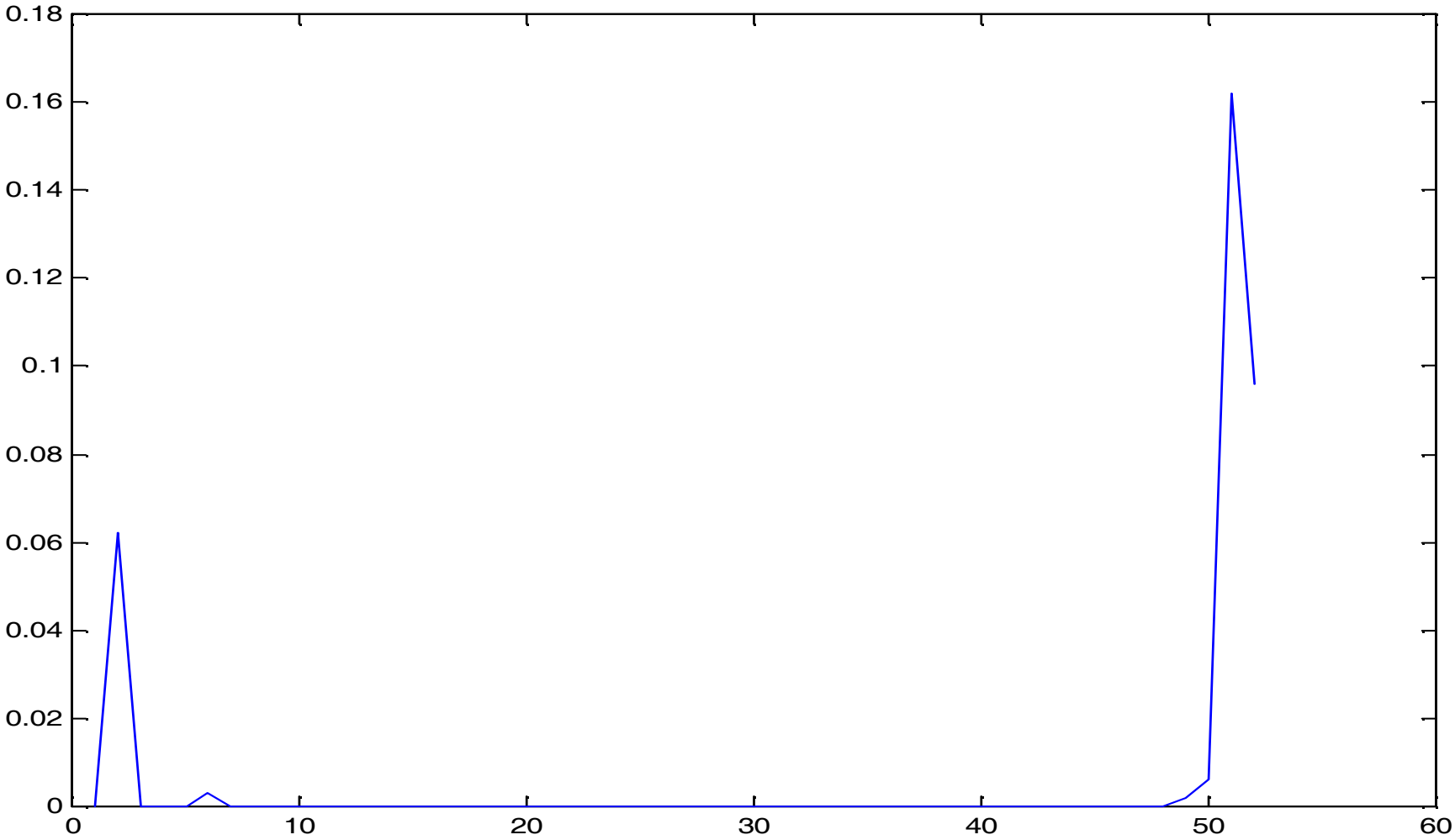


Aplicación a la monitorización - 2

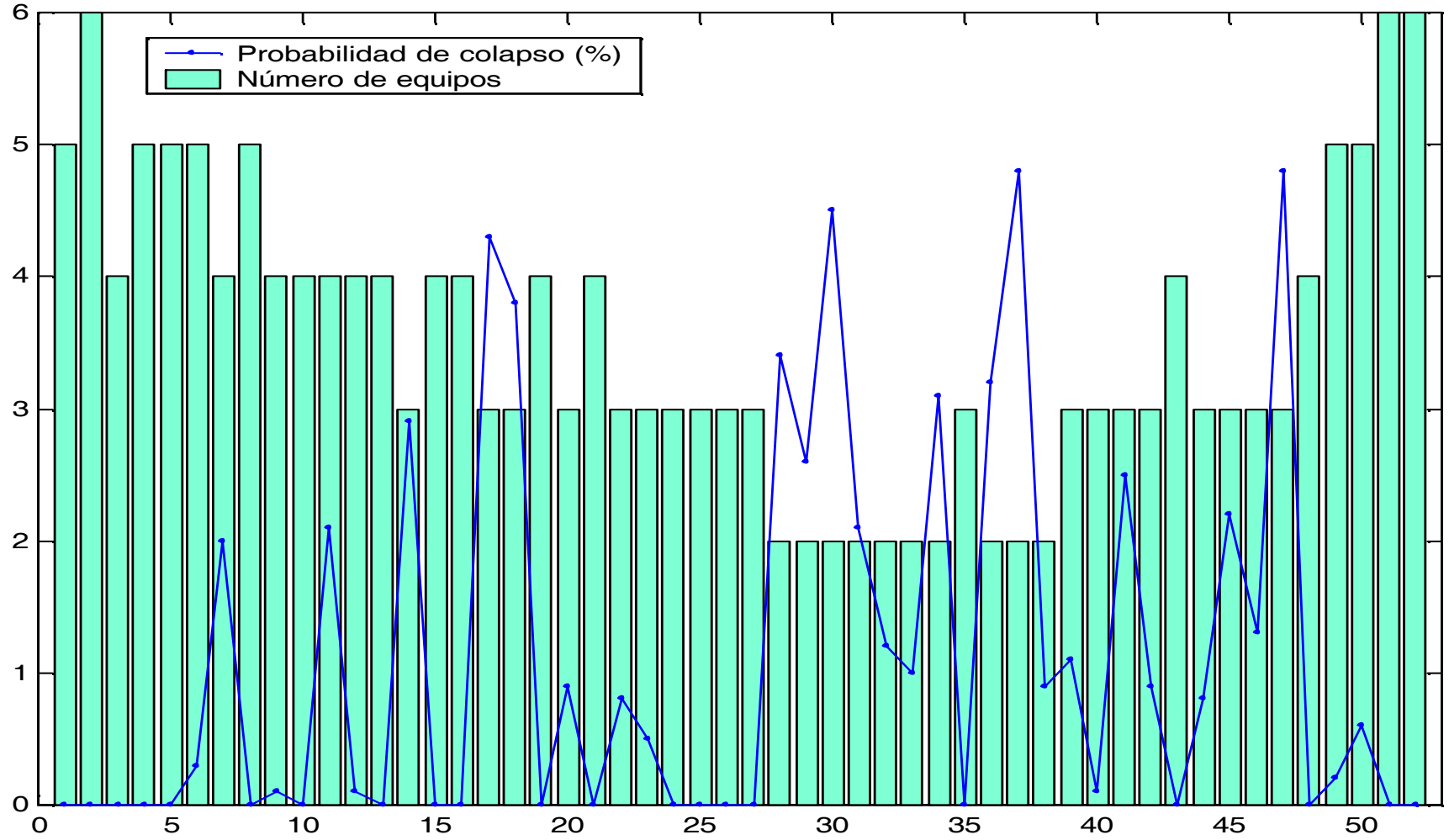
Predicción del número de atenciones hospitalarias por EPOC:



Probabilidades de “colapso” del sistema (100 atenciones por EPOC):



Número de equipos de atención a EPOC (20 atenciones/equipo) tal que la probabilidad de colapso sea inferior al 95 %:



Estructura

1. Introducción.

2. Métodos de remuestreo para datos i.i.d.

3. Métodos de remuestreo para series temporales.
   - Métodos basados en modelos.
   - Métodos no basados en modelos.

4. Aplicación a la monitorización de variables ambientales y epidemiológicas.

5. Conclusiones y Extensiones.

Métodos de remuestreo en series temporales

Métodos no basados en modelos

Elemento común. Utilización de bloques o subseries $\boldsymbol{X}_t^l = (X_t, \ldots, X_{t+l-1})$ de observaciones consecutivas como "aproximación" del modelo $\mathcal{P}$.

Referencias básicas.

• Künsch 1989 y Liu y Singh 1992:
a. Estimador de la varianza con el *jackknife por bloques móviles*, MBJ.
b. Estimadores de la varianza y la distribución con el *bootstrap por bloques móviles*, MBB.

• Politis y Romano 1994a: *submuestreo* para datos dependientes.

• Politis y Romano 1994b: *bootstrap estacionario* con bloques de tamaño aleatorio.

Métodos no basados en modelos

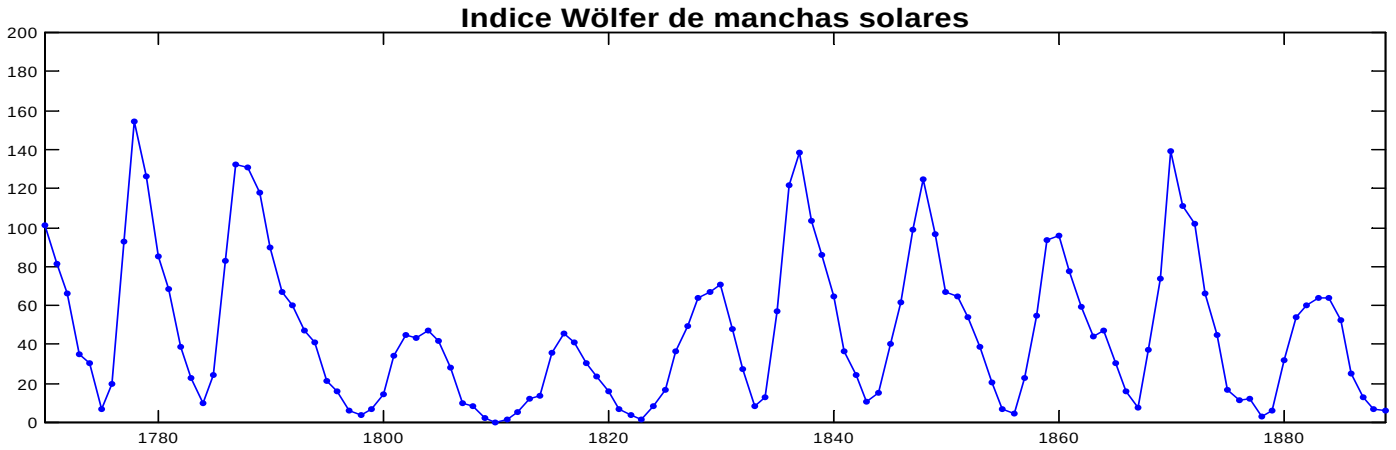
Problema. Sea $T_N = T(\rho_N^m)$, donde $\rho_N^m = n^{-1}\sum_{t=1}^{n}\delta_{\boldsymbol{X}_t}$ es la distribución empírica m-dimensional y $\boldsymbol{X}_t = (X_t, \ldots, X_{t+m-1})$ para $t = 1, \ldots, n = N - m + 1$ son los bloques de m observaciones consecutivas de $\boldsymbol{X}$.

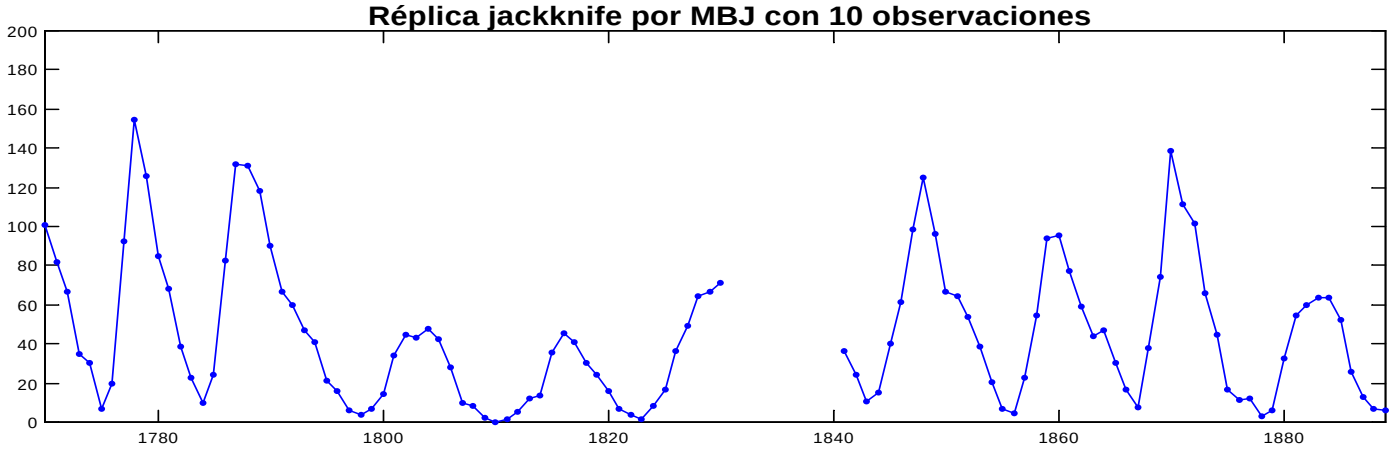
MBJ: Eliminar o infravalorar l bloques de "observaciones" consecutivas $(\boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_{j+l})$. Sea el estadístico $T_N^{(j)} = T(\rho_N^{m,(j)})$, donde $\rho_N^{m,(j)} = (n - \|w_n\|_1)^{-1}\sum_{t=1}^{n}(1 - w_n(t-j))\delta_{\boldsymbol{X}_t}$.

Un estimador de la varianza de $\sqrt{n}T_N$ es:

$$v_{MBJ} = \frac{(n - \|w_n\|_1)^2}{\|w_n\|_2^2(n-l+1)}\sum_{j=0}^{n-l}\left(T_N^{(j)} - T_N^{(\cdot)}\right)^2,$$

donde $T_N^{(\cdot)} = (n-l+1)^{-1}\sum_{j=0}^{n-l}T_N^{(j)}$, y los pesos verifican: $0 \leq w_n(i) \leq 1$ para $i \in \mathbb{Z}$, y $w_n(i) > 0$ para $1 \leq i \leq l$.


Indice Wölfer de manchas solares
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880


Réplica jackknife por MBJ con 10 observaciones
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880

Métodos de remuestreo propuestos en Alonso et al. 2003

Jackknife por bloques móviles omitidos (M^{2}BJ)

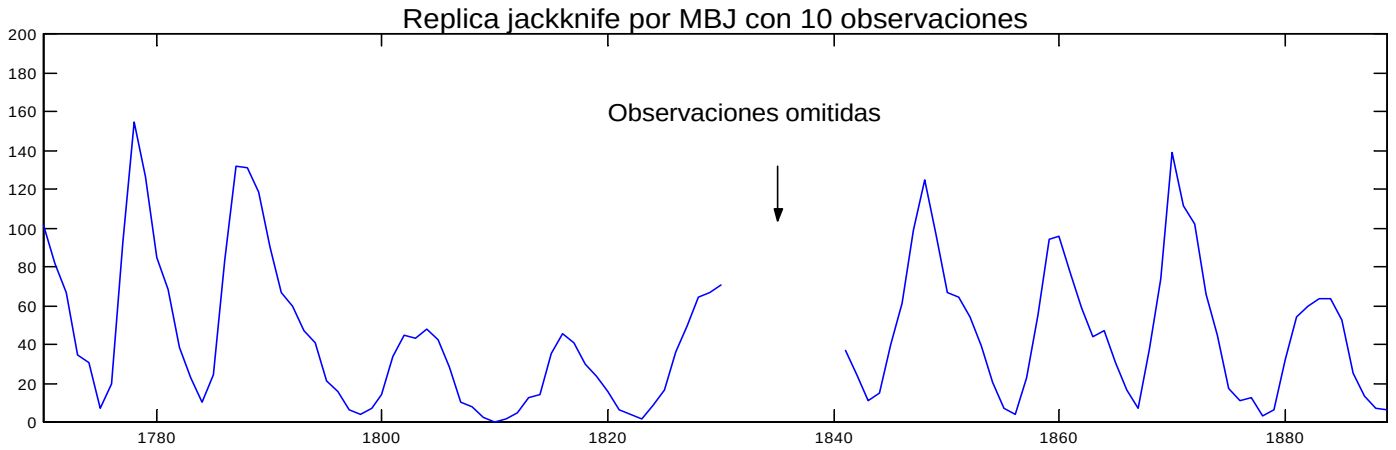
Solución por M^{2}BJ. Utilizar el estadístico $\widetilde{T}_{N,j} = T(\widetilde{\rho}_N^{m,(j)})$, donde $\widetilde{\rho}_N^{m,(j)} = n^{-1}\sum_{t=1}^{n}(1-w_n(t-j))\delta_{\boldsymbol{X}_t} + w_n(t-j)\delta_{\widehat{\boldsymbol{X}}_{t,j}}$ y $\widehat{\boldsymbol{X}}_{t,j}$ es una estimación de la "observación omitida" $\boldsymbol{X}_t$.

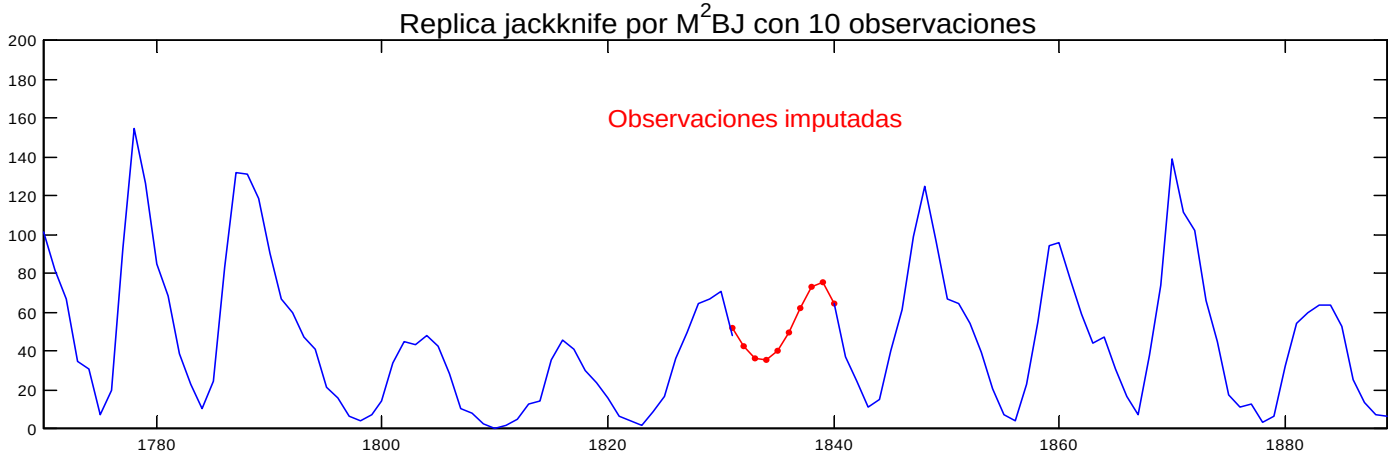
Un estimador M^{2}BJ de la varianza de $\sqrt{N}T_N$ es:

$$\widetilde{v}_{M^2BJ} = (n-l+1)^{-1}\|\boldsymbol{w}_n\|_2^{-2}\sum_{j=0}^{n-l}\left(\widetilde{T}_N^{(j)} - \widetilde{T}_N^{(\cdot)}\right)^2.$$

Un estimador M^{2}BJ de la distribución de $\tau_N(T_N - T(F))$ es:

$$\widetilde{H}_{M^2BJ}(x) = (n-l+1)^{-1}\sum_{j=0}^{n-l} I\left(\tau_l l^{-1}(n-l)(\widetilde{T}_N^{(j)} - T_N) \leq x\right).$$


Replica jackknife por MBJ con 10 observaciones
Observaciones omitidas
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880


Replica jackknife por M2BJ con 10 observaciones
Observaciones imputadas
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880

Métodos no basados en modelos

MBB: Suponemos que $n = kl$, y seleccionamos k bloques de l "observaciones" $\boldsymbol{Y}^*_{j_i} = (\boldsymbol{X}^*_{j_i+1}, \ldots, \boldsymbol{X}^*_{j_i+l})$ para $i = 1, \ldots, k$.

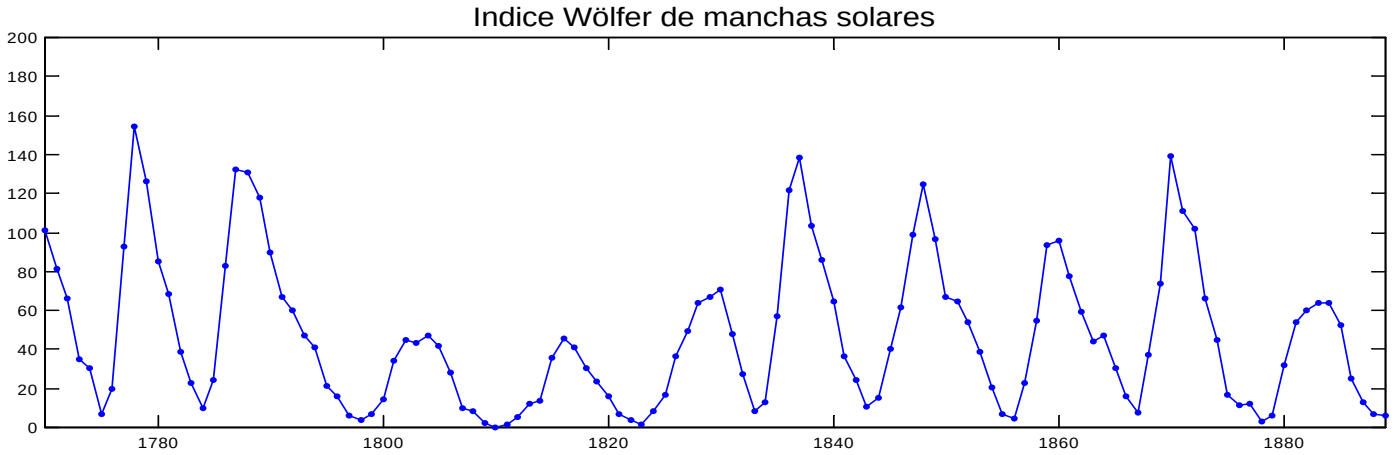
Sea $\boldsymbol{X}^* = (\boldsymbol{Y}^*_{j_1}, \ldots, \boldsymbol{Y}^*_{j_k}) = (\boldsymbol{X}^*_1, \boldsymbol{X}^*_2, \ldots, \boldsymbol{X}^*_n)$ la serie que resulta de unir los k bloques, $\rho_N^{m\,*} = n^{-1} \sum_{i=1}^{k} \sum_{t=1}^{l} \delta_{\boldsymbol{X}^*_{j_i+t}} = n^{-1} \sum_{t=1}^{n} \delta_{\boldsymbol{X}^*_t}$ es la distribución empírica m-dimensional bootstrap y $T^*_N = T(\rho_N^{m\,*})$ el estadístico análogo bootstrap.

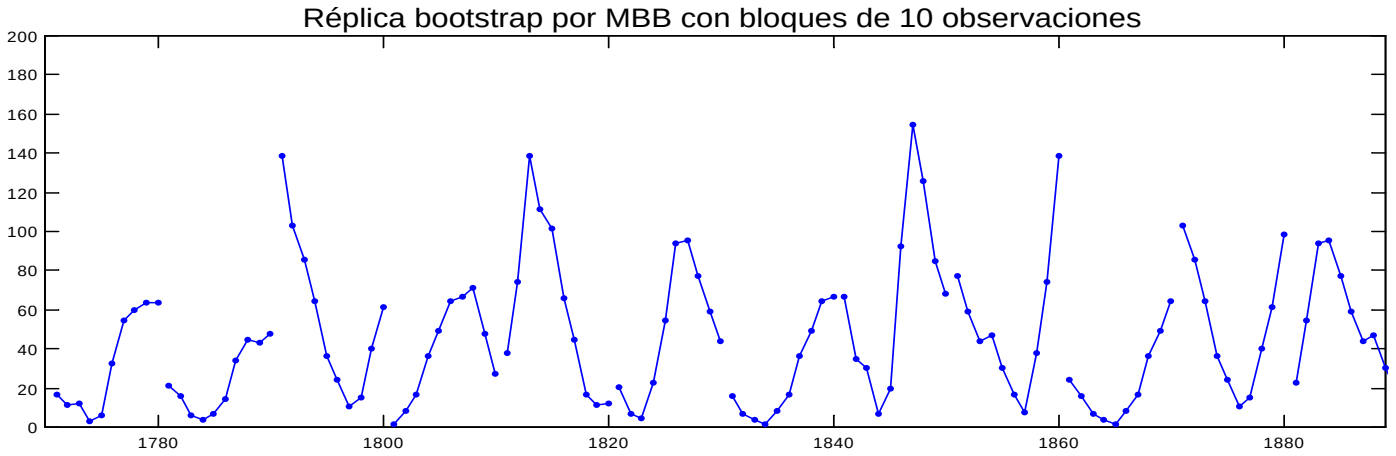
Un estimador de la varianza de $\sqrt{n} T_N$ es:

$$v_{MBB} = var^*(\sqrt{n} T(\rho_N^{m\,*})).$$

Un estimador de la distribución muestral H_N de $\sqrt{n}(T_N - T(\rho^m))$ es:

$$H_{MBB}(x) = \Pr^* \left\{ \sqrt{n}(T^*_N - \mathrm{E}_*[T^*_N]) \leq x \right\}.$$


Indice Wölfer de manchas solares
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880


Réplica bootstrap por MBB con bloques de 10 observaciones
200
180
160
140
120
100
80
60
40
20
0
1780
1800
1820
1840
1860
1880

Métodos de remuestreo propuestos en Alonso et al. 2003

Bootstrap por bloques móviles omitidos (M^{2}BB)

Réplica por M^{2}BB. Dada una réplica por MBB $\boldsymbol{X}^* = (\boldsymbol{Y}^*_{j_1}, \boldsymbol{Y}^*_{j_2}, \ldots, \boldsymbol{Y}^*_{j_k})$, obtenemos una réplica por M^{2}BB introduciendo bloques de observaciones entre cada dos bloques, i.e. $\widetilde{\boldsymbol{X}}^* = (\boldsymbol{Y}^*_{j_1}, \widehat{\boldsymbol{Y}}^*_{j_1}, \boldsymbol{Y}^*_{j_2}, \ldots, \boldsymbol{Y}^*_{j_k}, \widehat{\boldsymbol{Y}}^*_{j_k}) = (\widetilde{\boldsymbol{X}}^*_1, \ldots, \widetilde{\boldsymbol{X}}^*_n)$.

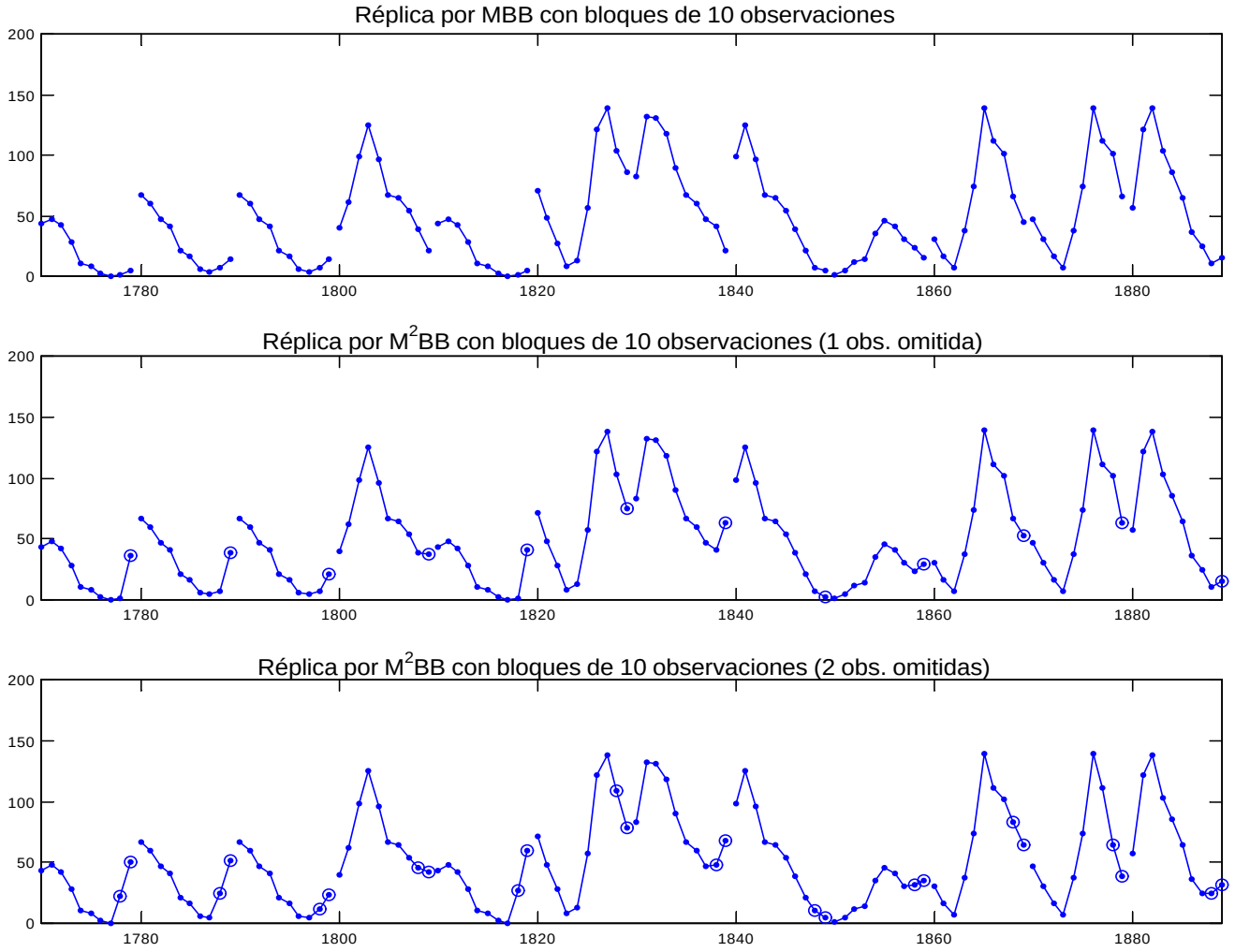
Solución por M^{2}BB. Utilizar el estadístico bootstrap $\widetilde{T}^*_N = T(\widetilde{\rho}^{m\,*}_N)$, donde $\widetilde{\rho}^{m\,*}_N = n^{-1}\sum_{t=1}^{n} \delta_{\widetilde{\boldsymbol{X}}^*_t}$ es la distribución empírica m-dimensional bootstrap.

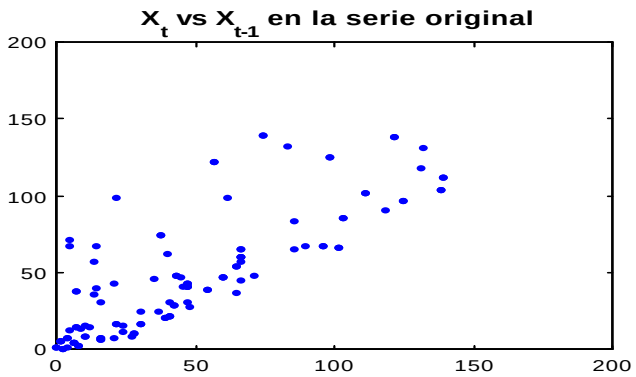
Un estimador de la varianza de $\sqrt{n}T_N$ es:

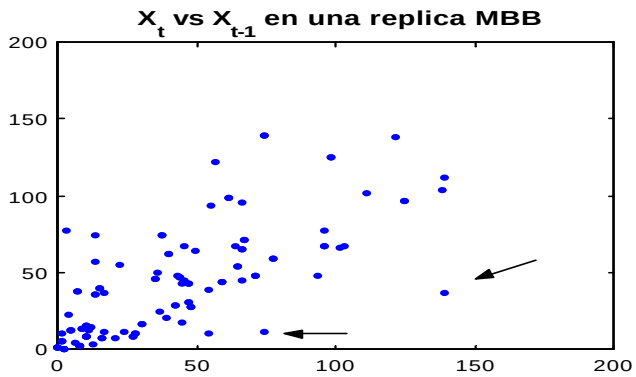
$$v_{M^2BB} = var^*(\sqrt{n}T(\widetilde{\rho}^{m\,*}_N)).$$

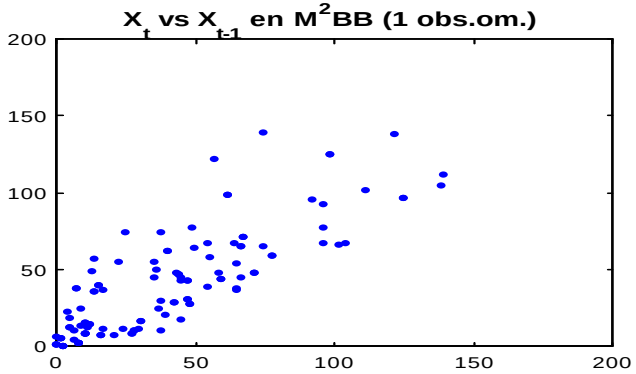
Un estimador de la distribución muestral H_N de $\sqrt{n}(T_N - T(\rho^m))$ es:

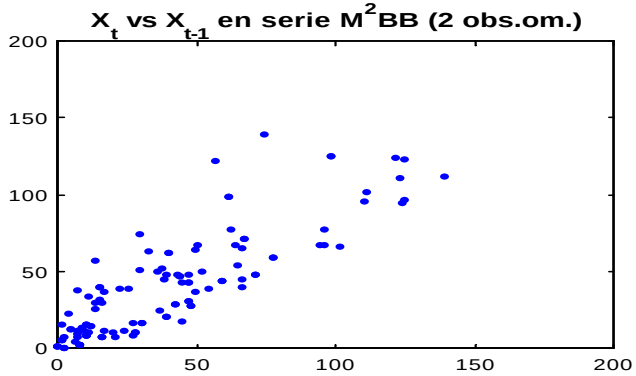
$$H_{M^2BB}(x) = \Pr{}^*\left\{\sqrt{n}(\widetilde{T}^*_N - \mathrm{E}_*[\widetilde{T}^*_N]) \leq x\right\}.$$


Réplica por MBB con bloques de 10 observaciones
Réplica por M2BB con bloques de 10 observaciones (1 obs. omitida)
Réplica por M2BB con bloques de 10 observaciones (2 obs. omitidas)
200
150
100
50
0
1780
1800
1820
1840
1860
1880


X_t vs X_{t-1} en la serie original
200
150
100
50
0
0
50
100
150
200


X_t vs X_{t-1} en una replica MBB
200
150
100
50
0
0
50
100
150
200

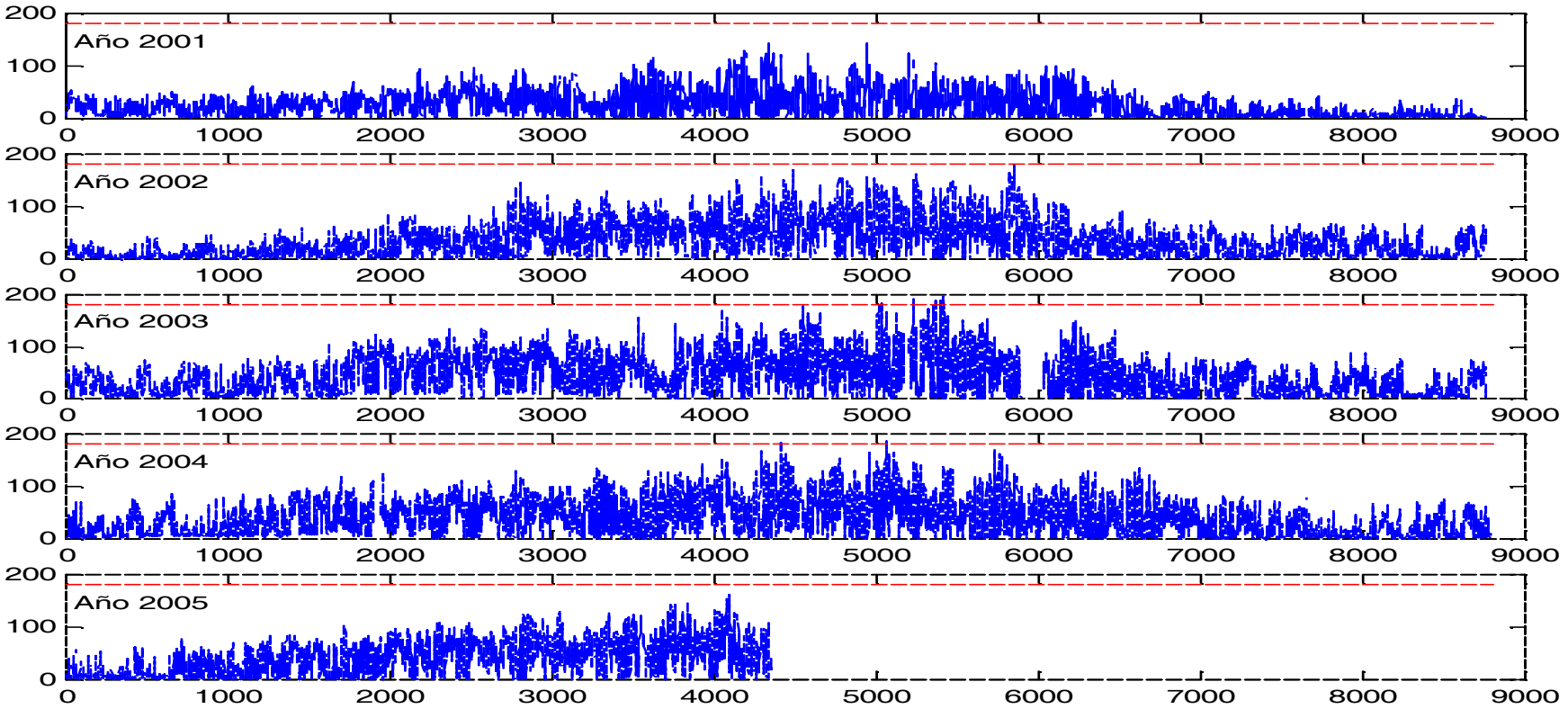

X_t vs X_{t-1} en M^2BB (1 obs.om.)
200
150
100
50
0
0
50
100
150
200


X_t vs X_{t-1} en serie M^2BB (2 obs.om.)
200
150
100
50
0
0
50
100
150
200

Aplicación a la monitorización - 1

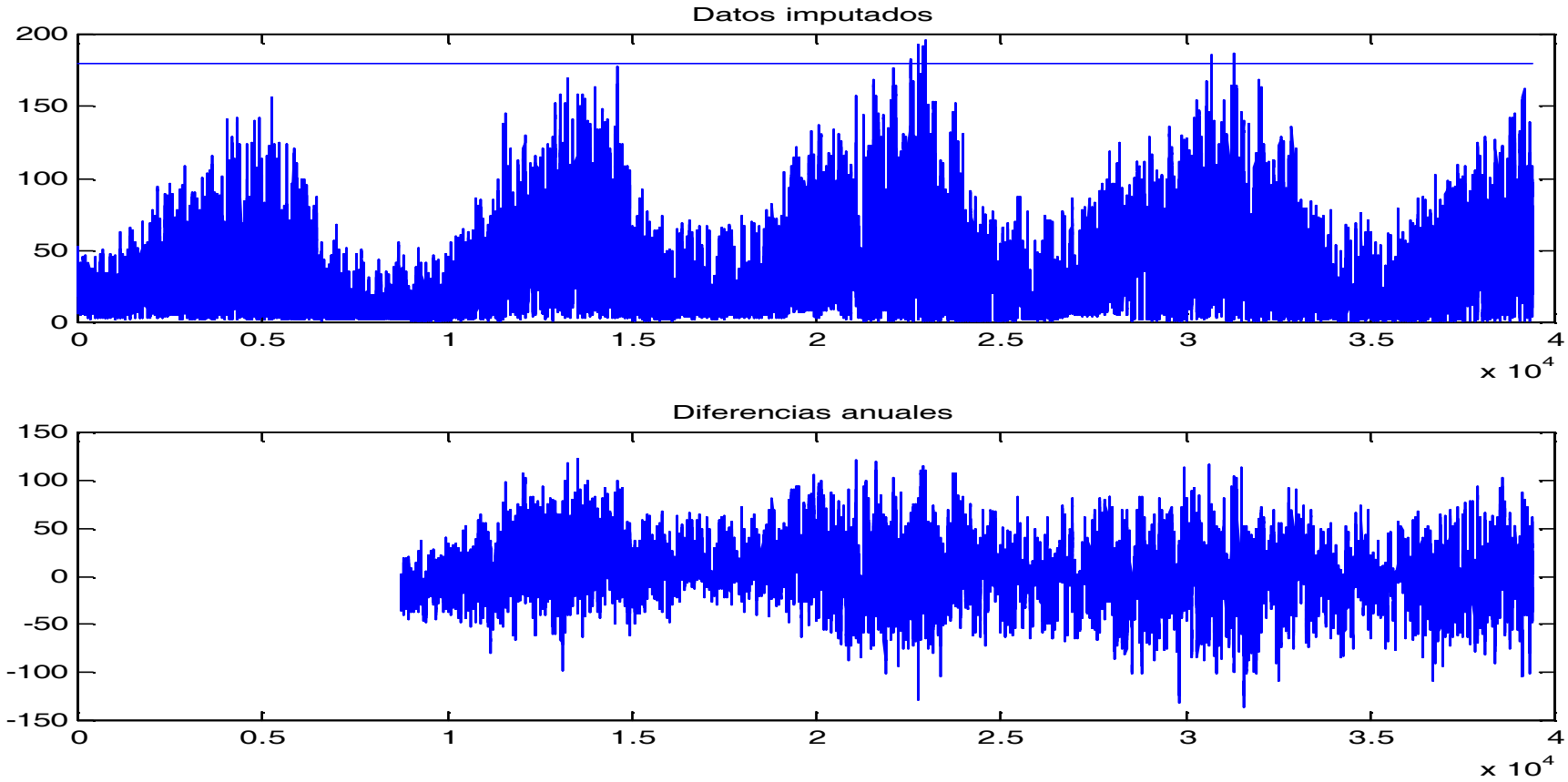
Datos horarios:

¿Qué modelo utilizar?



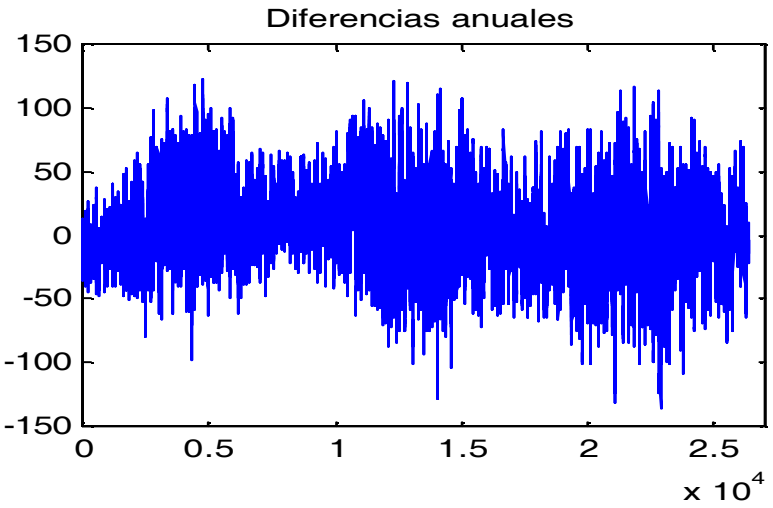
Las horas se representan en formato solar. Hora solar = Hora local -2 (en verano) o -1 (en invierno).

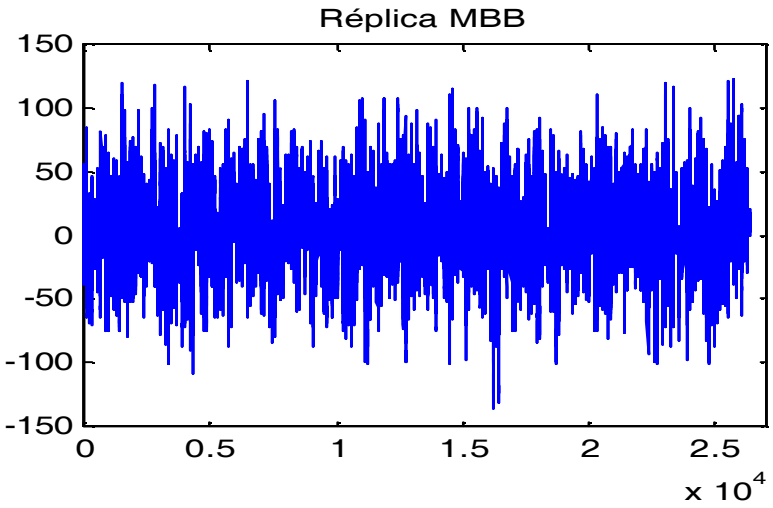
Diferencias anuales ($\sim 24 \times 365 = 8760$):

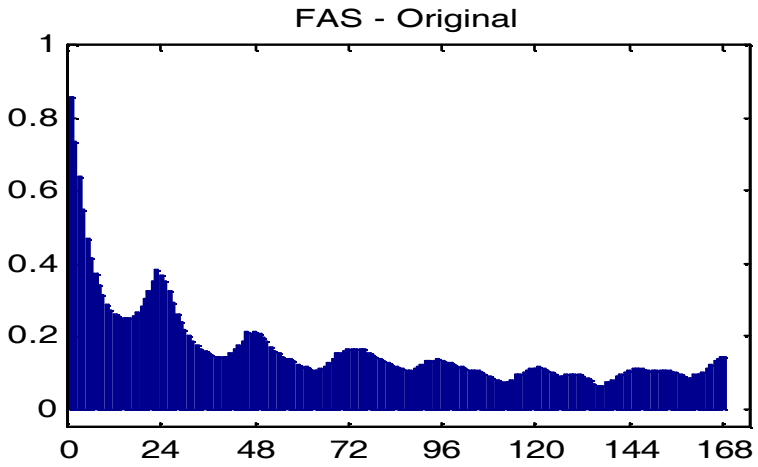


¿Qué modelo utilizar? ó ¿Nos interesa modelizar?

Réplicas MBB:







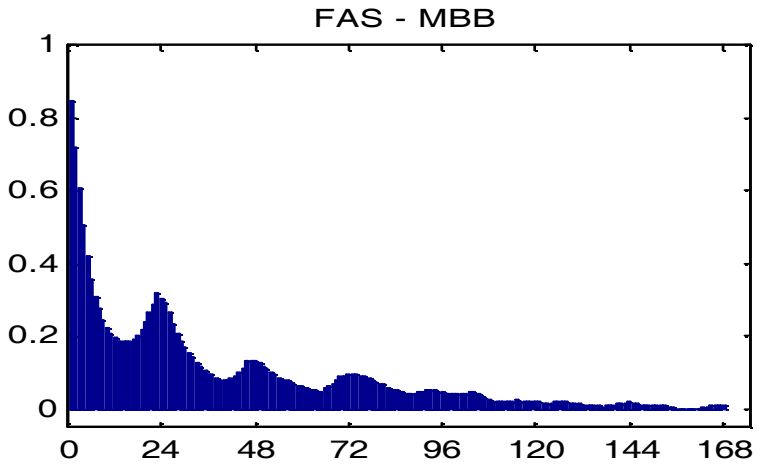


Gráfico de control Shewhart para las medias octo-horarias:

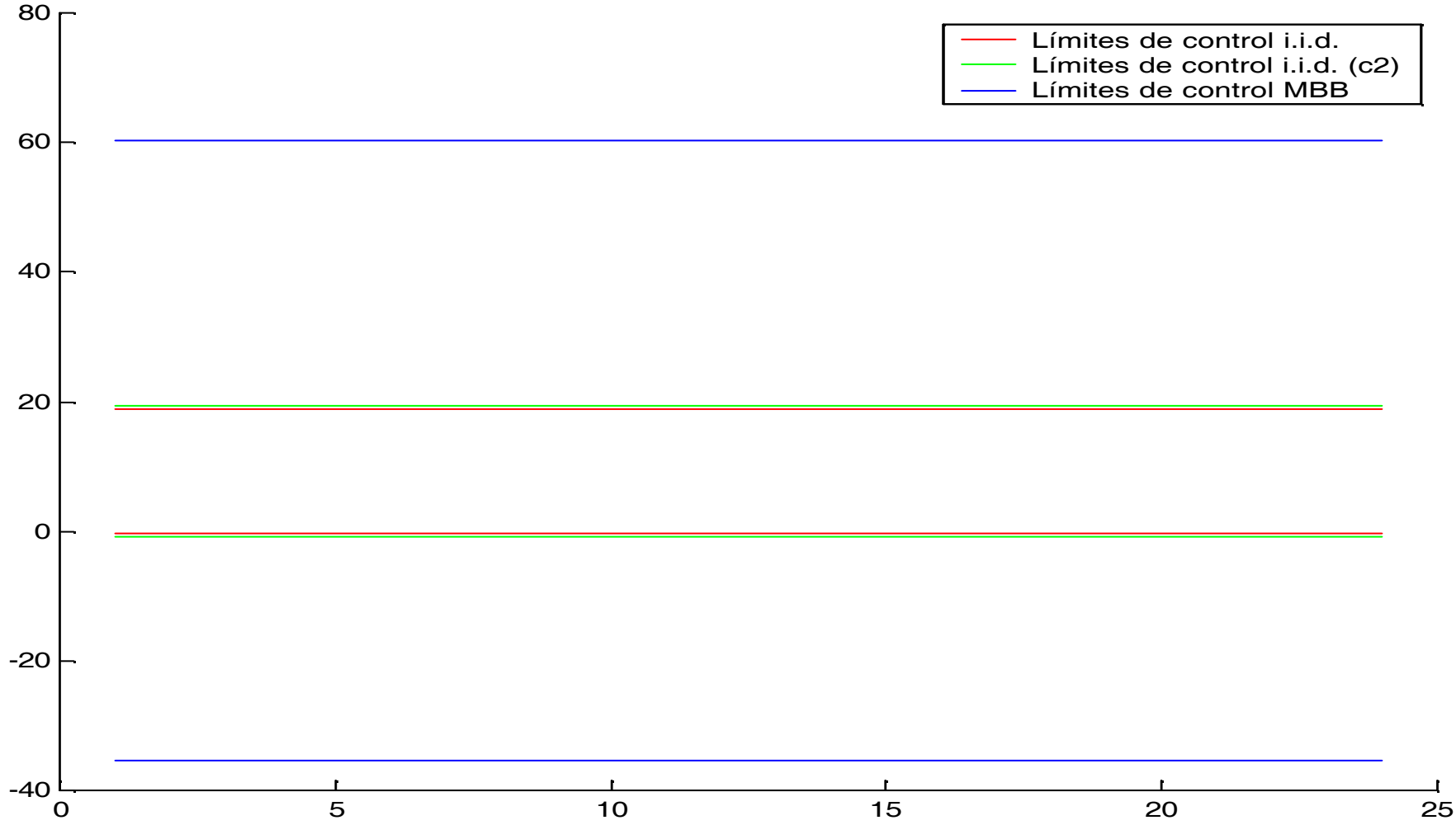
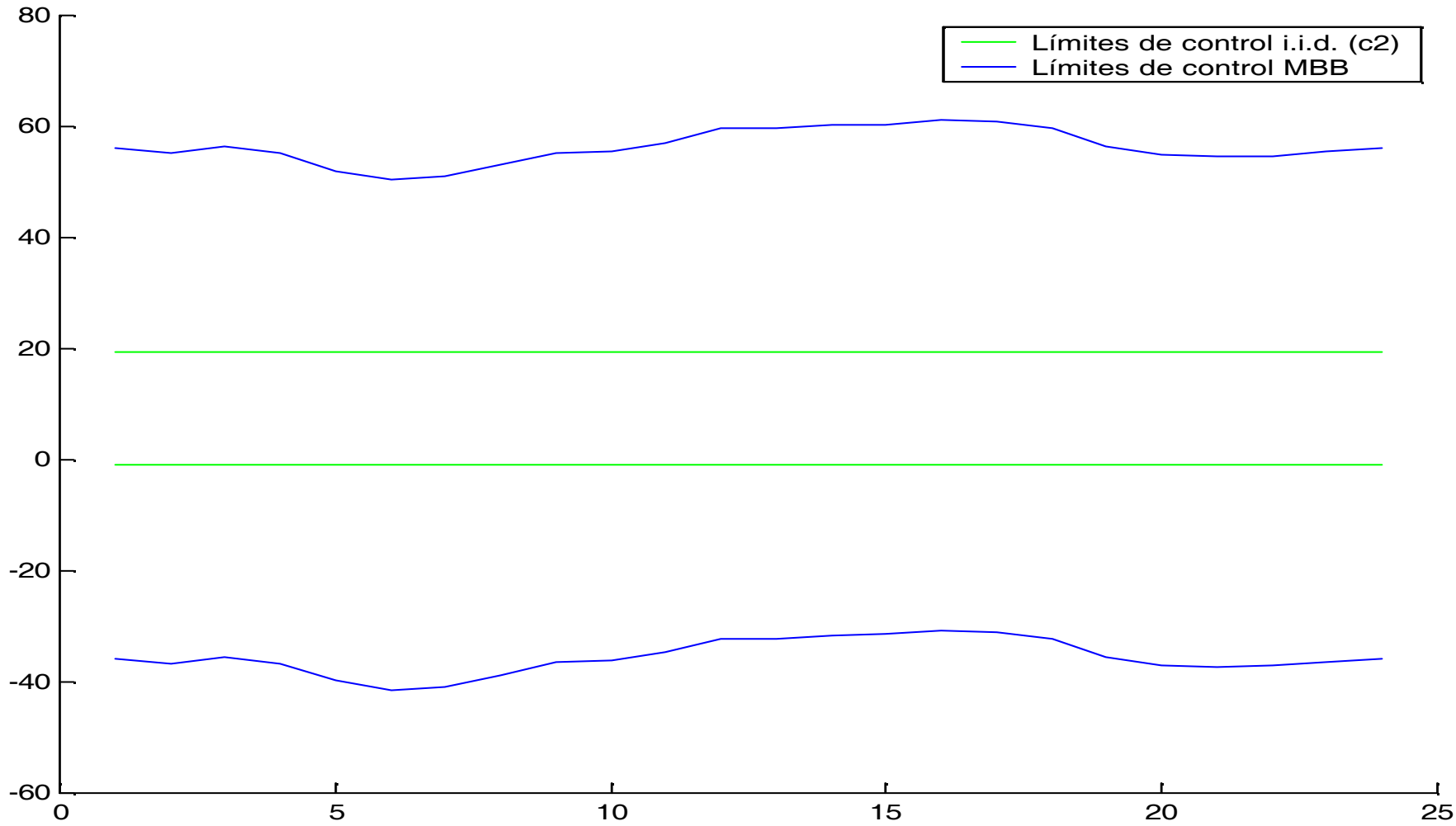


Gráfico de control Shewhart para las medias octo-horarias (diario):



Estructura

1. Introducción.

2. Métodos de remuestreo para datos i.i.d.

3. Métodos de remuestreo para series temporales.
   - Métodos basados en modelos.
   - Métodos no basados en modelos.

4. Aplicación a la monitorización de variables ambientales y epidemiológicas.

5. Conclusiones y Extensiones.

Conclusiones

▷ Problema de construir intervalos de predicción para procesos lineales.

- Se propone un procedimiento basado en el sieve bootstrap, se ilustra su comportamiento en muestras finitas (en Alonso et al. 2002).
- Se extiende el procedimiento anterior a series con comportamiento estacional y se ilustra su utilidad en datos de vigilancia epidemiológica (en Alonso y Romo 2005).

▷ Métodos de remuestreo por bloques móviles basados en técnicas de valores omitidos (en Alonso et al. 2003).

- **M^{2}BJ**: se propone una transición suave que tenga en cuenta la estructura de dependencia de la serie temporal.
- **M^{2}BB**: se propone un mecanismo para unir los bloques que corrige los efectos de la unión completamente aleatoria de los bloques del MBB.

Extensiones

- Estudio de procedimientos bootstrap en modelos más generales que permitan contemplar características presentes en datos diarios/horarios:
  - Estacionalidad múltiple: 365 y 7 (en diarios) 168 y 24 (en horarios).
  - Larga memoria, v.g., modelos ARFIMA.

- Estudio de procedimientos de remuestreo basados en bloques omitidos que permitan estacionalidad y para gráficos de control CUSUM y EWMA.
  - Utilización en datos de vigilancia epidemiológica y
  - Utilización en datos de vigilancia medioambiental.

Referencias

Alonso, A. M., Peña, D., y Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, **100**, 1–11.

Alonso, A. M., Peña, D., y Romo, J. (2003). Resampling time series by missing values techniques. *Annals of the Institute of Statistical Mathematics*, **55**, 765–796.

Alonso, A. M. y Romo, J. (2005). Forecast of the expected non-epidemic morbidity of acute diseases using resampling methods. *Journal of Applied Statistics*, **32**, 281–295.

Bickel, P. J. y Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9**, 1196–1217.

Bickel, P. J., Götze, F., y van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, **7**, 1–31.

Bose, A. (1990). Bootstrap in moving average models. *Annals of the Institute of Statistical Mathematics*, **42**, 753–768.

Cao, R., Febrero-Bande, M., Gonzalez-Manteiga, W., Prada-Sanchez, J.M., y Garcia-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. *Communications in Statistics. Simulation and Computation*, **26**, 961–978.

Davison, A. C. y Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. y Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54–77.

Efron, B. y Tibshirani, R. J. (1993). *Introduction to the bootstrap*. New York: Chapman & Hall.

Franke, J., Kreiss, J.-P., y Mammen, E. (1997). Bootstrap of kernel smoothing in nonlinear time series. Technical Report, Technische Universität Braunschweig, Braunschweig.

Freedman, D. A. (1984). On bootstrapping two-stage least-square estimates in stationary linear models. *The Annals of Statistics*, **12**, 827–842.

Kreiss, J.-P. y Franke, J. (1992). Bootstrapping stationary autoregressive moving average models. *Journal of Time Series Analysis*, **13**, 297–317.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, **17**, 1217–1241.

Liu, R. Y. y Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In: *Exploring the Limits of Bootstrap,* R. Lepage and L. Billard eds., 225–248. New York: Wiley.

Pascual, L., Romo, J., y Ruiz, E. (1998). Bootstrap predictive inference for ARIMA processes. Working Paper 98-86, Universidad Carlos III de Madrid, Madrid.

Politis, D. N. y Romano, J. F. (1994a). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, **22**, 2031–2050.

Politis, D. N. y Romano, J. F. (1994b). The stationary bootstrap. *Journal of the American Statistical Association*, **89**, 1303–1313.

Shao, J. y Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Shao, J. y Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, **17**, 1176–1197.

Thombs, L. A. y Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, **85**, 486–492.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, **18**, 1438–1452.