Trabajo Fin de Máster: Predicción en series de tiempo con Modelos Aditivos

Autor: Alejandro Calvo Rodríguez

Director: Juan Manuel Vilar

Máster en Técnicas Estadísticas

Universidade da Coruña

Curso 2008-09

Índice

Introducción

Una Serie de Tiempo no deja de ser un conjunto de observaciones sobre valores que toma una variable en diferentes momentos del tiempo.

Son muchas las materias en las que resulta útil conocer el comportamiento futuro de ciertos fenómenos con el fin de planificar y prevenir. La principal utilidad de las Series de Tiempo es la de predecir lo que ocurrirá con una variable en el futuro a partir del comportamiento de esa variable en el pasado y de otros factores que pueden influir. Por tanto, el estudio de métodos de análisis y predicción de Series de Tiempo es una temática de reciente investigación en ámbitos académicos y está en continuo desarrollo.

En este trabajo se presenta un método para hacer predicciones que se basa únicamente en lo sucedido anteriormente en la serie. El método utilizará la estadística no paramétrica de forma que sólo se trabaje con funciones unidimensionales, ya que se sabe que cuando se trabaja con dimensiones altas los métodos son costosos y, en ocasiones, dan problemas. Se presenta el método desde el punto de vista práctico principalmente, de modo que estará orientado al trabajo con el software estadístico R.

Como decíamos, es sabido que la predicción de futuras observaciones en series de tiempo es uno de los problemas más importantes en este campo. Más concretamente, si tenemos observados n instantes en una serie de tiempo, el problema es predecir lo que ocurrirá en el instante $n + p$, con $p \in \mathbb{Z}^+$. Una

forma de abordar este problema es considerar la serie de tiempo como un proceso autorregresivo de orden q: denotando por $S_1, \ldots, S_n$ los n instantes observados de la serie, suponemos que en el instante $t \geq n$,

$$S_t = m(S_{t-1}, S_{t-2}, \ldots, S_{t-q}) + \varepsilon_t,$$

considerando ε_t el error, independiente de los S_is anteriores. Con este planteamiento, lo primero que nos preguntamos es como construir la función $m(\cdot)$.

Una posibilidad es suponer que $m(\cdot)$ se puede construir con un modelo paramétrico, lo cual reduciría el problema a estimar un número finito de parámetros. Esto se podría hacer usando modelo ARIMA, que es muy conocido. El problema de esta posibilidad es que la mayoría de los procesos no se ajustan a un modelo paramétrico, lo que nos empuja a usar modelos no paramétricos, sin hacer ninguna suposición sobre la forma de $m(\cdot)$.

Los métodos no paramétricos para predecir en series de tiempo los podemos considerar un caso particular de la estimación no paramétrica de la regresión bajo dependencia. Podemos destacar como trabajos relevantes en este tema los siguientes artículos: Györfi et al (1989), Härdle and Vieu (1992), Hart (1991), Masry and Tjostheim (1995), Hart (1996), Härdle et al (1997), Härdle et al (1998), Bosq (1998) y Vilar-Fernández and Cao (2007).

Como ya mencionamos antes, en este trabajo se presenta un algoritmo, implementado posteriormente en R, para la predicción en series de tiempo y búsqueda de intervalos de confianza para dichas predicciones. El algoritmo utilizará modelos aditivos para obtener las predicciones. Cabe señalar que para todos los cálculos necesarios para generar nuevos ejemplos y modelos se utilizo el propio software R y, en alguna ocasión aislada, el paquete SPSS.

Comenzaremos con un capítulo de introducción a los modelos aditivos, ya que en ellos nos basaremos para implementar nuestro método de predicción.

Mostraremos una introducción teórica, veremos como funcionan desde el punto de vista práctico y como trabaja con ellos el software R.

En un segundo capítulo veremos como se desarrolla el proceso de implementar el método. Formularemos el problema con detalle, propondremos un algoritmo para elegir los autorregresores y explicaremos como calcular unas bandas de confianza para las estimaciones.

A continuación, en el tercer capítulo, adjuntaremos el código implementado en R acompañado de las pertinentes explicaciones de que es lo que se hace en cada momento.

Por último, dedicaremos un capítulo a comprobar el funcionamiento del método presentado y compararlo con otros métodos ya existentes. Probaremos 23 series de diferentes ámbitos y características y nos centraremos en dos de ellas para hacer un estudio detallado.

Capítulo 1

Modelos Aditivos

En este capítulo introducimos los modelos aditivos para regresión múltiple y el algoritmo "backfitting" para su estimación. También mostraremos los paquetes y funciones más usuales para trabajar en el entorno R con dichos modelos.

Los modelos aditivos son una generalización del modelo de regrisión lineal usual, por lo que es importante señalar las limitaciones del modelo lineal y los motivos por los que se generalizan con los modelos aditivos.

Regresión simple y modelos lineales

Vamos a centrarnos, en primer lugar, en el problema de regresión múltiple estándar.

Sea (X, Y) un vector aleatorio, donde Y es la variable respuesta, unidmensional, y X la variable independiente, p-dimensional. Supongamos que tenemos n observaciones de la variable Y, denotadas por $y = (y_1, \ldots, y_n)^t$, y otras n del vector X, denotadas por $x^i = (x_{i1}, \ldots, x_{ip})$ con $i = 1, \ldots, n$.

Nuestro objetivo ahora es modelar como depende Y de $X = (X_1, \ldots, X_p)$.

Las principales razones para querer llegar a esto son: buscar un modelo que nos permita saber más sobre el proceso que nos da Y y poder hacer predicciones concretas de Y.

La regresión lineal múltiple se basa en el siguiente modelo

$$Y = \alpha + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon \tag{1.1}$$

donde $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$. Este modelo asume que $E(Y)$ depende linealmente de cada uno de los predicotres, $X_1, \ldots, X_n$.

Si esta suposición es correcta en nuestro caso, este modelo es muy útil y conveniente. Ya que proporciona una descripción sencilla de los datos, resume la contribución de cada predictor con unos simples coeficientes y, además, proporciona un método muy sencillo para hacer nuevas predicciones.

Exigir la condición de linealidad es exigir una condición bastante fuerte, en la mayoría de los casos no se cumplirá. Luego tendremos que generalizar la fórmula 1.3 a un caso más general. Una posibilidad es pensar en un modelo de regresión no paramétrica como el siguiente

$$Y = f(X_1, \ldots, X_p) + \varepsilon \tag{1.2}$$

que se define de forma bastante intuitiva. Uno de los principales problemas que tiene esta solución es la conocida como "malidición de la dimensionalidad" cuando el número de predictores es grande.

Para solucionar el problema de la dimensionalidad podemos recurrir a la regresión por projection-pursuit, que consiste en utilizar un modelo de la forma

$$Y = \sum_{k=1}^{K} h_k(\alpha_k^T X) + \varepsilon \tag{1.3}$$

sabiendo que $\alpha_k^t X$ es una proyección unidimensional del vector X y h_k es una función de esa proyección. Las direcciones α_k y el número de términos

K se eligen de forma que las predicciones salgan lo mejor posible.

Con este método se soluciona el problema de la dimensionalidad, pero cuando K es grande este modelo es muy dificil de interpretar. Podríamos continuar desarrollando este método pero, por lo ya dicho, es preferible dejarlo aquí.

Desarrollo del Modelo Aditivo

Una de las principales características de los modelos lineales era que los efectos de cada predictor se iban añadiendo de forma aditiva para estimar la respuesta. Los modelos aditivos consevan esta característica, ya que son definidos de la siguiente forma

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon \tag{1.4}$$

donde los errores ε son independientes de los X_j, $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$. Las f_j son funciones arbitrarias univariantes, una para cada predictor. Por el momento, asumamos que las f_j son funciones "suaves´´ y cada una es estimada de forma individual.

No es estrictamente necesario que las f_j sean funciones "suaves´´ y univariantes, pero en el contexto que nosotros aplicaremos los modelos podemos asumir que si lo son.

Ahora necesitamos un algoritmo para construir los f_j de forma coherente. La esperanza condicionada nos proporciona una motivación intuitiva para elaborar el algoritmo. Asumiendo el modelo 1.4 tenemos que

$$E(Y - \alpha - \sum_{j \neq k} f_j(X_j)|X_k) = f_k(X_k) \qquad \forall k \in \{1, \ldots, p\}.$$

Esto sugiere un algoritmo para construir los f_j, conocido como "Algoritmo de Backfitting". Utilizaremos como suavizadores no paramétricos la estimación

núcleo de Nadaraya-Watson y para la selección del parámetro de suavizado utilizaremos la técnica de validación cruzada mínimo cuadrática. Dicho algoritmo funciona en tres pasos como mostramos a continuación

1. Fijamos $\alpha = \bar{y}$ y $f_j \equiv 1$ para cada $j = 1, \ldots, p$.

2. Tomando $j = 1, \ldots, p$ definimos las f_j de la siguiente manera

$$f_j = S_j(y - \alpha - \sum_{k \neq j} f_k | x_j).$$

3. Repetir el paso anterior hasta que las variaciones entre las nuevas funciones definidas y las anteriores sean pequeñas (menores que una tolerancia establecida).

Cuando utilizamos el operador S_j, en el paso 2, se elimina de la respuesta la parte estimada por otras variables en los pasos anteriores. El S_k es un operador que aplicado a la respuesta da la estimación, el cual es diferente según el contexto: en regresión lineal múltiple este operador es la matriz proyección formado por las columnas de la matriz de diseño, en regresión no paramétrica es la matriz suavizadora dada por el método núcleo, por los polinomios locales, splines, etc.

R y Modelos Aditivos

Actualmente el software estadístico R está siendo una herramienta muy utilizada dentro del mundo de la estadística. Uno de los motivos que impulsa a su uso es que se trata de software libre y de código abierto, lo que motiva a muchos investigadores a implementar librerías para el mismo.

Y es la mencionada popularidad la que nos impulsa a usar el software R en este trabajo. Existen varias librerías que permiten trabajar con modelos aditivos, aquí usaremos la librería `mgcv`. Esta librería fue creada por Simon

Wood, acompañando al libro "Wood S.N. (2006) *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC Press".

Dentro de esta librería las funciones que vamos a usar principalmente son las siguientes

- `gam`: Permite construir el modelo aditivo y lleva integrada la estimación del suavizado.
- `predict`: Realiza estimaciones usando el modelo deseado.
- `summary`: Recopila información acerca de un modelo.

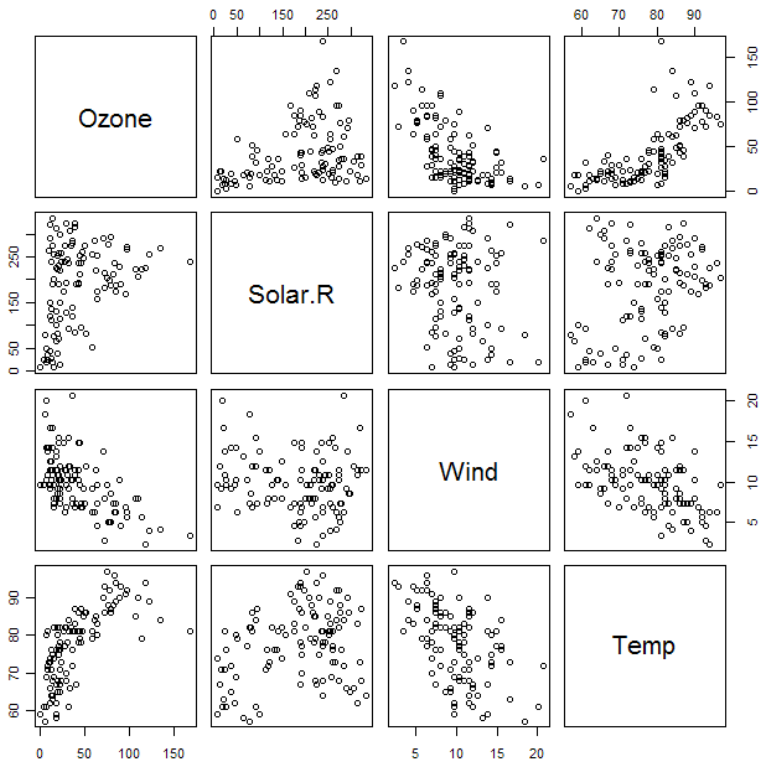

Figura 1.1: Dispersión entre las variables utilizadas.

A continuación presentamos un ejemplo en el que estimamos el nivel de Ozono en la atmósfera a partir de la influencia solar, el viento y la temperatura usando la libería `mgcv`. Los datos están dentro de los archivos de ejemplo de la propia librería. En la Figura 1.1 observamos como se relacionan las variables que intervienen.

Vamos a intentar buscar un modelo aditivo de la siguiente forma

$$E(Y|X_1, X_2, X_3) = \alpha + f_1(X_1) + f_2(X_2) + f_3(X_3) \tag{1.5}$$

donde

$$Y \equiv \text{Ozone}; \quad X_1 \equiv \text{Solar.R}; \quad X_2 \equiv \text{Wind}; \quad X_3 \equiv \text{Temp}.$$

Guardando los datos en la variable `aire` y usando los siguientes comandos en R

```
modelo.ozono<-gam(Ozone~s(Solar.R)+s(Wind)+s(Temp), data=aire)
summary(modelo.ozono)
```

obtenemos los detalles del modelo aditivo que estamos usando

```
Family: gaussian
Link function: identity

Formula:
Ozone ~ s(Solar.R) + s(Wind) + s(Temp)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.099      1.663   25.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
```

```
              edf Ref.df      F  p-value
s(Solar.R) 2.760  3.260  4.109  0.00698 **
s(Wind)    2.910  3.410 14.609 1.36e-08 ***
s(Temp)    3.833  4.333 12.786 7.46e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.723   Deviance explained = 74.7%
GCV score =  338.9   Scale est. = 306.83    n = 111
```

entre los cuales podemos ver el modelo, los grados de libertad en cada predictor, el p-valor del test de la influencia de cada predictor, etc. Utilizando `modelo.ozono$sp` podemos obtener los parámetros de suavizado estimados, en este caso: 0,0949, 0,2089 y 0,0345.

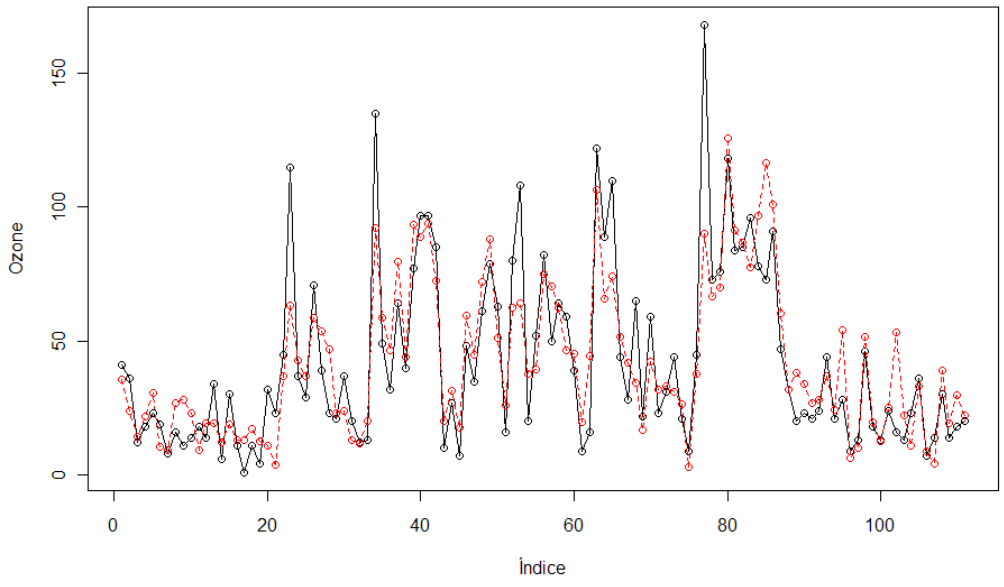

Figura 1.2: Datos reales y predichos del Ozono.

Una vez elaborado el modelo, nuestro interés se centra en hacer predicciones utilizando dicho modelo. Por ejemplo, si conocemos que en un determinado momento tenemos que

$$\text{Solar.R} = 127; \quad \text{Temp} = 74; \quad \text{Wind} = 12{,}1$$

podremos predecir que la variable `Ozone` será 12,44, utilizando el comando

```
predict(modelo.ozono,data.frame(Solar.R=127,Temp=74,Wind=12.1))
```

Con un comando similar podríamos ver como sería la estimación para cada uno de los datos. En la Figura 1.2 podemos observar los valores reales (linea continua) y los valores predichos por el modelos (línea discontinua).

En este capítulo hemos introducido los modelos aditivos de forma superfecial y las funciones básicas de R. Para un estudio más profundo se pueden ver los textos "Wood S.N. (2006) *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC Press" y "Hastie T.J. & Tibshirani R.J. (1990) *Generalized Additive Models.* Chapman and Hall".

Capítulo 2

Desarrollo del proceso

Formulación del Problema

Vamos a considerar una serie no estacional $\{S_t : t \in \mathbb{N}\}$. Supongamos que tenemos observado hasta el instante n. Podemos escribir la serie de la siguiente forma

$$S_t = m(S_{t-i_1}, S_{t-i_2}, \ldots, S_{t-i_p}) + \varepsilon_t \tag{2.1}$$

siendo $m(\cdot)$ la función de regresión usada para estimar S_t utilizando la información obtenida en $i_1, i_2, \ldots$ e i_p instantes antes y ε_t el error de estimación que es independiente de $S_1, S_2, \ldots, S_{t-1}$.

Para predecir S_{t+1} consideramos los datos

$$\{(Y_j, \widetilde{X}_j) : j \in \mathbb{N}\} \subseteq \mathbb{R} \times \mathbb{R}^p$$

donde $Y_j = S_{j+i_p}$ y $\widetilde{X}_j = (S_{j+i_p-i_1}, S_{j+i_p-i_2}, \ldots, S_j)$. Pero nos quedaremos con los datos que tenemos observados, es decir,

$$\{(Y_j, \widetilde{X}_j) : j = 1, 2, \ldots, n - i_p\}.$$

Fijándonos en la expresión (2.1), podemos considerar $\hat{m}$ un estimador de m, de forma que

$$\hat{m}(\widetilde{x}_{n-i_p+1}) = \hat{m}(s_{n+i_p-i_1}, s_{n+i_p-i_2}, \ldots, s_n).$$

La estimación usando modelos aditivos tiene ventajas ventajas respecto a los modelos paramétricos, ya que permitirán más flexibilidad, y ventajas respecto al modelo no paramétrico clásico, ya que al ser una suma de funciones univariantes no vamos a tener el problema que surgía al tener un número grande de autorregresores.

De esta manera podemos predecir también S_{t+l}, con $l \geq 1$, aunque a medida que aumentamos l también aumentamos el coste computacional: Cada vez que elijamos un l, los autorregresores serán diferentes y por tanto la muestra $\{(y_j, \widetilde{x}_j)\}$ también cambiará.

Existe un procedimiento alternativo, pero que proporcionará peores resultados por lo general, en el que sólo es necesario calcular una vez los autorregresores, sea cual sea el horizonte l, y por tanto tiene un coste computacional mucho menor. Éste consiste en implementar un algoritmo recursivo, de forma que si conocemos lo ocurrido hasta el instante n, estimamos lo ocurrido en $n+1$ y lo denotamos $\hat{S}_{n+1}$. Ahora consideramos que nuestros datos son

$$\{S_1, S_2, \ldots, S_n, \hat{S}_{n+1}\},$$

y, utilizando éstos, podremos obtener $\hat{S}_{n+2}$. Procediendo así, llegamos a estimar lo que ocurre en $n+l$ por $\hat{S}_{n+l}$. El principal problema de este método es que se apoya en estimaciones para hacer nuevas estimaciones, y así, a medida que avanzamos, el error cometido va a ser mucho mayor.

Sea cual sea el método que usemos el principal objetivo es estimar S_{n+1} y para ello debemos encontrar el estimador $\widehat{m}$ de m. Aquí utilizaremos un modelo aditivo y definiremos $\widehat{m}$ como

$$\hat{m}(s_{n+i_p-i_1}, s_{n+i_p-i_2}, \ldots, s_n) = f_1(s_{n+i_p-i_1}) + f_2(s_{n+i_p-i_2}) + \ldots + f_n(s_n) \quad (2.2)$$

donde las funciones se definen usando el "Algoritmo de Backfitting" visto en el capítulo "Modelos Aditivos". Una vez que sabemos como estimar S_{n+1} podremos buscar formas para estimar S_{n+l} siendo l cualquier horizonte.

Selección de los autorregresores

El objetivo de esta selección es escoger los elementos de la serie más adecuados para estimarla en un instante posterior en el que el valor de la serie es desconocido. Existen varias propuestas para solucionar este problema: Vieu (1994) y Yao y Tong (1994) proponen diversos métodos basados en la validación cruzada, mientras que Tjostheim y Auestad (1994) y Tscherning y Yang (2000) proponen uasr una versión no paramétrica del criterio del error final de predicción (FPE). Nosotros nos basaremos en este último, si bien haremos modificaciones.

La idea es buscar una medida que calcule el error que se comete al elegir unos u otros autorregresores, de forma que podamos elegir aquellos que nos den el error más pequeño. Si denotamos por Ω_S^p el conjunto de p autorregresores que estamos considerando, podemos definir $er(\Omega_S^p)$ como el error de estimación que se comete al usar los autorregresores del conjunto Ω_S^p. Los pasos del algoritmo que vamos a utilizar para seleccionar los autorregresores son los siguientes

Algoritmo 2.1.

1. Consideramos el conjunto de instantes, Ω, candidatos a ser elegidos como autorregresores. Consideraremos los 10-15 últimos instantes de los datos que tenemos, se podrían elegir más, incluso todos desde 1 hasta n, pero su coste computacional sería muy elevado.

2. Consideramos Ω_r^1, (con $r \in \Omega$) conjuntos formado por un único autorregresor r y calculamos $er(\Omega_r^1)$ para cada uno de ellos. Supongamos que r_1, sin pérdida de generalidad, es el autorregresor que tiene menor error. Entonces r_1 va a ser nuestro primero autorregresor escogido, pero puede haber más.

3. Consideramos $\Omega^2_{r_1,r}$ el conjunto que contiene el autorregresor ya escogido y otro tal que $r \in \Omega - \{r_1\}$. Ahora se presentan dos posibilidades

   a) $\forall r \in \Omega - \{r_1\} \quad er(\Omega^2_{r_1,r}) \geq er(\Omega^1_{r_1})$, entonces nos quedaremos con un único autorregresor r_1 y terminamos con el algoritmo.

   b) $\exists r_2 \in \Omega - \{r_1\}$ tal que $er(\Omega^2_{r_1,r_2}) < er(\Omega^1_{r_1})$, entonces consideramos $\Omega_R = \{r_1, r_2\}$ el conjunto de autorregresores y continuamos en el paso siguiente.

4. Consideramos $\Omega_R \cup \{r\}$ el conjunto que contiene los autorregresores ya escogidos y otro tal que $r \in \Omega - \Omega_R$. Ahora se presentan dos posibilidades

   a) $\forall r \in \Omega - \Omega_R \quad er(\Omega_R \cup \{r\}) \geq er(\Omega_R)$, entonces nos quedaremos con los autorregresor de Ω_R y terminamos con el algoritmo.

   b) $\exists r_0 \in \Omega - \Omega_R$ tal que $eer(\Omega_R \cup \{r_0\}) < er(\Omega_R)$, entonces añadiremos a Ω_R el autorregresor r_0 y volveremos al principio de este paso. Si se diera el caso de que $\Omega - \Omega_R =$ el algoritmo también terminaría considerando como autorregresores todos los elementos de Ω.

Ahora falta ver como definimos la función $er(\cdot)$ usada para calcular el error. En primer lugar consideramos el FPE, mencionado anteriormente, que se define como

$$FPE(\Omega^p_S) = \frac{1}{n'} \sum_{j=1}^{n'} \left(y_j - \hat{m}\left(\widetilde{x}_j\right)\right)^2 \; \frac{1 + (nh^p)^{-1} J^p B_p}{1 - (nh^p)^{-1} \left(2K^p(0) - J^p\right)} B_p \qquad (2.3)$$

teniendo en cuenta que

- K es el núcleo gaussiano.
- $K(0) = \frac{1}{\sqrt{2\pi}}$
- $\hat{m}\left(\widetilde{x}_j\right)$ es el estimador (2.1) a partir de la muestra $\{(y_j, \widetilde{x}_j)\}_{j=1}^{n'}$.

- $J = \int (K(x))^2 dx = 0{,}28209$
- $B_p = \frac{1}{n'} \sum_{j=1}^{n'} \frac{1}{\hat{f}(\widetilde{x}_j)}$; siendo $\hat{f}(\widetilde{x}_j)$ el estimador de Rosenblatt-Parzen de la densidad.

Otra posibilidad es limitarnos al error cuadrático medio, que sería

$$ECM(\Omega_S^p) = \frac{1}{n'} \sum_{j=1}^{n'} \left(y_j - \hat{m}\left(\widetilde{x}_j\right)\right)^2 \tag{2.4}$$

donde $\hat{m}\left(\widetilde{x}_j\right)$ es el estimador (2.1) a partir de la muestra $\{(y_j, \widetilde{x}_j)\}_{j=1}^{n'}$.

Despues de realizar unas pruebas decidimos quedarnos con el ECM en vez de con el FPE, ya que los autorregresores escogidos por uno y otro solían ser los mismos e implementar el ECM tenía mucho menos coste computacional.

Bandas de Confianza

Una vez estimado un S_t, vamos calcular un intervalo de confianza para el verdadero valor de S_t basándonos en un remuestreo bootstrap. El procedimiento que utilizaremos va a ser el siguiente.

Usando la muestra que ya teníamos,

$$\{(y_j, \widetilde{x}_j) : j = 1, 2, \ldots, n'\},$$

podemos definir los residuos de la forma

$$\hat{\varepsilon}_j = y_j - \hat{m}\left(\widetilde{x}_j\right), \quad j = 1, \ldots, n',$$

donde $\hat{m}(\cdot)$ es el estimador (2.2). Si denotamos por s_ε la desviación típica de los residuos, $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_{n'}$, podemos definir el parámetro de suavizado

$$g = \left(\frac{4}{3n'}\right)^{1/5} s_\varepsilon$$

que vamos a usar para calcular los residuos bootstrap suavizados. Fijando un número de réplicas bootstrap B del orden de 100-1000 y considerando que, para cualquier $i \in \{1, 2, \ldots, B\}$, I_i es una variable aleatoria uniforme discreta con soporte $\{1, \ldots, n'\}$ y Z_i una $N(0, 1)$ definimos

$$\hat{\varepsilon}_i^* = \hat{\varepsilon}_{I_i} + gZ_i, \quad i = 1, \ldots, B.$$

Una vez que tenemos los residuos bootstrap consideramos su conjunto ordenado

$$\{\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_B^*\},$$

fijamos un nivel de confianza $1-\alpha$ y elegimos los elementos en las posiciones $[(\alpha/2)\, B]$ y $[(1-(\alpha/2))\, B]$ del conjunto ordenado. Una vez hecho esto, el intervalo de confianza que buscábamos es

$$\left(\hat{m}\left(\tilde{\mathbf{u}}\right) + \hat{\varepsilon}_{[(\alpha/2)B]}^*, \hat{m}\left(\tilde{\mathbf{u}}\right) + \hat{\varepsilon}_{[(1-(\alpha/2))B]}^*\right). \tag{2.5}$$

Repitiendo esto para cada S_t que queramos estimar obtenemos una banda de confianza para la serie.

Capítulo 3

Desarrollo del Algoritmo y el Código

La idea es elaborar una función, dentro del entorno de R, que, dada una serie de tiempo, devuelva la estimación en los k instantes posteriores. Será necesario cargar el paquete `mgcv` para poder trabajar con los modelos aditivos, usaremos el comando

```
library(mgcv)
```

Para implementar la función mencionada, vamos a necesitar implementar otras funciones previas, más sencillas, que hagan los cálculos necesarios. La función `coloca` servirá para colocar los datos en forma de lista de vectores y así poder aplicar la regresión directamente: si tenemos observados $\{S_1, S_2, \ldots, S_n\}$ y queremos predecir S_{n+1} utilizando como autorregresores $\{S_{n-i_1}, S_{n-i_2}, \ldots, S_{n-i_p}\}$ los datos que tendremos serán

$$\{(y_j, \widetilde{x}_j)\}_{j=1}^{n-i_p} \subseteq \mathbb{R}^{p+1}$$

teniendo en cuenta que $y_j = S_j$ y $\widetilde{x}_j = (S_{j-i_1}, S_{j-i_2}, \ldots, S_{j-i_p})$. El código utilizado para implementar esta función fue el siguiente

```
coloca<-function(serie,retardos){
    retardos<-unique(sort(retardos))
```

```
    n<-length(serie)
    A<-matrix(1,n-max(retardos),length(retardos)+1)
    for (i in seq(n,max(retardos)+1,-1)){
        A[n-i+1,]<-c(serie[i],serie[i-retardos])
    }
colnames(A)<-c("Y",paste("X",1:(ncol(A)-1),sep=""))
SALIDA<-data.frame(A)
SALIDA
}
```

Ahora necesitamos una función que, dada una muestra de datos

$$\{(y_j, \widetilde{x}_j)\}_{j=1}^{n'} \subseteq \mathbb{R}^{p+1},$$

y un punto (o varios)

$$\vec{x}_0 := (x_1^0, \ldots, x_p^0) \in \mathbb{R}^p,$$

nos construya la función de regresión (2.2), usando un modelo aditivo, y la use para estimar lo que ocurre en el punto $\vec{x}_0$. De esa función de regresión nos interesan los parámetros de suavizado de cada una de las f_is del modelo, por lo que la función de R que implementamos devuelve una lista en la que aparecen el valor de $\hat{m}(\cdot)$ en cada punto en la primera columna y en la segunda los parámetros de suavizado. El código necesario para esto es el que mostramos a continuación

```
predice<-function(datos,punto){
    indices<-2:length(datos)
    n<-length(colnames(datos))
    k<-length(punto$X1)
    names<-colnames(datos)
    expresion<-as.formula(paste(paste(names[1],"~"),
        paste(paste("s(",names[indices],")"), collapse= "+")))

    datos.gam<-gam(expresion,data=datos)
```

```
    SALIDA<-matrix(0,max(k,n-1),2)
    SALIDA[1:k,1]<-predict(datos.gam,punto)
    SALIDA[1:n-1,2]<-datos.gam$sp
    SALIDA<-data.frame(SALIDA)
    colnames(SALIDA)<-c("pred","smooths")
    SALIDA
}
```

Por otro lado, es necesaria una función que nos calcule el error cometido por el método al estimar. Las dos posibilidades que habíamos mencionado eran el error cuadrático medio (ECM) y el error final de predicción (FPE). El ECM se implementa de manera sencilla, sólo necesita los datos observados y los estimados. Podemos ver a continuación su código

```
ecm<-function(datos,estimados){
    npr<-length(datos$Y)
    sum((datos$Y-estimados)^2)/npr
}
}
```

En cambio, el FPE necesita también la serie que estamos manejando, un parámetro de suavizado y el número de autorregresores que usamos. Podemos observar que su expresión más compleja conlleva también un código más extenso

```
fpe2<-function(datos,estimados,serie,h,p){
    n<-length(serie)
    npr<-length(datos$Y)
    f.RP<-0
    for (i in 1:npr){
        z<-as.vector(datos[i,2:length(datos)],mode="numeric")
        f.RP<-f.RP+sum(apply(dnorm(t((t(as.matrix(
            datos[,2:length(datos)]))-z))/h),1,prod))
```

```
    }
    f.RP<-f.RP/(n*h)
    Bp<-sum(1/f.RP)/npr
    J<-0.28209
    Omega<-(1+(n*h^p)^(-1)*Jp*Bp)/
        (1-(n*h^p)^(-1)*(2*dnorm(0)^p-J^p)*Bp)

    sum((datos$Y-estimados)^2)/npr*Omega
}
```

Ahora trataremos de implementar el Algoritmo 2.1 para buscar los autorregresores. Utilizaremos como medida del error el ECM para agilizar su ejecución. La función de R necesitará la serie y el conjunto de instantes en los que vamos a buscar los autorregresores. Este último lo determinaremos dando el número de instantes del final donde no vamos a buscar (`salto`) y la cantidad de instantes que vamos a usar (`num`). El código será el siguiente

```
selector<-function(num,salto,serie){
    a<-vector()
    pos<-vector()
    A<-(1:num)+salto
    for (i in A){
        dat<-coloca(serie,i)
        a[i-salto]<-ecm(dat,predice(dat,dat)$pred)
        #PR<-predice(dat,dat); p<-1; h<-PR$smooth[1];
            a[i-salto]<-fpe2(dat,PR$pred,serie,h,p)

        pos[1]<-which.min(a)+salto
    }
    minimo<-min(a)
    fuera<-0
    lugar<-2
    while (fuera==0){
```

```
        fuera<-1
        for (i in A){
            if (sum(pos==i)==0){
                dat<-coloca(serie,c(pos,i))
                a[i-salto]<-ecm(dat,predice(dat,dat)$pred)
                #PR<-predice(dat,dat); p<-length(unique(c(pos,i)));
                    h<-mean(PR$smooth[1:p]);
                    a[i-salto]<-fpe2(dat,PR$pred,serie,h,p)

            }
        }
        if (min(a)<minimo){
            pos[lugar]<-which.min(a)+salto
            lugar<-lugar+1
            fuera<-0
            minimo<-min(a)
        }

    }
sort(pos)
}
```

Notemos que en un momento de la función ponemos a modo de comentario lo que deberíamos cambiar para usar el FPE en vez del ECM como medida del error.

Con todas estas funciones implementadas ya podemos implementar la función que planteábamos desde un principio: dada una serie y un número de instantes en los que se quiere estimar la serie. También utilizaremos esta función para calcular las bandas de confianza dadas en (2.5), por lo que también tendremos que darle el nivel de confianza $1 - \alpha$.

Podíamos intentar hacer la estimación de forma recursiva, aunque los resultados no sean muy fiables, con el siguiente código

```
estima.rec<-function(serie,pasos){
    ret<-selector(8,0,serie)
    datos<-coloca(serie,ret)
    indices<-2:length(datos)
    n<-length(colnames(datos))
    names<-colnames(datos)
    expresion<-as.formula(paste(paste(names[1],"~"),
        paste(paste("s(",names[indices],")"), collapse= "+")))

    datos.gam<-gam(expresion,data=datos)
    SALIDA<-serie
    for (i in 1:pasos){
        a<-t(matrix(SALIDA[length(SALIDA)-ret+1]))
        colnames(a)<-paste("X",1:(ncol(a)),sep="")
        SALIDA[length(SALIDA)+1]<-predict(datos.gam,data.frame(a))
        rm(a)
    }
SALIDA
}
```

Tengamos en cuenta que buscamos los autorregresores entre los 8 instantes anteriores para agilizar las ejecuciones. Ya que no va a ser el método que usemos definitivamente, tampoco implementamos aquí las bandas de confianza.

A continuación mostraremos el código para hacer las estimaciones usando el método directo que mencionábamos en la Formulación del Problema. Además también calcularemos varias medidas del error

```
estima<-function(serie,pasos,alfa){
  SALIDA<-serie[1:(length(serie)-pasos)]
```

```
n<-length(SALIDA)
lim.inf<-vector()
lim.sup<-vector()
for (i in 1:pasos){
  ret<-selector(8,i-1,SALIDA)
  dat<-coloca(SALIDA,ret)
  eps<-dat$Y-predice(dat,dat)$pred
  g<-(4/(3*length(dat$Y)))^(1/5)*sd(eps)
  B<-1000
  eps.boot<-sort(sample(eps,rep=T,B)+g*rnorm(B))
  a<-t(matrix(SALIDA[length(SALIDA)-ret+1]))
  colnames(a)<-paste("X",1:(ncol(a)),sep="")
  b<-predice(dat,data.frame(a))[1,1]
  SALIDA[length(SALIDA)+1]<-b
  lim.inf[i]<-b+eps.boot[B*alfa/2]
  lim.sup[i]<-b+eps.boot[B*(1-alfa/2)]
}
med<-c("RMSE","MAE","MAPE","SMAPE",sample("-",pasos-4,rep=T))
rmse<-sqrt(sum((SALIDA[(n+1):length(SALIDA)]-
  serie[(n+1):length(SALIDA)])^2)/pasos)

mae<-sum(abs(SALIDA[(n+1):length(SALIDA)]
  -serie[(n+1):length(SALIDA)]))/pasos

mape<-sum(abs((SALIDA[(n+1):length(SALIDA)]-serie[(n+1):
  length(SALIDA)])/serie[(n+1):length(SALIDA)]))*100/pasos

smape<-sum(abs((SALIDA[(n+1):length(SALIDA)]-serie[(n+1):
  length(SALIDA)])/(SALIDA[(n+1):length(SALIDA)]+serie[(n+1):
  length(SALIDA)])))*200/pasos
```

```
  data.frame(predichos=SALIDA[(n+1):length(SALIDA)],lim.inf=
    lim.inf,lim.sup=lim.sup,Medida=med,Valor=
    c(rmse,mae,mape,smape,sample(0,pasos-4,rep=T)))
}
```

Hay que destacar que esta función no estima lo que ocurre en los p instantes siguientes a los que tiene la serie que introducimos, sino que considera los p últimos instanstes de la serie desconocidos y los estima. Así podremos ver como se ajusta la predicción a la realidad.

Capítulo 4

Estudio Comparativo

Una vez desarrollado el modelo, nos interesa probar como funciona en la práctica. Para ello utilizamos 23 series obtenidas en su mayoría de los libros Box and Jenkins (1976), Brockwell and Davis (1987), Abraham and Ledolter (1983), Pankratz (1983), Makridakis et al (1998) y Tong (1990). Usaremos series no estacionales, que son para las que está implementado el método. Al final del capítulo mostramos un apéndice en el que comentamos cada una de las series utilizadas. Los pronósticos los hicimos usando el estimador no paramétrico de Nadaraya-Watson y el modelo aditivo.

Para comparar los resultados obtenidos con cada método utilizaremos varias medidas del error que denotaremos por sus iniciales originales del inglés. Tengamos en cuenta que denotamos por r el máximo horizonte de predicción (que será 8 por lo general).

- Raiz del Error Cuadrático Medio ($RMSE$)

  $$RMSE = \sqrt{MSE}.$$

  donde MSE es el error cuadrático medio

  $$MSE = \frac{1}{r}\sum_{l=1}^{r}\left(\hat{z}_n(l) - z_{n+l}\right)^2$$

  el cual ya usamos anteriormente para la selección de autorregresores.

- Error Medio Absoluto (MAE)

$$MAE = \frac{1}{r}\sum_{l=1}^{r} \left|\hat{z}_n(l) - z_{n+l}\right|.$$

- Error Medio Absoluto en Porcentaje ($MAPE$)

$$MAPE = \frac{1}{r}\sum_{l=1}^{r} \left|\frac{\hat{z}_n(l) - z_{n+l}}{z_{n+l}}\right| 100.$$

- Error Medio Absoluto Simétrico en Porcentaje ($SMAPE$)

$$SMAPE = \frac{1}{r}\sum_{i=1}^{r} \frac{\left|\hat{z}_n(l) - z_{n+l}\right|}{\left(\hat{z}_n(l) + z_{n+l}\right)/2} 100,$$

  que es la medida de error recomendada por Makridakis & Hibon (2000).

En las siguiente tabla vamos a mostrar, por este orden, las relaciones $\frac{RMSE_{NWD}}{RMSE_{BJ}}$, $\frac{RMSE_{NWR}}{RMSE_{BJ}}$, $\frac{RMSE_{MA}}{RMSE_{BJ}}$, $\frac{MAE_{NWD}}{MAE_{BJ}}$, $\frac{MAE_{NWR}}{MAE_{BJ}}$ y $\frac{MAE_{MA}}{MAE_{BJ}}$, donde

- $RMSE_{NWD}$, $RMSE_{NWR}$, $RMSE_{MA}$ y $RMSE_{BJ}$ son la raiz del error cuadrático medio calculado usando un modelo no paramétrico de Nadaraya-Watson directo, el recursivo, un modelo aditivo y un modelo Box-Jenkis respectivamente.

- MAE_{NWD}, MAE_{NWR}, MAE_{MA} y MAE_{BJ} son el error absoluto medio usando un modelo no paramétrico de Nadaraya-Watson directo, el recursivo, un modelo aditivo y un modelo Box-Jenkis respectivamente.

Hay que tener en cuenta que donde la relación sea mayor que 1 tendremos que el modelo Box-Jenkis es mejor que los otros planteados, si por el contrario es menor que 1 éstos mejorarán al Box-Jenkis, en este caso el coeficiente menor indicará que modelo es mejor en cada caso.

	RMSE			**MAE**		
Serie	NWD	NWR	MA	NWD	NWR	MA
1	1,1076	0,9945	1,1158	1,1243	0,9681	1,0882
2	1,4078	1,7342	1,1043	1,3871	1,8613	1,2672
3	0,9327	0,9979	0,7473	1,0186	1,0444	0,7691
4	1,1024	1,2997	1,6571	1,1068	1,2804	1,3840
5	0,9970	0,9665	1,0782	0,9985	0,9733	1,1027
6	0,9145	1,0628	0,9970	0,8746	1,0427	0,0606
7	0,9938	1,0007	1,1478	1,0668	0,9972	0,6937
8	1,3678	0,4323	2,2863	1,2961	0,3669	2,1661
9	1,0484	0,9021	1,3215	0,8921	0,8291	1,1392
10	0,6321	1,5126	0,5624	0,6419	1,5455	0,6434
11	0,9366	0,8333	0,9253	0,9265	0,7860	1,0081
12	0,9098	0,9259	1,3587	0,9414	0,9473	1,3707
13	1,2212	1,6452	1,7880	1,1568	1,6022	1,8179
14	1,1025	1,0362	2,9255	1,0165	0,9769	2,0030
15	0,8854	0,8524	0,9963	0,8486	0,8075	0,9067
16	0,8900	0,6174	1,0429	0,8638	0,6024	1,0306
17	1,0394	0,7835	0,8809	1,1557	0,9364	1,0667
18	1,0024	1,1260	1,0656	0,9694	1,0823	1,1319
19	0,9576	0,8882	0,9076	0,9430	0,9220	0,8930
20	0,9305	0,8712	1,0717	0,9119	0,8363	1,0882
21	0,9056	1,4640	1,2687	0,9617	1,5396	1,1940
22	0,3294	0,4020	0,9676	0,2979	0,3464	1,0287
23	0,9729	0,9849	0,9845	0,9142	0,9345	1,0079

Podemos apreciar que el modelo lineal sólo es capaz de mejorar la estimación en algunas de las series paramétricas (1-14), en el resto de paramétricas y en las no paramétricas (15-23) el modelo que ya teníamos de Nadaraya-Watson es más efectivo. En la mayoría de las series el modelo aditivo mejora al modelo Box-Jenkis si el modelo de Nadaraya-Watson lo mejora también.

Para distinguir si las series son lineales o no usamos un test sencillo propuesto en McLeod and Li (1983).

Estudio completo de dos series

Serie 17

Esta serie es conocida en el mundo de la estadística no paramétrica. Sus datos son los records anuales, desde 1821 hasta 1934, de linces canadienses cazados en el río Mackenzie (Distrito Noroeste de Canadá). Es una serie con 114 observaciones y que fue muy usada como ejemplo en otros estudios. En la Figura 4.1 podemos observar el aspecto que tiene dicha serie.

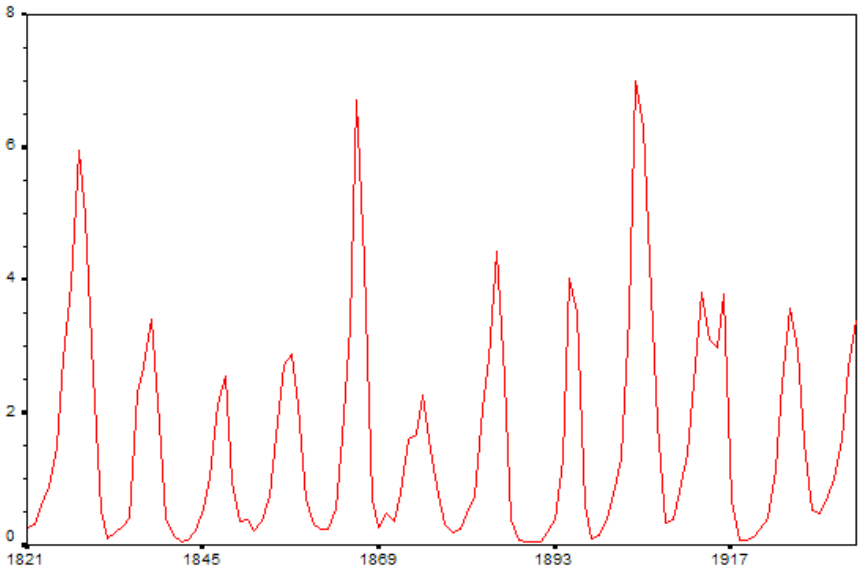

Figura 4.1: Datos sobre los linces (Serie 17).

Es conveniente hacer un estudio descriptivo de dicha serie, para ello mostraremos en la Figura 4.2 la tabla con sus principales estadísticos descriptivos.

Estadístico	Valor
Media	1538,018
Mediana	771
Varianza	2514900,9
Desviación Típica	1585,844
Asimetría	1,368
Curtosis	1,583
Mínimo	39
Máximo	6991

Figura 4.2: Tabla de Estadísticos Descriptivos.

Retardo	Autocorrelación	Error típico	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	,711	,092	59,129	1	,000
2	,214	,092	64,557	2	,000
3	-,189	,092	68,791	3	,000
4	-,433	,091	91,383	4	,000
5	-,502	,091	121,983	5	,000
6	-,400	,090	141,608	6	,000
7	-,148	,090	144,315	7	,000
8	,218	,090	150,264	8	,000
9	,501	,089	181,864	9	,000
10	,514	,089	215,445	10	,000
11	,283	,088	225,760	11	,000
12	-,029	,088	225,868	12	,000
13	-,303	,087	237,893	13	,000
14	-,450	,087	264,658	14	,000
15	-,461	,087	293,018	15	,000
16	-,346	,086	309,178	16	,000

b. Basado en la aproximación chi cuadrado asintótica.

Figura 4.3: Test de Box-Ljung.

Observando la Figura 4.1 parece que hay dependencia en la serie, para contrastarlo estadísticamente haremos un test de Box-Ljung (Figura 4.3), un gráfico de las autocorrelaciones (Figura 4.4) y las autocorrelaciones parciales (Figura 4.5).

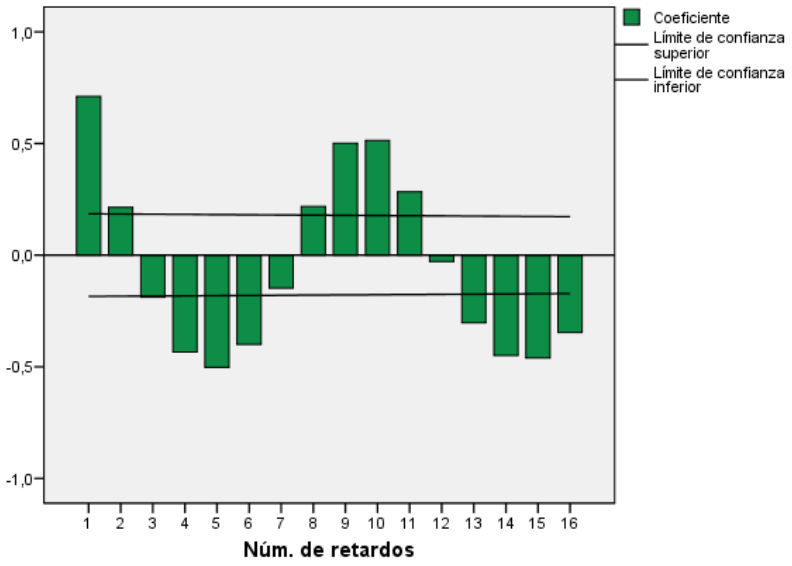

Figura 4.4: Autocorrelaciones de la serie.

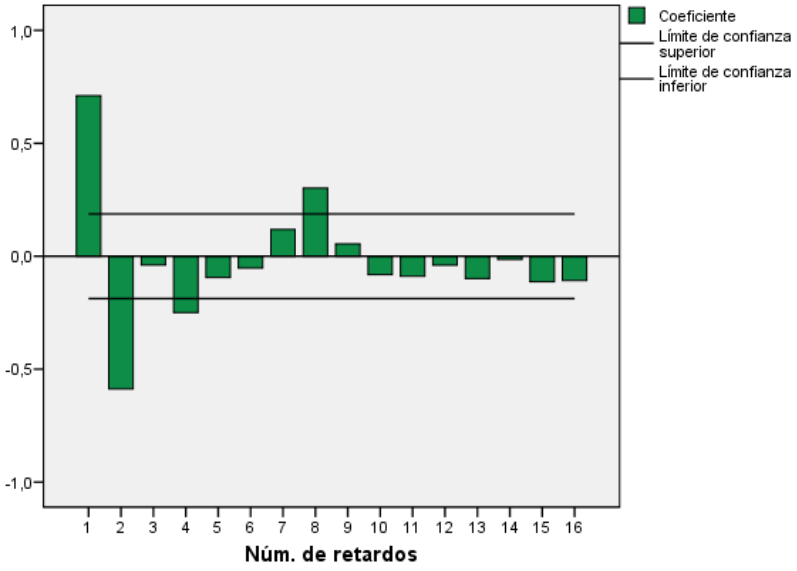

Figura 4.5: Autocorrelaciones parciales de la serie.

Como ya habíamos dicho, usaremos las 106 primeras observaciones para ajustar el modelo y dejaremos las 8 restantes para comprobar el ajuste del modelo.

Utilizando el modelo Box-Jenkis obtuvimos que el mejor ajuste es un ARMA(2,2). Podemos observar, en la Figura 4.6, como se ajusta el ARMA(2,2): la línea discontinua representa el modelo y la continua (roja) los datos reales. Notemos que el ajuste se hace también en los últimos instantes, los cuales no se usaron para buscar el modelo.

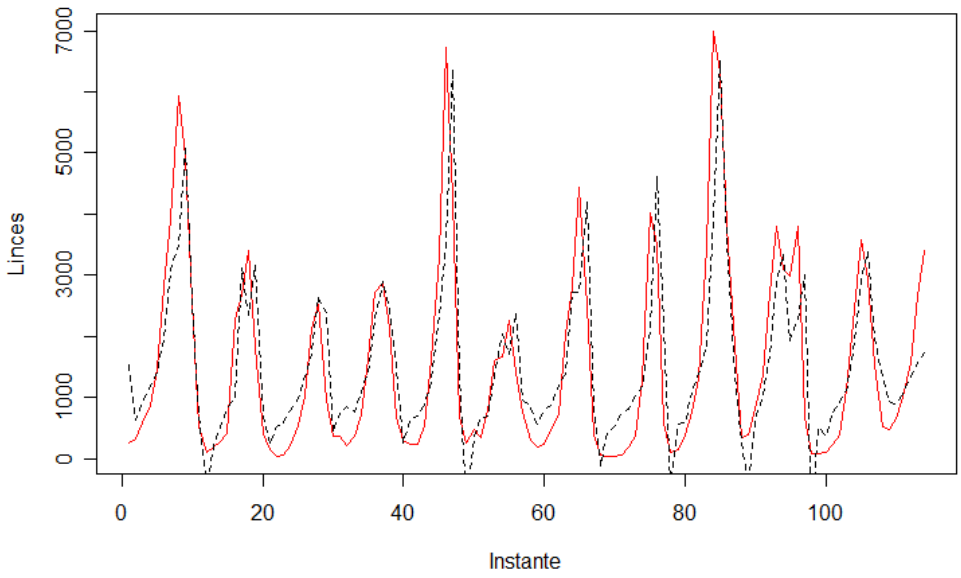

Figura 4.6: Ajuste del modelo mediante un ARMA(2,2)

Para comprobar la bondad de ajuste del modelo ARMA(2,2) vamos a estudiar sus residuos, la diferencia en cada instante entre el valor real y el valor estimado de la serie. En la Figura 4.7 tenemos los residuos del modelo ARMA(2,2) en los 106 primeros instantes, los que utilizamos para estimar el modelo.

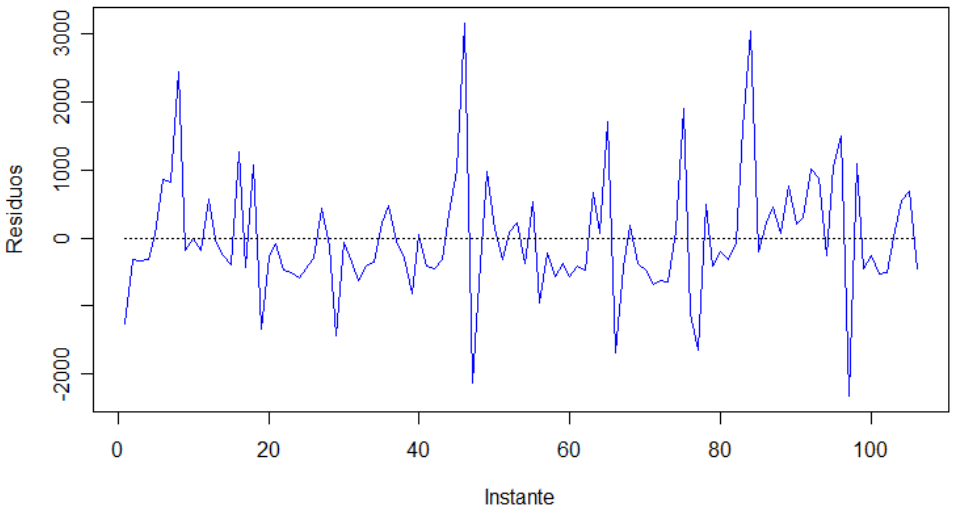

Figura 4.7: Residuos utilizando el ARMA(2,2).

Estadístico	**Valor**
Media	−7,537
Mediana	−210,607
Varianza	791517,2
Desviación Típica	889,673
Asimetría	0,864
Curtosis	2,556
Mínimo	−2330,99
Máximo	3167,152

Figura 4.8: Tabla de Estadísticos Descriptivos de los Residuos.

En la Figura 4.8 veremos la tabla con los principales estadísticos descriptivos para completar la idea intuitiva de como serán esos residuos.

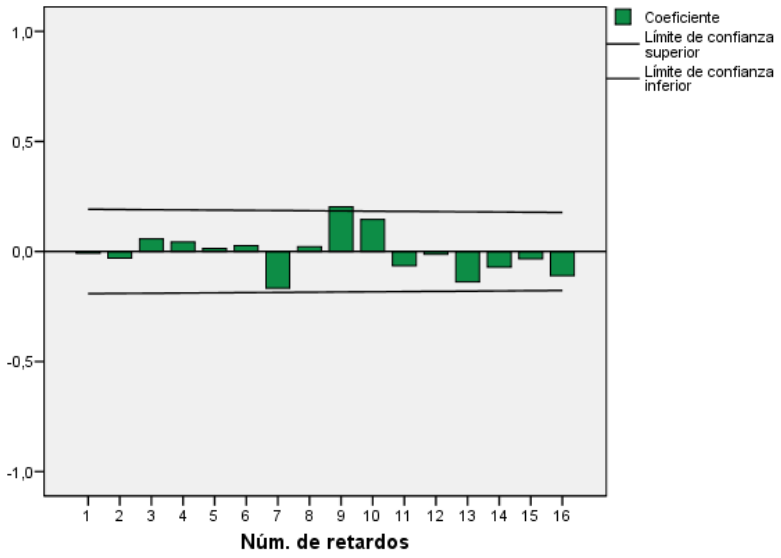

Figura 4.9: Autocorrelaciones de los residuos.

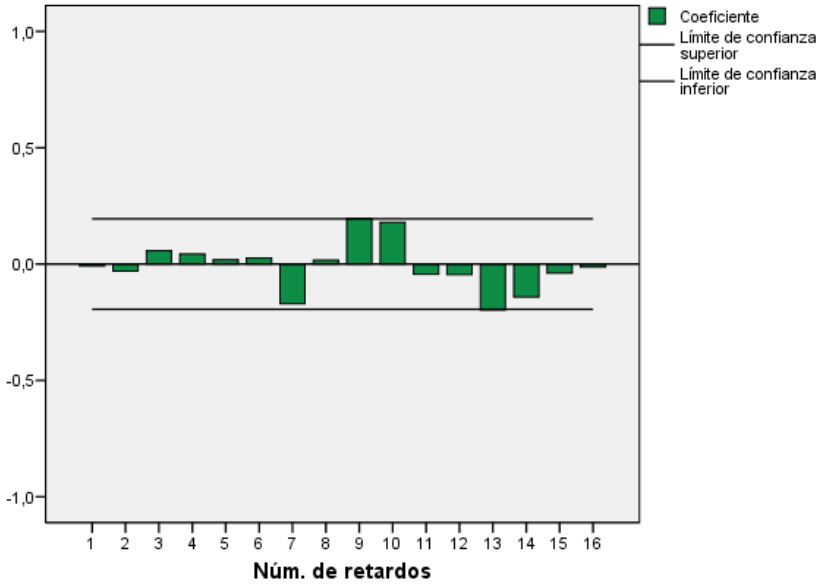

Figura 4.10: Autocorrelaciones parciales de los residuos.

En las Figuras 4.9 y 4.10 podemos ver el gráfico de autocorrelaciones y autocorrelaciones parciales de los residuos, respectivamente. Mirando eso y la tabla de la Figura 4.11, en la que se muestran los p-valores del contraste de independencia de Box-Ljung, podemos concluir que los residuos son independientes, propiedad deseable en un modelo ARMA.

Retardo	Autocorrelación	Error típico[a]	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	-,008	,096	,008	1	,930
2	-,031	,095	,110	2	,946
3	,057	,095	,478	3	,924
4	,043	,094	,686	4	,953
5	,014	,094	,708	5	,983
6	,026	,093	,786	6	,992
7	-,166	,093	3,985	7	,781
8	,022	,093	4,040	8	,854
9	,203	,092	8,906	9	,446
10	,145	,092	11,416	10	,326
11	-,065	,091	11,921	11	,370
12	-,013	,091	11,943	12	,450
13	-,139	,090	14,309	13	,352
14	-,072	,090	14,948	14	,382
15	-,033	,089	15,089	15	,445
16	-,110	,089	16,631	16	,410

a. El proceso subyacente asumido es la independencia (ruido blanco).

b. Basado en la aproximación chi cuadrado asintótica.

Figura 4.11: Test de Box-Ljung para los residuos.

Los residuos deberían seguir una distribución normal, pero vamos a ver que eso no se cumple. Observando el histograma (Figura 4.12), apreciamos un ligero ajuste a la curva normal, aunque en algunos momentos se aleja bastante. Utilizando las pruebas de normalidad Kolmogorov-Smirnov y Shapiro-Wilk obtenemos p-valores prácticamente nulos, con lo que se puede rechazar, con toda seguridad, la hipótesis de normalidad. Usando un gráfico Q-Q (Figura 4.13) es evidente también que no existe normalidad.

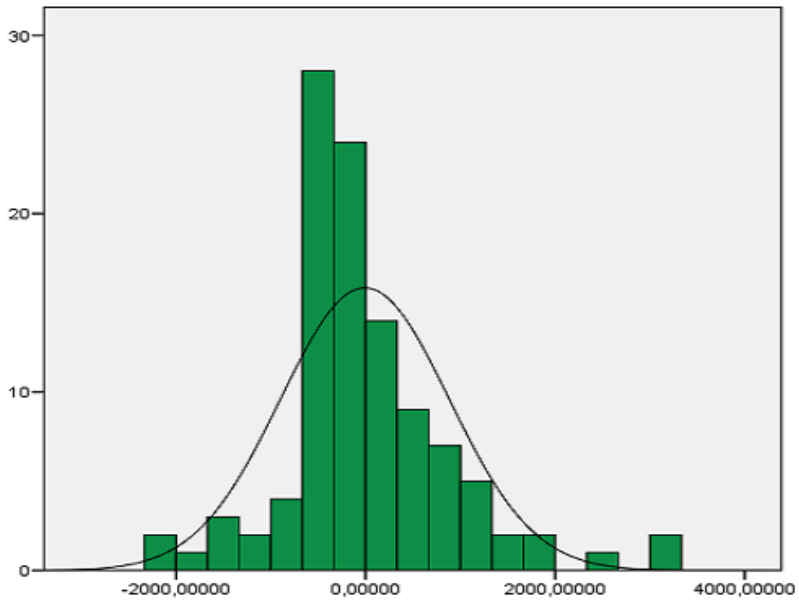

Figura 4.12: Histograma de los residuos.

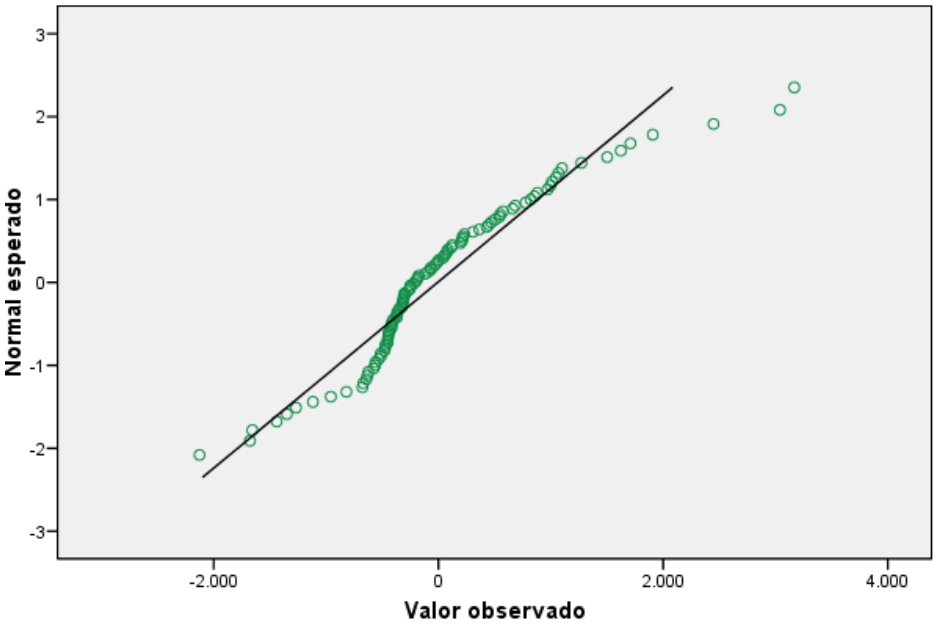

Figura 4.13: Gráfico Q-Q de normalidad.

Vamos, ahora, a estudiar lo que ocurre usando el modelo aditivo. Nos limitaremos, como ya habíamos hecho en otras ocasiones, a las 106 primeras observaciones para ajustar el modelo y utilizar las 8 restantes para comprobar el parecido de la realidad con lo estimado.

En la Figura 4.14 podemos observar la estimación por el modelo aditivo (línea roja) y los datos reales (línea azul) en los últimos 8 instantes. En esa misma figura mostramos la banda de confianza, usando el procedimiento del Algoritmo 2.1 al nivel de confianza $\alpha = 0{,}1$, delimitada por las líneas discontinuas. Podemos apreciar que los datos reales de la serie quedan dentro de los márgenes de confianza. Con una simple observación se puede apreciar un buen ajuste.

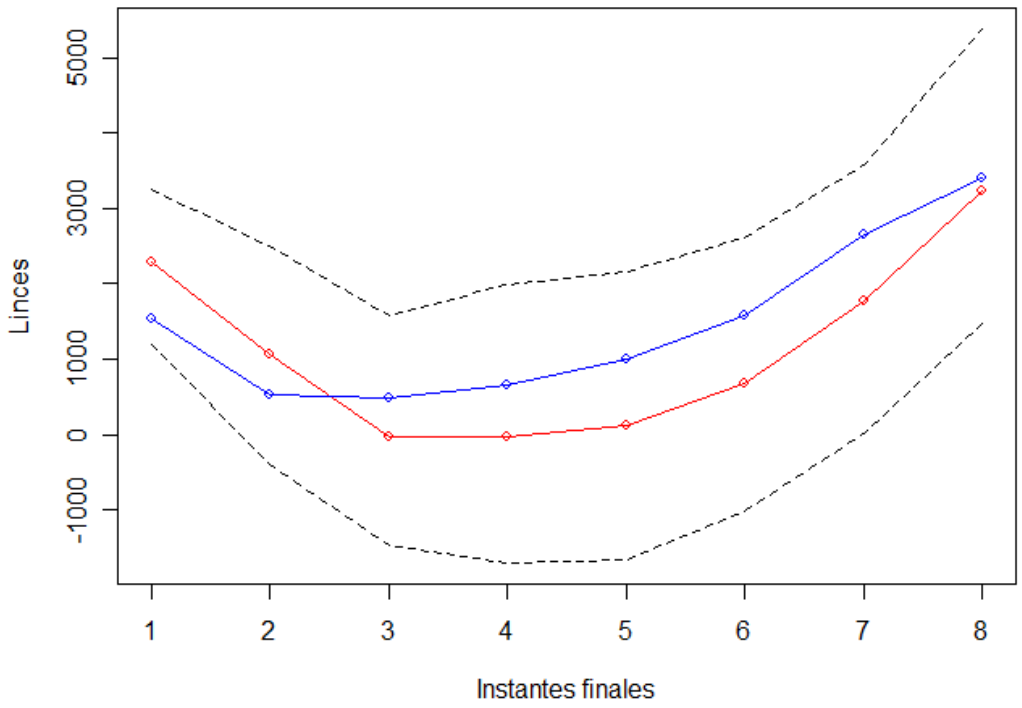

Figura 4.14: Estimación de los 8 últimos instantes de la serie usando el modelo aditivo.

En la Figura 4.15, para comprobar la bondad del ajuste del modelo de forma analítica en relación con otros modelos, vamos a comparar, mostrándolo en una tabla, distintas medidas del error de estimación según el método usado.

	RMSE	**MAE**	**MAPE**	**SMAPE**
Box Jenkins	796,81	621,10	52,07	42,85
Nadaraya-Watson Directo	828,19	717,81	54,67	58,09
Nadaraya-Watson Recursivo	624,26	581,59	53,47	47,43
Modelo Aditivo	701,92	662,50	67,52	101,99

Figura 4.15: Tabla de Medidas del error comparando varios modelos.

Según esto, el método Nadaraya-Watson Recursivo es el mejor para esta serie ya que sus 4 medidas del error son inferiores a las del resto. El modelo aditivo ganaría al Box Jenkins si nos fijamos en el RMSE, pero no si nos fijamos en el resto de medidas del error.

Serie 3

Esta serie está formada por el número de usuarios que se conectan (o desconectan) al minuto en cierto servidor de internet. Es decir, si en un minuto se conectan 35 usuarios y se desconectan 30 tendremos 5 conexiones en ese minuto. Tenemos los datos de una hora y 39 minutos, por lo que la serie tiene 99 observaciones. Podemos ver el aspecto que tiene la serie, en esos 99 instantes, si nos fijamos en la Figura 4.16.

Comenzaremos haciendo un estudio descriptivo de la serie, en la Figura 4.17 tenemos una tabla con los principales estadísticos descriptivos. De ahí podemos concluir que los datos se distribuyen en torno al 1, no muy dispersos y de forma bastante simétrica.

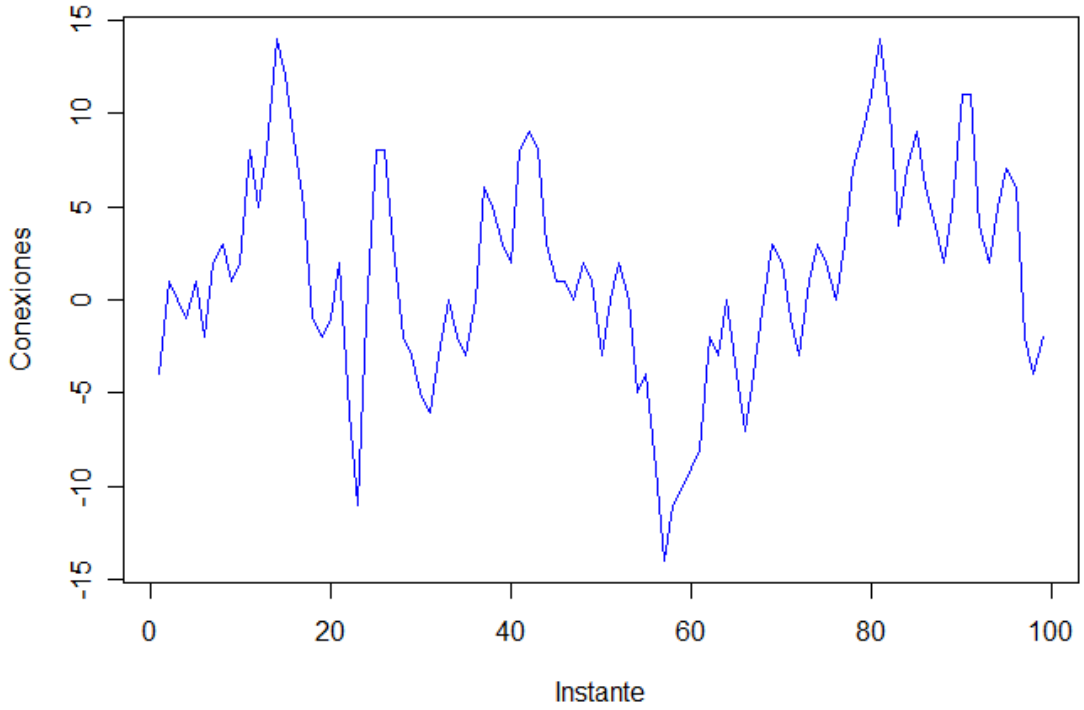

Figura 4.16: Datos sobre los usuarios que se conectan/desconectan (Serie 3).

Estadístico	**Valor**
Media	1,33
Mediana	1
Varianza	32,184
Desviación Típica	5,673
Asimetría	−0,117
Curtosis	0,001
Mínimo	−14
Máximo	14
Rango	28

Figura 4.17: Tabla de Estadísticos Descriptivos.

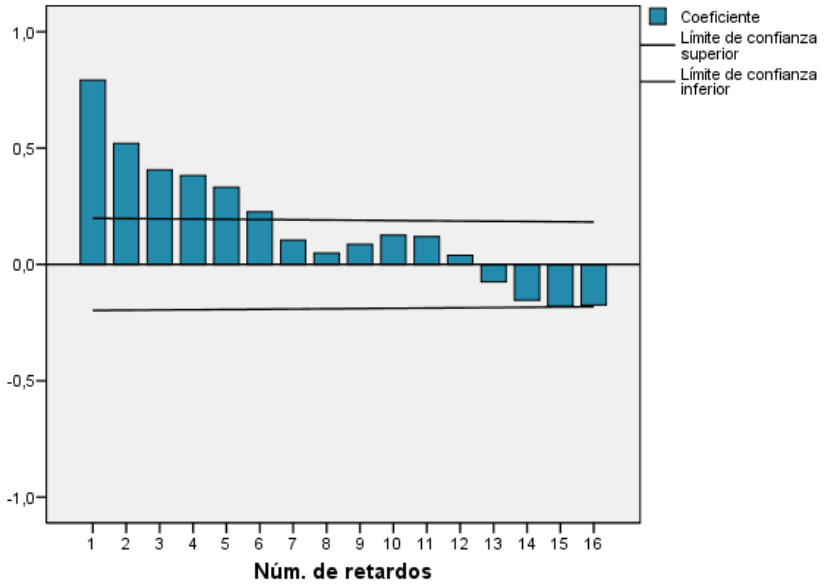

Figura 4.18: Autocorrelaciones de la serie.

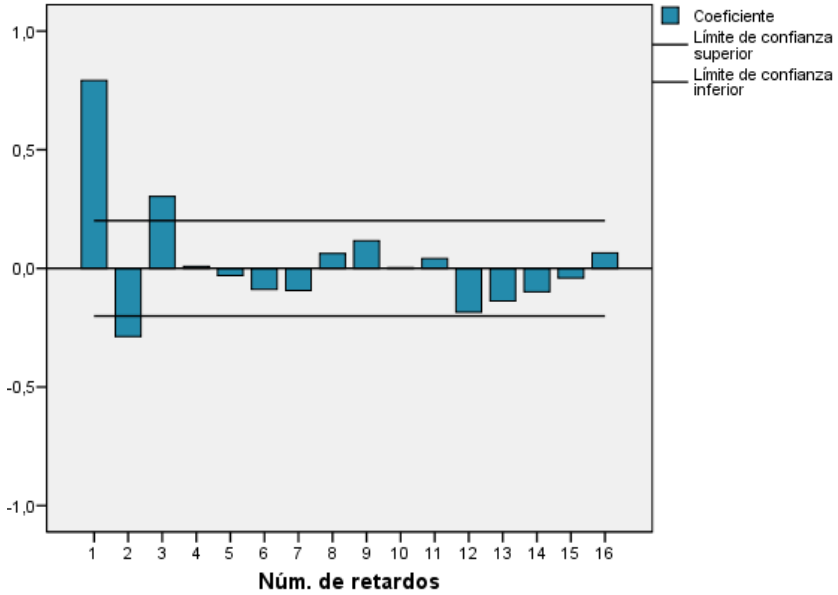

Figura 4.19: Autocorrelaciones parciales de la serie.

Retardo	Autocorrelación	Error típico[a]	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	,792	,099	63,962	1	,000
2	,520	,098	91,814	2	,000
3	,406	,098	108,995	3	,000
4	,382	,097	124,356	4	,000
5	,332	,097	136,050	5	,000
6	,226	,096	141,545	6	,000
7	,104	,096	142,724	7	,000
8	,049	,095	142,987	8	,000
9	,086	,095	143,815	9	,000
10	,126	,094	145,585	10	,000
11	,120	,094	147,212	11	,000
12	,039	,093	147,390	12	,000
13	-,075	,093	148,045	13	,000
14	-,154	,092	150,825	14	,000
15	-,179	,092	154,654	15	,000
16	-,175	,091	158,324	16	,000

a. El proceso subyacente asumido es la independencia (ruido blanco).

b. Basado en la aproximación chi cuadrado asintótica.

Figura 4.20: Test de Box-Ljung.

Si nos fijamos en las autocorrelaciones y autocorrelaciones parciales (Figuras 4.18 y 4.19) podemos apreciar que la serie tiene dependencia. Para contrastarlo estadísticamente usaremos el test de Box-Ljung (Figura 4.20), viendo que los p-valores son todos próximos a cero podemos rechazar con toda seguridad la hipótesis de independencia.

Para modelar la serie, vamos a considerar los 91 primeros instantes como datos y dejaremos los 8 finales para comprobar como funciona el modelo.

El mejor ajuste que se obtiene mediante el modelo de Box-Jenkis es el que nos da un ARMA(3,0). Este ajuste lo podemos ver en la Figura 4.21, donde la línea continua (azul) son los datos reales de la serie y la línea discontinua (negra) los datos generados por el modelo ARMA(3,0).

Ahora trabajaremos con los residuos para comprobar si es bueno el ajuste por un ARMA(3,0). En primer lugar, observaremos el gráfico de los residuos, en la Figura 4.22.

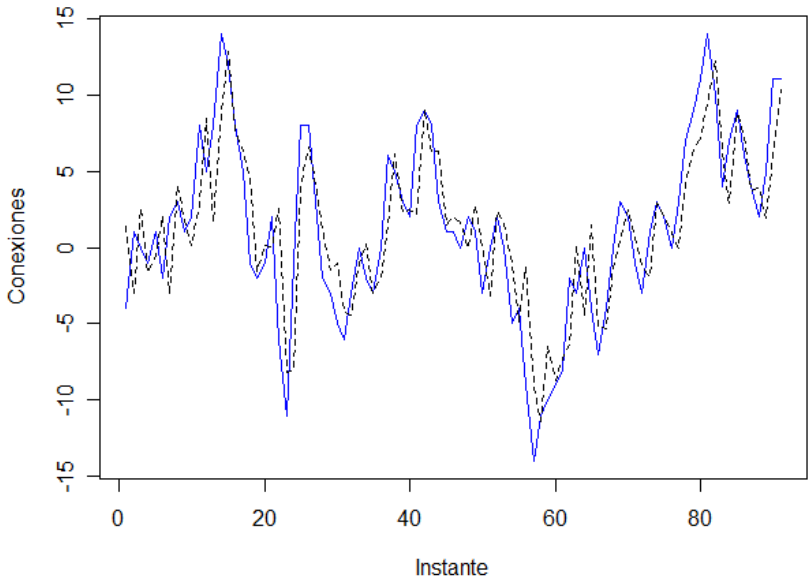

Figura 4.21: Ajuste del modelo mediante un ARMA(3,0)

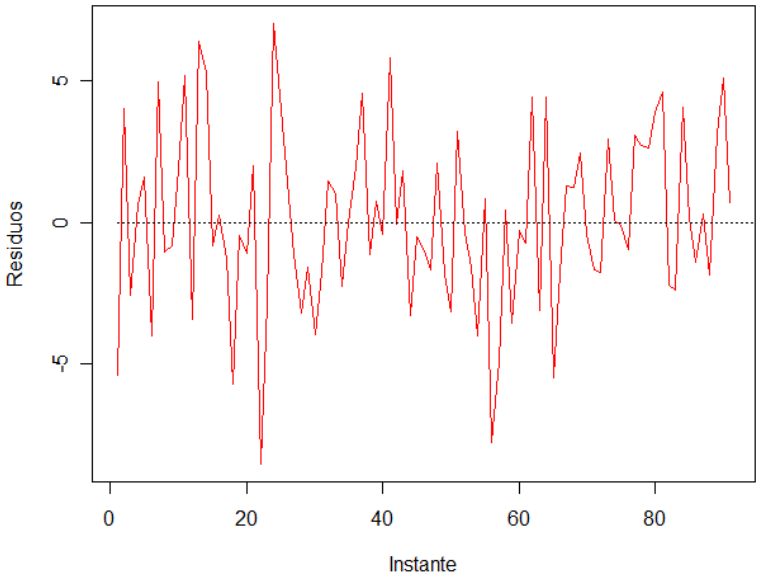

Figura 4.22: Residuos utilizando el ARMA(3,0).

Estadístico	Valor
Media	0,0518
Mediana	−0,2915
Varianza	9,927
Desviación Típica	3,151
Asimetría	−0,071
Curtosis	−0,069
Mínimo	−8,5113
Máximo	7,0072

Figura 4.23: Tabla de Estadísticos Descriptivos de los Residuos.

La tabla con los principales estadísticos descriptivos la podemos ver en la Figura 4.23. Observamos que estan centrados casi en cero, estan distribuidos bastante simétricos y tienen poca variabilidad.

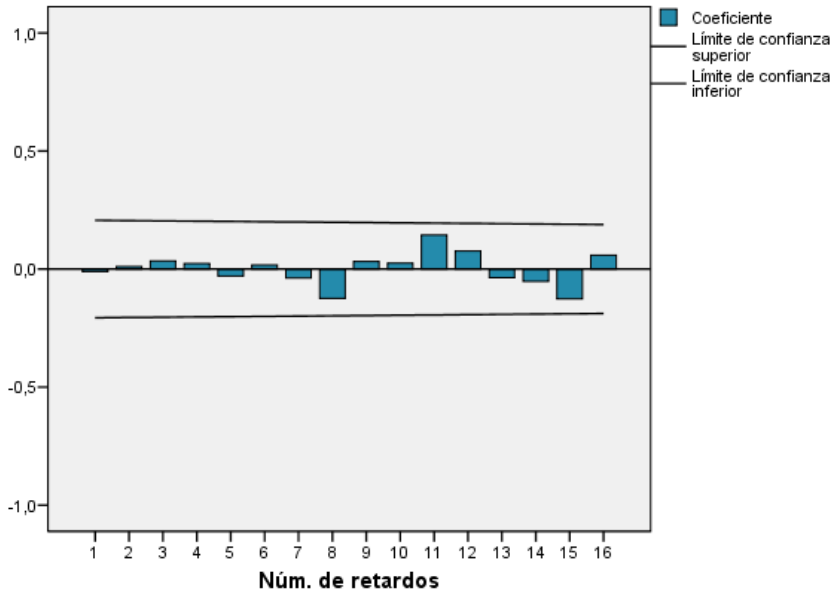

Figura 4.24: Autocorrelaciones de los residuos.

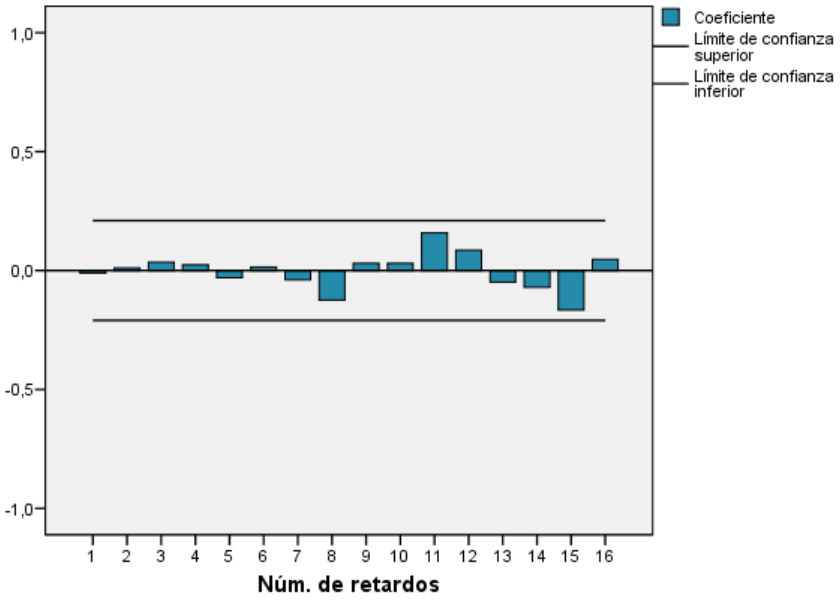

Figura 4.25: Autocorrelaciones parciales de los residuos.

Retardo	Autocorrelación	Error típico	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	-,010	,103	,010	1	,921
2	,011	,103	,021	2	,990
3	,035	,102	,136	3	,987
4	,023	,101	,189	4	,996
5	-,030	,101	,276	5	,998
6	,017	,100	,305	6	,999
7	-,038	,100	,450	7	1,000
8	-,124	,099	2,022	8	,980
9	,032	,098	2,125	9	,989
10	,025	,098	2,190	10	,995
11	,144	,097	4,381	11	,957
12	,076	,097	4,999	12	,958
13	-,036	,096	5,141	13	,972
14	-,052	,095	5,439	14	,979
15	-,127	,095	7,230	15	,951
16	,059	,094	7,621	16	,959

b. Basado en la aproximación chi cuadrado asintótica.

Figura 4.26: Test de Box-Ljung para los residuos.

Los residuos deberían ser independientes, sólo con fijarnos en las autocorrelaciones (Figura 4.24) y las autocorrelaciones parciales (Figura 4.25) nos hacemos una idea de que esto es así. Para confirmarlo utilizamos el test de Box-Ljung, cuyos resultados aparecen en la Figura 4.26, y observamos que todos los p-valores son altos, lo que nos da evidencias a la hipótesis de independencia.

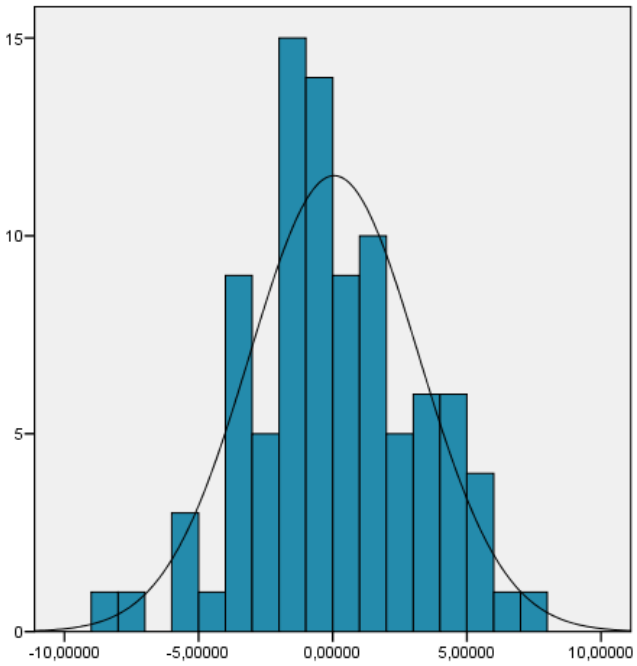

Figura 4.27: Histograma de los residuos.

En esta ocasión parece que la hipótesis de normalidad va a ser cierta, sólo con mirar el histograma (Figura 4.27) ya apreciamos cierto ajuste con la curva normal. Además los contraste de normalidad de Kolmogorov-Smirnov y de Shapiro-Wilk nos dan p-valores altos que le dan credibilidad a la hipótesis de normalidad. Para más evidencias de este hecho tenemos, en la Figura 4.28, el gráfico Q-Q de normalidad en el que se aprecia claramente.

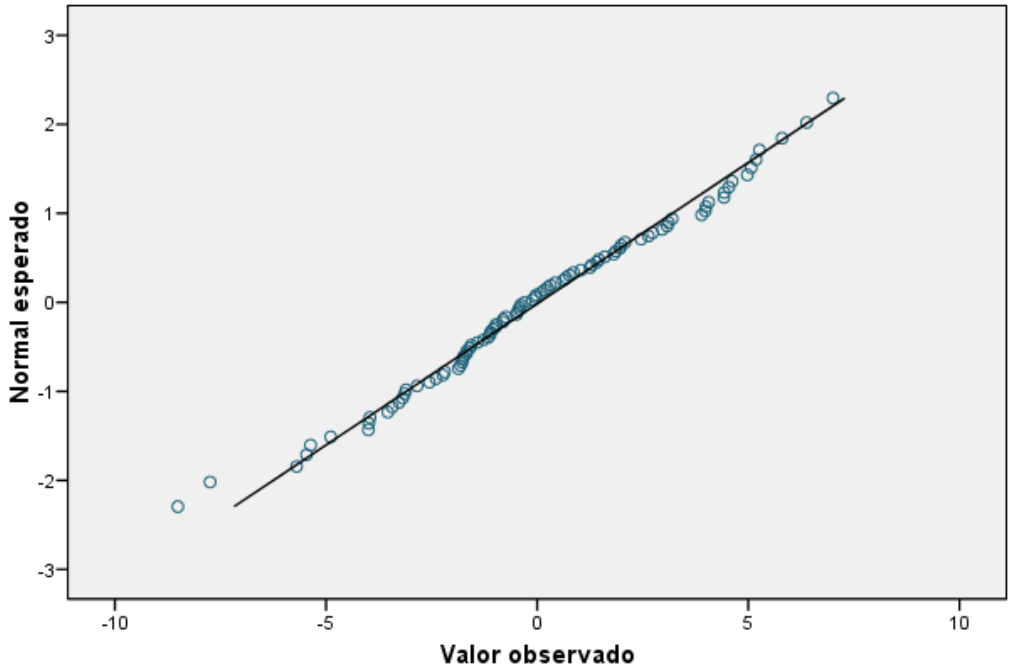

Figura 4.28: Gráfico Q-Q de normalidad.

Una vez hecho todo esto, nos interesa ver como funcionará el modelo aditivo para estimar esta serie. Como ya dijimos, los datos de los 91 primeros instantes los usaremos para determinar el modelo y el resto para comprobar su funcionamiento.

En la Figura 4.29 podemos observar la estimación por el modelo aditivo (línea roja) y los datos reales (línea azul) en los últimos 8 instantes. En esa misma figura mostramos la banda de confianza, usando el procedimiento del Algoritmo 2.1 al nivel de confianza $\alpha = 0{,}1$, delimitada por las líneas discontinuas. Podemos apreciar que los datos reales de la serie quedan dentro de los márgenes de confianza. Con una simple observación se puede apreciar un buen ajuste.

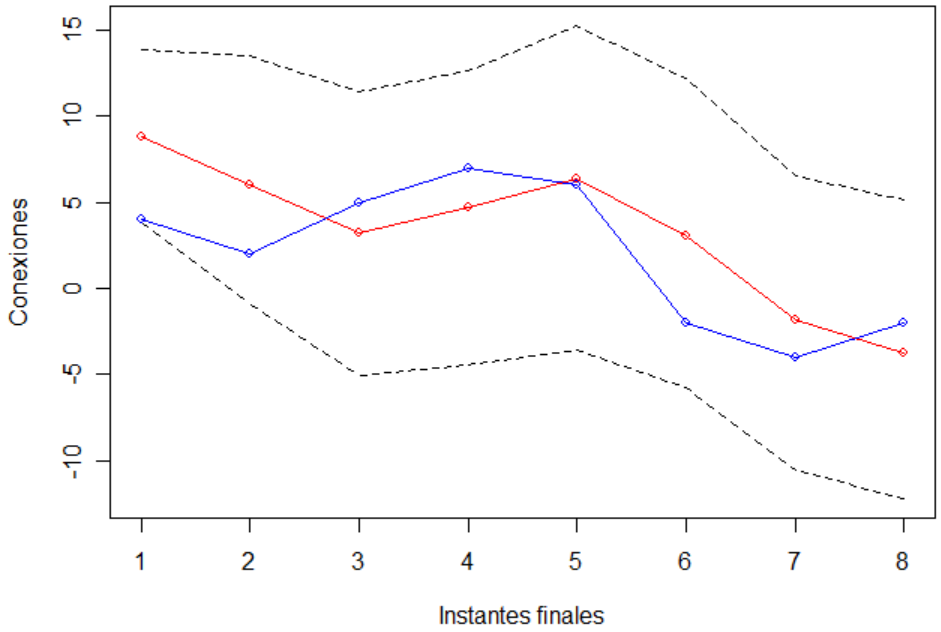

Figura 4.29: Estimación de los 8 últimos instantes de la serie usando el modelo aditivo.

Vamos a comparar ahora, analíticamente, el funcionamiento del modelo aditivo con otros métodos. Para ello veamos la tabla donde se muestran medidas del error de varios métodos (Figura 4.30)

	RMSE	**MAE**
Box Jenkins	4,2615	3,611
Nadaraya-Watson Directo	3,974	3,678
Nadaraya-Watson Recursivo	4,253	3,771
Modelo Aditivo	3,1847	2,7769

Figura 4.30: Tabla de Medidas del error comparando varios modelos.

En este caso apreciamos que el modelo aditivo es más eficiente que el resto, ya que consigue una medidas de error inferiores a las del resto de modelos.

Conclusión

El modelo aditivo sirve para solucionar el problema de dimensionalidad que presentaba el método de Nadaraya-Watson, propuesto en Vilar-Fernández and Cao (2007), pero los resultados obtenidos por este método son, por lo general, peores.

Apéndice de las series usadas

1. Concentraciones de cierta sustancia en un proceso químico (medición cada 2 horas). Box & Jenkins. $n = 197$.
2. Viscosidad de cierta sustancia en un proceso químico (medición cada hora). Box & Jenkins. $n = 309$.
3. Número de usuarios conectados, cada minuto, a un servidor de Internet. Makridakis et al. $n = 99$.
4. Porcentaje económico. Pankratz. $n = 100$.
5. Producción de carbón. Pankratz. $n = 96$.
6. Hipotecas y préstamos de gobierno. Abraham & Ledolter. $n = 159$.
7. Demanda mensual de reparaciones de equipos en Iowa entre 1972 y 1979. Abraham & Ledolter. $n = 94$.
8. Serie simulada E-923. Brockwell & Davis. $n = 200$.
9. Nivel del agua del Lago Hurón (en pies, reducido 570) entre 1875-1972. Brockwell & Davis. $n = 98$.
10. Serie gaussiana simulada AR(2), E921. Brockwell & Davis. $n = 192$.
11. Serie gaussiana simulada MA(1), E1042. Brockwell & Davis. $n = 152$.
12. Serie de Cauchy simulada MA(1), E1251. Brockwell & Davis. $n = 192$.

13. Serie de Cauchy simulada AR(1), E1252. Brockwell & Davis. $n = 192$.

14. Unidades privadas de alojamiento, APPC. Brockwell & Davis. $n = 136$.

15. Sunspot. Box & Jenkins. $n = 100$.

16. Datos del vuelo de moscas. Makridakis et al. $n = 261$.

17. Datos sobre el número de linces en cierta región de Canadá. Tong. $n = 114$.

18. Crecimiento trimestral de los porcentajes de ingresos no agrícolas en Iowa. Abraham & Ledolter. $n = 126$

19. Precios al cierre de las acciones de IBM. Tong. $n = 218$.

20. Proceso químico. Box & Jenkins. $n = 217$.

21. Serie no lineal simulada. $n = 100$.

22. Serie TAR simulada. $n = 100$.

23. Serie ARCH simulada. $n = 100$.

Bibliografía

- Abraham, B. and Ledolter, J. (1983). *Statistical methods for forecasting*, Wiley.
- Bosq, D. (1998). *Nonparametric statistics for stochastic processes,* Springer.
- Box, G.E.P and Jenkins, G.M. (1976). *Time series analysis: forecasting and control,* Holden-Day.
- Brockwell, P.J. and Davis, R.A. (1987). *Time series. Theory and methods*, Springer.
- Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1989). *Nonparametric curve estimation from time series*, Springer.
- Härdle, W., Lütkepohl, H. and Chen, R. (1997). A review of nonparametric time series analysis, *International Statistical Review,* 65:49-72.
- Härdle, W., Tsybakov, A. and Yang, L. (1998). Nonparametric vector autoregression, *Journal. of Statistical Planning and Inference,* 68:22-245.
- Härdle, W., and Vieu, P. (1992). Kernel regression smoothing of time series, *Journal of the Time Series Analysis*, 13:209-232.
- Hart, J.D. (1991). Kernel regression estimation with time series errors, *Journal of the Royal Statistical Society, B*, 53(1):173-187.

- Hart, J.D. (1996). Some automated methods of smoothing time-dependent data, *Journal of Nonparametric Statistics*, 6:115-142.
- Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models,* Chapman and Hall.
- Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implementations. *International Journal of Forecasting* , 16:451–476.
- Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). *Forecasting, methods and applications*, Wiley.
- Masry, E. and Tjostheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series, *Econometric Theory*, 11:258-289.
- McLeod, A.I. and Li, W.K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations, *Journal of Time Series Analysis*, 17:571–599.
- Pankratz, A. (1983).*Forecasting with univariate Box-Jenkins models,* Wiley.
- Tjostheim, D. and Auestad, B.H. (1994). Nonparametric identification of nonlinear time series: selecting significant lags, *Journal of the American Statistical Association*, 89:1410-1419.
- Tscherning, R. and Yang, L. (2000). Nonparametric lag selection for time series, *Journal of Time Series Analysis*, 21:457-487.
- Tong, H. (1990). *Non-linear time series,* Oxford.
- Vieu, P. (1994). Order choice in nonlinear autoregressive models, *Statistics*, 24:1-22.

- Vilar-Fernández, J.M. y Cao, R. (2007). Nonparametric Forecasting in Time Series–A Comparative Study, *Communications in Statics–Simulation and Computation*, 36:311-334.
- Wood, S.N. (2006)*Generalizes Additive Models. An introduction with R*, Chapman and Hall.
- Yao, Q. and Tong, H. (1994). On subset selection in non-parametric stochastic regression, *Statistica Sinica*, 4:51-70.