Proyecto fin de Máster

Modelos semiparamétricos para la predicción de la superficie quemada en los incendios forestales.

Aplicación a diez años de historia de Galicia

Curso 2011-2012

Alumno:
Miguel Boubeta Martínez

Codirectores:
Wenceslao González Manteiga
Manuel F. Marey Pérez

29 de Junio de 2012

Índice general

Índice general I

Agradecimientos IV

Resumen V

1. Introducción 1

1.1. La situación forestal en Galicia 2

1.2. La base de datos 3

1.3. Los incendios en Galicia 5

2. El modelo semiparamétrico 7

2.1. Modelización ARMA posterior a la suavización 7

2.1.1. Introducción 7

2.1.2. Modelo 1: Definición y características 8

2.1.3. Intervalos de predicción bootstrap 9

2.2. Suavización posterior a la modelización ARMA 11

2.2.1. Introducción 11

2.2.2. Modelo 2: Definición y características 12

3. Un estudio de simulación 14

3.1. Simulación 1 15

3.1.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización . . . 16

3.1.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA . . . 23

3.1.3. Análisis comparativo de la simulación 1 . . . 27

3.2. Simulación 2 . . . 29

3.2.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización . . . 29

3.2.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA . . . 36

3.2.3. Análisis comparativo de la simulación 2 . . . 39

4. Aplicación a datos reales **41**

4.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización . . . 42

4.1.1. Estimación no paramétrica de la función de regresión . . . 43

4.1.2. Modelización ARMA de los residuos . . . 45

4.1.3. Predicciones del modelo semiparamétrico 1 . . . 49

4.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA . . . 51

4.2.1. Modelización ARMA de la serie . . . 51

4.2.2. Estimación no paramétrica de la función de regresión . . . 52

4.2.3. Predicciones del modelo semiparamétrico 2 . . . 54

4.3. Análisis comparativo . . . 56

5. Conclusiones **57**

A. Base de datos **59**

A.1. Variables de localización . . . 59

A.2. variables de fechas . . . 60

A.3. variables de inicio y causas . . . 61

A.4. variables de terreno 63
A.5. variables de tipo 63
A.6. variables de medios 64
A.7. variables de partes 65
A.8. variables de consecuencias 65
A.9. variables de clima 66
A.10.variables de índices de riesgo 67
A.11.Algunos resultados exploratorios relevantes 67

B. Los modelos Box-Jenkins 70

B.1. Procesos autorregresivos (AR) 70
B.2. Procesos de medias móviles (MA) 71
B.3. Procesos ARMA 71
B.3.1. Procesos ARMA estacionales 72
B.3.2. Procesos ARMA estacionales multiplicativos 72

C. Regresión no paramétrica 73

C.1. Regresión tipo núcleo 74
C.2. Regresión lineal local 75
C.3. Regresión B-Splines 76
C.4. Regresión P-Splines 77

D. Software 78

Bibliografía 80

Agradecimientos

En primer lugar deseo expresar mi más sincero agradecimiento a los profesores Wenceslao González Manteiga y Manuel F. Marey Pérez por la confianza depositada en mí al permitirme formar parte del grupo de investigación y sus horas de dedicación, ya que sin su apoyo este Trabajo Fin de Máster no sería posible.

Finalmente me gustaría agradecer también la ayuda proporcionada por el profesor Manuel Febrero Bande, por sus consejos y apoyo incondicional, especialmente en el aspecto computacional.

Resumen

Partiremos de un proceso estocástico en tiempo discreto y espacios de estados continuo, $\{X_t\}_{t\in\mathbb{Z}}$, del que hemos observado parte de su trayectoria, es decir una serie temporal

$$(X_1, \ldots, X_n)\,.$$

El principal objetivo consiste en obtener predicciones para la misma en instantes futuros de tiempo, existiendo para ello tres corrientes: la paramétrica, la no paramétrica y la semiparamétrica.

La clase de modelos ARMA sigue siendo hoy en día, la familia paramétrica más utilizada. Las razones de su éxito se deben a la generalidad, la relativa facilidad de implementación y la habilidad de proporcionar predicciones lineales óptimas. Sin embargo, existen situaciones donde las familias paramétricas no pueden ser adoptadas con seguridad. En tales situaciones los modelos no paramétricos ofrecen una alternativa idónea.

Los modelos semiparamétricos que se presentan en este trabajo ofrecen una modelización alternativa de las series temporales. En ellos, se descompone la predicción en dos componentes: una no paramétrica y otra paramétrica, modelizable a través de la metodología Box-Jenkins. Para analizar cuál de ellos proporciona mejores resultados, se realizará un análisis comparativo a través de varios estudios de simulación, donde consideraremos muestras de tamaño $n = 500$, $\{x_1, \ldots, x_n\}$. De estas $n = 500$ observaciones, seleccionaremos las $T = 450$ primeras para construir los modelos (*muestra de entrenamiento*), dejando las $n - T = 50$ restantes para la posterior comprobación (mediante diversos criterios de error).

Finalmente, se aplicarán los modelos presentados a la serie temporal de la superficie semanal quemada en los incendios forestales gallegos desde el año 1999 hasta el 2008. En este caso se tomará como muestra de entrenamiento la correspondiente al periodo $1999 - 2007$, esto es $\{s_1, \ldots, s_{468}\}$, dejando el año 2008 $\{s_{469}, \ldots, s_{520}\}$, para la validación.

Capítulo 1

Introducción

Un *incendio forestal* es el fuego que se extiende sin control en terreno forestal y afectando a combustibles vegetales (Molina et al., 2009). También puede definirse como el fuego que se expande sin control sobre especies arbóreas, arbustivas, de matorral o herbáceas, siempre que no sean características del cultivo agrícola o fueren objeto del mismo y que no tengan calificación de terrenos urbanos, afectando a vegetación que no estaba destinada para la quema. En la literatura científica se utiliza el término inglés *wildfire* aunque también puede aparecer referenciado como *brush fire*, *bushfire*, *forest fire*, *desert fire*, *grass fire*, *hill fire*, *peat fire*, *vegetation fire* o *veldfire*, en función del tipo de vegetación sobre el que se desarrolle.

El fuego es un componente importante en muchos ecosistemas forestales (Moritz, 2003), con una gran influencia en las consecuencias ecológicas (Minnich and Bahre, 1995) y las funciones económicas del bosque (Hardy, 2005). Estos fenómenos, al igual que otros como avalanchas, terremotos, tormentas de arena, etc., tienen la propiedad de que cuando ocurren por encima de un determinado umbral producen una cascada de actividad ambiental, social y económica (Malamud et al., 1998). A su vez experimentan comportamientos muy distintos según los diferentes ámbitos en los que se puedan localizar (Reed and McKelvey, 2002), variando en función de variables naturales y socioeconómicas ligadas a los paisajes forestales, y presentando comportamientos difícilmente parametrizables.

En los países Mediterráneos, el fuego se ha convertido en un problema muy grave durante las últimas tres décadas y es actualmente la principal causa de destrucción de los bosques, con una media de área quemada de 500.000 hectáreas por año (Vélez, 2002). En Galicia han ardido 475.940 ha. entre los años 1996 y 2011, sobre un total de 2.060.500 ha. de superficie forestal. El máximo anual se alcanzó en el año 2006, en el que se quemaron un total de 95.945 ha. (DGCN, 2006). Por ello los incendios forestales son percibidos como el principal problema medioambiental de la región (Alonso-Betanzos et al., 2003; Chas, 2007).

1.1. La situación forestal en Galicia

Según el tercer y último Inventario Forestal Nacional de España (1997-2007) (MARM, 2008), la superficie forestal de Galicia representa el 69 % de toda la superficie gallega. La superficie forestal española representa el 51.9 % del total y la de la Unión Europea (UE), el 43.7 %. España en superficie forestal solo se ve superada en la UE15 por Suecia con el 66.9 % y por Finlandia con el 67.3 %.

La superficie forestal gallega es de 2.039.575 ha.; y la arbolada, que es de 1.4 millones de ha. (el 10 % de la superficie forestal arbolada española y el 2.5 % de la europea), ha aumentado un 34 % en los últimos diez años.

Galicia es una potencia forestal al ser la mayor productora de madera de España. En el año 2010 alcanzó la cifra de 6.868.500 m^3, lo que supone el 50 % de la producción maderera española y del orden del 4.5 % de la europea (Fearmaga, 2011).

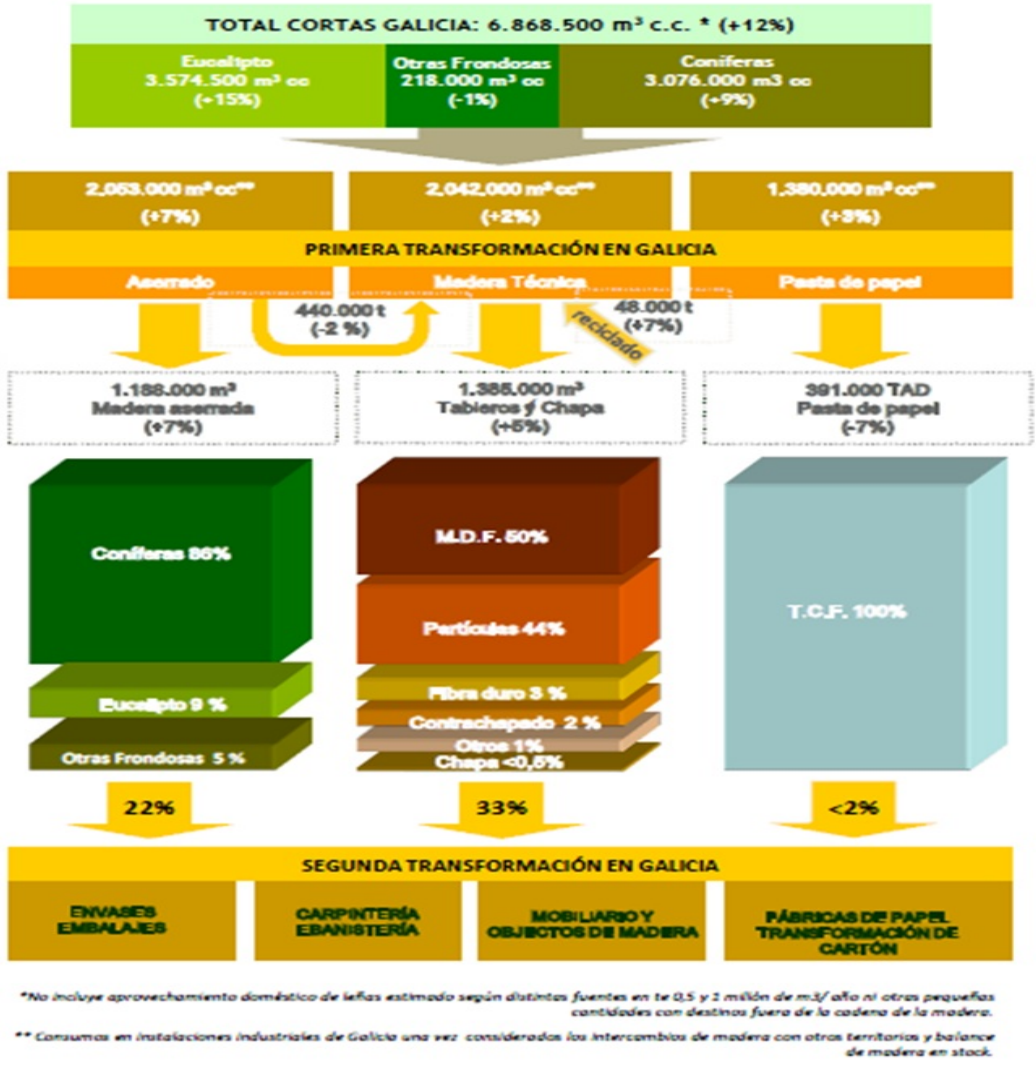

Figura 1.1: Fuente de información: *Resultados Industria de la Madera de Galicia 2010*

Desde 1973 al 2002 la producción maderera gallega se multiplicó por 3.2 mientras que el total de España se multiplicó por 1.4. De esta producción el pino gallego representa el 27.29 %, el roble ("carballo") el 13.36 %, el eucalipto el 12.39 % y el carballo cerquiño o rebolo (otra modalidad del roble) el 7.15 %.

La figura (1.2) representa gráficamente el peso del sector forestal en las diferentes comunidades autónomas [1].

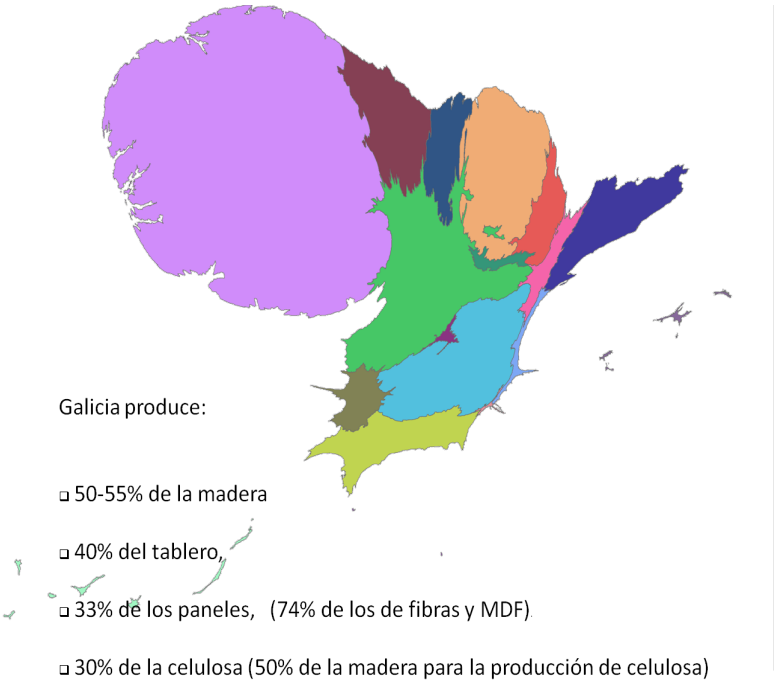

Figura 1.2: Producción de madera en España

La propiedad del monte gallego es minifundista, pues, hay del orden de 600.000 propietarios (la población gallega es de 2.76 millones de habitantes en el 2006) y 2.860 comunidades de montes (Marey-Pérez and Rodríguez-Vicente, 2009). Hay 645.380 ha. de superficie forestal que son montes en mano común y que suponen un tercio de la superficie forestal de Galicia (Marey-Pérez et al., 2006). Tampoco conviene olvidar que en Galicia, en los 315 ayuntamientos hay 30.000 núcleos de población, es decir, la mitad de los núcleos de población de España [2].

1.2. La base de datos

En el año 1956 se creó el Servicio de Incendios Forestales. A partir de su creación se inició la sistematización de los datos estadísticos referentes a estos siniestros, que hasta entonces se habían recogido de modo irregular por los servicios provinciales

[1] Fuente de información: Martín (2010)
[2] Fuente de información: INE (2011)

de los Distritos y el Patrimonio Forestal (Enríquez-Alcalde, 2010). En el año 1967 comienza el proceso de informatización de los datos y la elaboración de un nuevo modelo de Parte de Incendio (formulario utilizado para el acopio de datos), recogiéndose la información a partir del año 1968 de esta forma (MARM, 2008).

Tras la transferencia de competencias en la gestión de los montes a las Comunidades Autónomas durante los años 1984 y 1985, en 1992 se creó la Comisión Técnica de Normalización, cuyas funciones posteriormente fueron asumidas por el Comité de Lucha contra Incendios Forestales ($CLIF$), tras la promulgación en 1994, del Real Decreto que creaba la Comisión Nacional de Protección de la Naturaleza. Esto ha permitido disponer de la información de los incendios forestales en formato digital desde 1968, conformando la Estadística General de Incendios Forestales ($EGIF$) (Enríquez-Alcalde, 2010).

En la actualidad el Área de Defensa contra Incendios Forestales ($ADCIF$) del actual Ministerio de Agricultura Alimentación y Medio Ambiente ($MAGRAMA$) es el organismo encargado de la elaboración y publicación de esta estadística con carácter nacional, a partir de la información que remiten las Comunidades Autónomas, de cada uno de los siniestros forestales que ocurren en el Estado.

La información recogida por la Estadística General de Incendios Forestales a través del Parte de Incendio se ha visto modificada con el paso de los años, adaptándose a las necesidades marcadas por la evolución del fenómeno del incendio forestal y de los medios utilizados para su detección y extinción, manteniendo sin embargo la necesaria continuidad y compatibilidad de dicha información. La última actualización del Parte de Incendio entró en vigor en el año 2005 y la información recogida hace referencia a los siguientes aspectos [3]:

- Localización.
- Tiempos de llegada de los medios, control y extinción del siniestro.
- Detección y lugar de inicio.
- Causalidad y motivaciones de la intencionalidad.
- Condiciones de peligro en el inicio del incendio.
- Tipo de fuego.
- Medios utilizados para la extinción.
- Técnicas de extinción.
- Pérdidas: víctimas, superficies afectadas y efectos ambientales.
- Incidencias de protección civil.
- Afección a espacios naturales protegidos y a reforestación de tierras agrarias.

[3] Una revisión más detallada puede contemplarse en el apéndice A.

1.3. Los incendios en Galicia

En las últimas décadas, el estudio de los incendios forestales en Galicia ha suscitado gran interés dada su alarmante evolución.

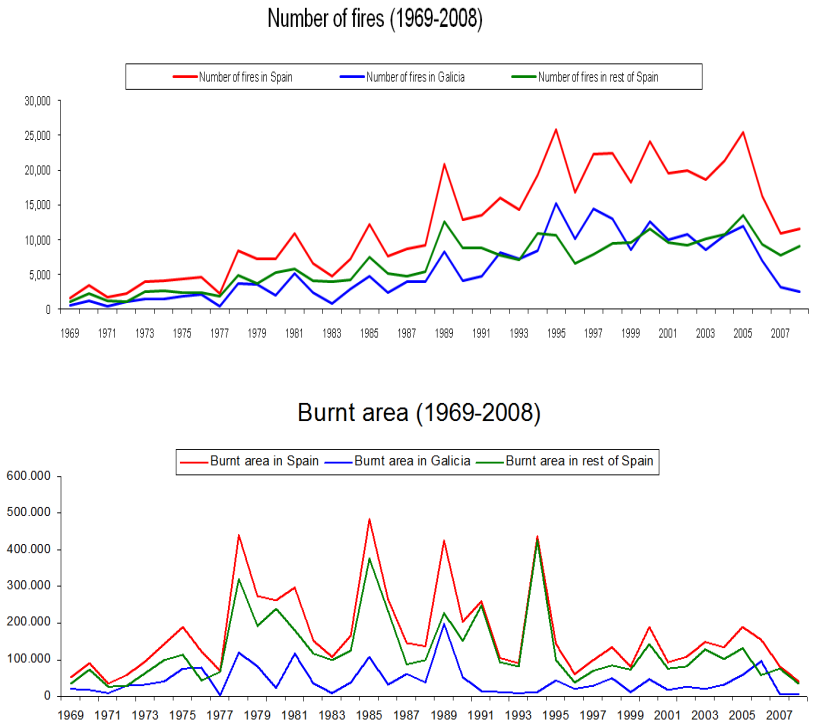

Figura 1.3: Gráficas comparativas del número de incendios y área quemada

Como se puede apreciar en la figura (1.3), la evolución del número de incendios en esta comunidad muestra un comportamiento fundamentalmente creciente desde el año 1969 hasta el 1997, seguido de un decrecimiento en los años posteriores. Además, obsérvese como entre 1995 y 2002, el número de incendios ocurridos en la comunidad gallega supera claramente a los registrados en el resto de España. En cuanto a la superficie quemada, se observa un comportamiento oscilatorio de amplitud considerablemente grande entre los años 1978 y 1990, contemplando el máximo histórico en el año 1989. En relación a lo ocurrido en España, se observa como en los años 1976 y 2006, el área quemada registrada en Galicia supera a la del resto del país.

La base de datos [4] analizada en este trabajo recoge los incendios registrados en Galicia desde el año 1999 al 2008. El total de incendios registrados en estos 10 años es de 85.134, distribuidos por provincias según se muestra a continuación:

[4]Fuente de información: Ministerio de Agricultura Alimentación y Medio Ambiente

Distribución de los incendios					
Año	A Coruña	Lugo	Ourense	Pontevedra	Total
1999	1907	1111	3148	2396	8562
2000	2283	1912	5284	3019	12498
2001	2636	922	3539	2829	9926
2002	2326	1241	3887	3261	10715
2003	1917	1587	2688	2321	8513
2004	2993	1030	3340	3225	10588
2005	2631	1125	4270	3883	11909
2006	2361	871	1601	2071	6904
2007	538	321	1437	757	3053
2008	342	280	1329	515	2466
Total	19934	10400	30523	24277	85134

Tabla 1.1: Distribución de los incendios por provincias

Como se puede apreciar en la tabla (1.1), *Ourense* es la provincia gallega en la que se produce un mayor número de incendios (excepto en el año 2006 donde *A Coruña* registra una mayor cantidad). Además, podemos afirmar que *Pontevedra* y *A Coruña* poseen ciertas similitudes en cuanto a la distribución del número de eventos, mientras que Lugo se presenta como la provincia gallega donde existen menos incendios. Relativo a la distribución temporal de los incendios, se observa una tendencia decreciente en los tres últimos años, siendo 2000 y 2005 los que presentan mayores cantidades.

Diferentes autores han realizado trabajos con el objetivo de relacionar las componentes espaciales y temporales de los incendios forestales en variables como el número de incendios o la superficie quemada. Trabajos interesantes en es tema son los propuestos por (Lee et al., 2006), que relaciona el área quemada y el número de incendios ocurridos en Korea entre los años 1970 y 2003, o (Li et al., 2003) donde se modeliza una variable importante para la presencia de incendios, como es el *índice de sequedad del suelo* (SDI) a través de modelos ARMA. Autores como (Beckage and Platt, 2003), para el caso de los incendios en Florida (EEUU), y (Riaño et al., 2007) para los incendios en África Tropical, modelizan el área quemada por medio de modelos ARIMA. En todos estos trabajos la metodología desarrollada permite anticiparse en el tiempo y activar los sistemas de prevención y lucha contra el fuego.

El objetivo del trabajo es la implementación y validación de un modelo de análisis temporal con capacidad predictiva a un año para variables tales como, número de incendios o superficie forestal quemada en Galicia.

Capítulo 2

El modelo semiparamétrico

Los *modelos semiparamétricos* que se presentan a continuación permiten modelizar series temporales y calcular predicciones para las mismas en instantes futuros de tiempo. Estos modelos suponen una alternativa a la metodología Box-Jenkins; pero a diferencia de estos últimos en los que la modelización es únicamente paramétrica, en los modelos semiparamétricos se descompone la predicción en una componente no paramétrica y otra paramétrica, modelizable a través de un $ARMA$ [1].

Además, es frecuente acompañar las predicciones por los correspondientes intervalos de predicción, siendo éstos típicamente más informativos. Si la hipótesis de normalidad sobre los residuos no se cumple, los intervalos de confianza clásicos no pueden garantizar su nivel de confianza. Por este motivo se han incorporado los intervalos de predicción bootstrap.

2.1. Modelización ARMA posterior a la suavización

Este modelo semiparamétrico propuesto por (García Jurado et al., 1995), se basa a grosso modo, en descomponer la predicción en una componente no paramétrica que estima la tendencia y una predicción *Box-Jenkins* de la serie de los residuos.

2.1.1. Introducción

Sea (Z_l, Y_l), con $l = 0, \pm 1, \pm 2, \ldots$ una serie estrictamente estacionaria, donde Z_l es una serie r-dimensional y Y_l es una serie respuesta unidimensional. El objetivo consiste en estimar

$$\varphi\left(z_l^0\right) = \varphi\left(F\left(.|Z_l = z_l^0\right)\right),$$

[1] Una breve revisión de los modelos ARMA puede contemplarse en el Apéndice B

donde $F(.|Z_l = z_l^0)$ denota la distribución condicional de Y_l dado $Z_l = z_l^0$, usando una serie temporal $\{(Z_l, Y_l)\}$ de longitud n. Frecuentemente, φ es la media o mediana funcional. En particular, cuando $Y_l = X_{l+k}$, con $k \geq 1$ y $Z_l = (X_l, \ldots, X_{l-r+1})$, siendo X_l una serie estacionaria, estamos estimando la función de autorregresión de orden k,

$$\varphi\left(x_1^0, \ldots, x_r^0\right) = \mathbb{E}\left[X_{l+k} | \left(X_l, \ldots, X_{l-r+1}\right) = \left(x_1^0, \ldots, x_r^0\right)\right], \tag{2.1}$$

usando una muestra $\{X_{t-m+1}, \ldots, X_t\}$ de tamaño m.

Como ya se ha mencionado anteriormente este enfoque está basado en procedimientos no paramétricos, estimando la función $\varphi(z_l^0) = \mathbb{E}[Y_l | Z_l = z_l^0]$ sin realizar ninguna hipótesis paramétrica sobre ella.

En general, dada una muestra $\{(Z_i, Y_i)\}_{i=1}^n$, se puede expresar el estimador mediante:

$$\widehat{\varphi}_n\left(z_l^0\right) = \sum_{i=1}^{n} W_{ni}\left(z_l^0, (Z_1, Y_1), \ldots, (Z_n, Y_n)\right) Y_i, \tag{2.2}$$

siendo $\{W_{ni}\}$ una sucesión de pesos.

En este tema, podemos resaltar el artículo propuesto por (Yakowitz, 1985), en el cuál se predice $Y_t = X_{t+1}$ a partir de $Z_t = X_t$ usando una muestra $\{X_1, \ldots, X_m\}$ de un modelo Markoviano estacionario. La función $\varphi(x) = \mathbb{E}[X_{l+1} | X_l = x]$ es estimada a partir de (2.2) usando pesos tipo kernel,

$$W_{ni}\left(x, (X_1, X_2), \ldots (X_n, X_{n+1})\right) = \frac{K\left(\dfrac{x - X_i}{h_n}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right)}, \quad i = 1, \ldots, n - m - 1, \tag{2.3}$$

donde K es la *función kernel* y h_n es el *parámetro ventana*. En este artículo, *Yakowitz* propone la predicción $\widehat{\varphi}_n(X_t)$ como una buena alternativa a los modelos ARMA.

También cabe resaltar el artículo propuesto por (Yakowitz, 1987), en el que se predice $Y_t = X_{t+1}$ a partir de $Z_t = (X_t, X_{t-1}, \ldots, X_{t-p_1+1}, U_t, \ldots, U_{t-p_2+1})$, donde $p_1 + p_2 = p$, X_t es una serie estacionaria, U_t es una serie estacionaria exógena y la secuencia de pesos es del tipo del K_n-*vecino más próximo*,

$$W_{ni}\left(z_l^0, (Z_1, Y_1), \ldots, (Z_n, Y_n)\right) = \frac{\mathbb{1}_{\left\{\|z_l^0 - Z_i\| \leq R(n)\right\}}}{k_n}, \quad i = 1, \ldots, n, \tag{2.4}$$

construidos usando una muestra de tamaño m de X_l y la correspondiente exógena U_l con $p_2 \leq p_1$ y $n = m - p_1$.

2.1.2. Modelo 1: Definición y características

El modelo semiparamétrico 1 (Modelización ARMA posterior a la suavización) es presentado como una alternativa a la metodología *Box-Jenkins*, generalizando el

procedimiento no paramétrico al modelo semiparamétrico. Consideremos el modelo

$$Y_l = \varphi(Z_l) + e_l, \tag{2.5}$$

donde e_l tiene una estructura $ARMA(q,s)$ independiente de Z_l. El objetivo es la obtención de la predicción de Y_l una vez observada la serie Y_l hasta el tiempo $t-k$ y Z_l hasta el tiempo t. En particular, usando la muestra $\{(Z_{t-n+1-k}, Y_{t-n+1-k}), \ldots, (Z_{t-k}, Y_{t-k})\}$ de tamaño n, la predicción $\widehat{Y}_t$ de Y_t es definida mediante

$$\widehat{\varphi}_n(Z_l) + \widehat{e}_l, \tag{2.6}$$

donde $\widehat{\varphi}_n$ es la estimación no paramétrica dada por (2.2), considerando por ejemplo pesos del tipo (2.3) o (2.4) y $\widehat{e}_t$ es la predicción *Box-Jenkins* a k retardos construida a partir de la componente ARMA estimada de la serie,

$$\widehat{e}_t = Y_t - \widehat{\varphi}_n(Z_t). \tag{2.7}$$

Las predicciones obtenidas con este modelo semiparamétrico suelen ser mejores que las obtenidas utilizando los modelos no paramétricos, pues en estos últimos, los residuos pueden no ser ruido blanco y en consecuencia podríamos estar dejando información sin tratar.

2.1.3. Intervalos de predicción bootstrap

Los intervalos de predicción bootstrap se presentan como alternativa a los intervalos de predicción *clásicos*, ya que el comportamiento de estos últimos empeora cuando las hipótesis sobre los residuos no se verifican.

Para simplificar el problema, consideremos el modelo (2.5) sin componente no paramétrica, esto es $\varphi = 0$, y con $\{e_t\}$ siguiendo una estructura $AR(q)$. En tal situación (Tombs and Schucany, 1990) proponen un mecanismo bootstrap para aproximar la distribución condicional de e_t dado $e_{t-k}, e_{t-k-1}, \ldots, e_{t-(n+k)+1}$. En este contexto se propone un mecanismo más general que puede ser usado en modelos $ARI(q,d)$. Supongamos que $\{e_t\}$ en (2.5) sigue una estructura $ARI(q,d)$, es decir, admite la expresión

$$\phi(B)(1-B)^d e_t = a_t,$$

donde $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_q B^q$, B denota al operador retardo (definido por $BX_t = X_{t-1}$) y $\{a_t\}$ es ruido blanco. Es evidente que la serie diferenciada $\tilde{e}_t = \nabla^d e_t$, sigue un modelo $AR(q)$ y en consecuencia se puede calcular $\tilde{e}^*_{t+i-k}$ ($i = -n+1+d, \ldots, -q$ e $i = 1, \ldots, k$) de forma análoga a lo hecho en (Tombs and Schucany, 1990), cuyo algoritmo para una serie $\{X_1, \ldots, X_n\}$, modelizada a través de un $AR(p)$, puede resumirse en:

1. Construir los residuos hacia atrás:

$$\widehat{e}_i = X_i - \widehat{\phi}_1 X_{i+1} - \widehat{\phi}_2 X_{i+2} - \ldots - \widehat{\phi}_p X_{i+p}, \;\; i = n-p, n-p-1, \ldots, 1$$

y calcular su versión corregida, $\widehat{e}_i'$, donde $\widehat{e}_i' = \widehat{e}_i - \overline{e}$, con

$$\overline{e} = \frac{1}{n-p} \sum_{i=p+1}^{n} \widehat{e}_i.$$

2. Arrojar errores bootstrap hacia atrás, $\widehat{e}_i^*$, de la función de distribución empírica de los residuos hacia atrás corregidos.

3. Definir réplicas bootstrap hacia atrás:

$$X_i^* = \widehat{\phi}_1 X_{i+1}^* + \widehat{\phi}_2 X_{i+2}^* + \ldots + \widehat{\phi}_p X_{i+p}^* + \widehat{e}_i^*, \;\; i = n-p, n-p-1, \ldots, 1.$$

4. Calcular versiones bootstrap de los estimadores, $\widehat{\phi}_1^*, \widehat{\phi}_2^*, \ldots, \widehat{\phi}_p^*$.

5. Construir residuos hacia adelante:

$$\widehat{a}_i = X_i - \widehat{\phi}_1 X_{i-1} - \widehat{\phi}_2 X_{i-2} - \ldots - \widehat{\phi}_p X_{i-p}, \;\; i = p+1, p+2, \ldots, n$$

y su versión corregida $\widehat{a}_i'$.

6. Arrojar errores bootstrap hacia adelante, $\widehat{a}_i^*$, de la función de distribución empírica de los residuos hacia adelante corregidos.

7. Definir las réplicas bootstrap hacia adelante:

$$X_{n+j}^* = \widehat{\phi}_1^* X_{n+j-1}^* + \widehat{\phi}_2^* X_{n+j-2}^* + \ldots + \widehat{\phi}_p^* X_{n+j-p}^* + \widehat{a}_{n+j}^*, \;\; j = 1, 2, \ldots, k.$$

Luego, la serie bootstrap puede ser producida por:

$$\left\{ \tilde{e}_{t-(n-d)-k+1}^*, \ldots, \tilde{e}_{t-q-k}^*, \tilde{e}_{t-q-k+1}, \ldots, \tilde{e}_{t-k}, \tilde{e}_{t-k+1}^*, \ldots, \tilde{e}_t^* \right\}.$$

Replicando el proceso bootstrap muchas veces se obtiene el siguiente intervalo de predicción aproximado (a k retardos) para e_t:

$$\left(z_t^{*(\alpha/2)}, z_t^{*(1-\alpha/2)} \right),$$

donde $z_t^{*(\alpha/2)}$ y $z_t^{*(1-\alpha/2)}$ denotan los cuantiles $\alpha/2$ y $1-\alpha/2$ de la distribución bootstrap de e_t^*. De este modo, dado el modelo (2.5) se puede dar un intervalo semiparamétrico de predicción para Y_t usando el intervalo de predicción bootstrap, expresado mediante:

$$\left(\widehat{\varphi}_n(Z_t) + z_t^{*(\alpha/2)}, \widehat{\varphi}_n(Z_t) + z_t^{*(1-\alpha/2)} \right), \tag{2.8}$$

donde los cuantiles bootstrap $\widehat{z}^*$ son obtenidos de la componente $ARMA$, $\{\widehat{e}_t\}_t$. La consistencia del intervalo $\left(z_t^{*(\alpha/2)}, z_t^{*(1-\alpha/2)} \right)$ es una consecuencia del bootstrap para modelos ARI, y su demostración puede verse en (García Jurado et al., 1995).

2.2. Suavización posterior a la modelización ARMA

Este modelo semiparamétrico introducido en (Dabo Niang et al., 2010) proporciona predicciones más eficientes de un proceso estrictamente estacionario admitiendo una representación *ARMA*.

El procedimiento está basado en la estimación de la representación *ARMA*, seguida por una regresión no paramétrica donde los residuos del modelo *ARMA* son usados como variables explicativas. Comparado con los métodos de regresión no paramétricos estándar, el número de variables explicativas puede ser reducido de forma considerable.

2.2.1. Introducción

Después de décadas de modelos de series temporales no lineales, la clase de modelos *ARMA* sigue siendo hoy en día, la familia paramétrica más utilizada. Las razones de su éxito se deben a la generalidad, la relativa facilidad de implementación y la habilidad de proporcionar predicciones lineales óptimas.

Si un proceso estacionario $\{X_t\}$ es un modelo *ARMA*, su predictor lineal óptimo,

$$\mathbb{E}L\left[X_t|\left(X_u, u<t\right)\right] = \sum_{i=1}^{\infty} a_i X_{t-i}, \tag{2.9}$$

es obtenido a partir del modelo *ARMA*. Sin embargo, los modelos *ARMA* también tienen importantes inconvenientes, como por ejemplo el desvanecimiento de la generalidad cuando se hacen fuertes hipótesis sobre el ruido. En tal caso, el predictor lineal óptimo no coincide con el predictor óptimo,

$$\mathbb{E}\left[X_t|\left(X_u, u<t\right)\right] = \phi\left(X_{t-1}, X_{t-2}, \ldots\right). \tag{2.10}$$

En situaciones donde las familias paramétricas no pueden ser adoptadas con seguridad, los modelos no paramétricos ofrecen una alternativa idónea. La regresión tipo núcleo puede ser una elección atractiva, pues tiene por objetivo la estimación de la función de regresión del proceso X_t en sus valores pasados $X_{t-1}, \ldots X_{t-d}$, es decir

$$r(X_{t-1}, \ldots X_{t-d}) = \mathbb{E}\left[X_t|X_{t-1}, \ldots X_{t-d}\right], \tag{2.11}$$

sin el requerimiento de fuertes hipótesis sobre el proceso generador de los datos. Ahora bien, la elección del número d es crucial, pues si d es demasiado pequeño, es probable que las predicciones no paramétricas dejen de ser optimales (incluso cuando el número de observaciones aumenta). Por otra parte, si d es demasiado grande el método está sujeto al llamado *problema de la dimensionalidad* (el estimador núcleo converge a una tasa de $n^{2/(4+d)}$, la cuál se hace pequeña si d es grande).

2.2.2. Modelo 2: Definición y características

Consideraremos una tercera clase de predictores de series temporales combinando técnicas paramétricas y no paramétricas. La idea consiste en utilizar los residuos ARMA como regresores en el enfoque no paramétrico. Más precisamente, se consideran dos enfoques. En el primero se usa

$$\tilde{r}(X_{t-1}, \ldots, X_{t-\ell}, \epsilon_{t-1}, \ldots, \epsilon_{t-m}) \tag{2.12}$$

como aproximación del predictor optimal dado en (2.10), donde

$$\epsilon_t = X_t - \mathbb{E}L\left[X_t | \left(X_u, u < t\right)\right]$$

denota la innovación lineal del proceso estacionario X_t. El uso de la regresión no paramétrica tiene como objetivo captar la estructura no lineal de X_t mientras que el uso de las innovaciones lineales alivia los efectos mencionados anteriormente del problema de la dimensionalidad. Dado que los $\epsilon_t's$ no son observables, éstos serán reemplazados por los residuos $\widehat{\epsilon_t}$ de un modelo $ARMA$.

El segundo enfoque utiliza la descomposición del predictor optimal dado en (2.10) como suma del predictor optimal lineal y el predictor optimal (no lineal) de las innovaciones del proceso:

$$\mathbb{E}\left[X_t | \left(X_u, u < t\right)\right] = \mathbb{E}L\left[X_t | \left(X_u, u < t\right)\right] + \mathbb{E}\left[\epsilon_t | \left(\epsilon_u, u < t\right)\right]. \tag{2.13}$$

La idea consiste en estimar el primer término de forma paramétrica y el segundo de forma no paramétrica. Al igual que sucedía en el caso anterior, las innovaciones no son conocidas por lo que serán reemplazadas por los residuos del modelo $ARMA$.

Además, en (Dabo Niang et al., 2010) se establece la consistencia y normalidad asintótica para los estimadores propuestos en ambos enfoques.

Estimadores tipo núcleo aplicados a los residuos ARMA

Muchos procesos no lineales admiten representaciones $ARMA$. Esto no debe sorprendernos pues por el *teorema de Wold*, si el proceso estocástico $\{X_t\}$ es estacionario y no contiene componentes deterministas, entonces admite una representación del tipo

$$X_t = c + \psi_0 a_t + \psi_1 a_{t-1} + \ldots, \tag{2.14}$$

con $\psi_0 = 1$ y $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

Luego el resultado anterior garantiza que cualquier proceso estacionario o bien es lineal o bien puede ser transformado para que lo sea (sin más que sustraerle la componente determinista). Además la representación MA infinita anterior puede ser

aproximada mediante modelos $ARMA$ con órdenes finitos. El ruido de esa representación $ARMA$ es la innovación lineal de $\{X_t\}$ y no sería una secuencia de variables iid (pues de otra forma $\{X_t\}$ sería un proceso lineal). Estas representaciones se definen como representaciones $ARMA$ *débiles* en contraposición a las representaciones $ARMA$ *estándar* en las que se supone que el ruido es iid.

El comportamiento de los residuos $ARMA$ *débiles*, $\widehat{\epsilon}_t$, se muestra en la siguiente proposición:

Proposición 2.2.1 Bajo condiciones de regularidad (véase Dabo Niang et al. (2010)), se tiene que

$$\widehat{\epsilon}_t = \epsilon_t + s_t + O_P(n^{-1/2}), \quad con\, |s_t| \leq C\rho^t, \tag{2.15}$$

donde las constantes $\rho \in (0,1)$ y C sólo dependen de los valores iniciales. Es más

$$\sum_{t=1}^{n} |\epsilon_t - \widehat{\epsilon}_t| = O_P(n^{-1/2}).$$

Si $\tilde{r}(x) = \tilde{r}\left(\widehat{\epsilon}_t | (\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d}) = x\right)$ denota el estimador tipo núcleo de la regresión de $\widehat{\epsilon}_t$ sobre $\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d}$, evaluado en $x = (x_1, \ldots, x_d)$, entonces es legítimo el uso de

$$\widehat{X}_t^L + \tilde{r}\left(\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d}\right),$$

pues bajo condiciones de regularidad se tiene que:

$$\tilde{r}\left[\widehat{\epsilon}_t | (\widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-d}) = x\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[\epsilon_t | (\epsilon_{t-1}, \ldots, \epsilon_{t-d}) = x\right]$$

en probabilidad.

Implementación

Por simplicidad asumiremos que hemos observado $X_1, \ldots, X_n$ y consideremos la predicción a un retardo, X_{n+1}. Pueden ser investigados tres predictores de X_{n+1}:

1. El estimador puramente no paramétrico,

$$\widehat{X}_{n+1}^{NP} = \widehat{r}\left(X_n, \ldots, X_{n-d+1}\right), \tag{2.16}$$

   donde $\widehat{r}$ denota el estimador de Nadaraya-Watson de la función de regresión evaluado en $X_n, \ldots, X_{n-d+1}$.

2. El estimador puramente paramétrico, $\widehat{X}_{n+1}^L$.

3. El predictor mixto,

$$\widehat{X}_{n+1}^M = \widehat{X}_{n+1}^L + \tilde{r}\left(\widehat{\epsilon}_n, \ldots, \widehat{\epsilon}_{n-d+1}\right), \tag{2.17}$$

   donde los $\widehat{\epsilon}_t's$ son los residuos del modelo ARMA.

Capítulo 3

Un estudio de simulación

En este capítulo realizaremos la simulación de varios modelos teóricos, aplicaremos los modelos semiparamétricos vistos en el capítulo (2) y realizaremos un estudio comparativo entre ambos.

En todos los ejemplos que se mostrarán a continuación realizaremos la simulación de $n = 500$ observaciones de un modelo teórico conocido. De estas $n = 500$ observaciones, nos quedaremos con $T = 450$ para construir los modelos (muestra de entrenamiento), dejando las $n - T = 50$ restantes para la posterior comprobación (muestra de validación).

Recuérdese que los modelos semiparamétricos considerados descomponen la predicción en una componente paramétrica (modelizable a través de la metodología Box-Jenkins) y otra no paramétrica. Esta última estimación se realizará eligiendo el mejor modelo de entre los siguientes: regresión tipo núcleo, lineal local, B-Splines y P-Splines [1]. La selección del modelo de regresión óptimo se llevará a cabo mediante el siguiente algoritmo:

i) Dada la muestra de entrenamiento de tamaño T, $\{x_1, \ldots, x_T\}$, consideraremos la *matriz histórica* generada por esta muestra:

$$MH = \{(x_t, x_{t+1})\}, \text{ con } t = 1, \ldots, T-1.$$

ii) De la matriz histórica anterior (MH), seleccionamos un 75 % de las observaciones al azar (matriz histórica de entrenamiento), dejando el 25 % restante para la posterior validación (matriz histórica de validación).

iii) Con la matriz histórica de entrenamiento construimos las correspondientes estimaciones no paramétricas de la función de regresión.

iv) Calculamos los errores sobre la matriz histórica de validación.

[1] Una revisión teórica de estos métodos puede contemplarse en el Apéndice (C)

v) Repetimos los pasos *ii*)-*iv*) $M = 1000$ veces.

Para el cálculo de los errores del paso *iv*) se han considerado 4 posibilidades:

- Error cuadrático medio: $ECM = \frac{1}{K}\sum_{i=1}^{K}(x_i - \widehat{x}_i)^2$
- Error absoluto: $EA = \frac{1}{K}\sum_{i=1}^{K}|x_i - \widehat{x}_i|$
- Error relativo cuadrático: $\frac{1}{K}\sum_{i=1}^{K}\left(\frac{x_i - \widehat{x}_i}{x_i}\right)^2$
- Error relativo absoluto: $\frac{1}{K}\sum_{i=1}^{K}\left|\frac{x_i - \widehat{x}_i}{x_i}\right|$

siendo K el tamaño de la matriz histórica de validación.

Dado que los criterios de error anteriores pueden proporcionar varios modelos de regresión óptimos, seleccionaremos como método de regresión a utilizar en el modelo semiparamétrico aquél que proporcione un menor error cuadrático medio (ECM).

3.1. Simulación 1

El primer modelo teórico que simularemos se corresponde con un $ARMA(1,1)$, el cuál admite la siguiente expresión:

$$X_t = c + \phi_1 X_{t-1} + a_t + \theta_1 a_{t-1}, \tag{3.1}$$

donde c, ϕ_1 y θ_1 son constantes. De este modelo simularemos una muestra de tamaño $n = 500$, con $c = 0$, $\phi_1 = -0.6$ y $\theta_1 = -0.6$.

A la muestra simulada anteriormente $\{x_1, \ldots, x_n\}$ le extraeremos la muestra de entrenamiento, esto es $\{x_1, \ldots, x_T\}$, con $T = 450$ y dejaremos la parte restante $\{x_{T+1}, \ldots, x_n\}$ para la validación.

El gráfico secuencial de la serie simulada es:

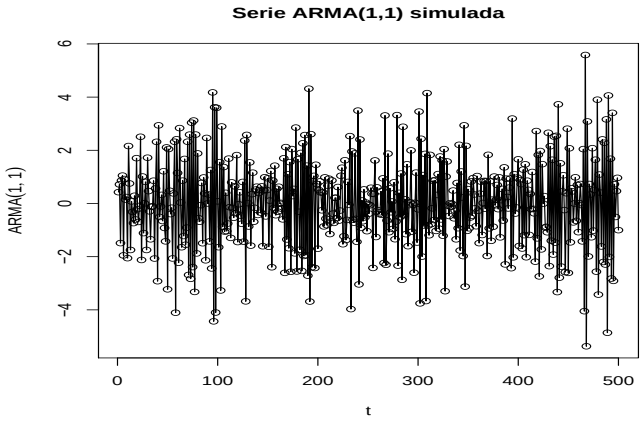

Figura 3.1: Modelo ARMA(1,1) simulado

3.1.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización

Recordemos en primer lugar la expresión matemática de este modelo:

$$\widehat{Y}_t = \widehat{\varphi}_n(Z_t) + \widehat{e}_t, \tag{3.2}$$

donde en lo que resta de trabajo se tomará $Y_t = X_{t+1}$ y $Z_t = X_t$. Consideremos ahora la matriz histórica asociada a la muestra de entrenamiento,

$$\{(x_t, x_{t+1})\}, \text{ con } t = 1, \ldots, T-1$$

y mostremos su diagrama de dispersión

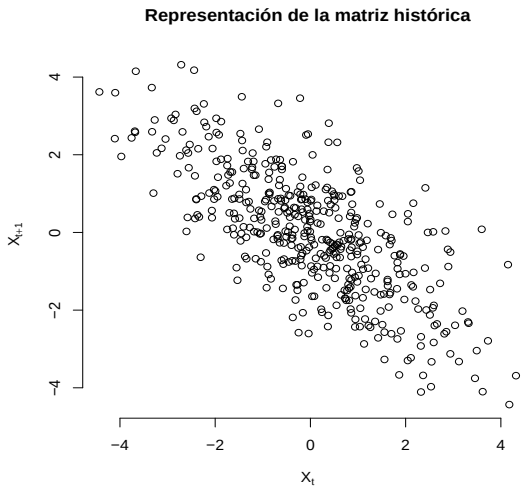

Figura 3.2: Matriz histórica asociada al modelo ARMA(1,1) simulado

Estimación no paramétrica de la función de regresión

Estimaremos la función de regresión de X_{t+1} sobre X_t,

$$m(x) = \mathbb{E}\left[X_{t+1}|X_t = x\right], \tag{3.3}$$

mediante regresión tipo núcleo, lineal local, B-Splines y P-Splines (sobre una matriz histórica de entrenamiento elegida al azar). En los dos primeros casos utilizaremos como función kernel la densidad gaussiana y seleccionaremos el parámetro ventana usando el método de *validación cruzada*. En lo que se refiere a los métodos B-Splines y P-Splines, los grados de libertad se han elegido de acuerdo al criterio GCV. Una vez realizadas las cuatro estimaciones de la función de regresión, seleccionaremos como mejor modelo no paramétrico aquél que proporcione un menor error cuadrático medio (sobre la matriz histórica de validación), tal y como se ha explicado en el algoritmo de la página 14.

Los resultados obtenidos en una iteración particular del algoritmo son:

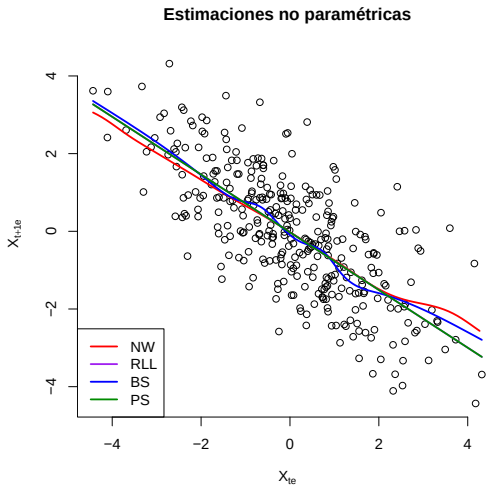

Figura 3.3: Estimaciones no paramétricas de la función de regresión

La figura (3.3) muestra las cuatro estimaciones consideradas de la función de regresión. En este caso, las estimaciones proporcionadas por la regresión lineal local y P-Splines prácticamente coinciden.

Ejecutadas las $M = 1000$ iteraciones obtenemos:

Contadores			
NW	RLL	BS	PS
94	590	79	237

Tabla 3.1: Contadores del mejor estimador no paramétrico de la función de regresión en cada iteración

La tabla (3.1) nos sugiere que el mejor estimador no paramétrico de la función de regresión es el *lineal local*, pues es el que mejores resultados proporciona en 590 de las $M = 1000$ iteraciones. Luego consideraremos la regresión lineal local como la mejor estimación no paramétrica de las consideradas y aplicaremos los modelos ARMA sobre los residuos proporcionados por este modelo.

Además podemos representar los ECM de los cuatro estimadores.

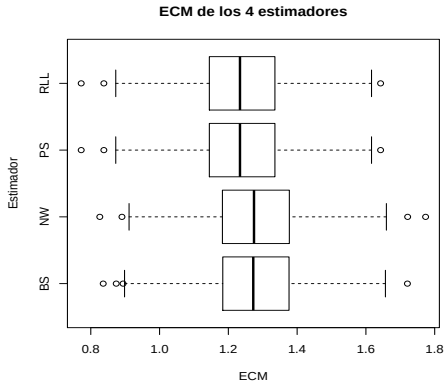

Figura 3.4: Boxplots de los ECM

Obsérvese como la distribución de los ECM en la regresión lineal local y P-Splines se comportan de modo similar.

Modelización ARMA de los residuos

La serie de los residuos obtenida al aplicar el modelo de regresión lineal local es:

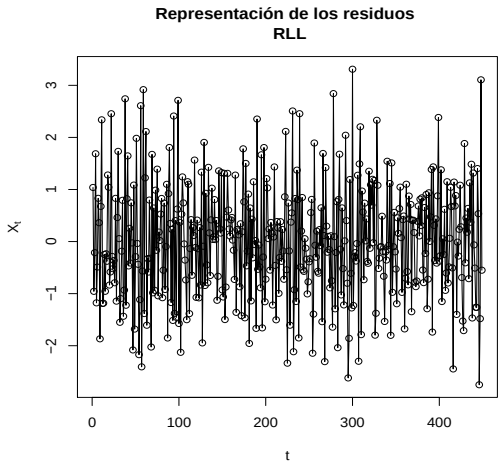

Figura 3.5: Gráfico secuencial de la serie temporal de los residuos (RLL)

Partimos entonces del conocimiento de la serie de los residuos y trataremos de descubrir algún proceso estocástico sencillo susceptible de haber generado la serie. Esto es, trataremos de identificar dicho proceso, estimar sus parámetros y chequear el modelo estimado. Luego, basándonos en tal proceso podemos comprender la dinámica de la serie de tiempo y realizar predicciones de futuros valores.

Trataremos de modelizar nuestra serie a través de un proceso ARMA, ya que esta clase conforma una familia muy flexible, capaz de modelizar gran variedad de series generadas por procesos estacionarios. Para proponer los órdenes del proceso nos basaremos en la información que nos suministra la *fas* y la *fap* muestrales.

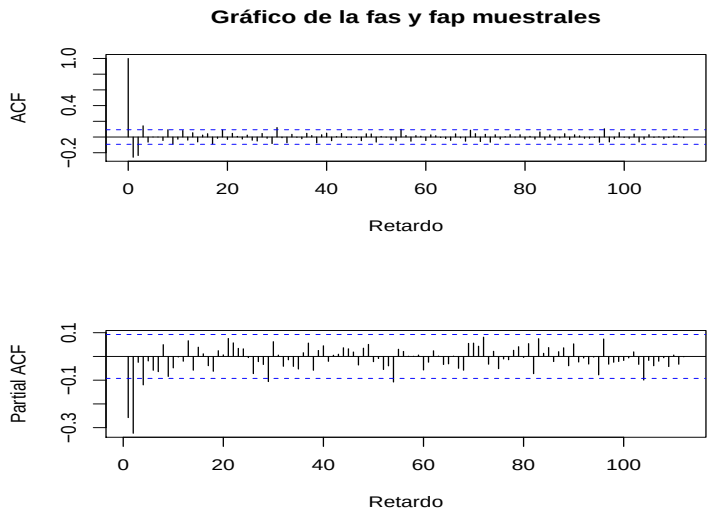

Figura 3.6: fas y fap muestrales de la serie temporal de los residuos (RLL)

Los gráficos anteriores sugieren que la serie proviene de un proceso

$$AR(2) \qquad \text{ó} \qquad MA(3).$$

A continuación utilizaremos el criterio BIC (*criterio de información bayesiana*), con órdenes máximos de cuatro para p y q, con el fin de construir otros modelos tentativos. Este criterio propone seleccionar aquel modelo ARMA que minimice el valor de la función

$$BIC = -2\, lnL(\varphi_{k+1}) + (k+1)\, ln(T), \tag{3.4}$$

donde k denota la cantidad de coeficientes del modelo ARMA ($k = p + q$), φ_{k+1} es el vector formado por dichos coeficientes y por σ_a^2, y $L(\varphi_{k+1})$ denota la función de verosimilitud.

El modelo tentativo propuesto según este criterio es un $AR(2)$ con los siguientes parámetros estimados:

Parámetros estimados		
AR(2)	ϕ_1	ϕ_2
coef.	-0.3407	-0.3285
s.e.	0.0446	0.0450

Tabla 3.2: Parámetros del modelo AR(2) estimados por mínimos cuadrados

La etapa siguiente consiste en comprobar que las hipótesis básicas realizadas sobre el modelo se verifican. Esta etapa se conoce como *diagnosis o chequeo del modelo* y en ella analizaremos si los residuos proceden de un proceso de ruido blanco gaussiano. Para ello empezaremos mostrando el gráfico de los residuos frente al tiempo, pues éste puede ayudarnos a detectar de manera visual y rápida la presencia de: tendencia, componente estacional, variabilidad no constante o dependencia lineal. Cualquiera de estas situaciones invalidaría la hipótesis de ruido blanco.

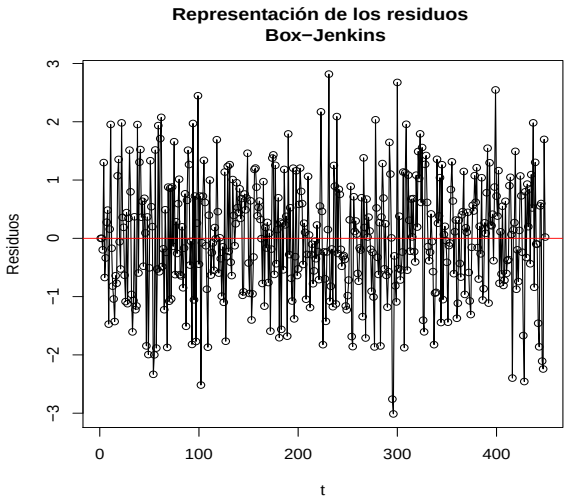

Figura 3.7: Gráfico secuencial de los residuos

Obsérvese que la figura (3.7) no muestra indicios para rechazar la hipótesis.

Para contrastar la hipótesis nula $\mu = 0$ utilizaremos el *t test*. Éste nos proporciona un p-valor de 0.959 y entonces podemos decir que no existen evidencias significativas para rechazar la hipótesis anterior.

El contraste de la independencia será realizado mediante *Ljung-Box*, y dado que los p-valores calculados (véase figura (3.8)) son lo suficientemente grandes, no tenemos evidencias significativas para rechazar la hipótesis. Esto es, asumimos que los errores son independientes.

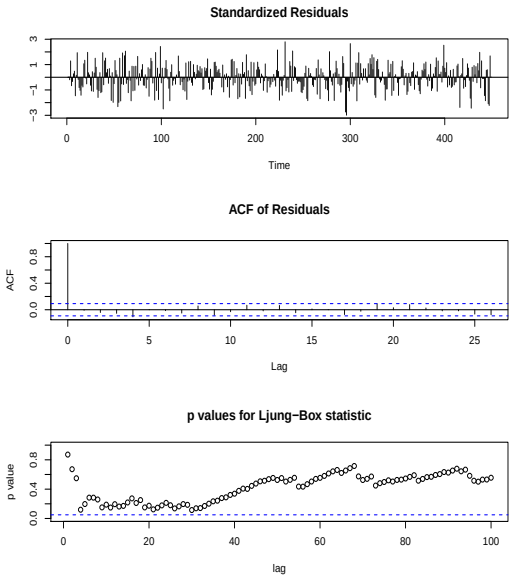

Figura 3.8: Contraste de Ljung-Box

Para finalizar la etapa de diagnosis, contrastaremos la normalidad de los residuos mediante el contraste de *Shapiro-Wilk*, obteniendo un p-valor de 0.891. Luego podemos asumir que los datos provienen de una distribución normal.

En consecuencia podemos afirmar que el modelo ajustado $AR(2)$ sin constante puede ser utilizado como generador de la serie de los residuos, ya que ha pasado con éxito las pruebas de diagnosis. Además las innovaciones son gaussianas por lo que podremos calcular los intervalos de confianza clásicos para las predicciones de los residuos,

$$\left(\widehat{e}_T(k) \pm 1.96\sqrt{Var\left(\epsilon_T(k)\right)}\right), \tag{3.5}$$

siendo $\epsilon_T(k)$ el error de predicción a k retardos, es decir,

$$\epsilon_T(k) = e_{T+k} - \widehat{e}_T(k).$$

Predicciones del modelo semiparamétrico 1

Una vez calculada la componente no paramétrica y paramétrica del modelo semiparamétrico 1 (3.2), podemos obtener las predicciones proporcionadas por el mismo. Para ello optaremos por la estrategia de realizar predicciones a un retardo actualizando los valores de la serie. Esto es, seguiremos el algoritmo:

1. Observada la serie temporal $\{X_1 \ldots, X_T\}$, calculamos:
2. Los residuos del modelo de regresión lineal local, $\{e_1, \ldots, e_{T-1}\}$.
3. La predicción Box-Jenkins a 1 retardo, $\widehat{e}_T$.

4. La predicción del modelo semiparamétrico 1,

$$\widehat{X}_{T+1} = \widehat{m}(X_T) + \widehat{e}_T.$$

5. El residuo del modelo semiparamétrico 1,

$$res = X_{T+1} - \widehat{X}_{T+1}.$$

6. Finalmente actualizamos los valores de la serie, $\{X_1, \ldots, X_T, X_{T+1}\}$,

y repetimos el algoritmo anterior $n - T = 50$ iteraciones. Una vez hechas las cuentas anteriores podemos comparar los valores obtenidos por el modelo semiparamétrico 1 con los futuros valores observados de la serie (muestra de validación).

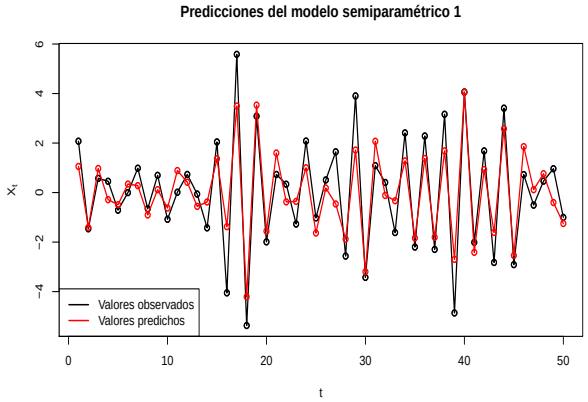

Figura 3.9: Predicciones proporcionadas por el modelo semiparamétrico 1 y futuros valores de la serie observados

Como se puede apreciar en la figura (3.9), las predicciones obtenidas por el modelo semiparamétrico 1 y los futuros valores de la serie son bastante similares. Además podemos calcular la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
1.028	0.826	2522.662	8.959

Tabla 3.3: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Finalmente se ha calculado el intervalo de predicción bootstrap dado en (2.8), en el cuál, en vez de utilizar el método bootstrap propuesto por *Thombs y Schucany (1990)*, se considerará un método alternativo propuesto por Cao, Febrero-Bande, González-Manteiga, Prada-Sánchez y García-Jurado (1997), que es computacionalmente más rápido y también consistente. Éste puede resumirse en los siguientes pasos:

1. Construir la distribución empírica de los residuos hacia adelante corregidos, $F_n^{\widehat{a}'}$.
2. Generar $\widehat{a}_i^*$ con distribución $F_n^{\widehat{a}'}$.
3. Construir réplicas bootstrap futuras,

$$X_{n+j}^* = \widehat{\phi}_1 X_{n+j-1}^* + \widehat{\phi}_2 X_{n+j-2}^* + \ldots + \widehat{\phi}_p X_{n+j-p}^* + \widehat{a}_{n+j}^*, \quad j = 1, 2, \ldots, k,$$

donde $X_i^* = X_i$, para $i = n, n-1, \ldots, n-p+1$.

Además, en este caso también consideraremos el intervalo de predicción clásico, dado que los residuos de la serie Box-Jenkins siguen una distribución normal. Su expresión para el modelo semiparamétrico 1 viene dada por:

$$\left(\widehat{\varphi}_n(X_t) + \widehat{e}_t \pm 1.96\sqrt{Var\left(\epsilon_T\right)}\right) \tag{3.6}$$

siendo ϵ_T el error de predicción de la serie de los residuos.

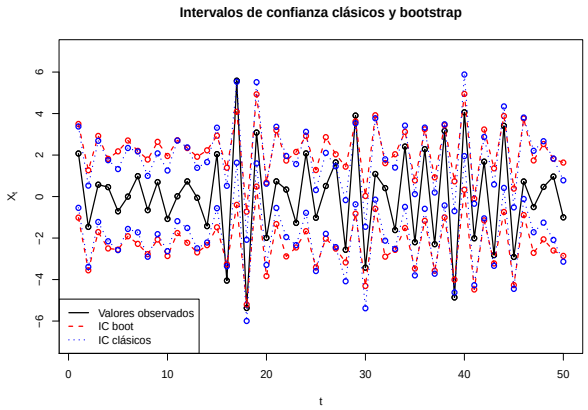

Figura 3.10: Intervalos de predicción clásicos y bootstrap para el modelo semiparamétrico 1

Obsérvese como los intervalos de predicción propuestos por ambos métodos muestran resultados similares.

3.1.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA

Empezaremos recordando la expresión matemática de este modelo:

$$\widehat{X}_{T+1}^M = \widehat{X}_{T+1}^L + \tilde{r}(\widehat{\varepsilon}_T, \ldots, \widehat{\varepsilon}_{T-d+1}), \tag{3.7}$$

donde los $\varepsilon_t's$ representan los residuos ARMA y

$$\tilde{r}(x) = \tilde{r}\left[\varepsilon_t \middle| \left(\varepsilon_{T-1}, \ldots, \varepsilon_{T-d}\right) = x\right].$$

En este trabajo consideraremos $d = 1$.

Modelización ARMA de la serie

Empezaremos realizando un análisis temporal de la serie (3.1),

$$(X_1, \ldots, X_T)\,,$$

utilizando la metodología Box-Jenkins. Para ello utilizaremos el criterio de información bayesiana (3.4), con órdenes máximos de tres, para p y q. El modelo tentativo propuesto por este método es un $ARMA(1,1)$, lo cuál no debe sorprendernos ya que precisamente este es el modelo simulado. Si ajustamos este modelo obtenemos la siguiente estimación de los parámetros:

Parámetros estimados		
ARMA(1,1)	ϕ_1	θ_1
coef.	-0.493	-0.598
s.e.	0.049	0.047

Tabla 3.4: Parámetros del modelo ARMA(1,1) estimados por mínimos cuadrados

Estimación no paramétrica de la función de regresión

El ajuste del modelo ARMA(1,1) nos proporciona los residuos $(\varepsilon_1, \ldots, \varepsilon_T)$, y en consecuencia podemos calcular la matriz histórica asociada:

$$\{(\varepsilon_t, \varepsilon_{t+1})\}\,, \quad \text{con} \quad t = 1, \ldots, T-1.$$

Seguidamente estimaremos la función de regresión de ε_{t+1} sobre ε_t. Es decir,

$$m(x) = \mathbb{E}\left[\varepsilon_{t+1} | \varepsilon_t = x\right], \tag{3.8}$$

mediante regresión tipo núcleo, lineal local, B-Splines y P-Splines. Los resultados obtenidos en una iteración particular del algoritmo son:

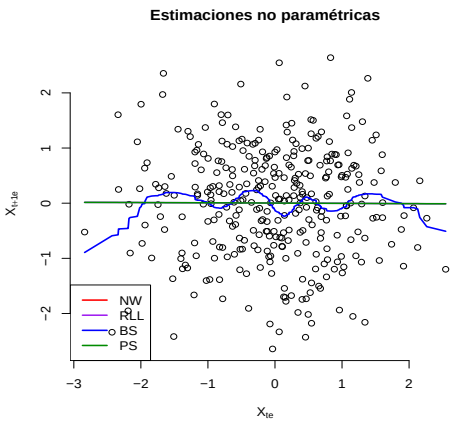

Figura 3.11: Estimaciones no paramétricas de la función de regresión

En los dos primeros casos utilizamos como función kernel la densidad gaussiana y seleccionamos el parámetro ventana usando el método de *validación cruzada*. Para los dos últimos, se han seleccionado los grados de libertad de acuerdo al criterio GCV.

La figura (3.11) muestra las cuatro estimaciones de la función de regresión consideradas. En este caso, las estimaciones obtenidas son bastante similares, salvo la proporcionada por B-Splines.

Escogeremos como modelo no paramétrico óptimo, aquél que nos proporcione un menor error cuadrático medio (sobre la matriz histórica de validación), tal y como se ha explicado en el algoritmo de la página 14.

Ejecutadas las $M = 1000$ iteraciones obtenemos:

Contadores			
NW	RLL	BS	PS
869	18	101	12

Tabla 3.5: Contadores del mejor estimador de la función de regresión en cada iteración

La tabla (3.5) nos sugiere que el mejor estimador de la función de regresión es *Nadaraya-Watson*, pues es el que mejores resultados proporciona en 869 de las $M = 1000$ iteraciones. Luego consideraremos la regresión tipo núcleo como la mejor estimación no paramétrica de las consideradas.

Además podemos representar los ECM de los cuatro estimadores no paramétricos.

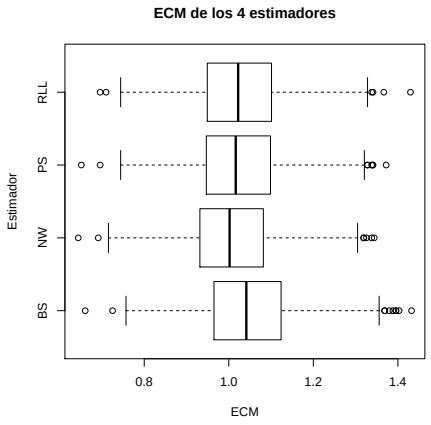

Figura 3.12: Boxplots de los ECM

Obsérvese que el modelo que proporciona mejores resultados es el estimador de Nadaraya-Watson.

Predicciones del modelo semiparamétrico 2

Una vez calculada la componente paramétrica y no paramétrica del modelo semiparamétrico 2 (3.7), podemos calcular las predicciones proporcionadas por el mismo. Para ello optaremos por realizar las predicciones a un retardo, actualizando los valores de la serie. Esto es, seguiremos el algoritmo:

1. Consideramos la serie temporal $\{X_1 \ldots, X_T\}$.
2. Ajustamos el modelo $ARMA(1,1)$ y obtenemos la predicción, $\widehat{X}_{T+1}$.
3. Dados los residuos del modelo ARMA, $\{\varepsilon_1, \ldots, \varepsilon_T\}$, construimos la matriz histórica asociada, $\{(\varepsilon_t, \varepsilon_{t+1})\}_t$.
4. Calculamos la predicción de los residuos por Nadaraya-Watson, $\widehat{\varepsilon}_{T+1}$.
5. Obtenemos la predicción del modelo semiparamétrico 2,
$$\widehat{X}^M_{T+1} = \widehat{X}^L_{T+1} + \widehat{\varepsilon}_{T+1},$$
y el error, $err = X_{T+1} - \widehat{X}^M_{T+1}$.
6. Finalmente actualizamos los valores de la serie, $\{X_1, \ldots, X_T, X_{T+1}\}$,

y repetimos el algoritmo $n-T = 50$ iteraciones. Una vez hechas las cuentas anteriores podemos comparar los valores obtenidos por el modelo semiparamétrico 2 con los futuros valores observados de la serie (muestra de validación).

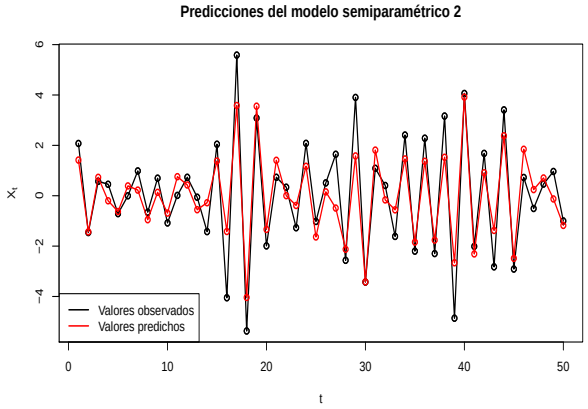

Figura 3.13: Predicciones proporcionadas por el modelo semiparamétrico 2 y futuros valores de la serie observados

Además, podemos calcular la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
0.996	0.791	3182.3	9.632

Tabla 3.6: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Finalmente hemos propuesto intervalos de predicción bootstrap para este modelo semiparamétrico 2, de forma análoga a los dados en (2.8), donde al igual que hemos hecho en el caso anterior utilizaremos el método propuesto por Cao, Febrero-Bande, González-Manteiga, Prada-Sánchez y García-Jurado (1997). Además, dado que los residuos del modelo ARMA son normales (el p-valor obtenido en el contraste de Shapiro-Wilk es 0.76), también se han considerado los intervalos de predicción clásicos.

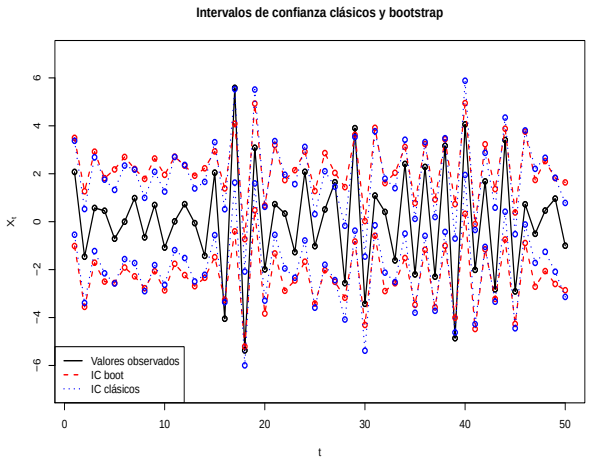

Figura 3.14: Intervalos de predicción clásicos y bootstrap para el modelo semiparamétrico 2

En la figura (3.14) se muestran los dos tipos de intervalos de predicción comentados anteriormente y como se puede apreciar, los resultados son bastante similares, si bien las longitudes en el método clásico parecen menores, en términos generales.

3.1.3. Análisis comparativo de la simulación 1

Ajustados los dos modelos semiparamétricos vistos en este trabajo, podemos realizar un análisis comparativo entre ambos a través de los cuatro criterios de error considerados. Además hemos añadido la predicción paramétrica proporcionada por Box-Jenkins, donde se ha considerado el modelo ARMA(1,1) propuesto por el criterio BIC.

	Criterios de error			
Modelo	ECM	EA	ERC	ERA
BJ	0.992	0.790	3597.234	10.158
SP1	1.028	0.826	2522.662	8.959
SP2	0.996	0.791	3182.300	9.632

Tabla 3.7: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Obsérvese que el modelo Box-Jenkins es el que mejores resultados proporciona para los criterios del Error Cuadrático Medio y del Error Absoluto, mientras que para los errores relativos, es el modelo semiparamétrico 1 el que aporta menor error.

Si deseamos realizar un análisis exploratorio entre los tres métodos podemos representar gráficamente los tres ajustes.

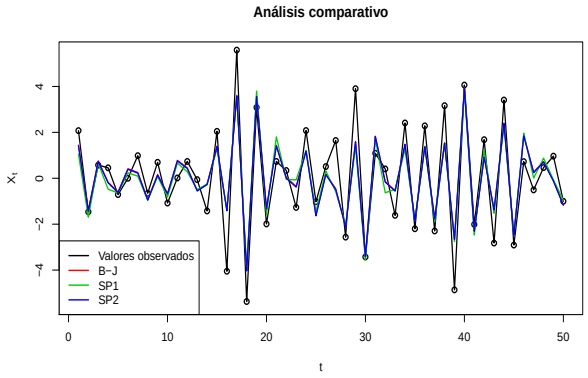

Figura 3.15: Predicciones obtenidas por Box-Jenkins y los dos modelos semiparamétricos

Finalmente, y ya para concluir este primer estudio de simulación, hemos incorporado los intervalos de predicción clásicos al modelo paramétrico (Box-Jenkins),

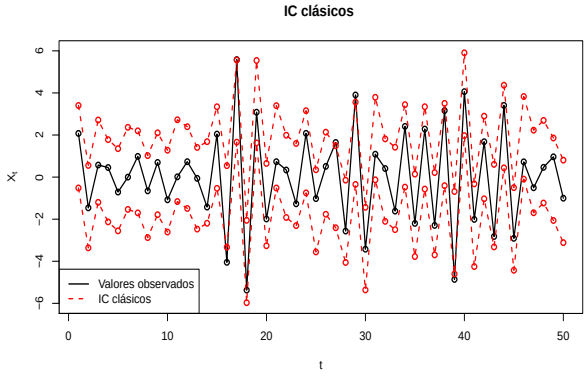

Figura 3.16: Intervalos de predicción clásicos para el modelo Box-Jenkins, ARMA(1,1)

pues los residuos obtenidos por el ARMA(1,1), se distribuyen según una normal (siendo el p-valor del contraste de Shapiro-Wilk, 0.76).

3.2. Simulación 2

El segundo modelo teórico que simularemos admite la siguiente expresión:

$$X_t = S_t + P_t, \tag{3.9}$$

donde $S_t = \sqrt{t}$ y P_t denota un modelo $ARMA(1,1)$. De este modelo simularemos una muestra de tamaño $n = 500$, $\{X_1, \ldots, X_n\}$, con parámetros $\phi_1 = 0.4$ y $\theta_1 = 0.5$. A ésta le extraeremos la muestra de entrenamiento $\{X_1, \ldots, X_T\}$, con $T = 450$, dejando la parte restante, $\{X_{T+1}, \ldots, X_n\}$, para la posterior validación.

A continuación (véase figura (3.17)) se representa el gráfico secuencial de este segundo modelo, el cuál permite observar la tendencia proporcionada por S_t.

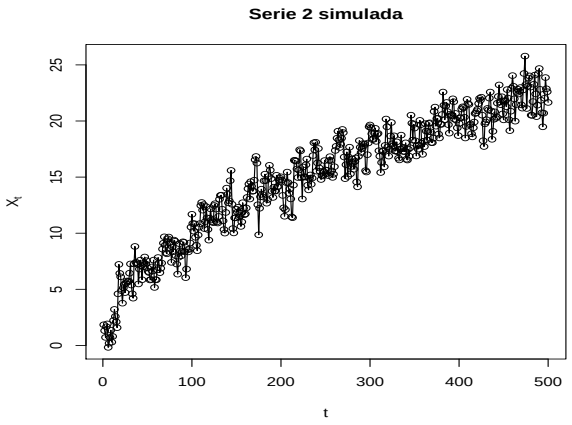

Figura 3.17: Modelo X_t simulado

3.2.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización

Dada la expresión matemática de este modelo semiparamétrico (véase (3.2)), empezaremos considerando la matriz histórica asociada a la muestra de entrenamiento,

$$\{(x_t, x_{t+1})\} \text{ con } t = 1, \ldots, T-1,$$

y mostremos su diagrama de dispersión:

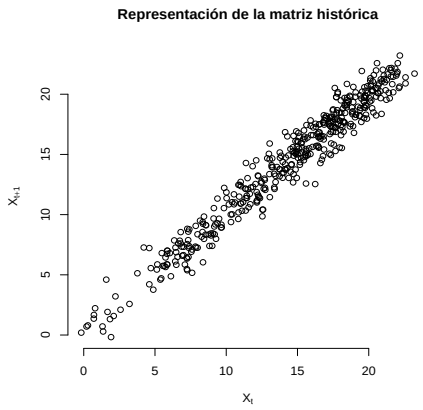

Figura 3.18: Matriz histórica asociada al modelo X_t simulado

Obsérvese como la figura (3.18) muestra una clara dependencia positiva entre las variables.

Estimación no paramétrica de la función de regresión

Estimaremos la función de regresión,

$$m(x) = \mathbb{E}\left[X_{t+1} \mid X_t = x\right],$$

mediante los cuatro estimadores no paramétricos considerados, donde la elección del parámetro ventana y de los grados de libertad se realiza de forma análoga a lo hecho en la simulación 1 (véase página 17). Seleccionaremos como modelo óptimo aquél que proporcione un menor error cuadrático medio (sobre la matriz histórica de validación). [2]

Los resultados obtenidos en una iteración particular del algoritmo son:

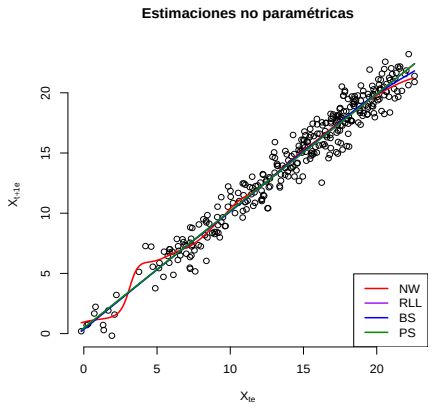

Figura 3.19: Estimaciones no paramétricas de la función de regresión

[2] véase el algoritmo descrito en la página 14

La figura (3.19) muestra las cuatro estimaciones no paramétricas consideradas de la función de regresión, y como se puede apreciar, todas ellas muestran resultados bastante similares, lo cual no debe sorprendernos pues el gráfico de dispersión muestra una clara dependencia lineal positiva.

Ejecutadas las $M = 1000$ iteraciones tenemos:

Contadores			
NW	RLL	BS	PS
297	284	54	365

Tabla 3.8: Contadores del mejor estimador no paramétrico de la función de regresión en cada iteración

Entonces consideraremos la regresión *P-Splines* como mejor estimación no paramétrica de las consideradas y aplicaremos los modelos ARMA sobre los residuos proporcionados por esta regresión.

Además podemos representar los ECM de los cuatro estimadores mediante gráficos de boxplots, obteniendo los siguientes resultados:

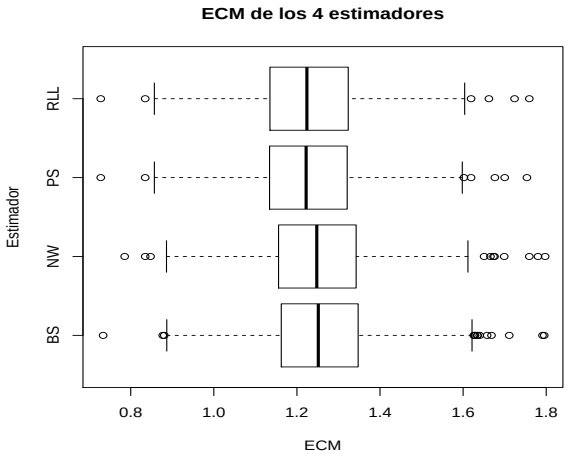

Figura 3.20: Boxplots de los ECM

La figura anterior (3.20) sugiere que los modelos que mejor se comportan (tomando como criterio el ECM), son la regresión lineal local y P-Splines.

Modelización ARMA de los residuos

La serie de los residuos obtenida al aplicar el modelo P-Splines es:

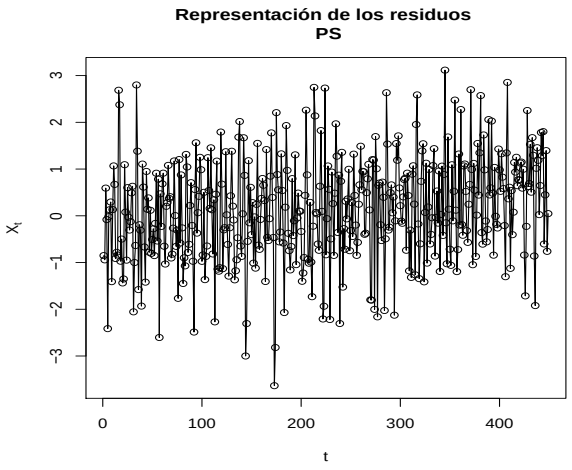

Figura 3.21: Gráfico secuencial de la serie temporal de los residuos (PS)

La figura (3.21) muestra que la componente no paramétrica del modelo semiparamétrico 1 ha absorbido el efecto de la tendencia en la serie original, pues como bien se puede apreciar, la serie de los residuos tiene nivel constante. A continuación trataremos de modelizar esta nueva serie a través de los modelos ARMA y para proponer los órdenes del proceso nos basaremos en la información que nos suministra la *fas* y *fap* muestrales.

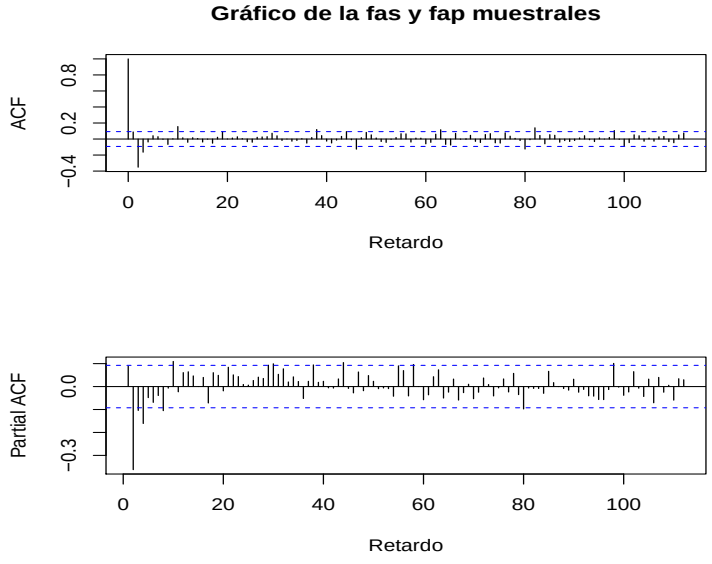

Figura 3.22: fas y fap muestrales de la serie temporal de los residuos

La figura anterior nos sugiere que un proceso

$$AR(4) \qquad \text{ó} \qquad MA(3),$$

puede ser admitido como posible generador de la serie. Además, podemos utilizar el criterio *BIC*, con órdenes máximos de 4 para p y q, con el fin de construir otros modelos tentativos. Así obtenemos que el modelo tentativo propuesto según este criterio es un $ARMA(2,3)$, con los siguientes parámetros estimados:

	Parámetros estimados				
ARMA(2,3)	ϕ_1	ϕ_2	θ_1	θ_2	θ_3
coef.	1.2537	-0.4383	-1.2362	0	0.3478
s.e.	0.0710	0.0524	0.0545	0	0.0344

Tabla 3.9: Parámetros del modelo ARMA(2,3) estimados por mínimos cuadrados

Dado el modelo tentativo anterior, la siguiente etapa consiste en comprobar que las hipótesis básicas del modelo se verifican, es decir analizaremos si los residuos proceden de un proceso de ruido blanco gaussiano. Para ello empezaremos mostrando el gráfico de los residuos frente al tiempo, pues éste puede ayudarnos a detectar de manera visual y rápida la presencia de: tendencia, componente estacional o variabilidad no constante. Cualquiera de estas situaciones invalidaría la hipótesis de ruido blanco.

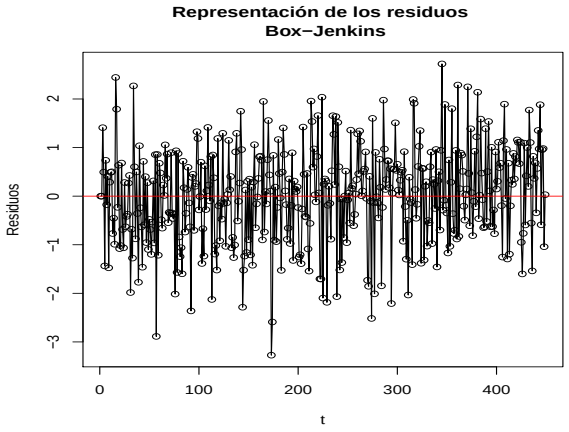

Figura 3.23: Gráfico secuencial de los residuos

Obsérvese que la figura (3.23) no muestra indicio alguno para rechazar la hipótesis de ruido blanco.

El contraste de la hipótesis nula $\mu = 0$ será resuelto utilizando el t *test*. Éste nos da un p-valor de 0.689 y por lo tanto podemos afirmar que no existen evidencias significativas para rechazar la hipótesis anterior.

En el estudio de la hipótesis de independencia podemos utilizar el contraste de Ljung-Box, y dado que los p-valores obtenidos (véase figura (3.24)) son lo suficientemente

grandes, no tenemos evidencias significativas para rechazar la hipótesis de independencia. Esto es, asumimos que los errores son independientes.

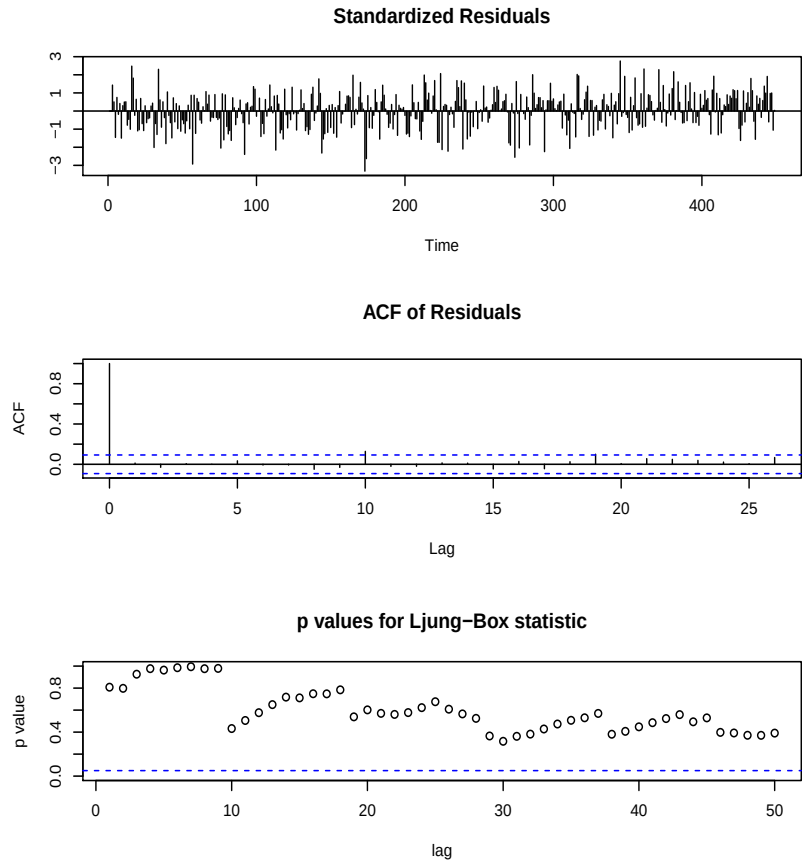

Figura 3.24: Contraste de Ljung-Box

Para finalizar la etapa de diagnosis del modelo, contrastaremos la normalidad de los residuos mediante el contraste de *Shapiro-Wilk*, obteniendo un p-valor de 0.337, y por tanto no tenemos evidencias significativas para rechazar que los datos provengan de una distribución normal.

Predicciones del modelo semiparamétrico 1

Calculada la estimación no paramétrica (por P-Splines) y la paramétrica (mediante modelos ARMA), estamos en condiciones de calcular las predicciones que el modelo semiparamétrico 1 aporta a este ejemplo simulado. Para ello seguiremos el algoritmo visto en la página 21, repitiéndolo tantas veces como instantes de tiempo queramos predecir (en este caso $n - T = 50$ iteraciones). Una vez hechas las cuentas anteriores podemos comparar los valores obtenidos por el modelo semiparamétrico 1 con los futuros valores de la serie observados (muestra de validación).

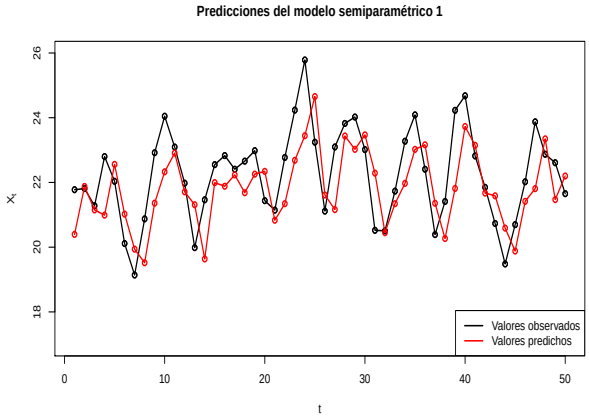

Figura 3.25: Predicciones proporcionadas por el modelo semiparamétrico 1 y futuros valores de la serie observados

Como se puede apreciar en la figura (3.25), las predicciones obtenidas por el modelo semiparamétrico 1 y los futuros valores de la serie observados son bastante similares. Además este modelo nos permite calcular la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
1.302	0.967	0.003	0.043

Tabla 3.10: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Finalmente se han calculado los intervalos de predicción bootstrap, dados en (2.8), y los clásicos. El uso de estos últimos es lícito, dado que los residuos de la serie Box-Jenkins pertenecen a una distribución normal. Los resultados obtenidos son:

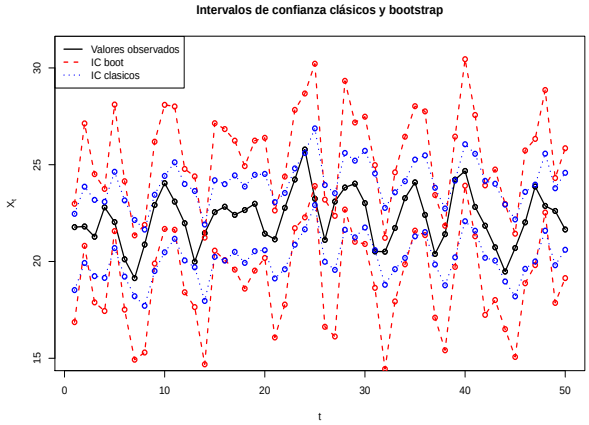

Figura 3.26: Intervalos de predicción clásicos y bootstrap para el modelo semiparamétrico 1

Obsérvese en este caso las diferencias existentes entre los intervalos propuestos por la teoría asintótica y la bootstrap.

3.2.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA

Dada la expresión matemática de este modelo, (véase (3.7)), empezaremos estimando la primera componente a través de los modelos ARMA.

Modelización ARMA de la serie

Empezaremos realizando el análisis temporal de la serie $(X_1, \ldots, X_T)$ del modelo definido en (3.9). Para ello utilizaremos el criterio BIC, con órdenes máximos de 4 para p y q, con el fin de construir modelos tentativos. El modelo propuesto por este criterio es un $ARMA(1,3)$, con los siguientes parámetros estimados:

	Parámetros estimados				
ARMA(1,3)	ϕ_1	θ_1	θ_2	θ_3	*intercept*
coef.	0.9954	-0.0919	-0.6006	-0.2445	23.3590
s.e.	0.0007	0.0438	0.0423	0.0452	1.5118

Tabla 3.11: Parámetros del modelo ARMA(1,3) estimados por mínimos cuadrados

Estimación no paramétrica de la función de regresión

El ajuste del modelo $ARMA(1,3)$ nos proporciona los residuos $(\varepsilon_1, \ldots, \varepsilon_T)$ y en consecuencia podemos calcular la matriz histórica asociada,

$$\{(\varepsilon_t, \varepsilon_{t+1})\}, \quad \text{con} \quad t = 1, \ldots, T-1.$$

Luego, podemos estimar la función de regresión de ε_{t+1} sobre ε_t, utilizando para ello las cuatro estimaciones no paramétricas consideradas en este trabajo. La selección del parámetro ventana y los grados de libertad se realizará de forma análoga a lo hecho en la simulación 1 [3]. Escogeremos como modelo no paramétrico óptimo aquél que nos proporcione un menor error cuadrático medio (sobre la matriz histórica de validación), tal y como se ha explicado en la página 14.

Los resultados obtenidos en una iteración particular del algoritmo son:

[3] Para más información consúltese la página 17

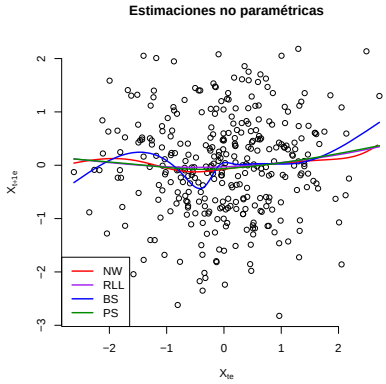

Figura 3.27: Estimaciones no paramétricas de la función de regresión

Obsérvese como todas las estimaciones proporcionan aproximaciones similares, excepto la de B-Splines, donde se observa un comportamiento menos suave.

Ejecutadas las $M = 1000$ iteraciones tenemos:

Contadores			
NW	RLL	BS	PS
569	70	262	99

Tabla 3.12: Contadores del mejor estimador de la función de regresión

La tabla (3.12) sugiere que el mejor estimador no paramétrico de la función de regresión es *Nadaraya-Watson*, pues es el que mejores resultados proporciona en 569 de las $M = 1000$ iteraciones. Luego consideraremos este estimador como el que mejor estima la función de regresión, siendo el utilizado para obtener las posteriores predicciones.

Además podemos representar los ECM de los cuatro estimadores no paramétricos.

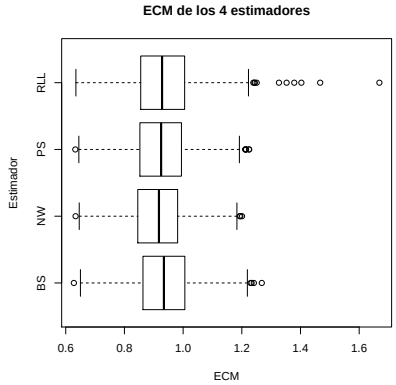

Figura 3.28: Boxplots de los ECM

Obsérvese como la distribución de los ECM que proporciona mejores resultados es la correspondiente al estimador de Nadaraya-Watson.

Predicciones del modelo semiparamétrico 2

Una vez calculada la componente paramétrica y no paramétrica del modelo semiparamétrico 2 (3.7), podemos calcular las predicciones proporcionadas por el mismo. Para ello se opta por realizar las predicciones a un retardo actualizando los valores de la serie [4], repitiendo el algoritmo tantas veces como instantes queramos predecir (en este caso $n - T = 50$ iteraciones). Una vez hechas las cuentas anteriores podemos comparar los valores obtenidos por el modelo semiparamétrico 2 con los futuros valores de la serie observados (muestra de validación).

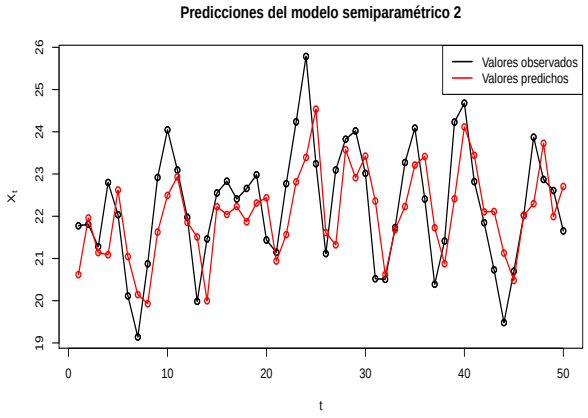

Figura 3.29: Predicciones proporcionadas por el modelo semiparamétrico 2 y futuros valores de la serie observados

Y podemos calcular la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
1.128	0.889	0.002	0.040

Tabla 3.13: Tabla de errores: Error cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Además hemos calculado los intervalos de predicción clásicos (dado que los residuos siguen una distribución normal, siendo el p-valor obtenido en el test de *Shapiro-Wilk* de 0.629) y bootstrap.

[4] tal y como se ha explicado en el algoritmo de la página 26

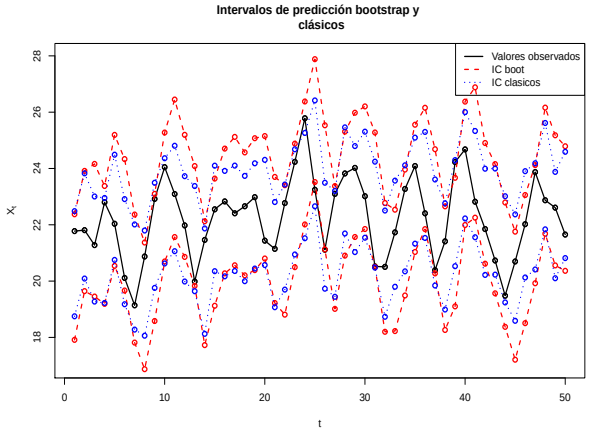

Figura 3.30: Intervalos de predicción clásicos y bootstrap para el modelo semiparamétrico 2

Obsérvese como en este caso los intervalos de predicción clásicos y bootstrap muestran un comportamiento similar.

3.2.3. Análisis comparativo de la simulación 2

Ajustados los dos modelos semiparamétricos vistos en este trabajo, podemos realizar un análisis comparativo entre ambos a través de los cuatro criterios de error considerados. Además hemos añadido la predicción paramétrica proporcionada por Box-Jenkins, donde se ha considerado un modelo $ARIMA(0,1,3)$.

	Criterios de error			
Modelo	ECM	EA	ERC	ERA
BJ	1.131	0.888	0.002	0.040
SP1	1.302	0.967	0.003	0.043
SP2	1.128	0.889	0.002	0.040

Tabla 3.14: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

Obsérvese que el modelo $ARIMA(0,1,3)$ y el semiparamétrico 2 son los que ofrecen mejores resultados en todas las mediciones del error, y téngase en cuenta también, que el ECM es el criterio que muestra mayores diferencias entre los modelos considerados, mientras que en los restantes, los resultados son bastante similares.

Si deseamos realizar un análisis exploratorio entre los tres métodos podemos representar gráficamente los tres ajustes.

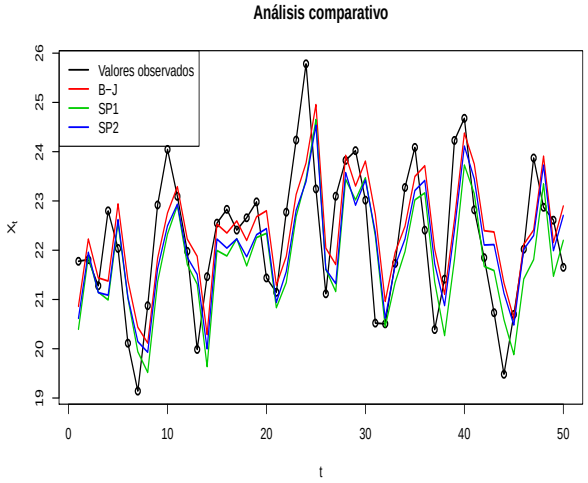

Figura 3.31: Predicciones obtenidas por Box-Jenkins y los dos modelos semiparamétricos

Finalmente, y para concluir este segundo estudio de simulación, se han incorporado los intervalos de predicción clásicos al modelo paramétrico (Box-Jenkins), pues los residuos obtenidos por el modelo $ARIMA(0,1,3)$ se distribuyen según una normal (siendo el p-valor obtenido en el contraste de Shapiro-Wilk de 0.339).

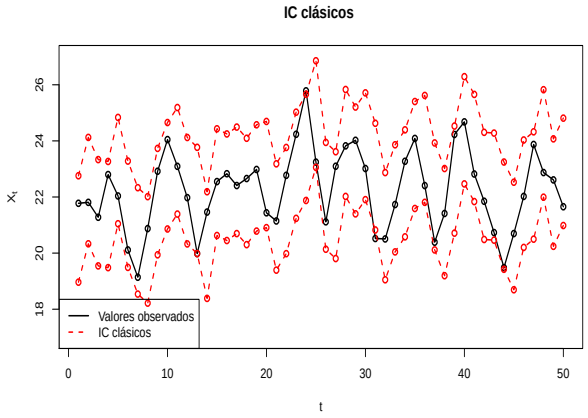

Figura 3.32: Intervalos de predicción clásicos para el modelo $ARIMA(0,1,3)$

Capítulo 4

Aplicación a datos reales

Predecir la superficie quemada en instantes de tiempo futuros es de vital importancia, pues como ya se ha comentado, esto permite una mejora de la gestión en los servicios forestales. Por este motivo, el principal objetivo del presente capítulo es el de calcular estimaciones de la superficie quemada en los incendios forestales gallegos y para ello emplearemos los modelos semiparamétricos definidos en el capítulo 2.

Partiremos de la observación de las superficies semanales quemadas desde el año 1999 hasta el 2008. Es decir, tenemos

$$(S_1, \dots S_{520}) \,, \tag{4.1}$$

donde S_i con $i = 1, \dots, 520$ releja la superficie total quemada en la semana i.

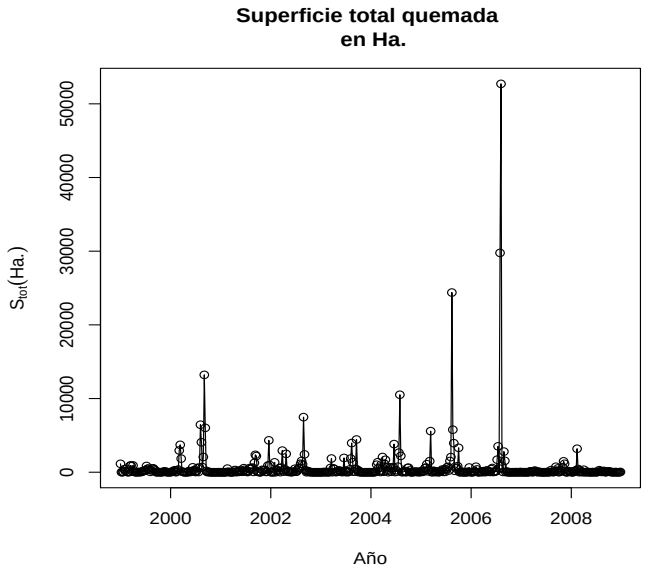

Figura 4.1: Serie de tiempo de la superficie semanal quemada

El comportamiento de nuestra serie en estudio puede decirse que es aproximadamente constante, incorporando bruscos incrementos debidos fundamentalmente al periodo estival (véase 4.1).

Con el fin de evaluar la calidad de las predicciones obtenidas y de realizar un análisis comparativo entre ambos modelos semiparamétricos, consideraremos la serie de tiempo anterior (4.1), pero sin las observaciones del último año. Es decir, nuestra serie temporal observada será:

$$(S_1, \dots S_T)\,, \text{ con } T = 468. \tag{4.2}$$

Los modelos serán construidos a partir de esta nueva serie y dejaremos las observaciones correspondientes al año 2008 $(S_{T+1}, \dots, S_{520})$, para la validación.

4.1. Modelo semiparamétrico 1: Modelización ARMA posterior a la suavización

Dada la expresión matemática de este modelo,

$$\widehat{S}_{t+1} = \widehat{\varphi}(S_t) + \widehat{e}_t, \tag{4.3}$$

empezaremos considerando la matriz histórica asociada a la muestra de entrenamiento $(S_1, \dots, S_T)$,

$$\{(s_t, s_{t+1})\}\,, \text{ con } \; t = 1, \dots, T-1,$$

la cuál admite el siguiente diagrama de dispersión:

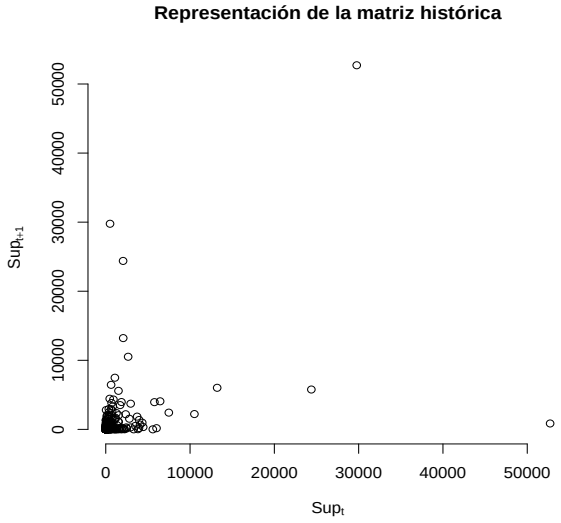

Figura 4.2: Matriz histórica asociada a la distribución temporal de los incendios en Galicia (1999-2007)

La figura (4.2) sugiere presencia de heterocedasticidad, pues se observa un incremento de la variabilidad a medida que aumentan los valores de S_t. Luego proponemos una transformación logarítmica con el fin de estabilizar la variabilidad.

Efectuada la transformación logarítmica, obtenemos la siguiente representación de la matriz histórica:

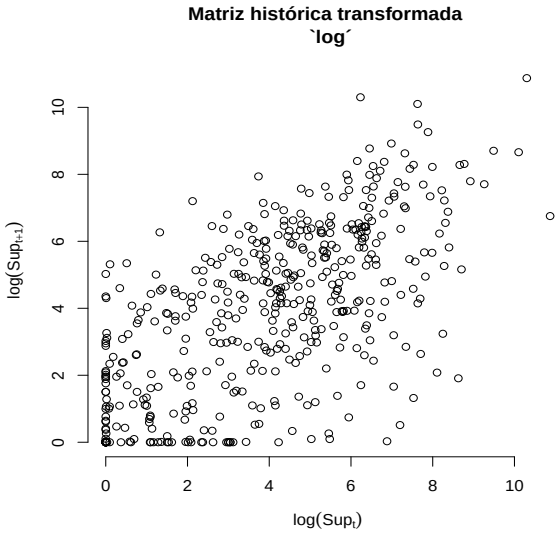

Figura 4.3: Matriz histórica transformada (mediante un logaritmo) asociada a la distribución temporal de los incendios en Galicia (1999-2007)

Obsérvese que la transformación logarítmica ha estabilizado la variabilidad por lo que a partir de ahora trabajaremos sobre las variables transformadas.

4.1.1. Estimación no paramétrica de la función de regresión

Estimaremos la función de regresión de S_{t+1} sobre S_t,

$$m(x) = \mathbb{E}\left[S_{t+1} \mid S_t = x\right], \tag{4.4}$$

mediante los cuatro estimadores no paramétricos de la función de regresión considerados (Nadaraya-Watson, lineal local, B-Splines y P-Splines), sobre una matriz histórica de entrenamiento elegida al azar. En los dos primeros casos utilizaremos como función kernel la densidad gaussiana y seleccionaremos el parámetro ventana usando el método de validación cruzada, mientras que en los dos últimos elegiremos los grados de libertad mediante el criterio GCV. Una vez realizadas las cuatro estimaciones de la función de regresión, seleccionaremos como mejor modelo no paramétrico aquél que proporcione un menor error cuadrático medio (sobre la matriz histórica de validación), tal y como se ha explicado en el algoritmo de la página 14.

Los resultados obtenidos en una iteración particular del algoritmo son:

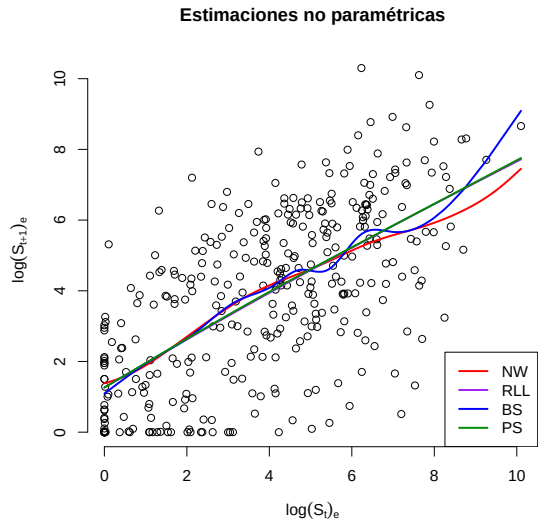

Figura 4.4: Estimaciones no paramétricas de la función de regresión

La figura (4.4) muestra las cuatro estimaciones consideradas de la función de regresión. En este caso, las estimaciones proporcionadas por la regresión lineal local y P-Splines prácticamente coinciden.

Ejecutadas las $M = 1000$ iteraciones obtenemos:

Contadores			
NW	RLL	BS	PS
209	207	98	486

Tabla 4.1: Contadores del mejor estimador no paramétrico de la función de regresión en cada iteración

La tabla (4.1) sugiere que el mejor estimador no paramétrico de la función de regresión es *P-Splines*, pues es el que mejores resultados proporciona en 486 de las $M = 1000$ iteraciones. Luego consideraremos esta regresión como la mejor estimación no paramétrica de las consideradas y aplicaremos los modelos ARMA sobre los residuos proporcionados por este modelo.

Además podemos representar los ECM de los cuatro estimadores mediante boxplots, observando como su distribución en la regresión lineal local y P-Splines se comporta de modo similar. Obsérvese también como en los casos restantes (Nadaraya-Watson y B-Splines), la distribución de los ECM toma valores mayores, por lo que su uso no es aconsejable en este caso.

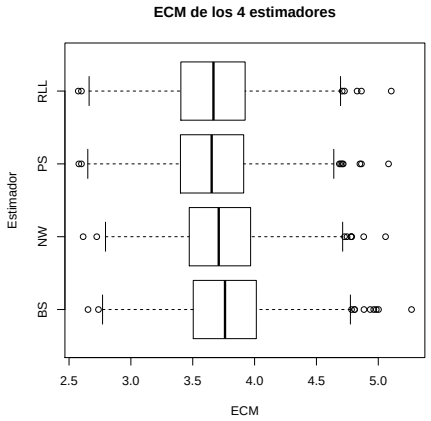

Figura 4.5: Boxplots de los ECM

4.1.2. Modelización ARMA de los residuos

La serie de los residuos obtenida al aplicar el modelo P-Splines, sobre la matriz histórica transformada mediante un logaritmo es:

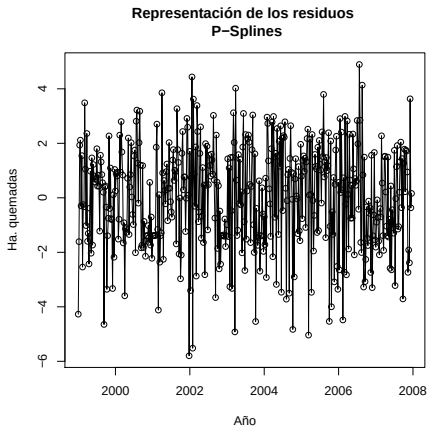

Figura 4.6: Gráfico secuencial de la serie temporal de los residuos (P-Splines)

Partimos entonces del conocimiento de la serie de los residuos (P-Splines) y trataremos de descubrir algún proceso estocástico sencillo susceptible de haberla generado. Para ello es lícito el uso de los modelos ARMA, pues como puede contemplarse en la figura (4.6), la serie observada es compatible con un proceso estacionario, dado que la variabilidad es constante y no posee tendencia ni componente estacional. Para el estudio de esta última afirmación tendremos en cuenta el gráfico secuencial y la función de autocorrelación simple (fas).

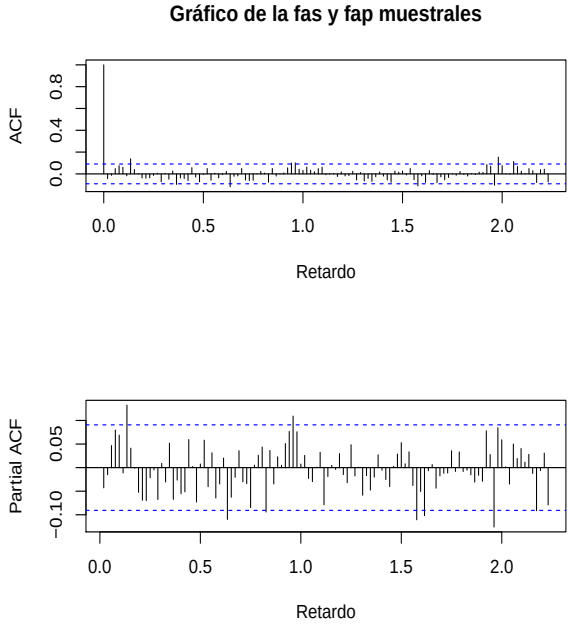

Figura 4.7: fas y fap muestrales de la serie de los residuos (P-Splines)

Pues la componente estacional suele ser delatada por:

1. El gráfico secuencial de la serie.
2. La fas muestral:
   - Presencia de fuerte correlación positiva en el retardo estacional (y, posiblemente en sus múltiplos).
   - Convergencia lenta a cero a medida que el retardo crece.
   - Presencia de periodicidad del mismo periodo que la serie.

Obsérvese como la fas muestral converge a cero rápidamente a medida que el retardo crece (véase figura (4.7)), por tanto desechamos la presencia de componente estacional.

En consecuencia podemos modelizar la serie anterior a través de un proceso ARMA. Para proponer los órdenes del proceso nos basaremos en la información que nos suministra la fas y la fap muestrales. El gráfico anterior sugiere que la serie de los residuos ha sido generada por un proceso ARMA estacional multiplicativo, con periodo estacional $s = 50$. Ahora utilizaremos el criterio BIC (con órdenes máximos de 10, para (p, q) y 3 para (P, Q)), con el fin de construir un modelo tentativo que podamos tomar como generador de la serie. El modelo propuesto por este criterio es un

$$ARMA(9,9) \times (2,0)_{50},$$

con los siguientes parámetros estimados para la parte regular:

	Parámetros estimados (parte regular)								
Parte AR	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9
coef.	-0.0941	0	-0.3504	0	0	0	0	0	0.4938
s.e.	0.0399	0	0.0686	0	0	0	0	0	0.0744
Parte MA	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
coef.	0	0	0.4143	0	0.1112	0	0.1930	0.1561	-0.5123
s.e.	0	0	0.0655	0	0.0392	0	0.0341	0.0390	0.0631

Tabla 4.2: Parámetros estimados del modelo $ARMA(9,9) \times (2,0)_{50}$ (parte regular)

y para la parte estacional:

Parámetros estimados (parte estacional)		
Parte estacional	Φ_1	Φ_2
coef.	0.1024	0.0965
s.e.	0.0461	0.0464

Tabla 4.3: Parámetros estimados del modelo $ARMA(9,9) \times (2,0)_{50}$ (parte estacional)

Posteriormente comprobaremos si las hipótesis básicas realizadas sobre el modelo se verifican, esto es, analizaremos si los residuos proceden de un proceso de ruido blanco gaussiano (etapa de diagnosis). Para ello empezaremos mostrando el gráfico de los residuos frente al tiempo, pues éste puede ayudarnos a detectar de manera visual y rápida la presencia de: tendencia, componente estacional o variabilidad no constante.

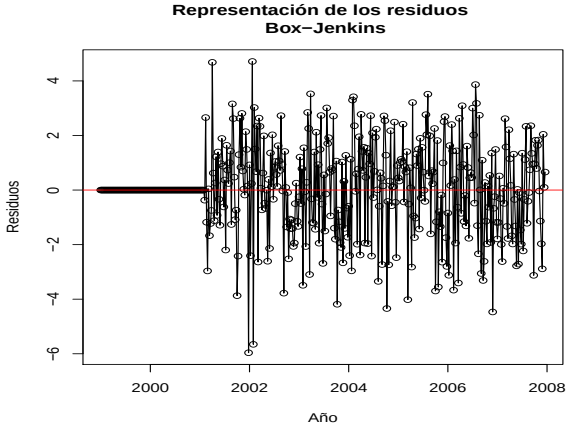

Figura 4.8: Gráfico secuencial de los residuos

Obsérvese que la figura (4.8) no muestra indicios para rechazar la hipótesis de ruido blanco.

El estudio de la hipótesis de independencia puede ser abordado utilizando el contraste de Ljung-Box.

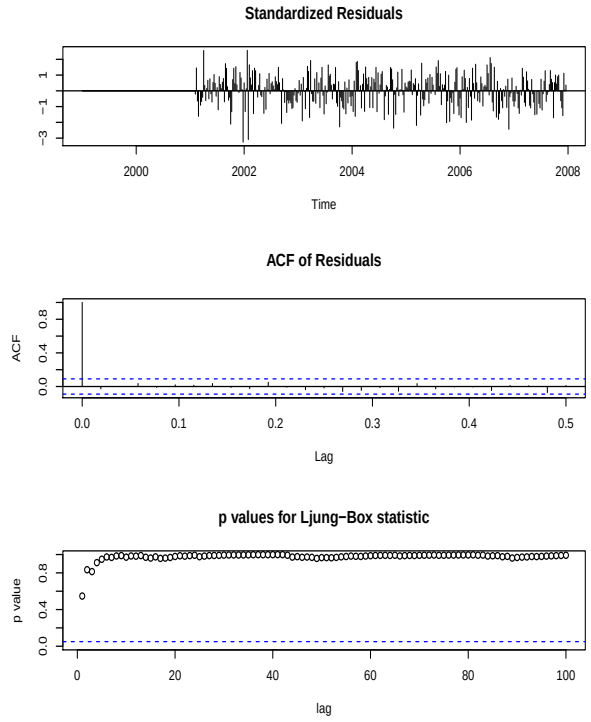

Figura 4.9: Contraste de Ljung-Box

Dado que los p-valores calculados son lo suficientemente grandes, no tenemos evidencias significativas para rechazar la hipótesis de independencia. Es decir, asumimos que los errores son independientes.

Para contrastar la hipótesis nula de que la media de los residuos es cero, utilizaremos el *t test* y obtenemos un p-valor de 0.741. Luego, podemos decir que no existen evidencias significativas para rechazar la hipótesis anterior.

Finalmente contrastaremos la normalidad de los residuos mediante el test de *Shapiro-Wilk*, obteniendo un p-valor menor que 0.01, y en consecuencia rechazamos la normalidad.

Luego, podemos afirmar que un modelo $ARMA(9,9)\times(2,0)_{50}$ sin constante y con innovaciones no gaussianas resulta adecuado como generador de la serie de los residuos, pues ha pasado con éxito las pruebas de diagnosis.

4.1.3. Predicciones del modelo semiparamétrico 1

Calculada la componente no paramétrica y paramétrica del modelo semiparamétrico 1 (4.3), podemos obtener sus predicciones. Para ello seguiremos un algoritmo análogo al dado en la página 21, repitiéndolo tantas veces como instantes queramos predecir (en este caso, deseamos obtener las predicciones del año 2008, por lo que $n - T = 52$ semanas). Una vez hecho esto, podemos comparar los valores obtenidos por el modelo semiparamétrico con los futuros valores observados de la serie (muestra de validación), transformados mediante el logaritmo.

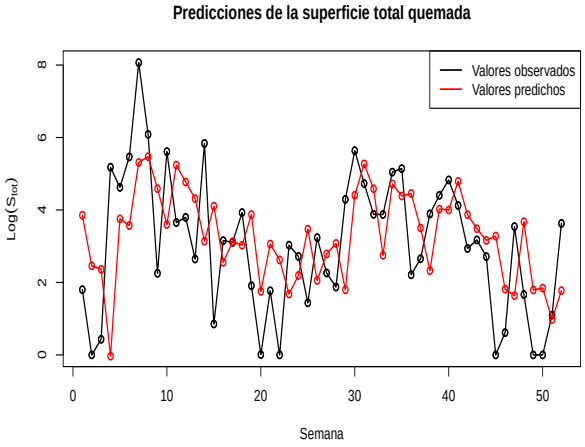

Figura 4.10: Predicciones proporcionadas por el modelo semiparamétrico 1 y futuros valores de la serie observados (para las variable transformada mediante el logaritmo)

Como se puede apreciar en la figura (4.10), las predicciones obtenidas por el modelo semiparamétrico 1 y los futuros valores de la serie son bastante similares. Además podemos calcular la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
5068.218	45.736	5333.057	25.336

Tabla 4.4: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto, para las variables transformadas

Además se han incorporado los intervalos de predicción bootstrap dados en (2.8), en donde, al igual que se hacía en los estudios de simulación, se considerará el método propuesto por Cao, Febrero-Bande, González-Manteiga, Prada Sánchez y García Jurado, como alternativa al método bootstrap de Thombs y Schucany.

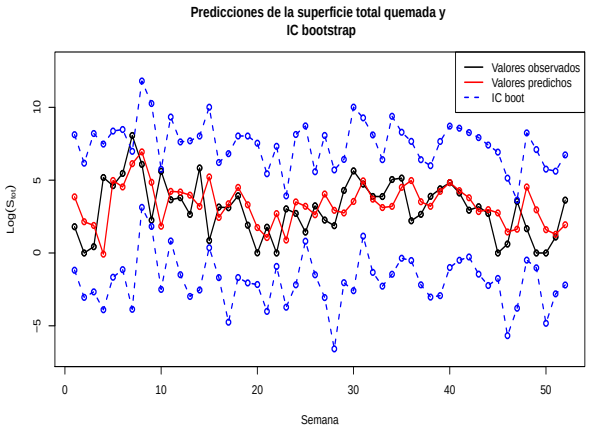

Figura 4.11: Intervalos de predicción bootstrap para el modelo semiparamétrico 1 (para las variables transformadas)

Finalmente, si deseamos obtener la predicción para la variable original (superficie total), debemos deshacer los cambios hechos anteriormente aplicando una exponencial. Hechos los cálculos anteriores obtenemos:

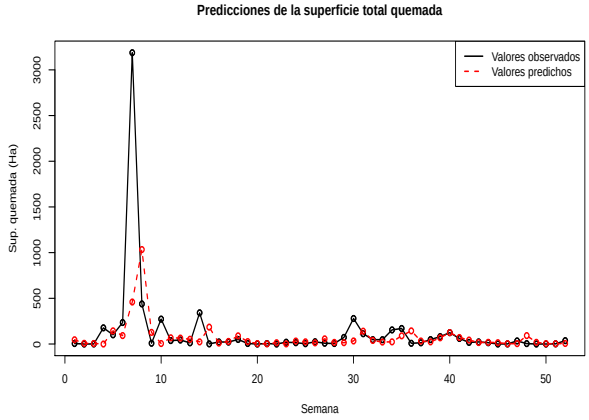

Figura 4.12: Predicciones proporcionadas por el modelo semiparamétrico 1 y futuros valores de la serie observados

Como se puede apreciar en la figura (4.12), el modelo semiparamétrico 1 proporciona unas aproximaciones razonablemente buenas de la superficie total quemada en los incendios forestales gallegos. Además obtenemos la siguiente tabla de errores:

Criterios de error			
ECM	EA	ERC	ERA
179283.7	113.688	4882.217	13.454

Tabla 4.5: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

4.2. Modelo semiparamétrico 2: Suavización posterior a la modelización ARMA

Consideremos la expresión matemática de este modelo,

$$\widehat{X}_{T+1}^{M} = \widehat{X}_{T+1}^{L} + \tilde{r}(\widehat{\varepsilon}_T), \tag{4.5}$$

donde la primera componente se estima mediante técnicas paramétricas (metodología Box-Jenkins), mientras que la segunda se lleva a cabo de forma no paramétrica.

4.2.1. Modelización ARMA de la serie

Empezaremos realizando la estimación paramétrica de la serie temporal:

$$(S_1, \ldots, S_T), \quad \text{con } T = 468.$$

Para ello, dado que la serie (4.1) presenta variabilidad no constante, hemos decidido aplicar una transformación logarítmica.

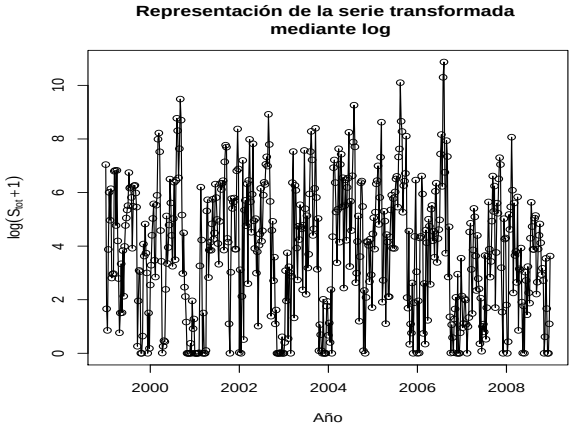

Figura 4.13: Gráfico secuencial de la serie temporal transformada mediante el logaritmo

Es decir, aplicaremos la metodología Box-Jenkins sobre

$$\{log(S_1 + 1), \ldots, log(S_T + 1)\}, \quad \text{con } T = 468.$$

Para la selección del modelo ARMA óptimo aplicaremos el criterio BIC, con órdenes máximos de 3 para p y q. El modelo óptimo según este criterio es un $AR(1)$, con los siguientes parámetros estimados:

Parámetros estimados		
AR(1)	ϕ_1	*intercept*
coef.	0.644	3.917
s.e.	0.035	0.247

Tabla 4.6: Parámetros del modelo AR(1) estimados por mínimos cuadrados

Este ajuste nos proporciona los residuos $(\varepsilon_1, \ldots, \varepsilon_T)$, que serán usados para calcular la componente no paramétrica del modelo semiparamétrico (4.5).

4.2.2. Estimación no paramétrica de la función de regresión

Dados los residuos obtenidos anteriormente $(\varepsilon_1, \ldots, \varepsilon_T)$, calcularemos su matriz histórica asociada,

$$\{(\varepsilon_t, \varepsilon_{t+1})\}, \quad \text{con} \quad t = 1, \ldots, T-1$$

y estimaremos la función de regresión de ε_{t+1} sobre ε_t ,

$$m(x) = \mathbb{E}\left[\varepsilon_{t+1} \mid \varepsilon_t = x\right], \tag{4.6}$$

mediante los cuatro estimadores no paramétricos de la función de regresión considerados (Nadaraya-Warson, lineal local, B-Splines y P-Splines).

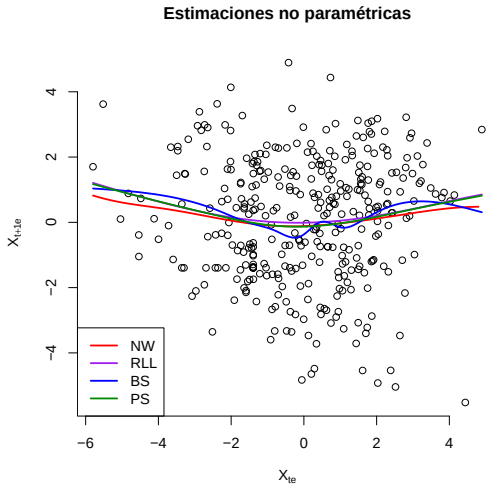

Figura 4.14: Estimaciones no paramétricas de la función de regresión

En los dos primeros casos utilizaremos como función kernel la densidad gaussiana y seleccionaremos el parámetro ventana usando el criterio de *validación cruzada*, mientras que en los dos últimos se seleccionarán los grados de libertad de acuerdo al

criterio GCV. Se elegirá como modelo no paramétrico óptimo aquél que nos proporcione un menor error cuadrático medio (sobre la matriz histórica de validación), tal y como se ha explicado en el algoritmo de la página 14.

La figura (4.14) muestra las cuatro estimaciones de la función de regresión consideradas. Obsérvese como todas la estimaciones son bastante similares, excepto la proporcionada por B-Splines que muestra un comportamiento menos suave.

Ejecutadas las $M = 1000$ iteraciones obtenemos:

Contadores			
NW	RLL	BS	PS
231	179	64	526

Tabla 4.7: Contadores del mejor estimador de la función de regresión en cada iteración

La tabla anterior (4.7) sugiere que el mejor estimador no paramétrico de la función de regresión es *P-Splines*, pues es el que mejores resultados proporciona en 526 de las $M = 1000$ iteraciones. Luego consideramos esta regresión como la mejor estimación no paramétrica de las consideradas y representamos los ECM de los cuatro estimadores:

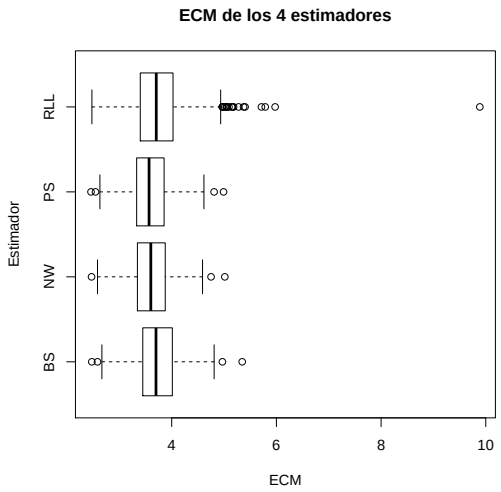

Figura 4.15: Boxplots de los ECM

La figura (4.15) nos muestra que los ECM obtenidos en la regresión lineal local incluyen más valores atípicos superiores que los demás estimadores, por lo que su uso no parece adecuado en este ejemplo. Obsérvese también como la distribución de los ECM con Nadaraya-Watson y P-Splines se comportan de modo similar.

4.2.3. Predicciones del modelo semiparamétrico 2

Una vez calculada la componente paramétrica y no paramétrica del modelo semiparamétrico 2 (4.5), podemos calcular las predicciones proporcionadas por el mismo. Para ello utilizaremos un algoritmo análogo al dado en la página 26, con las salvedades que ahora el modelo ARMA considerado es un $AR(1)$ y las estimaciones no paramétricas son obtenidas por P-Splines. Este algoritmo se repetirá $n - T = 52$ iteraciones, obteniendo así las predicciones que el modelo semiparamétrico 2 otorga al año 2008. Hechos los cálculos anteriores, podemos comparar las estimaciones con los futuros valores observados de la serie (muestra de validación), transformados mediante el logaritmo.

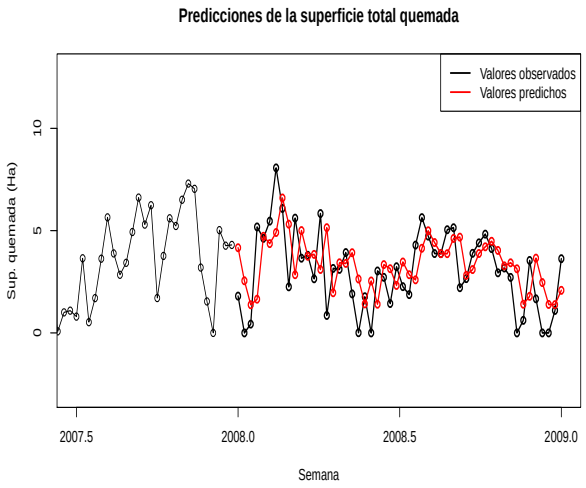

Figura 4.16: Predicciones proporcionadas por el modelo semiparamétrico 2 y futuros valores de la serie observados (para la variable transformada mediante el logaritmo)

Como se puede apreciar en la figura (4.16), las predicciones proporcionadas por el modelo semiparamétrico 2 y los futuros valores de la serie (transformados mediante el logaritmo) son bastante similares. Si calculamos los 4 tipos de errores considerados tenemos:

Criterios de error			
ECM	EA	ERC	ERA
2.998	1.388	1483.167	6.157

Tabla 4.8: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto, para las variables transformadas

Además, hemos construido los intervalos de predicción bootstrap para este modelo semiparamétrico de forma análoga al dado en (2.8). Los resultados obtenidos para la serie transformada son:

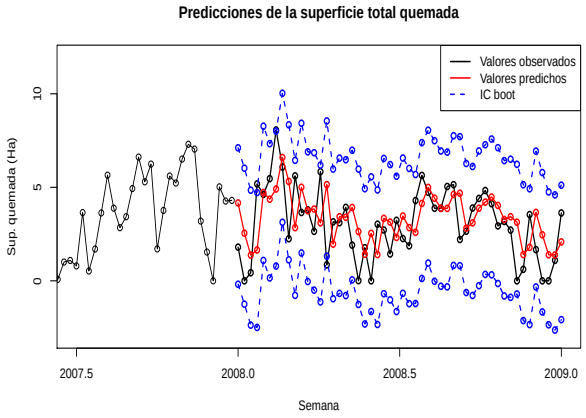

Figura 4.17: Intervalos de predicción bootstrap para el modelo semiparamétrico 2 (serie transformada)

Finalmente, si deseamos calcular la predicción para la serie original (superficie total), debemos deshacer los cambios anteriores aplicando una exponencial y obtenemos:

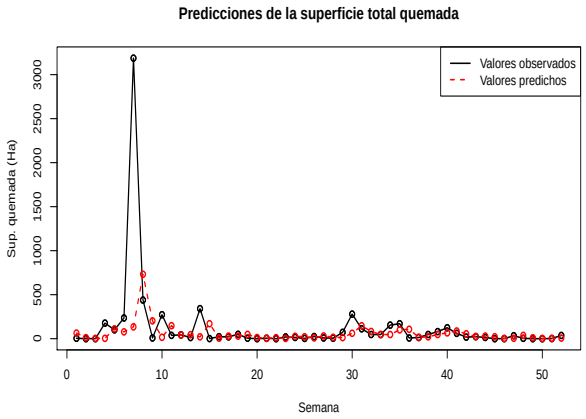

Figura 4.18: Predicciones del modelo semiparamétrico 2 y futuros valores de la serie observados

Como se puede apreciar en la figura (4.18), las predicciones proporcionadas por el modelo semiparamétrico 2 son razonablemente buenas. Los errores obtenidos son:

Criterios de error			
ECM	EA	ERC	ERA
188610.400	115.620	36040.950	32.468

Tabla 4.9: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto

4.3. Análisis comparativo

Ajustados los dos modelos semiparamétricos vistos en este trabajo, podemos realizar una análisis comparativo entre ambos con el fin de estudiar cuál de ellos proporciona mejores predicciones. Para ello utilizaremos los cuatro tipos de errores considerados, y además, tomaremos como modelo referencia el puramente paramétrico (Box-Jenkins), que ha resultado ser un $ARIMA(3,0,0) \times (2,1,3)_{52}$.

Una apreciación visual de las tres estimaciones puede contemplarse en la siguiente gráfica:

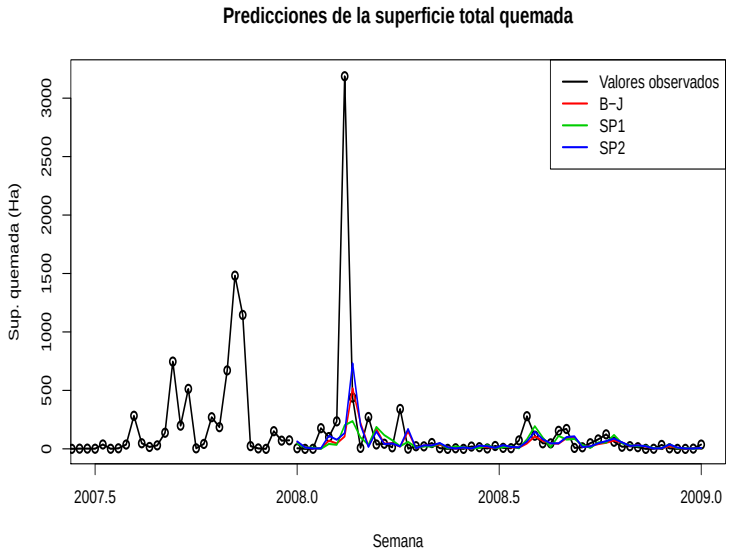

Figura 4.19: Predicciones obtenidas por Box-Jenkins y los dos modelos semiparamétricos

Finalmente, se muestra una tabla comparativa de los tres modelos de predicción, en donde se reflejan los diversos errores considerados.

	Criterios de error			
Modelo	ECM	EA	ERC	ERA
BJ	190913.60	111.94	32120.26	30.40
SP1	179283.70	113.68	4882.21	13.45
SP2	188610.40	115.62	36040.95	32.47

Tabla 4.10: Tabla de errores: Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto (para las tres predicciones)

La tabla anterior (4.10) sugiere que el modelo semiparamétrico 1 es el que proporciona mejores resultados, pues es el que aporta menores errores (excepto en el Error Absoluto donde el modelo puramente paramétrico proporciona mejores resultados).

Capítulo 5

Conclusiones

Los incendios forestales son una de las mayores amenazas ambientales, económicas y sociales de muchas zonas del planeta. En el Sur de Europa se han convertido en el principal problema para las autoridades ambientales. En el caso de Galicia presentan una especial incidencia en determinadas zonas de la comunidad autónoma y que debido a su origen provocado suponen un desafío para el futuro de la región.

El desarrollo de metodologías de cualquier tipo, pero en especial de aquellas contrastadas por la evidencia de los datos, permitirá una más eficaz organización y planificación de la lucha contra el fuego, lo que redundará en menor número de superficie quemada y menor riesgo para vidas y bienes.

En este trabajo se ha realizado una revisión de dos modelos de predicción semiparamétricos para series temporales, los cuales descomponen la predicción en dos componentes. Una de ellas es estimada mediante técnicas de regresión no paramétrica, mientras que la otra se lleva a cabo con modelos Box-Jenkins.

El primer modelo, (García Jurado et al., 1995), descompone la predicción en una componente no paramétrica que estima la tendencia seguida de una estimación Box-Jenkins de los residuos, mientras que el segundo ,(Dabo Niang et al., 2010), está basado en la estimación de la representación ARMA, seguida por una regresión no paramétrica para capturar la parte no lineal.

El comportamiento de estos sistemas de predicción ha resultado ser competitivo en comparación con otros modelos de predicción en series temporales, tales como los modelos no paramétricos o la metodología Box-Jenkins. Esto puede observarse en los diferentes estudios de simulación del capítulo 3, o en la aplicación a datos reales del capítulo 4.

Finalmente, hemos considerado también los intervalos de predicción bootstrap para el modelo semiparamétrico 1 y se han propuesto, de forma análoga, otros para el modelo semiparamétrico 2. Éstos han mostrado un comportamiento competitivo respecto a los intervalos de predicción clásicos, con la ventaja añadida de su generalidad, ya que

no precisan de fuertes hipótesis realizadas sobre los residuos.

Apéndice A

Base de datos

La base de datos [1] analizada en este trabajo recoge los incendios registrados en Galicia, en el periodo que va desde el año 1999 al 2008. El total de incendios registrados en estos 10 años es de 85134, a los cuáles se le han registrado 158 variables, clasificadas en: localización, fechas, inicio y causas, terreno, tipo, medios, partes, consecuencias, clima e índices de riesgo, y que pasaremos a comentar a continuación.

A.1. Variables de localización

Hacen referencia al lugar donde ha ocurrido el incendio, conteniendo las coordenadas geográficas del incendio, provincia, distrito, ayuntamiento o parroquia (entre otros). Así, por ejemplo la distribución de los incendios por provincias resulta.

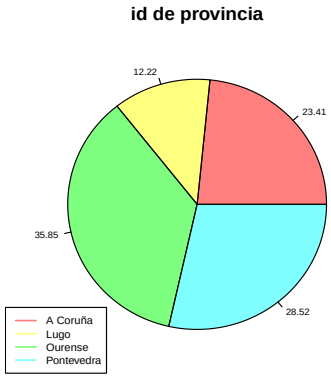

Figura A.1: id de provincia

[1] Fuente de información: Ministerio de Agricultura Alimentación y Medio Ambiente

La figura (A.1) muestra que *Ourense* es la provincia gallega con mayor frecuencia de ignición ya que el 35.85 % de los incendios tienen lugar ahí. Seguidamente *Pontevedra*, *A Coruña* y *Lugo* con porcentajes de ignición de 28.52 %, 23.41 % y 12.22 %, respectivamente.

Otra división atractiva desde el punto de vista forestal es el de los distritos. De esta forma obtenemos una división de la comunidad en 19 distritos.

Figura A.2: División territorial por distritos

Analizando la distribución de los incendios forestales de esta forma obtenemos que los distritos con mayor porcentaje de ignición son: Valdeorras-Trives (8.49 %), O condado-A Paradanta (8.48 %), Verín-Viana (8.47 %), Caldas-O Salnés (7.985 %), A Limia (7.48 %) y Miño-Arnoia (6.97 %). Estos resultados corroboran la tendencia provincial comentada anteriormente, observando como gran parte de los distritos mencionados pertenecen a *Ourense*. Sin embargo la tendencia de ignición está cambiando en los últimos años, aumentando el número de incendios en las zonas costeras, especialmente en épocas veraniegas. Debido a este cambio, tenemos que el distrito de *Caldas-O Salnés* se sitúa entre los más castigados en cuanto a número de incendios.

A.2. variables de fechas

Las variables del conjunto *fechas* hacen referencia a registros temporales de variables relacionadas con los incendios, como *fecha* o *hora* del inicio y fin de los incendios, así como del control. Además también están registradas las mediciones de los instantes de tiempo donde se producen las intervenciones de los diversos medios de extinción, tanto aéreos como terrestres. Estas variables han servido para calcular otras de duración del incendio, o duración desde que se produce el incendio hasta que es controlado, y que tienen su importancia pues miden de alguna forma la eficacia con la que actúan los medios de extinción.

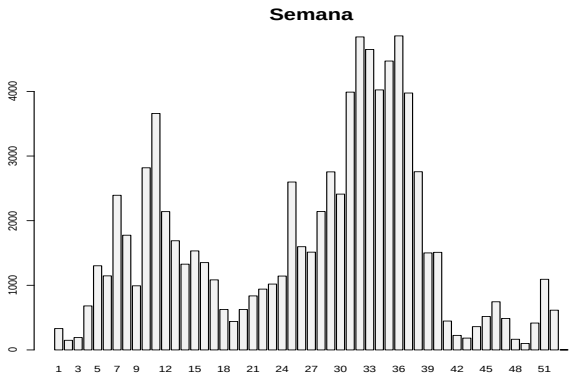

Figura A.3: Distribución de los incendios por semana

La figura (A.3) muestra la frecuencia de incendios según la semana del año. Posee un aspecto bimodal, con una moda sobre la semana 10 (Marzo) y otra más acentuada en Agosto-Septiembre. La primera se debe en gran parte a las quemas de rastrojos en el inicio de la primavera mientras que la segunda se debe fundamentalmente a acciones humanas, intencionadas o negligentes en condiciones de altas temperaturas y sequedad propias del periodo estival.

A.3. variables de inicio y causas

El conjunto de variables de *inicio y causas* hacen referencia a la ignición, esto es, el comienzo del incendio. Aquí, se cuenta con variables categóricas sobre el lugar del incendio, quién lo ha detectado y cómo se ha producido. Además, en el caso de que el incendio fuese intencionado, se refleja cuál ha sido la motivación para provocarlo.

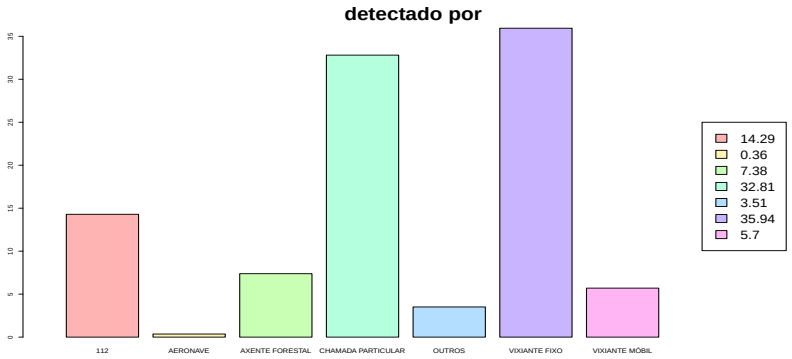

Figura A.4: Porcentajes de la detección

La figura (A.4) revela como en la mayoría de los casos, los incendios son detectados gracias a la colaboración de viajantes fijos (35.94 %), llamadas de particulares

(32.81 %) y en menor medida, por el 112 (14.29 %).

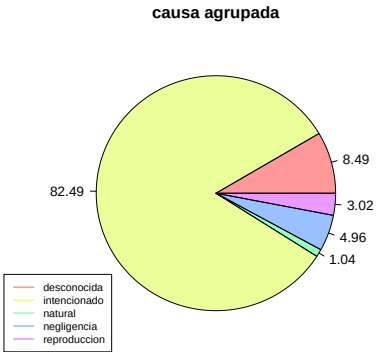

Figura A.5: Distribución de los incendios su causa

La figura (A.5) muestra la triste realidad en relación a los incendios gallegos, y es que el 82.49 % de los incendios que se producen en esta comunidad son intencionados. Otras causas importantes son negligencia (4.96 %), reproducción (3.02 %) y natural (1.04 %). Téngase en cuenta también que la segunda categoría con más importancia es la de *desconocido* (8.49 %) y aunque no se sabe a ciencia cierta qué es lo que ha provocado estos incendios, pensamos que gran parte de este porcentaje pueda deberse a actos intencionados.

Una vez estudiado cuál es la causa de los incendios, un análisis importante a posteriori sería ver qué ha motivado a estos individuos para que llegasen a plantar fuego. La tabla (A.1) nos muestra las principales motivaciones en los incendios intencionados.

Estudio de la motivación	
Motivacion	**Porcentaje**
Agricultures eliminar matogueira	44.46
Animais	1.48
Baixar prezo madeira	0.10
Cazadores	2.16
Gandeiros nacemento pasto	16.10
Obter modificación uso do solo	0.67
Outras motivacións	20.04
Pirómanos (enfermos mentais)	12.09
Vandalismo	1.14
Vinganzas	0.74

Tabla A.1: Incendios intencionados. Motivación

A.4. variables de terreno

Estas variables se corresponden con el tipo de terreno sobre el que ha tenido lugar el incendio, es decir, si es matoguera, bosque, pasteros ...

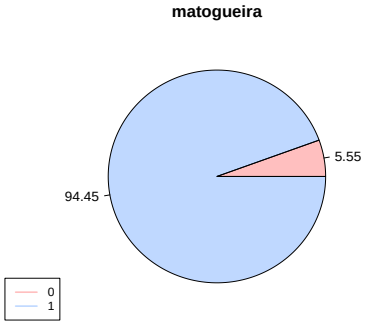

Figura A.6: Distribución de los incendios según el terreno

La figura anterior (A.6) nos muestra que en el 95.45 % de los incendios se quema matoguera.

A.5. variables de tipo

Las variables de *tipo* tienen en cuenta por donde se propaga el incendio. Existen tres posibilidades.

i) **Fuego de subsuelo:** El fuego se propaga por la materia orgánica en descomposición y las raíces. Casi siempre se queman despacio y en combustión incandescente (poca o ausencia de llama) al no disponer de suficiente oxígeno.

ii) **Fuego de superficie:** El incendio se propaga por el combustible que encuentra sobre el suelo, incluye la hojarasca, hierbas, arbustos y madera caída pero no inmersa en la hojarasca en descomposición.

iii) **Fuego de copas:** Existen 3 tipos

- *Antorcheo:* Paso de fuego de superficie a fuego de copas, pero solo de forma puntual en algunos pies.
- *Copas pasivo:* Es el fuego que avanza por las copas de los árboles, acoplado y dependiente de un fuego de superficie, si se extingue este se detiene el de copas.

- *Copas activo:* Es el fuego que avanza por las coronas de los árboles independientemente de la superficie. Solo se puede atacar de forma indirecta y suele necesitar un viento mayor de 30 km/h y proximidad de copas (alta densidad aparente de copas y largas copas).

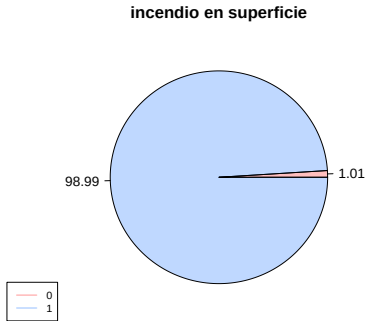

Figura A.7: Distribución de los incendios según el tipo

La figura (A.7) sugiere que en el 98.99 %, la propagación del fuego se produce a través de la superficie.

A.6. variables de medios

Este tipo de variables involucran todo tipo de medios que han trabajado en la extinción del incendio, tanto sean humanos como maquinaria, terrestres o aéreos. Estas variables miden todo el despliegue que se produce ante un incendio y permite, de alguna forma, evaluar la gestión de la extinción de los incendios.

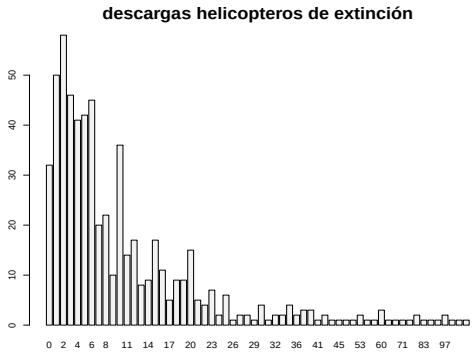

Figura A.8: Distribución de las descargas de los helicópteros de extinción

Como se puede observar, la figura (A.8) muestra como se distribuye el número de descargas efectuadas por los helicópteros de extinción. Apréciese que el número habitual de descargas oscila entre 2 y 7.

A.7. variables de partes

Las variables de *partes* informan cuál ha sido la forma escogida para atajar el fuego. En la mayoría de los casos se opta por actitudes directas que consisten en intentar neutralizar el fuego utilizando tanto medios terrestres como aéreos. El agua toma un papel importante en este modo de actuación.

Existen otras formas de extinción donde se adopta una actitud indirecta ante el fuego. Aquí podemos mencionar por ejemplo, la apertura de *cortafuegos* que consiste en abrir un espacio de terreno que no posea ningún tipo de combustible, y así los incendios forestales no se pueden esparcir. De esta forma no actuamos directamente sobre el fuego pero estas actividades pueden ayudar a apagarlo. Otra posibilidad sería provocar un *contrafuego*, es decir el fuego promovido voluntariamente, apoyado en una línea de defensa suficientemente segura, que se propaga en dirección contraria al avance natural del incendio que se combate. El propósito es hacer que el contrafuego "choque" con el incendio, lo cuál impedirá momentáneamente el aporte de oxígeno, debilitando al incendio.

El análisis realizado sobre la base de datos muestra que aproximadamente en el 99.6 % de los incendios producidos se adoptan actitudes de ataque directo sobre el fuego.

A.8. variables de consecuencias

Los incendios forestales ocasionan consecuencias en el paisaje, precisando del transcurso de décadas para paliar los efectos que estos sucesos ocasionan en el medio ambiente. Estas consecuencias producidas por el fuego aparecen reflejadas en las diferentes variables de este conjunto. Además a las consecuencias materiales hay que sumar, en ocasiones heridos o muertes, fundamentalmente personas del servicio de extinción.

También se contemplan los daños ocasionados en infraestructuras (carreteras, telefonía, ferrocarril, viviendas, ...) y mediciones de la superficie quemada en función del tipo de terreno. Estas últimas variables son de gran importancia y permiten desarrollar el objetivo fundamental de este trabajo, que no es otro que el de realizar un análisis temporal exhaustivo sobre la superficie total quemada y obtener predicciones

en instantes de tiempo futuros [2].

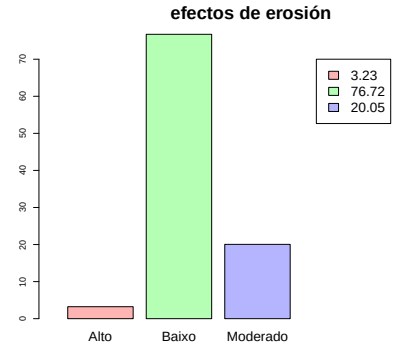

Figura A.9: Porcentajes de los efectos de erosión

La figura (A.9) muestra que en el 76.72 % de los casos, el efecto erosivo producido por los incendios es bajo, moderado en el 20.05 % y alto en el 3.23 % restante.

A.9. variables de clima

Las variables climatológicas juegan un papel importante en el origen y transcurso de los incendios. Por este motivo hay que realizar un estudio riguroso sobre ellas, ya que condicionan el comportamiento y duración del mismo. Estas variables conciernen altitud, temperatura, humedad, velocidad y dirección del viento, ...

A continuación (véase figura A.10) se muestra la distribución de la temperatura corregida [3]. Como se puede apreciar, ésta presenta un comportamiento ligeramente asimétrico hacia la derecha, con moda en torno a 25 °C.

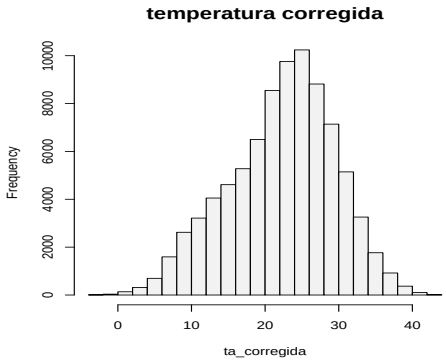

Figura A.10: Distribución de la temperatura corregida

[2] Véase capítulo 4

[3] Temperatura asociada al incendio, medida a través de las estaciones de Meteogalicia y corregida mediante un factor de altitud

A.10. variables de índices de riesgo

En la literatura forestal existen cantidad de índices de riesgo desarrollados en diferentes países. Ahora bien, las causas del origen de los incendios es diferente de un país a otro, y por consiguiente un buen índice tiene que incorporar estos factores diferenciadores.

En la base de datos aparecen implementados varios índices de riesgo pero muchos de ellos incorporan únicamente factores climáticos por lo que su resultado no es del todo bueno, pues como se ha estudiado anteriormente (véase A.5), en Galicia el 82.48 % de los incendios son intencionados y en consecuencia no parece adecuado considerar índices de riesgo en los que sólo intervengan variables naturales.

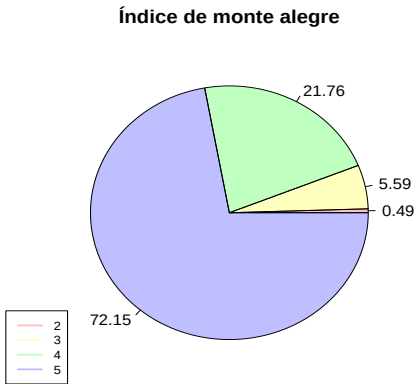

Figura A.11: Índice de Monte Alegre

La figura (A.11) muestra el índice de *Monte Alegre*, que resultó ser el que mejores resultados ha proporcionado de todos los implementados. Obsérvese que dicho índice proporciona el valor 5 (riesgo muy alto), en el 72.15 % de los incendios y el valor 4 (riesgo alto), en el 21.76 %. El gráfico sugiere entonces que este índice proporciona buenos resultados (riesgo alto o muy alto) condicionado a instantes en los que se ha producido incendios. Además se ha comprobado el funcionamiento de este índice en los días sin incendios y los resultados obtenidos han sido bastante buenos.

A.11. Algunos resultados exploratorios relevantes

En el estudio de los incendios forestales se pueden analizar múltiples características, si bien, algunas de las más importantes son: el número de eventos (véase 1.1) y la superficie quemada.

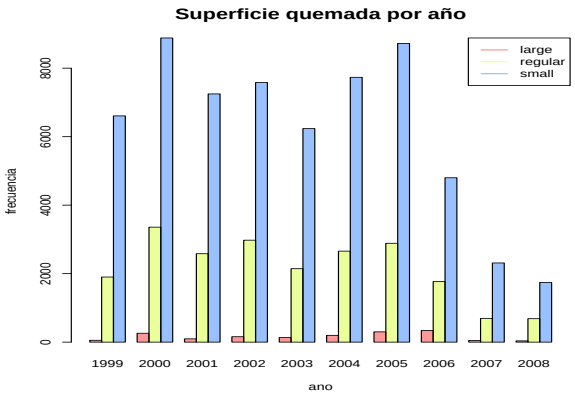

Figura A.12: Superficie quemada. El convenio utilizado es el siguiente: *small* si el valor de la superficie total quemada (S) verifica $S < 1$ Ha., *regular* si $1 \leq S < 25$ Ha. y *large* si $S \geq 25$ Ha.

La figura (A.12) proporciona información acerca de la tipología de los incendios gallegos; y es que esta comunidad se caracteriza por sufrir un gran número de incendios pero de poca extensión, pues la mayoría de los incendios han sido catalogados como *pequeños*.

Otro aspecto a tener muy en cuenta sería el lugar de inicio del fuego ya que esto puede informarnos acerca de cuál es la causa que lo ha llegado a provocar.

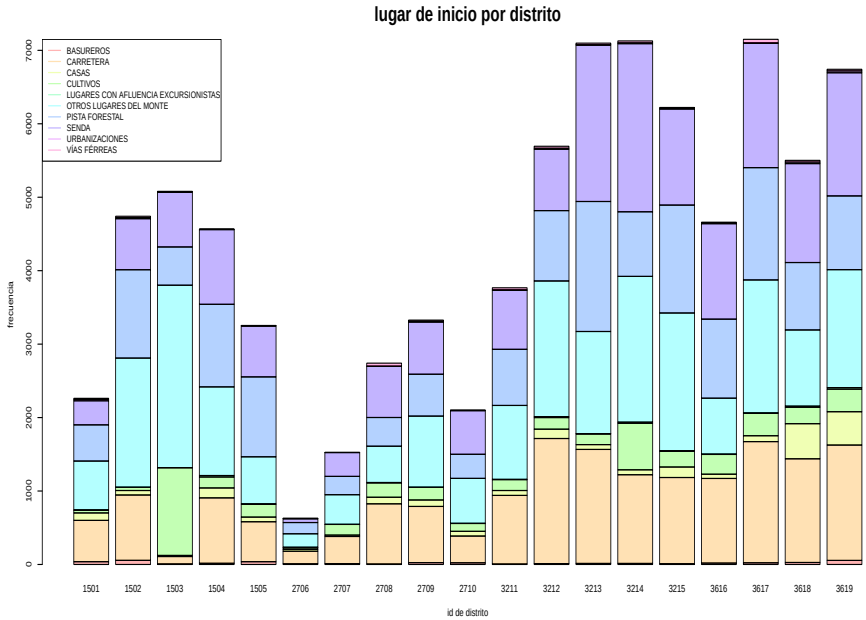

Figura A.13: Lugar de inicio por distritos

La figura (A.13) muestra dónde tiene lugar el inicio del fuego por distritos. Obsérvese que la información suministrada en el gráfico nos permite pensar en la intencionalidad

de los incendios pues los lugares más habituales de inicio son junto a carreteras, pistas forestales y sendas. Veamos como en el distrito de *Santiago-Meseta interior* (1503) existe una mayor proporción de incendios iniciados junto a cultivos y esto puede deberse a quemas de rastrojos en las actividades agrícolas. Por otra parte, en los distritos pontevedreses de *Vigo-Baixo Miño* y *Caldas-O Salnés* (códigos 3618 y 3619, respectivamente) se observa una mayor proporción de incendios iniciados cerca de las casas.

Finalmente, nos podemos preguntar dado el carácter intencionado de los incendios, si existe alguna diferencia motivacional de unos distritos a otros.

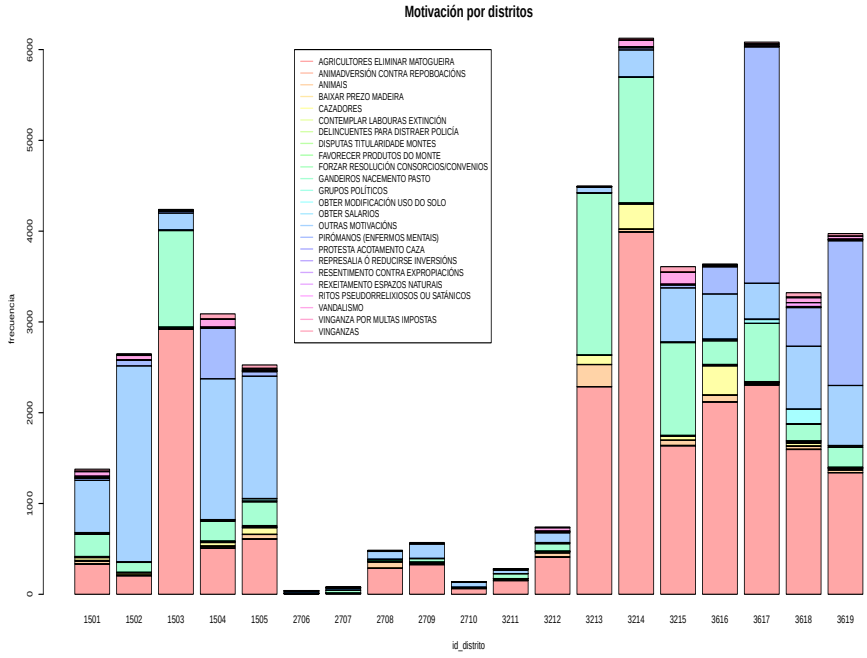

Figura A.14: Lugar de inicio por distritos

En la figura (A.14) observamos como la motivación principal de los incendios intencionados es la eliminación de las matogueras por parte de los agricultores. Un aspecto a resaltar en este gráfico es la baja proporción de incendios intencionados en la provincia luguesa (códigos 2706-2710). Véase también el alto grado de incendios provocados por pirómanos (enfermos mentales), especialmente en O Condado-A Paradanta (3617) y Caldas-O Salnés (3619).

Apéndice B

Los modelos Box-Jenkins

Consideremos un proceso estocástico, esto es, un conjunto de variables aleatorias definidas sobre el mismo espacio de probabilidad, en tiempo discreto y espacios de estados continuo, $\{X_t\}_{t\in\mathbb{Z}}$, del que hemos observado parte de su trayectoria, es decir una muestra de datos dependientes,

$$x_1, \ldots, x_n. \tag{B.1}$$

El subíndice t de cada variable aleatoria representa el instante de tiempo en que es observada. El objetivo principal consistirá en construir un proceso estocástico sencillo que de manera razonable haya podido generar la serie (B.1). Una vez construido el modelo podremos comprender la dinámica de la serie de tiempo y obtener predicciones en valores futuros del tiempo.

B.1. Procesos autorregresivos (AR)

Se define el *proceso autorregresivo de orden p*, $AR(p)$, como el *proceso estacionario* [1] $\{X_t\}_{t\in\mathbb{Z}}$, que admite la siguiente representación:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t, \tag{B.2}$$

donde $c, \phi_1, \ldots, \phi_p$ son constantes y $\{a_t\}_t$ es ruido blanco (innovaciones).

El proceso autorregresivo de orden p, $AR(p)$, explica el valor actual (X_t) a través de una función lineal de p valores pasados $(X_{t-1}, \ldots, X_{t-p})$.

[1] Un proceso estacionario es aquél que verifica:

i) $\mu_t = \mu$, $\forall t$.

ii) $\sigma_t^2 = \sigma^2$, $\forall t$.

iii) $\gamma(t, t+k) = \gamma_k$, $\forall t, k$,

siendo $\gamma(t, t+k)$ la función de autocovarianzas, es decir, $\gamma(s,t) = Cov(X_s, X_t) = \mathbb{E}\left[(X_s - \mu_s)(X_t - \mu_t))\right]$.

B.2. Procesos de medias móviles (MA)

Se define el *proceso de medias móviles de orden q*, $MA(q)$, como aquél proceso $\{X_t\}_{t\in\mathbb{Z}}$ que admite la siguiente representación:

$$X_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}, \tag{B.3}$$

donde $c, \theta_1, \ldots, \theta_q$ son constantes y $\{a_t\}_t$ es ruido blanco (innovaciones).

El proceso de medias móviles de orden q, $MA(q)$, explica el valor actual (X_t) como una función lineal de q valores pasados de un proceso de ruido blanco $(a_{t-1}, \ldots, a_{t-q})$.

B.3. Procesos ARMA

La incorporación en un mismo proceso de una parte autorregresiva y otra de medias móviles da lugar a lo que se conoce como los modelos ARMA. Entonces se definen los *procesos ARMA(p,q)* como aquellos procesos estacionarios, $\{X_t\}_{t\in\mathbb{Z}}$, que admiten la representación:

$$\begin{aligned} X_t = \; & c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} \\ & + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} \; + \ldots + \theta_q a_{t-q}, \end{aligned} \tag{B.4}$$

donde $c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ son constantes y $\{a_t\}_t$ es ruido blanco (innovaciones).

Los procesos $ARMA$ generales explican el presente a través de una función lineal de p observaciones y/o q innovaciones consecutivas ocurridas en el pasado inmediato (*dependencia regular*).

Observación B.3.1 Es intuitivamente claro que:

i) ARMA(p,0) $\Leftrightarrow$ AR(p).

ii) ARMA(0,q) $\Leftrightarrow$ MA(q).

La ecuación que define el proceso ARMA (B.4) puede escribirse de forma compacta mediante la expresión:

$$\phi(B) X_t = c + \theta(B) a_t, \tag{B.5}$$

donde

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \\ \theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q, \end{aligned}$$

con B denotando el operador retardo, definido por:

$$B X_t = X_{t-1}.$$

La clase de procesos $ARMA$ que se acaban de presentar es una familia de procesos estacionarios muy flexible, pues permiten modelizar una gran variedad de series temporales generadas por procesos estacionarios.

B.3.1. Procesos ARMA estacionales

En el caso particular en que sean nulos los coeficientes ϕ_i y θ_j con subíndice no múltiplo de s, tenemos procesos que explican el presente a través de una función lineal de observaciones e/o innovaciones ocurridas en instantes múltiplos del periodo estacional s (*dependencia estacional*).

Podemos definir entonces los procesos *ARMA estacionales*, que denotaremos por $ARMA(P,Q)_s$, como aquellos procesos estacionarios, $\{X_t\}_{t\in\mathbb{Z}}$, que admiten la siguiente representación:

$$\begin{aligned} X_t = c \ &+ \Phi_1 X_{t-s} + \Phi_2 X_{t-2s} + \ldots + \Phi_P X_{t-Ps} \\ &+ a_t + \Theta_1 a_{t-s} \ + \Theta_2 a_{t-2s} + \ldots + \Theta_Q a_{t-Qs}, \end{aligned} \tag{B.6}$$

donde $c, \Phi_1, \ldots, \Phi_P, \Theta_1, \ldots \Theta_Q$ son constantes y $\{a_t\}_t$ es ruido blanco (innovaciones).

Observación B.3.2 Es claro observar que se verifican las siguientes relaciones:

i) $ARMA(P,0)_s \Leftrightarrow AR(P)_s$.

ii) $ARMA(0,Q)_s \Leftrightarrow MA(Q)_s$.

La ecuación que define el proceso ARMA (B.6) puede escribirse de forma compacta mediante la expresión:

$$\Phi(B^s)X_t = c + \Theta(B^s)a_t, \tag{B.7}$$

donde

$$\begin{aligned} \Phi(B) &= 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_p B^{Ps}, \\ \Theta(B) &= 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_q B^{Qs}, \end{aligned}$$

con B denotando el operador retardo estacional, definido por

$$B^s X_t = X_{t-s}.$$

B.3.2. Procesos ARMA estacionales multiplicativos

Aquellos procesos que modelizan conjuntamente la dependencia regular y estacional, se conocen como *procesos ARMA estacionales multiplicativos*. Éstos admiten la siguiente expresión:

$$\phi(B)\Phi(B^s)X_t = c + \theta(B)\Theta(B^s)a_t. \tag{B.8}$$

Denotaremos estos modelos por $ARMA(p,q)\times(P,Q)_s$ y son, en particular, procesos $ARMA(p+sP, q+sQ)$ con muchos coeficientes nulos.

Esta clase de procesos ARMA (que incluye a los estacionales y a los estacionales multiplicativos) que acabamos de presentar es una familia muy flexible de procesos estacionarios.

Apéndice C

Regresión no paramétrica

Los modelos de regresión permiten establecer la relación entre la covariable X y la respuesta Y de acuerdo a la relación matemática,

$$y = m(x) + \varepsilon, \tag{C.1}$$

donde $m(.)$ denota una función desconocida que representa el valor medio de Y, condicionado a los valores de X y ε un término de error,

$$m(x) = \mathbb{E}\left[Y|X = x\right]. \tag{C.2}$$

Los modelos de regresión paramétricos establecen una forma paramétrica para m y son de gran utilidad ya que toda la información se resume en los parámetros del modelo y la interpretación suele ser muy sencilla. Sin embargo, si la parametrización elegida es demasiado rígida, y no ajusta bien a los datos, entonces las conclusiones obtenidas pueden ser erróneas.

Ante este tipo de situaciones es necesario el desarrollo y aplicación de modelos más generales y flexibles que permitan una correcta modelización matemática. En los últimos años ha surgido una linea de investigación en el campo de la estadística funcional no paramétrica que ha permitido el desarrollo y aplicación de modelos más generales.

En los modelos de regresión no paramétrica no se supone ningún modelo conocido sobre la función de regresión, m. A lo sumo se pide alguna condición para la misma como diferenciabilidad. La estimación no paramétrica de m se obtiene mediante técnicas de suavizado aplicadas localmente a los pares de observaciones

$$(x_i, y_i), \quad i = 1, 2, \ldots, n.$$

El procedimiento es similar al utilizado en la estimación de funciones de densidad. El valor medio condicional para un intervalo pequeño de x se estima, no sólo con las observaciones de dicho intervalo, sino con las de intervalos adyacentes. Esta información se pondera en forma decreciente a medida que es mayor la distancia de la observación respecto al centro del intervalo.

Entre los posibles métodos de suavización, podemos destacar:

- Suavizadores kernel (Nadaraya, 1964; Watson,1964).
- Suavizadores polinómico locales.
- Cubic smoothing splines (Wahba, 1990; Hastie e Tibshirani, 1990).
- Splines de regresión penalizados (Ruppert et al., 2003; Wood, 2006).

C.1. Regresión tipo núcleo

El estimador tipo núcleo de *Nadaraya-Watson* de la función de regresión m (C.2) viene dado por:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}, \tag{C.3}$$

donde K es una *función kernel* (normalmente una densidad simétrica en torno al cero), h es un parámetro de suavizado, llamado *ventana* que regula el tamaño del entorno que se usa para llevar a cabo la estimación y n es el tamaño muestral.

El estimador tipo núcleo definido anteriormente puede reescribirse como:

$$\widehat{m}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} Y_i = \sum_{i=1}^{n} W_{hi}(x) Y_i,$$

donde

$$W_{hi}(x) = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)},$$

y en consecuencia el estimador tipo núcleo de la función de regresión es una media (local) ponderada de los valores observados de la variable Y, donde

$$\sum_{i=1}^{n} W_{hi}(x) = 1.$$

Una posibilidad para la elección del parámetro de suavizado sería usar el método de validación cruzada. Para medir la bondad de ajuste que se consigue con la ventana

h se podría usar el error medio,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{m}(x_i)\right)^2.$$

Esta medida de error global aproximaría el error de predicción. Sin embargo, la aproximación sería un tanto optimista ya que estaríamos usando el valor Y_i dos veces: una a la hora de calcular el error, y otra a la hora de construir el estimador. Para evaluar mejor el error de predicción se suele usar el mismo criterio de error, pero eliminando el dato i-ésimo cuando calculamos el error de predicción para Y_i.

Así la función de validación cruzada se define como:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{m}_{-(i)}(X_i)\right)^2, \tag{C.4}$$

donde $\widehat{m}_{-(i)}$ denota el estimador de Nadaraya-Watson construido a partir de la muestra original después de eliminar el par (x_i, y_i). La idea sería tomar aquel h que haga que $CV(h)$ sea mínimo.

Teorema C.1.1 La función de validación cruzada del estimador de Nadaraya-Watson se puede escribir de la siguiente forma:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{m}(x_i)}{1 - H_{ii}}\right)^2, \tag{C.5}$$

donde H_{ii} es el elemento i-ésimo de la diagonal de la matriz de suavizado H necesaria para calcular el estimador en los puntos $(x_1, \ldots, x_n)$. Es decir,

$$H_{ii} = \frac{K_h(0)}{\sum_{k=1}^{n} K_h(x_i - x_k)}, \tag{C.6}$$

donde $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$.

C.2. Regresión lineal local

El objetivo general de la regresión polinomial local es aproximar localmente los datos por un polinomio de grado p determinado . Aquí consideraremos el caso de la regresión lineal local y en consecuencia el polinomio considerado será de grado $p = 1$. Esto es, dado $h > 0$ podemos proponer un modelo lineal válido sólo en el entorno $(x - h, x + h)$.

La idea subyacente es que una función continua puede aproximarse razonablemente bien por un polinomio de grado bajo. La aproximación lineal puede expresarse mediante:

$$m(x_i) \approx \alpha(x) + \beta(x)x_i, \quad x_i \in (x - h, x + h). \tag{C.7}$$

Se ajustaría entonces por mínimos cuadrados los parámetros del modelo usando sólo los datos del entorno $(x-h, x+h)$. Al igual que sucede en el caso anterior (C.1), no parece del todo razonable que todas las observaciones tengan los mismos pesos en el intervalo $(x-h, x+h)$ independientemente de su proximidad a x. Luego ajustaríamos los parámetros por mínimos cuadrados ponderados, es decir:

$$\sum_{i=1}^{n} \left(y_i - \alpha(x) - \beta(x)x_i\right)^2 K_h(x - x_i). \tag{C.8}$$

En consecuencia, el estimador lineal local en el punto x vendrá dado por,

$$\widehat{m}_{LL}(x) = a(x) + b(x)x, \tag{C.9}$$

donde $a(x), b(x)$ denotan los valores que minimizan la suma de cuadrados ponderada (C.8).

C.3. Regresión B-Splines

En este contexto, la función parcial m tiene la siguiente estructura:

$$m(x) = a_1 B_1(x) + \ldots + a_K B_K(x), \tag{C.10}$$

donde

- K indica el número bases.
- $a_1, \ldots, a_K$ son parámetros desconocidos.
- $B_1, \ldots, B_K$ son funciones conocidas que dependen únicamente de la posición de los llamados nodos.

Por lo tanto, en la regresión *spline* se reduce un problema de regresión no paramétrica a un problema paramétrico. Únicamente será necesario estimar los coeficientes $a_1, \ldots, a_K$ ajustando un modelo de regresión lineal.

Un *spline* es un conjunto de polinomios (de orden m) definidos en subintervalos, construidos de tal forma que deben unirse "suavemente" en cada uno de los nodos interiores, siendo éstos los puntos de corte de los subintervalos $C = \{c_l\}_{l=0}^{L}$. Es decir, el final del polinomio en un subintervalo debe coincidir con el inicio del polinomio en el siguiente subintervalo (hasta la derivada $m-2$).

Los *B-Splines* (bases de Splines) se calculan fácilmente con el algoritmo de Boor:

i) B-Splines de grado m=0,

$$B_j^0(x) = \mathbb{1}_{[c_j, c_{j+1}]}(x).$$

ii) B-Splines de orden superior,

$$B_j^m(x) = \frac{x - c_j}{c_{j+1} - c_j} B_j^{m-1}(x) + \frac{c_{j+m+1} - x}{c_{j+m+1} - c_{j+1}} B_{j+1}^{m-1}(x).$$

Entre ellos los más utilizados son los cúbicos.

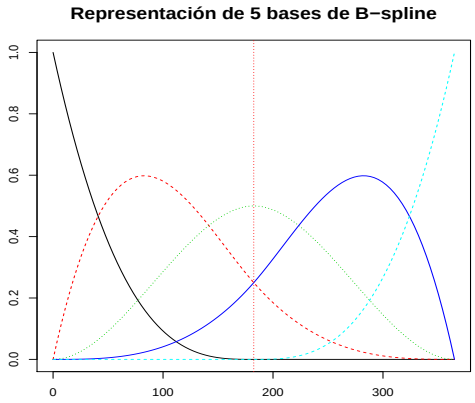

Figura C.1: Bases de B-Splines

La figura (C.1) muestra la representación de 5 bases de B-Splines en el rango $[0, 365]$. El grado de los polinomios es 4 y en consecuencia hay únicamente un nodo interior ($c_1 = 182.5$), dado que el número de parámetros necesarios para definir una función Spline, K, es el número de nodos interiores $(L - 1)$ más el orden del polinomio (m). Obsérvese que en nuestro ejemplo $c_0 = 0$, $c_1 = 182.5$ y $c_2 = 365$.

C.4. Regresión P-Splines

Los *P-Splines* permiten solucionar el problema de selección del número y ubicación de los nodos en el modelo de regresión anterior.

Los P-Splines son suavizadores de bajo rango usando bases de B-Splines, usualmente definidas en nodos uniformemente espaciados, añadiendo un término de penalización aplicado directamente sobre los parámetros, a_i, para controlar la variabilidad de la función m.

Los P-Splines utilizan el siguiente término de penalización relacionado con la curvatura de la función obtenida:

$$pen(m) = \int m''(x)^2 dx. \tag{C.11}$$

Apéndice D

Software

Para la realización de este trabajo se ha utilizado en software estadístico *R*.

En los estudios vistos en el Capítulo 3 se ha utilizado la función ***arima.sim***, que permite simular observaciones de un modelo ARIMA.

Para programar el algoritmo de la página 14 se han hecho uso de las siguientes funciones:

- ***gam*** de la librería **mgcv**, que ha permitido calcular las regresiones de B-Splines y P-Splines.
- ***npreg*** de la librería **np** para calcular la estimación proporcionada por la regresión lineal local.
- Una función propia para calcular el estimador de Nadaraya-Watson y los diversos criterios de error considerados (Error Cuadrático Medio, Error Absoluto, Error Relativo Cuadrático y Error Relativo Absoluto).

En el análisis paramétrico (ARMA) de las series temporales se han utilizado las siguientes funciones:

- ***adf.test*** de la librería **tseries** para contrastar la estacionariedad de la serie temporal.
- ***acf*** y ***pacf*** para calcular los gráficos de autocorrelaciones simples y parciales, respectivamente.
- ***mejor.arma***, programada por el profesor Germán Aneiros Pérez, que permite calcular modelos ARMA tentativos óptimos para algún criterio (en este trabajo se ha utilizado el criterio BIC).
- ***arima*** que permite ajustar un modelo ARIMA a un conjunto de datos.

- ***tsdiag***, ***t.test*** y ***shapiro.test*** para la etapa de diagnosis de la serie mediante Box-Jenkins.

Finalmente se ha programado la función ***IC.boot*** que permite calcular los intervalos de confianza bootstrap de los modelos semiparamétricos, utilizando para ello el algoritmo propuesto por Cao, Febrero-Bande, González-Manteiga, Prada-Sánchez y García-Jurado (1997), visto en la página 22.

Bibliografía

Alonso-Betanzos, A., Fontenla-Romero, O., Guijarro-Berdiñas, B., Hernández-Pereira, E., Paz Andrade, M., Jiménez, E., Legido Soto, J., and Carballas, T. (2003). An intelligent system for forest fire risk prediction and fire fighting management in galicia. *Expert Systems with Applications*, 25:545–554.

Beckage, B. and Platt, W. J. (2003). Predicting severe wildfire years in the florida everglades. *Front Ecol Environ*, 1.

Cao Abad, R., Bande Febrero, M., García Jurado, I., González Manteiga, W., and Prada Sánchez, J. M. (1994). Un estudio de simulación comparativo de técnicas no paramétricas, semiparamétricas y box-jenkins para la predicción con datos dependientes. *Estadística Española*, 36:5–20.

Chas, M. (2007). Forest fires in galicia (spain): Threats and challenges for the future. *Journal of Forest Economics*, 13:1–5.

Cryer, J. and Chan, K.-S. (2008). *Time Series Analysis. With Applications in R.* Springer.

Dabo Niang, S., Francq, C., and Zakoian, J. M. (2010). Combining nonparametric and optimal linear time series predictions. *JASA*.

DGCN (2006). *Los incendios forestales en España. Decenio 1996-2005, Dirección General de Conservación de la Naturaleza.* MMA, Madrid.

Enríquez-Alcalde, E. (2010). *Informes Técnicos de los Grandes Incendios, Contenidos para Elaborar una Base de Datos. Lecciones aprendidas en los grandes incendios forestales.* Ministerio de Medio Ambiente, Junta de Andalucía, Universidad de Córdoba y Sociedad Española de Ciencias Forestales. Madrid.

Fearmaga, M. I. (2011). *Resultados de la Industria de la madera de Galicia 2010.*

García Jurado, I., González Manteiga, W., Prada Sánchez, J. M., Febrero Bande, M., and Cao, R. (1995). Predicting using box-jenkins, nonparametric, and bootstrap techniques. *Technometrics*, 37.

González, M. and del Puerto García, I. (2009). *Series temporales.* Colección manuales uex-60.

Hardy, C. (2005). Wildland fire hazard and risk: Problems, definitions and context. *Forest Ecology and Management*, 211:73–82.

Lee, B., Park, P. S., and Chung, J. (2006). Temporal and spatial characteristics of forest fires in south korea between 1970 and 2003. *International Journal ofWildland Fire*, 15.

Li, Y., Campbell, E., Haswell, D., Sneeuwjagt, R., and Venables, W. (2003). Statistical forecasting of soil dryness index in the southwest of western australia. *Forest Ecology and Management*, 183.

Malamud, B., Morein, G., and Turcotte, D. (1998). Forest fires: an example of self-organized criticality. *Science*, 281:1840–1842.

Marey-Pérez, M. and Rodríguez-Vicente, V. (2009). Forest transition in northern spain: Local responses on large-scale programmes of field-afforestation. *Land Use Policy*, 26:139–156.

Marey-Pérez, M., Rodríguez-Vicente, V., and Crecente-Maseda, R. (2006). Using gis to measure changes in the temporal and spatial dynamics of forestland: experiences from north-west spain. *Forestry*, 79:409–423.

MARM (2008). *Los Incendios Forestales en España, 1968-2007*. Ministerio de Medio Ambiente, Rural y Marino.

Martín, J. P. (2010). Forest fire impacts in galician forest economy. *Ceida*.

Minnich, R. A. and Bahre, C. J. (1995). Wildland fire and chaparral succesion along the california-baja california boundary. *International Journal of Wildland Fire*, 5.

Molina, D., Blanco, J., Galan, M., Pous, E., García Jurado, I.a, J., and García Jurado, I.a, D. (2009). *Incendios Forestales: Fundamentos, Lecciones Aprendidas y Retos de Futuro*. Editorial AIFEMA, Granada (España).

Moritz, M. (2003). Spatiotemporal analysis of controls on shrubland and fire regimes: age dependency and fire hazard. *Ecology*, 84:351–361.

Peña, D. (2005). *Análisis de Series Temporales*. Alianza Editorial.

Reed, W. J. and McKelvey, K. S. (2002). Power-law behaviour and parametric models for the size-distribution of forest fires. *fires.Ecological Modelling*, 150.

Riaño, D., Moreno Ruiz, J., Barón Martínez, J., and Ustin, S. (2007). Burned area forecasting using past burned area records and southern oscillation index for tropical africa (1981-1999). *Remote Sensing of Environment*, 107.

Shumway, R. and Stoffer, D. (2006). *Time Series Analysis and Its Applications. With R Examples*. Springer.

Tombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, 85.

Vélez, R. (2002). Causes of fires in the mediterranean basin. *EFI proceedings*, 45:35–42.

Wood, S. (2006). *Generalized Additive Models. An introduction with R.* Chapman-Hall.

Yakowitz, S. J. (1985). Nonparametric density estimation, prediction and regression for markov sequences. *Journal of the American Statistical Association*, 339.

Yakowitz, S. J. (1987). Nearest-neighbour methods for time series analysis. *Journal of Time Series*, 2.